Monika Maciejewska

ANALIZA DANYCH
W CZUJNIKOWYCH POMIARACH ZANIECZYSZCZEŃ POWIETRZA

Oficyna Wydawnicza Politechniki Wrocławskiej
Wrocław 2012

Recenzenci
Andrzej DZIEDZIC
Roman ZARZYCKI

Projekt okładki
Marcin ZAWADZKI

OFICYNA WYDAWNICZA POLITECHNIKI WROCŁAWSKIEJ
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław
http://www.oficyna.pwr.wroc.pl
e-mail: oficwyd@pwr.wroc.pl
zamawianie.ksiazek@pwr.wroc.pl

ISBN 978-83-7493-724-5

Drukarnia Oficyny Wydawniczej Politechniki Wrocławskiej. Zam. nr 1046/2012.

Spis treści

Ważniejsze skróty i oznaczenia*

R – macierz czujnikowych danych pomiarowych
n – liczba czujników w matrycy
i – numer czujnika w matrycy (1, …, n)
m – liczba punktów czasowych sygnału czujnika
j – numer punktu czasowego sygnału czujnika, 1, …, m
r_{ij} – wartość sygnału czujnika w punkcie czasowym
X – macierz danych wielowymiarowych
x – wektor cech
x – cecha
k – liczba klas; liczba składników mieszaniny; krotność walidacji
c – stężenie gazu
y – zmienna objaśniana
ε – składnik losowy
MCR – udział błędnych klasyfikacji
RMSE – średni błąd kwadratowy predykcji
MRE – średni błąd względny predykcji
PCA – analiza składowych głównych
LDA – liniowa analiza dyskryminacyjna
k-NN – metoda *k*-najbliższych sąsiadów
CART – drzewa klasyfikacji i regresji
PLS – metoda cząstkowych najmniejszych kwadratów
ANN – sztuczne sieci neuronowe

*Objaśnienia użytych skrótów pochodzących z języka angielskiego znajdzie Czytelnik w tekście monografii.

1. Wprowadzenie

We współczesnym świecie zasadniczą rolę odgrywa informacja. Głównym źródłem informacji są dane. Dlatego podstawowe znaczenie dla społeczeństw informacyjnych mają metody i narzędzia umożliwiające przetwarzanie danych w informację [1]. Zjawisko to jest widoczne także w obszarze pozyskiwania informacji o środowisku.

Nowoczesne metody i techniki obserwacji dostarczają dane, których cechą charakterystyczną jest duża złożoność. Badane obiekty są reprezentowane za pomocą wielu zmiennych. Najczęściej wykorzystuje się dodatkowo więcej niż jedną realizację tych zmiennych. Te celowe zabiegi prowadzą do uzyskania danych o dużej zawartości informacji odnoszącej się do wielu właściwości określonego obiektu. Kluczową rolę w przekształceniu złożonych danych pomiarowych w użyteczną informację odgrywają systemy analizy danych. Angażowane są w tym celu różnorodne metody i techniki przetwarzania oraz analizy danych, odpowiednie dla rodzaju poszukiwanej informacji.

Połączenie źródła złożonych danych pomiarowych i systemu ich analizy jest propozycją metody pozyskiwania kompleksowej informacji o zanieczyszczeniu środowiska [2].

W tym kontekście coraz więcej uwagi poświęca się metodom i technikom czujnikowym, zwłaszcza z perspektywy zanieczyszczenia powietrza. W czujnikowych pomiarach gazów złożone dane pomiarowe są uzyskiwane na przykład dzięki zastosowaniu matryc składających się z czujników o zróżnicowanej częściowej selektywności oraz czułości [3]. Szereg czynników związanych z szeroko rozumianą budową sensorów oraz trybem ich pracy decyduje o dużej zawartości informacyjnej czujnikowych danych pomiarowych. Rozwój techniki sensorowej zmierza w kierunku dostarczania danych tego typu niewielkim kosztem, w czasie rzeczywistym. Ponadto poszerza się zakres dostępnych informacji [4]. Z drugiej strony bardzo dobre rezultaty przynosi zastosowanie metod z obszaru rozpoznawania wzorców (ang. *pattern recognition*) [5] jako strategii analizy złożonych danych uzyskiwanych w pomiarach wykonywanych matrycami czujników [6]. Metody czujnikowe stwarzają możliwość jakościowego oraz ilościowego określania gazów. W połączeniu z techniką rozpoznawania wzorców powstają obiecujące podwaliny pod opracowywanie nowoczesnych systemów pozyskiwania informacji o gazowych zanieczyszczeniach środowiska.

Istotna kategoria problemów leżąca w zakresie możliwości takiego rozwiązania dotyczy pozyskiwania informacji o środowisku w warunkach braku jasno zdefiniowanego przedmiotu poszukiwań. W obliczu rosnącej liczby danych pochodzących z ciągłej obserwacji ziemi zadania tego rodzaju będą się pojawiały coraz częściej. Celem analizy danych jest wówczas ujawnienie rodzaju informacji występującej w danych pomiarowych. Istnieje duża grupa metod ułatwiających realizację takiego zadania. Są to metody eksploracji danych (ang. *exploratory analysis*, *data mining*) określane również jako metody rozpoznawania wzorców bez nadzoru (ang. *unsupervised pattern recognition*) [7]. Ich podstawowym zadaniem jest wykrywanie prawidłowości, jakie ujawniają się w danych i powiązanie ich z prawidłowościami, które dotyczą właściwości badanych obiektów. Znacznie szersza kategoria obejmuje problemy, w których rodzaj poszukiwanej informacji o środowisku jest znany. Adekwatna metodologia analizy danych jest określana mianem rozpoznawania wzorców z nadzorem (ang. *supervised pattern recognition*).

Jeżeli przedmiotem uwagi są właściwości środowiska o charakterze jakościowym, to zadaniem analizy danych jest rozpoznanie przynależności danych do jednej z wcześniej określonych kategorii. Zadania kategoryzacji danych realizują tzw. metody klasyfikacji (ang. *classification methods*) [8]. W podejściu tradycyjnym kategorie są na ogół związane z tożsamością chemiczną substancji zanieczyszczających [9, 10]. Możliwe jest jednak zdefiniowanie kategorii w odniesieniu do innych właściwości zanieczyszczeń, takich jak np. toksyczność, charakter odorotwórczy [11, 12], źródło pochodzenia [13, 14]. Bardzo interesująca jest perspektywa posługiwania się kategoriami zanieczyszczenia jako stanu, np. w zakresie jakości powietrza wewnętrznego (ang. *indor air quality*) [15, 16]. Jest to obecnie jeden z najbardziej aktualnych problemów badawczych w dziedzinie inżynierii środowiska. Dane z pomiarów czujnikowych są zasobnym źródłem informacji o różnych właściwościach gazów.

Jeżeli zanieczyszczenie podlega ocenie pod względem ilościowym, to zadaniem analizy danych jest znalezienie przekształcenia, które odwzorowuje dane pomiarowe w zmienne ciągłe, reprezentujące właściwości ilościowe. Problemy tego typu są rozwiązywane z zastosowaniem analizy regresji. W tradycyjnym rozumieniu właściwością o charakterze ilościowym jest stężenie substancji zanieczyszczającej [17, 18]. Jednak posługiwanie się tą miarą ma sens, jeżeli w powietrzu występuje tylko kilka istotnych substancji zanieczyszczających. Zawodzi ono dla bardziej skomplikowanych układów, złożonych z kilkunastu lub kilkudziesięciu substancji. W takich warunkach powstają trudności oceny faktycznego stanu zanieczyszczenia powietrza, gdyż łączne oddziaływanie zanieczyszczeń nie jest prostą sumą oddziaływań pojedynczych substancji. Przykładem takiego problemu jest zanieczyszczenie powietrza, zwłaszcza wewnętrznego, lotnymi związkami organicznymi (LZO), które zazwyczaj występują w wieloskładnikowych mieszaninach. Potrzebne są zatem propozycje alternatywnych miar zbiorczych oraz metod określania takich miar [19, 20]. W tym zakresie dane

z pomiarów czujnikowych charakteryzują się bardzo dużym potencjałem jako źródło informacji.

Przegląd rodzajów informacji dostępnych w wyniku analizy danych z czujnikowych pomiarów gazów pokazuje, że systemy takie otwierają możliwość przewartościowania spojrzenia na zanieczyszczenie środowiska z perspektywy pomiarowej. Powstaje metodologia dostarczania informacji o środowisku komplementarna do tradycyjnych metod chemii analitycznej, pozwalająca odnieść się do problemu zanieczyszczenia również w innych aspektach, niż ściśle chemiczny.

Celem monografii było wykazanie, że analiza danych umożliwia pozyskanie informacji o zanieczyszczeniu powietrza w szerokim zakresie na podstawie czujnikowych danych pomiarowych. Warunkiem realizacji tego zadania jest dobór odpowiedniego zestawu metod i technik analizy danych. Cel pracy realizowano ze świadomością, że postęp techniki czujnikowej oraz rozwój metod analizy danych wysuwa połączenie matrycy czujników i systemu analizy danych na czoło propozycji systemów pozyskiwania kompleksowej informacji o środowisku.

Praca składa się z dwóch części. Pierwszą z nich, o charakterze opisowym, oparto na doniesieniach literaturowych z zakresu analizy danych w czujnikowych pomiarach gazów z uwzględnieniem prac powstałych z udziałem autorki.

Podstawowa koncepcja analizy danych w czujnikowych pomiarach gazów wywodzi się z dziedziny rozpoznawania wzorców. Dlatego część teoretyczną pracy (rozdział drugi) rozpoczęto od przedstawienia zasad opracowywania systemu rozpoznawania wzorców.

Przydatność tego podejścia wynika ze sposobu przenoszenia informacji przez dane czujnikowe. Dane takie zawierają informację o badanych gazach dzięki zdolności czujników do przetwarzania informacji chemicznej w sygnał użyteczny analitycznie. Ze względu na właściwości pomiarowe sensorów pozyskanie informacji wymaga posłużenia się wielowymiarową reprezentacją badanych gazów. Zagadnienia te omówiono w rozdziale trzecim.

System rozpoznawania wzorców pracuje z danymi. Mają one z reguły charakter złożony. W rozdziale czwartym pracy omówiono powiązanie struktur złożonych danych pomiarowych z pomiarów czujnikowych z ich pojemnością informacyjną.

Złożone dane pomiarowe z definicji charakteryzuje nadmiarowość. Zasady i metody budowy skondensowanej reprezentacji badanych gazów przedstawiano w rozdziale piątym. Operacje te służą efektywnemu pozyskaniu informacji dzięki redukcji wymiarowości przestrzeni cech, w której następuje odczyt.

Systemy analizy danych realizują odczyt informacji o badanych gazach przez rozwiązywanie zadań eksploracji, klasyfikacji lub regresji dla danych pomiarowych. Wybrane metody z tego zakresu przedstawiono w rozdziale szóstym. Obejmuje on metody ogólnego zastosowania, które zyskały sobie ugruntowaną pozycję w analizie danych z czujnikowych pomiarów gazów, lecz również takie, dla których istnieją dopiero pojedyncze przykłady zastosowań, natomiast metody są perspektywiczne.

Część teoretyczną pracy zamyka ocena systemu analizy danych w czujnikowych pomiarach gazów, przedstawiona w rozdziale siódmym.

Lektura części teoretycznej powinna uzmysłowić czytelnikowi, że analiza danych w czujnikowych pomiarach gazów, a zwłaszcza zanieczyszczeń gazowych, jest procesem złożonym, którego efekt zależy od właściwego skomponowania elementów realizujących poszczególne etapy. Odpowiedni dobór i połączenie metod konstrukcji najlepszych przestrzeni cech oraz metod eksploracji, klasyfikacji bądź regresji operujących w tych przestrzeniach umożliwia efektywne pozyskanie różnorodnych informacji o zanieczyszczeniu środowiska na podstawie czujnikowych danych pomiarowych.

Druga część pracy ma charakter badawczy i dotyczy pozyskiwania różnych rodzajów informacji o zanieczyszczeniach powietrza z zastosowaniem analizy czujnikowych danych pomiarowych. Jako przedmiot analizy wybrano dane stanowiące wyniki czujnikowych pomiarów lotnych związków organicznych. Umożliwiło to analizę możliwości pozyskiwania wszelakich informacji o zanieczyszczeniu powietrza.

Dane czujnikowe poddano analizie pod względem rozpoznania rodzaju przenoszonej przez nie informacji oraz odczytu konkretnych informacji o zanieczyszczeniu. Detekcji rodzaju informacji, jakościowemu rozpoznawaniu zanieczyszczeń oraz określaniu zanieczyszczeń pod względem ilościowym poświęcono osobne rozdziały.

Pod względem metodologicznym zakres przedstawionej analizy danych został w dużej mierze zdefiniowany przez: i) przyjętą koncepcję cechy jako pojedynczej wartości sygnału czujnika, ii) postulat, że każdy problem określania gazów należy rozwiązywać w najlepszej (odmiennej) dla tego celu przestrzeni cech, iii) dążenie do maksymalnej redukcji wymiaru przestrzeni cech, iv) wybór metod klasyfikacji i regresji, parametryzowanych w ramach pojedynczej prezentacji uczącego zbioru danych.

Z rezultatów przeprowadzonej analizy danych wynika, że założenia te pozwalają zbudować efektywne systemy analizy danych na potrzeby systemów czujnikowych do pomiarów zanieczyszczeń.

Podstawową część obliczeń, których wyniki przedstawiono w pracy, wykonano, korzystając z zasobów obliczeniowych Wrocławskiego Centrum Sieciowo-Superkomputerowego. Autorka dziękuje za ich udostępnienie oraz życzliwość Zespołu Wsparcia.

2. Rozpoznawanie wzorców jako koncepcja analizy danych w czujnikowych pomiarach zanieczyszczeń

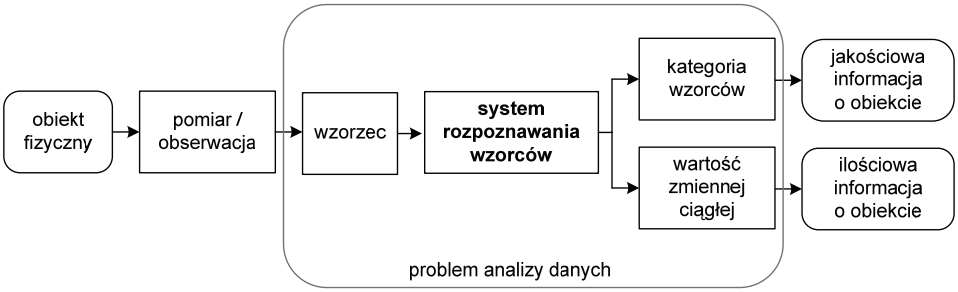
Metody analizy danych w czujnikowych pomiarach zanieczyszczeń należą do większej grupy metod rozwijanych na potrzeby analizy danych w czujnikowych pomiarach gazów. Podstawowe ramy koncepcyjne analizy danych z takich pomiarów są zgodne z podejściem wypracowanym w dziedzinie rozpoznawania wzorców (ang. *pattern recognition*) [21–23].

Rozpoznawanie wzorców zajmuje się przypisywaniem obiektów do różnego rodzaju kategorii lub klas, inaczej mówiąc klasyfikacją obiektów. W zakresie zainteresowania tej dyscypliny naukowej znajdują się również problemy przypisywania obiektom wartości zmiennej rzeczywistej ciągłej, tzw. problemy regresji. Wspomniane obiekty określane są ogólnym terminem wzorce (ang. *patterns*). Mogą nimi być na przykład dane pochodzące z pomiaru lub obserwacji obiektów fizycznych w formie liczb, lecz również w postaci obrazu lub dźwięku [24]. Wzorce na ogół w sposób naturalny odsyłają do obiektów fizycznych. Zawierają informację o nich. Rozpoznawanie wzorców jest w zasadzie sposobem pozyskiwania tej informacji i operowania nią. Koncepcja wzorca pozwala przedstawić i rozwiązać problemy identyfikacji obiektów fizycznych pod względem jakościowym bądź ilościowym jako problemy analizy danych (rys. 2.1).

Zaletą metody rozpoznawania wzorców jest dopuszczenie pewnego stopnia rozmycia wzorców, tzn. przyjęcia, że wzorce wskazujące na ten sam obiekt mogą w pewnym stopniu różnić się od siebie, np. ze względu na niedoskonałość procesu pomiaru czy obserwacji. Przystawalność tego założenia do rzeczywistości jest przyczyną dużej praktycznej przydatności opisywanej metody. Jej fizyczną realizacją są systemy rozpoznawania wzorców.

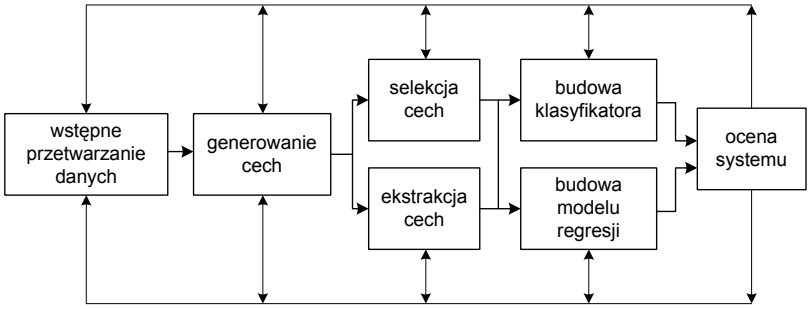
Sprzyjające warunki dla praktycznego zastosowania rozpoznawania wzorców powstały wraz z dynamicznym rozwojem techniki komputerowej, która umożliwia szybkie przetwarzanie dużych zbiorów danych. Stanowi ono integralną część większości systemów ze sztuczną inteligencją realizujących procesy podejmowania decyzji. Od-

grywa istotną rolę w wielu dziedzinach, takich jak np. widzenie maszynowe, rozpoznawanie mowy, rozpoznawanie pisma i diagnostyka wspomagana komputerowo [5].



Rys. 2.1. Pozyskiwanie informacji o obiekcie fizycznym na podstawie analizy danych pomiarowych z wykorzystaniem systemu rozpoznawania wzorców

Rozpoznawanie wzorców składa się z kilku etapów, które zarazem odpowiadają fazom budowy systemu rozpoznawania wzorców, jak pokazano na rys. 2.2. Etapy te nie są niezależne. Ze względu na ich wzajemne powiązanie optymalne funkcjonowanie systemu zależy od właściwego doboru metod realizujących zadania typowe dla każdego etapu. Proces budowy systemu może się odbywać iteracyjnie i polegać na przeprojektowywaniu poszczególnych etapów do momentu uzyskania najlepszego rezultatu ogólnego. Dostępne są również metody, które umożliwiają powiązanie etapów ze sobą i łączną optymalizację całego systemu.



Rys. 2.2. Fazy budowy systemu rozpoznawania wzorców

Pierwszym etapem budowy systemu rozpoznawania wzorców jest zidentyfikowanie mierzalnych wielkości, które pozwalają wyróżnić grupy wzorców podobnych, tzw. klasy lub kategorie wzorców. Wielkości te nie są znane z wyprzedzeniem. W zasadzie należy je wyłonić indywidualnie dla każdego problemu rozpoznawania wzorców. Określa się je mianem cech (ang. *feature*), a etap ich wyłaniania to generowanie cech. W ogólnym przypadku k cech x_i, $i = 1, 2, \ldots, k$, wektor $\mathbf{x} = [x_1, x_2, \ldots, x_k]^T$ stanowi wektor cech. Wektor cech wskazuje na wzorzec, który jest wektorem danych. Wzo-

rzec jest jednoznacznie związany z określonym realnym obiektem fizycznym. W praktyce zbiór wygenerowanych cech jest na ogół większy niż zbiór cech koniecznych do rozpoznawania wzorców. Uważa się je zatem za cechy kandydujące, z których część zostanie wyeliminowana.

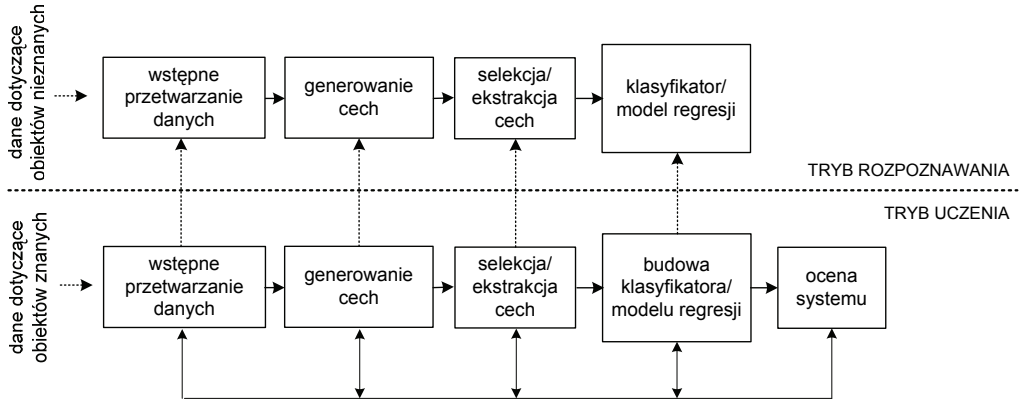
Drugi etap budowy systemu rozpoznawania wzorców służy do zredukowania zestawu cech kandydujących. Składowe uzyskanego wektora cech powinny utworzyć przestrzeń cech (ang. *feature space*), w której różne klasy wzorców są jak najbardziej odległe. Dzięki temu poprawia się rozróżnialność klas. Ze wzgl,ędów praktycznych istotne jest znalezienie jak najmniejszego zbioru takich cech. W tym celu stosowane są dwie podstawowe metody: ekstrakcja cech i selekcja cech. Skład ostateczny wektora cech może być uzyskany na drodze filtracji (ang. *filtration*), tj. bez udziału klasyfikatora w ocenie przydatności cech. Inne sposoby jego określania włączają klasyfikator w proces selekcji cech jako zewnętrzne kryterium oceniające – jest to tzw. podejście opakowane (ang. *wrapper*), lub jako kryterium wewnętrzne, uwzględniane w procesie budowy klasyfikatora – jest to tzw. podejście wbudowane (ang. *embedded*).

Zadaniem trzeciego etapu budowy systemu rozpoznawania wzorców jest opracowanie metody przypisania wzorców do klas, do których one należą. Zasadniczo dostępne są tu dwa rodzaje podejść. Pierwszy rodzaj nosi nazwę uczenia z nadzorem. Realizujący je klasyfikator lub model regresji jest budowany z wykorzystaniem informacji dostępnej *a priori*. W tym przypadku musi być zapewniony dostęp do bazy danych obejmującej wzorce, o których wiadomo, do jakich klas należą, lub jakie wartości odpowiedniej zmiennej ciągłej należy im przypisać. Są to tak zwane wzorce uczące (ang. *training patterns*), w przeciwieństwie do wzorców testowych, których przynależność do klas określa się po zbudowaniu klasyfikatora. Inny rodzaj podejścia nosi nazwę uczenia bez nadzoru. Jest ono stosowane wówczas, gdy nie ma informacji o przynależności klasowej danych uczących. W tym przypadku celem rozpoznawania wzorców jest ujawnienie podobieństw i różnic wśród wzorców. W rezultacie grupowania podobne wektory skupiają się blisko siebie, wskazując na klasę wzorców. Interpretacja klas wzorców wyłonionych w toku takiego postępowania niekoniecznie jest jednoznaczna i bywa trudna. Zazwyczaj jej opracowanie wymaga posłużenia się dodatkową informacją o pochodzeniu poszczególnych wzorców. Operacja ta jest jednak konieczna do zrozumienia podłoża podobieństwa w obrębie wyłonionych klas i zdobycia dodatkowej wiedzy o realnych obiektach, których dotyczą wzorce.

Kluczową cechą systemu rozpoznawania wzorców jest możliwość działania w dwóch trybach: uczenia i rozpoznawania (rys. 2.3).

W trybie uczenia są znajdowane najlepsze metody wstępnego przetworzenia danych, generowania cech, opracowywane metody selekcji lub ekstrakcji najlepszych zestawów cech oraz konstruowane modele matematyczne rozpoznawania, które odwzorowują przekształcenie wzorców w etykiety klas, do których wzorce te należą, lub w wartości zmiennej ciągłej, która ilościowo opisuje obiekt [5, 8]. W tym trybie działania system jest profilowany dla konkretnego problemu rozpoznawania. Poszukiwa-

nie najlepszych rozwiązań dla każdego modułu na ogół nie odbywa się w oderwaniu, lecz w powiązaniu z innymi modułami (linia sprzężeń zwrotnych). W efekcie jest optymalizowany cały system.



Rys. 2.3. Dwa podstawowe tryby działania systemu rozpoznawania wzorców, tryb uczenia i tryb rozpoznawania [25]

Podstawą uczenia systemu rozpoznawania wzorców jest statystycznie reprezentatywny zbiór danych. Na jego podstawie jest projektowana i uczona maszyna ucząca (ang. *learning machine*) [26]. Uzyskany model jest następnie testowany na osobnym zbiorze danych walidujących i/lub testujących pod względem oceny jego zdolności rozpoznawania wzorców. Należy mieć na uwadze, że powstały model ma status taki jak zmienna losowa. Oznacza to, że modele zbudowane dla różnych prób losowych danych nie są identyczne.

Uproszczoną wersją trybu uczenia jest tryb kalibracji. Jego zadaniem jest uaktualnienie systemu rozpoznawania ze względu na ewentualne przesunięcia klas w przestrzeni cech. W tym wypadku działania najczęściej koncentrują się na module klasyfikacji/regresji i polegają na reparametryzacji modelu matematycznego.

W trybie rozpoznawania system korzysta z rozwiązań opracowanych w trybie uczenia. Zgodnie z przyjętymi algorytmami dane pomiarowe dotyczące nieznanego obiektu są wstępnie przetwarzane, wyłaniany jest wektor cech i ostatecznie wektorowi danych zostaje przypisana klasa, do której przynależy, lub odpowiednia wartość zmiennej ciągłej.

Teoretyczne ramy badania problemu uczenia na podstawie danych zostały sformułowane w postaci teorii uczenia statystycznego (ang. *statistical learning theory*), głównie przez Vapnik [27] w latach dziewięćdziesiątych ubiegłego stulecia. Formalny zapis uczenia jest następujący [28]:

- Relację probabilistyczną między wejściem $\mathbf{x}$ i wyjściem $\mathbf{y}$ można wyrazić za pomocą (nieznanego) rozkładu prawdopodobieństwa $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{x}|\mathbf{y})$.

• Dany jest N-elementowy zbiór danych $S = \{(x_i, y_i)\}_{i = 1, \ldots, N}$, pochodzących z rozkładu $p(\mathbf{x}, \mathbf{y})$, gdzie $\mathbf{x} \in R^d$ i $\mathbf{y} \in \{1, \ldots, K\}$ w przypadku klasyfikacji oraz $\mathbf{y} \in R^k$ w przypadku regresji; $\mathbf{x}$ jest określane jako wektor cech lub wzorzec.

• Dla danego zestawu danych $\mathbf{S}$ algorytm uczący zwraca model (klasyfikator, funkcję aproksymującą), normalnie oznaczany literą h (lub $\hat{y}$), który należy do klasy hipotez H. Klasa jest parametryzowana wektorem parametrów, np. oznaczanym przez $\mathbf{w}$.

Z formalnego punktu widzenia uczenie nie obejmuje procesów wstępnego przetwarzania danych ani selekcji czy ekstrakcji cech. Zakłada się, że wektor cech, który stanowi wejście maszyny uczącej, jest właściwą reprezentacją informacji o badanym obiekcie. Z podanego opisu procesu uczenia wynika też, że pozwala on odwzorowywać zależności zarówno o charakterze jakościowym, jak i ilościowym. Formalny opis procesu uczenia określa teoretyczne postawy konstrukcji systemu rozpoznawania wzorców do pozyskiwania informacji jakościowej lub ilościowej o badanych obiektach.

Sformułowano szereg kryteriów, jakie powinien spełniać system rozpoznawania wzorców [29]. Do najistotniejszych należą:

• duża dokładność – liczba przypadków błędnych klasyfikacji (informacja jakościowa) oraz błąd określenia miary ilościowej (informacja ilościowa) powinny być jak najmniejsze;

• szybkość – aby mogły być stosowane w pomiarach typu *on-line*, algorytmy analizy wzorców powinny zapewniać uzyskanie rezultatu z jak najmniejszym opóźnieniem;

• łatwość uczenia – okresowo system będzie wymagał reparametryzacji na podstawie uaktualnionej bazy danych wzorców; liczba wymaganych wzorców powinna być jak najmniejsza, natomiast procedura uczenia powinna być łatwa do przeprowadzenia i możliwa do szybkiego zrealizowania;

• małe wymagania pod względem pamięci i mocy obliczeniowych – dotyczy to przede wszystkim możliwości realizacji rozwiązań zaimplementowanych w różnego rodzaju instrumentach;

• odporność na wzorce błędne – oczekuje się, że jeśli takie okoliczności zaistnieją, system będzie w stanie stwierdzić, że wzorzec skierowany do analizy jest zupełnie inny niż wzorce uczące i oceni go jako niejednoznaczny;

• określenie stopnia pewności rezultatu – dotyczy przede wszystkim rozpoznawania wzorców; istotna wartość dodana wiąże się z podaniem statystycznej miary określającej stopień poprawności podanego rezultatu, np. prawdopodobieństwa przynależności wzorca do danej klasy.

Metody i techniki rozpoznawania wzorców są powszechnie stosowane do określania obiektów o charakterze chemicznym. Ta dziedzina rozpoznawania wzorców określana jest jako rozpoznawanie wzorców chemicznych [30]. Jej początki sięgają lat dziewięćdziesiątych ubiegłego wieku i są związane z zaawansowanymi rozwiązaniami instrumentalnymi opracowywanymi na potrzeby chemii analitycznej [31]. Zasadniczą

rolę w rozwoju tego obszaru badań odegrało jednak zastosowanie techniki czujnikowej w pomiarach chemicznych, a zwłaszcza pomiary z użyciem czujników gazów wykazujących ograniczoną selektywność. W wyniku takich pomiarów otrzymuje się złożone dane, które zawierają różnorodne informacje o badanych gazach. Technika ta jest jednak niedostępna bez zastosowania odpowiednich metod analizy danych. Atrakcyjność techniki czujnikowej jako pozalaboratoryjnego rozwiązania pomiarowego stymulowała poszukiwania w obszarze metod pozyskiwania informacji z danych pochodzących z takich pomiarów. Odpowiednie strategie są rozwijane na podstawie osiągnięć w szeroko rozumianej dziedzinie rozpoznawania wzorców. Podejście to jest też atrakcyjne ze względu na możliwość adaptacji w czujnikowych pomiarach gazowych zanieczyszczeń powietrza.

3. Dane z czujnikowych pomiarów gazów jako źródło informacji

Według Unii Chemii Czystej i Stosowanej *czujnik chemiczny to urządzenie, które przekształca informację chemiczną* [...] *w analitycznie użyteczny sygnał. Czujniki chemiczne składają się z dwóch podstawowych elementów, które są połączone szeregowo: chemiczny (molekularny) system rozpoznawania (receptor) i przetwornik fizykochemiczny* [32].

Klasycznym założeniem przyjmowanym do opracowania czujnika chemicznego jest uzyskanie odpowiedzi tylko i wyłącznie na zadaną substancję, tj. osiągnięcie idealnej selektywności [21]. Jej zaletą jest wyeliminowanie wpływu innych substancji występujących w otoczeniu czujnika na wynik pomiaru. Perspektywę uzyskania czujników o dużej selektywności wiąże się dziś głównie z koncepcją *zamek–klucz*, realizowaną np. w bioczujnikach. Warstwa chemoczuła tego rodzaju sensorów zwiera miejsca receptorowe, tak zaprojektowane, że może się tam dostać i zostać wykryta jedynie ściśle określona cząsteczka. Bioczujniki są jednak ciągle niestabilne i nie spełniają wymagań co do zastosowania w technice. W praktyce również selektywność czujników istniejących i uważanych za selektywne jest najczęściej warunkowa. Oznacza to, że zachodzi, jeżeli z bardzo dużym prawdopodobieństwem można przyjąć, że poziom potencjalnych interferencji jest zaniedbywalny. Podstawową wadą koncepcji czujnika selektywnego jest konieczność opracowania osobnego rozwiązania dla każdej substancji interesującej z pomiarowego punktu widzenia.

Selektywność większości istniejących czujników chemicznych jest raczej mała. Przyczyny tego zjawiska mają charakter podstawowy i nie wynikają np. z niedociągnięć technologicznych. Informacja o badanym gazie pojawia się w sygnale czujnika elektrycznego w wyniku interakcji zachodzących między składnikami tego gazu a materiałem chemoczułym. Ze względu na rodzaj stosowanych materiałów źródłem informacji mogą być różne właściwości fizyczne oraz chemiczne badanego gazu. Należą tu np. masa cząsteczkowa, struktura cząsteczki, rozkład ładunku w cząsteczce, kinetyka reakcji, kinetyka adsorpcji i desorpcji gazu [33]. Ze względu na fakt, że wiele substancji chemicznych wykazuje podobieństwa pod względem wymienionych wła-

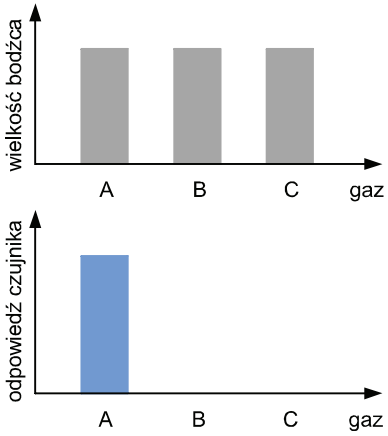
ściwości, odpowiedź czujnika chemicznego na gazy różniące się składem może być podobna. Wynikiem tego jest ograniczona selektywność pojedynczych czujników.

Częściowa selektywność jest jedną z podstawowych właściwości czujników gazów, którą zazwyczaj uznaje się za wadę. Wiele pomysłów na jej skompensowanie wiąże się z zastosowaniem odpowiednich metod analizy danych i zakłada uzyskanie tak zwanej selektywności obliczeniowej (ang. *computational selectivity*) lub elektronicznej (ang. *electronic selectivity*) [34]. Dość wcześnie dostrzeżono też przydatność tego pomysłu w pomiarach zanieczyszczeń powietrza [35].

3.1. Informacje o badanym gazie w odpowiedzi pojedynczego czujnika

Niezerowa wartość odpowiedzi czujnika selektywnego wskazuje na występowanie w badanym gazie substancji, w stosunku do której sensor wykazuje selektywność, w stężeniu przekraczającym próg detekcji. Przez odpowiedź czujnika rozumiemy wartość sygnału zmierzoną w ściśle określony sposób.

Zjawisko selektywności zilustrowano schematycznie na rys. 3.1. Selektywny czujnik odpowiada na gaz A, nie odpowiada zaś ani na gaz B, ani C. Przez sam fakt niezerowej odpowiedzi pojedynczy czujnik dostarcza jednoznaczną informację jakościową o występowaniu w gazie konkretnej substancji.

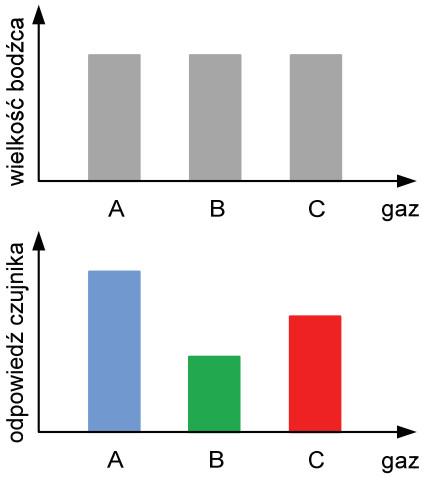


Rys. 3.1. Graficzna ilustracja selektywności czujnika

Łatwo zauważyć, że zestaw czujników selektywnie odpowiadających na różne gazy umożliwia bezpośrednie określenie składu chemicznego mieszaniny gazów.

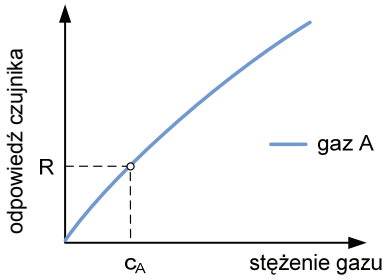
Niestety, informacja, którą można uzyskać z czujników selektywnych jest dość ograniczona. Dotyczy ona wyłącznie tych składników gazu, na które czujniki odpowiadają selektywnie. W praktyce oznacza to, co najwyżej kilka substancji, gdyż dotychczas opracowano czujniki selektywne dla niewielu gazów.

Większość czujników chemicznych ma charakter częściowo selektywny. Istotę częściowej selektywności zilustrowano na rys. 3.2. Jak pokazano, jeden i ten sam częściowo selektywny czujnik odpowiada zarówno na gaz A, B, jak i C. Nie jest zatem selektywny w stosunku do żadnego z nich. Ponieważ na wielkość sygnału czujnika wpływa zarówno rodzaj, jak i stężenie poszczególnych jego składników, więc nie ma możliwości jednoznacznego określenia rodzaju gazu na podstawie odpowiedzi jednego, częściowo selektywnego czujnika. Problem ten dotyczy pojedynczych substancji i w oczywisty sposób przenosi się na wieloskładnikowe mieszaniny gazów. Paradoksalnie jednak z zastosowaniem takich czujników wiąże się poszerzenie zakresu dostępnej informacji o badanych gazach w stosunku do czujników selektywnych. Zyskiwane są dzięki temu nowe możliwości pomiarowe.



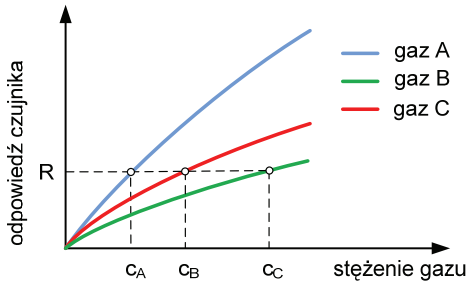
Rys. 3.2. Graficzna ilustracja
częściowej selektywności czujnika [21]

Odpowiedź czujnika selektywnego zmienia się monotonicznie w funkcji stężenia substancji, na którą odpowiada np. według zależności postaci przedstawionej na rys. 3.3. Obecność innych substancji w otoczeniu czujnika pozostaje bez wpływu na tę zależność. Na podstawie sygnału wyjściowego czujnika można zatem uzyskać informację ilościową o tym jednym, określonym składniku badanego gazu. Wymagana jest w tym celu znajomość jednowymiarowej zależności odwzorowującej stężenie gazu w odpowiedź czujnika.



Rys. 3.3. Konsekwencje selektywności czujnika
dla możliwości ilościowego oznaczania badanego gazu

Przełożenie odpowiedzi zestawu czujników selektywnych na informację o składzie mieszaniny jest stosunkowo proste. Należy w tym celu zbudować zestaw jednowymiarowych modeli, z których każdy odwzorowuje zależność odpowiedzi jednego czujnika od stężenia gazu, względem którego czujnik wykazuje selektywność.

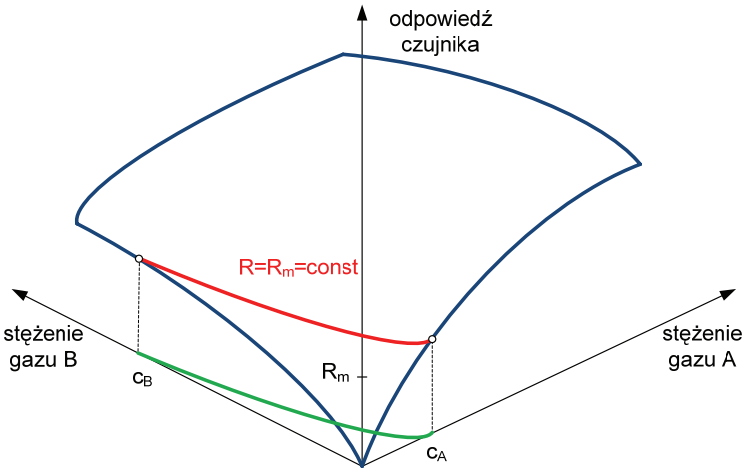


Rys. 3.4. Konsekwencje częściowej selektywności czujnika
dla możliwości ilościowego oznaczania badanych gazów

Odczyt informacji ilościowej zawartej w odpowiedzi czujnika częściowo selektywnego stwarza natomiast pewne problemy. Na rysunku 3.4 przedstawiono hipotetyczne charakterystyki czujnika tego typu dla trzech różnych substancji. Jak wynika z ilustracji, na podstawie odpowiedniej charakterystyki czujnika można określić stężenie każdej z tych substancji, jeżeli występuje ona w gazie pojedynczo. Konieczne jest jednak wcześniejsza identyfikacja substancji, aby było wiadomo, z której charakterystyki należy skorzystać. Na to pytanie nie można jednak odpowiedzieć, dysponując odpowiedzią jednego czujnika częściowo selektywnego, co pokazano wcześniej.

Możliwości pojedynczego czujnika nieselektywnego są jeszcze bardziej ograniczone w przypadku oznaczeń ilościowych wieloskładnikowych mieszanin gazów. Na rysunku 3.5 zilustrowano ten problem na najprostszym przykładzie, dotyczącym mieszaniny dwuskładnikowej [36]. Przedstawiono hipotetyczne charakterystyki czujnika nieselektywnego dla dwóch różnych substancji oraz powierzchnię odpowiedzi dla mieszaniny tych substancji. Łatwo zauważyć, że określona wartość odpowiedzi czuj-

nika może w tym przypadku oznaczać nieskończone spektrum składów mieszaniny wskazanych przez izolinię odpowiedzi R = const.



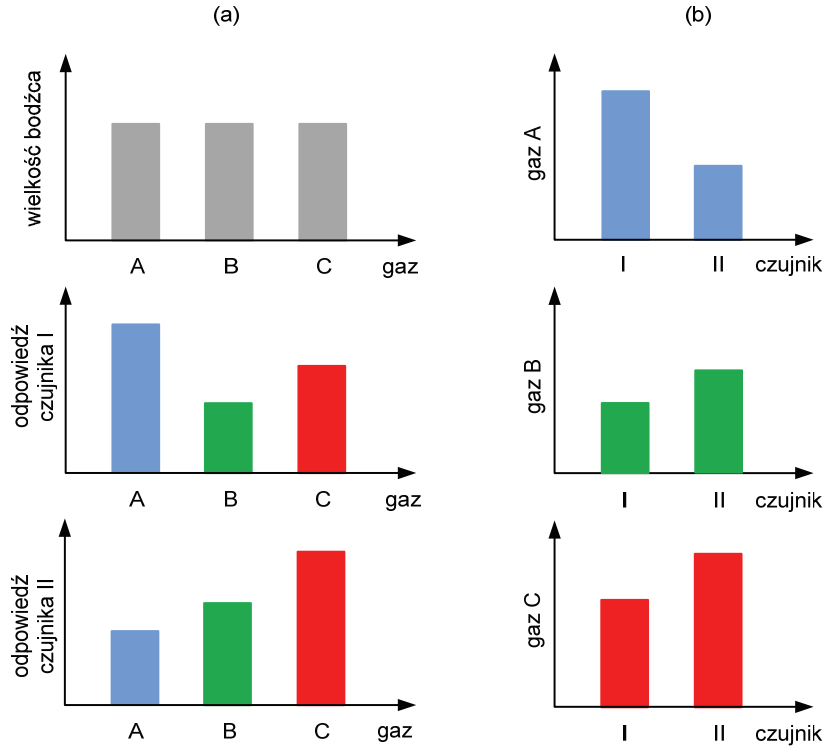
Rys. 3.5. Oznaczanie ilościowe dwuskładnikowej mieszaniny gazów na postawie odpowiedzi jednego czujnika częściowo selektywnego. Linia czerwona oznacza izolinię odpowiedzi czujnika. Linia zielona oznacza wszystkie kombinacje stężeń gazów A i B skutkujące odpowiedzią czujnika, którą pokazano linią czerwoną

3.2. Informacja o badanym gazie w odpowiedzi matrycy czujników

W połowie lat osiemdziesiątych XX wieku zaproponowano po raz pierwszy zastosowanie zestawu czujników częściowo selektywnych, tak zwanej matrycy czujników, do uzyskania jakościowej i ilościowej informacji o badanym gazie [37]. W ogólnej koncepcji tego rozwiązania elementami matrycy są czujniki, z których każdy odpowiada na wiele różnych substancji chemicznych lub klas substancji chemicznych.

Wśród czujników, które funkcjonują w ramach matrycy, powinno występować zróżnicowanie częściowej selektywności. Proporcje odpowiedzi poszczególnych sensorów na gazy z określonego zestawu powinny się różnić (rys. 3.6a). W matrycy czujników występuje zazwyczaj dodatkowo efekt tzw. *krosselektywności* [40]. Polega on na tym, że ten sam gaz wywołuje odpowiedź więcej niż jednego czujnika (rys. 3.6a). Na ogół zakresy czułości czujników stosowanych w matrycach nie są rozłączne, lecz powinny być inne. Dobrze, jeśli zestaw czujników charakteryzuje się jak największym zróżnicowaniem. Wówczas zakres rodzaju gazów, na które odpowiada matryca, może być bardzo szeroki [38, 39].

Odpowiedź pojedynczego czujnika cechuje ograniczona selektywność. Jednak zestaw czujników, z których każdy ma inne spektrum selektywności, dostarcza charakterystyczny dla danego gazu wzorzec sygnałów (rys. 3.6b), który jest odróżnialny od wzorca otrzymanego dla innego gazu. Analiza konkretnego wzorca stwarza możliwości identyfikacji gazu, na który system czujnikowy był eksponowany [41] podczas powstawania tego wzorca.



Rys. 3.6. Graficzna ilustracja: a) zróżnicowania częściowej selektywności i krosselektywności czujników, b) obrazu poszczególnych gazów w odpowiedzi matrycy składającej się z dwóch czujników

Rozwijane są co najmniej dwa kierunki wykorzystania właściwości matrycy czujników. Jeden z nich wiąże się ze wspomnianą ideą selektywności obliczeniowej. Badany gaz można traktować jak mieszaninę substancji i dążyć do zidentyfikowania jej składu zarówno pod względem jakościowym, jak i ilościowym [34]. Jest to zbieżne z tradycyjnym podejściem charakterystycznym dla chemii analitycznej. Najwcześniejsze teoretyczne oszacowanie liczby czujników potrzebnych do zidentyfikowania określonej liczby mieszanin gazów, rozumianych jako rożne kombinacje składników, podano w pracy [42]. Wyraża je następująca zależność:

$$2^n - 1 \geq \sum_{k=1}^{K} \frac{u!}{(u-k)!k!} \tag{3.1}$$

gdzie n jest liczbą czujników, u jest liczbą różnych substancji, które mogą znaleźć się w mieszaninie k-składnikowej, K jest maksymalną liczbą składników mieszaniny gazów ($A \leq u$). W oszacowaniu tym przyjęto najprostsze, lecz nieco nierealistyczne założenie, że pojedynczy czujnik odpowiada dwupoziomowo, tj. reaguje w obecności bodźca i nie reaguje w warunkach jego braku. Częściowe urealnienie zależności podanej w równaniu (3.1) uzyskujemy, założywszy $p > 2$ poziomów odpowiedzi czujnika [37]:

$$p^n - 1 \geq \sum_{k=1}^{K} \frac{u!}{(u-k)!k!} \tag{3.2}$$

Według wzoru (3.1) do rozróżnienia dwuskładnikowych mieszanin gazów, których składnikami może być osiem różnych substancji, potrzeba co najmniej sześciu czujników odpowiadających dwupoziomowo. Przy bardziej realistycznym założeniu, że liczba poziomów odpowiedzi czujnika jest dowolnie większa, wymagana jest mniejsza liczba czujników zgodnie z równaniem (3.2). Przedstawione oszacowania pozwalają zauważyć przewagę zastosowania czujników nieselektywnych w stosunku do selektywnych ze względu na liczbę mieszanin gazów możliwych do oznaczenia jakościowego za pomocą jednego zestawu czujników.

Podejście analityczne, tzn. w kategoriach substancji chemicznych, jest w naturalny sposób obecne w czujnikowych pomiarach zanieczyszczeń. Metody czujnikowe często uważa się za alternatywę dla tradycyjnych technik analitycznych. Rzeczywiście liczne publikacje pokazują możliwości identyfikowania konkretnych substancji zanieczyszczających oraz określania ich stężeń [9, 14, 43–46] na podstawie pomiarów czujnikowych. Mimo to metody i techniki czujnikowe mają indywidualny charakter, a za ich pomocą można realizować inne cele pomiarowe.

Alternatywny kierunek myślenia wiąże się ze spostrzeżeniem, że odpowiedź matrycy czujników na dany gaz niejako z definicji przedstawia go w całości. Jest ona wynikiem złożenia właściwości wszystkich składników gazu oddziałujących na czujniki. Odpowiedź matrycy nie przedstawia gazu w podziale na składniki. Wobec tego obraz gazu uzyskany w wyniku pomiaru czujnikowego można zastosować do wypowiadania się o gazie *jako takim*. Innymi słowy, jest on przydatny do określenia właściwości gazu nie w rozumieniu zestawu komponentów, lecz jako całej mieszaniny. Możliwość ta jest bardzo cenna z perspektywy zastosowań systemów czujnikowych w pomiarach środowiskowych.

Największa liczba prac pozostających w tym nurcie przedstawia czujnikowe rozwiązania problemu oceny odorowej gazów [11, 20, 23, 47–49, 50, 51]. Na przestrzeni ostatnich kilkunastu lat wyłonił się jednak inny obszar zainteresowania, w którym taka

filozofia oceny poziomu zanieczyszczenia jest bardzo użyteczna. Są to szeroko rozumiane zagadnienia oceny jakości powietrza wewnętrznego, IAQ). Potrzeba wykonywania takiej oceny pozostaje dziś poza sferą dyskusji ze względu na konieczność zapewnienia zdrowia i komfortu ludziom spędzającym czas głównie w pomieszczeniach zamkniętych. Problem dotyczy całego kręgu cywilizacji zachodniej, lecz nie tylko. Wiadomo, że szereg symptomów chorobowych i/lub poczucie dyskomfortu u takich osób wynika z łącznego oddziaływania wielu czynników zanieczyszczających powietrze występujących w ich otoczeniu. Sytuację taką określa się jako niską jakość powietrza wewnętrznego. Zanieczyszczenia chemiczne, w szczególności gazowe [52], a wśród nich zwłaszcza lotne związki organiczne [53] należą do grupy najważniejszych przyczyn chorób związanych z budynkiem (ang. *building related ilness*, BRI) oraz syndromu chorego budynku (ang. *sick building syndrome*, SBS). Pomiary czujnikowe sprzężone z systemami analizy danych są uważane za realną propozycję metody pomiaru jakości powietrza wewnętrznego [54–59].

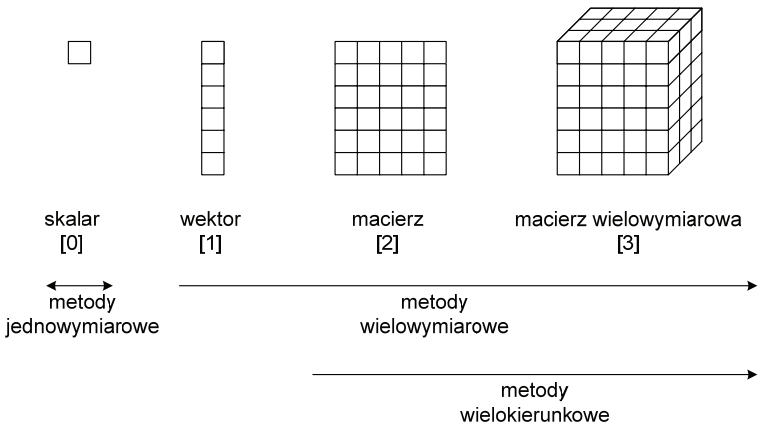
Podstawową zaletą matryc czujnikowych jest, że z definicji można je zastosować do pozyskania różnorodnych informacji o badanych gazach [36, 60, 39].

4. Dane złożone w czujnikowych pomiarach gazów

4.1. Struktury danych pomiarowych

Dane definiuje się jako zbiory wartości zmiennych ilościowych bądź poziomów zmiennych jakościowych odnoszące się do rozmaitych obiektów. Dane pochodzą najczęściej z pomiaru lub obserwacji, lecz mogą też być generowane sztucznie. Współcześnie charakteryzują się na ogół dużą złożonością.

Spójny aparat pojęciowy służący do opisu struktur danych pomiarowych zaproponowano w dziedzinie chemii analitycznej i na jej użytek [61]. Pochodzi on z algebry tensorowej i opiera się na pojęciu rzędu. W swobodnym rozumieniu rząd tensora jest równy minimalnej liczbie indeksów, które należy zastosować, by uporządkować dane w logiczny sposób. Skalar, wektor i macierz są tensorami rzędu zero, jeden i dwa. Kostka danych jest tensorem rzędu trzy.



Rys. 4.1. Dane pomiarowe różnych rzędów oraz grupy metod ich analizy [61]

Logika organizacji w strukturę jest ściśle związana z czynnikami wywołującymi zróżnicowanie w danych pomiarowych (rys. 4.1). Jeżeli takich czynników nie ma,

dane są skalarem i mają rząd zero. W przypadku istnienia jednego czynnika dane należy zorganizować w wektor. Poszczególne elementy wektora odpowiadają różnym poziomom tego czynnika. Dane mają wówczas rząd równy jeden. Jeżeli liczba czynników wywołujących zróżnicowanie danych wynosi dwa, najlepszym sposobem ich organizacji jest macierz. Porządek wierszy wyznacza wówczas kierunek zmienności jednego czynnika, innymi słowy każdy wiersz odpowiada innemu poziomowi tego czynnika. Porządek kolumn jest zaś związany z kierunkiem zmienności drugiego czynnika. Według tej zasady można konstruować coraz bardziej złożone struktury danych.

Przez określenie rzędu danych wskazuje się pośrednio na zasób zawartej w nich informacji. Jego wielkość wiąże się z czynnikami odpowiadającymi za poszczególne wymiary struktury danych, precyzyjniej za zmienność danych w poszczególnych kierunkach. Każdy z czynników umożliwia pojawienie się w danych innej informacji, zwiększając jej ogólną zawartość w danych przez fakt swojego istnienia i oddziaływania.

Rząd danych jest decydujący dla doboru metod ich analizy. Metody jednowymiarowe (ang. *univariate*) są przeznaczone do analizy danych rzędu zero. Dane rzędu jeden i wyższych są analizowane za pomocą metod wielowymiarowych (ang. *multivariate*). Metody analizy dotyczące danych o rzędzie wyższym niż jeden określa się jako wielokierunkowe (ang. *multiway*).

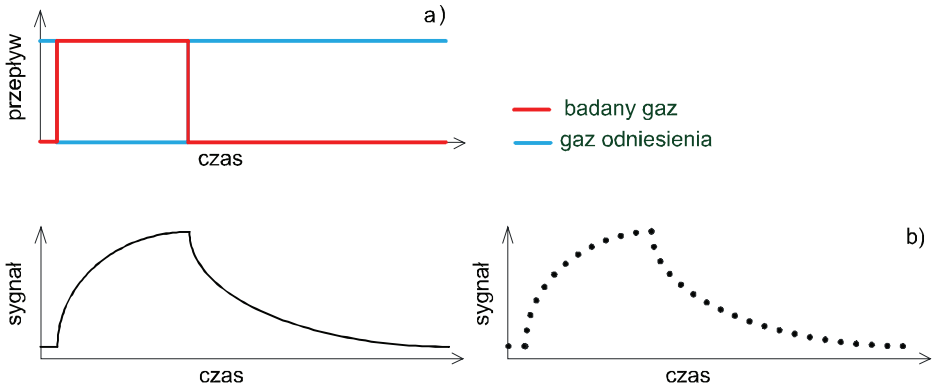
4.2. Struktura czujnikowych danych pomiarowych

System opisu struktur danych przyjęty w chemii analitycznej jest przejrzysty i praktyczny. Można go z powodzeniem zastosować w odniesieniu do danych pochodzących z pomiarów czujnikowych [3].

W wyniku ekspozycji czujnika na badany gaz powstaje użyteczny analitycznie sygnał. Według ogólnej definicji *sygnał jest funkcją zmiennej rzeczywistej t interpretowanej jako czas* [62]. Do obróbki cyfrowej sygnał analogowy o charakterze ciągłym jest poddawany konwersji analogowo-cyfrowej. W wyniku tego przekształcenia uzyskuje się sygnał o charakterze dyskretnym. Według definicji *sygnały dyskretne są funkcjami czasu dyskretnego. Ich wartości są również dyskretne i określone przez rozdzielczość systemu* [62]. Pozyskiwanie informacji z sygnału metodami analizy danych wymaga przedstawienia go w postaci cyfrowej. Obecnie operacja ta jest najczęściej standardowym elementem procedury pomiarowej.

Przykładową postać sygnału czujnika gazu przed i po konwersji analogowo-cyfrowej przedstawiono na rys. 4.2. Przebieg sygnału odpowiada typowej procedurze pomiarowej, gdzie po ustaleniu linii bazowej podczas ekspozycji w atmosferze gazu

nośnego następuje ekspozycja na badany gaz, a następnie regeneracja czujnika w gazie nośnym.



Rys. 4.2. Sygnał odpowiedzi czujnika w trakcie pomiaru:
a) analogowy, b) po konwersji cyfrowej [62]

Po konwersji analogowo-cyfrowej sygnał wyjściowy czujnika rejestrowany w dziedzinie czasu można zatem przedstawić jako wektor pionowy, $\mathbf{r} \in \mathbf{R}^m$, gdzie $j = 1, \ldots, m$ odnosi się do czasu:

$$\mathrm{r}^{\mathrm{T}} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix} \tag{4.1}$$

Dyskretny sygnał czujnika jest przykładem danych rzędu jeden. Łatwo zauważyć, że pojedyncza wartość sygnału i-tego czujnika, np. związana z dyskretnym momentem czasu t_s, jest strukturą danych rzędu zero.

$$\mathrm{r}_{j=t_s} = \begin{bmatrix} r_{t_s} \end{bmatrix} \tag{4.2}$$

Zestaw sygnałów pochodzących od matrycy n czujników to dane rzędu dwa:

$$\mathrm{R} = \begin{bmatrix} \mathrm{r}_1^{\mathrm{T}} & \mathrm{r}_2^{\mathrm{T}} & \ldots & \mathrm{r}_n^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix} \tag{4.3}$$

Kolumny macierzy **R** to wektory $\mathbf{r}_1$, $\mathbf{r}_{2,}$ ..., $\mathbf{r}_n$, będące sygnałami poszczególnych czujników.

Nasuwa się spostrzeżenie, że dane pomiarowe pochodzące od więcej niż jednego czujnika i związane z dyskretnym momentem czasu ekspozycji, $j = t_s$ (pojedynczy wiersz macierzy **R**) to również dane rzędu jeden:

$$\mathrm{R}(j = t_s) = \begin{bmatrix} r_{t_s 1} & r_{t_s 2} & \ldots & r_{t_s n} \end{bmatrix} \tag{4.4}$$

Jak wspomniano wcześniej, rząd danych jest decydujący dla doboru metod ich analizy ze względu na pozyskanie informacji użytecznej analitycznie. Jeśli badana próba gazów jest reprezentowana przez pojedynczą wartość sygnału jednego czujnika, to dane pomiarowe są rzędu zero zgodnie z równaniem (4.2). Dane takie zawierają informację ilościową o gazie, na który czujnik odpowiada. Na ich podstawie możliwe jest określenie stężenia gazu pod warunkiem, że badana próba nie zawiera substancji interferujących. Dane rzędu zero nie dają możliwości określenia badanej próby gazów pod względem jakościowym, w tym stwierdzenia, czy interferencje w ogóle występowały, czy też nie. Do analizy danych rzędu zero są stosowane metody jednowymiarowe. Opracowanie modeli matematycznych polega na odtworzeniu zależności ilościowej między wynikami pomiaru a miarą ilościową badanego gazu, np. stężeniem. Gotowe modele mogą być stosowane do określania stężeń prób gazu, jeśli nie zawiera on substancji interferujących ze składnikiem, dla którego zbudowano model. W ogólnym wypadku dane rzędu zero nie pozwalają na matematyczne rozpoznawania gazów.

Czujnikowe dane pomiarowe mają rząd jeden, jeżeli badaną próbę reprezentuje wektor wartości sygnałów wielu czujników, lecz zarejestrowanych w jednym punkcie czasu zgodnie z równaniem (4.4) lub wektor szeregu wartości pochodzących z sygnału jednego czujnika, jak pokazano w zapisie (4.1). W pierwszym z wymienionych przypadków zróżnicowanie elementów wektora wynika z różnych zakresów selektywności i krosselektywności czujników. Jest ono charakterystyczne dla badanego gazu. Dzięki temu dane rzędu jeden umożliwiają rozpoznawanie substancji gazowych, jak również określanie ich stężeń. Jest to tzw. *korzyść rzędu pierwszego*. Jeżeli dane rzędu jeden powstały z sygnału pojedynczego czujnika, zróżnicowanie ich wartości może być wywoływane celowo lub w sposób niezamierzony przez szereg czynników wpływających na odpowiedź czujnika w trakcie ekspozycji na badany gaz. Wskutek tego sygnał czujnika zyskuje selektywność obliczeniową. Analiza danych rzędu jeden umożliwia efektywne oznaczenie gazu, jeżeli w badanej próbie występują znane interferencje. Jest ona przeprowadzana z zastosowaniem metod wielowymiarowych. Istnieje możliwość opracowania modeli matematycznych służących do określania gazów pod względem jakościowym, jak również składu ilościowego na podstawie danych pomiarowych rzędu jeden. Uzyskane modele mogą być stosowane do określania prób gazu, które zawierają substancje interferujące ze składnikami, dla których zbudowano modele. Warunkiem adekwatności modeli jest uwzględnienie w danych uczących

wszystkich przewidywalnych przypadków interferencji. Wynik oznaczenia nie jest poprawny, jeżeli w badanej próbie znajdą się interferencje nieznane.

Pomiarowe dane czujnikowe drugiego rzędu powstają wówczas, gdy wynik pomiaru badanej próby jest przedstawiany za pomocą macierzy obejmującej sygnały czujników matrycy, zgodnie z zapisem (4.3). Podobnie jak elementy wektora danych rzędu pierwszego poszczególne elementy macierzy powinny wykazywać zróżnicowaną selektywność obliczeniową. Istotne jest jednak, by źródła zróżnicowania selektywności w wierszach i w kolumnach macierzy były niezależne. Dla pewnych założeń analiza danych rzędu dwa pozwala uzyskać poprawne oznaczenie ilościowe substancji nawet w warunkach występowania nieznanych interferencji. Właściwość ta jest znana pod nazwą *korzyści drugiego rzędu*. Najlepiej, jeżeli w takich okolicznościach zostaną zastosowane metody analizy wielokierunkowej [3]. Są one opracowywane dla danych rzędu dwa lub większych. Jedną z zalet analizy wielokierunkowej jest możliwość odtworzenia profili pomiarowych poszczególnych składników mieszaniny bez dostępu do danych pochodzących z pomiarów pojedynczych substancji. Model przeznaczony do określania składu mieszaniny może zatem powstać wyłącznie na podstawie danych pomiarowych dotyczących mieszaniny gazów. Dane drugiego rzędu można analizować metodami odpowiednimi dla analizy danych pierwszego rzędu pod warunkiem rozwinięcia macierzy do postaci wektorowej. Odpowiednie są tu również metody analizy danych rzędu zero, jeżeli uwaga skupi się na pojedynczych elementach macierzy danych. Przykładową analizę porównawczą możliwości czujnikowych danych pomiarowych rzędu jeden i dwa przedstawiono w [63].

Pomiary czujnikowe mogą stanowić źródło danych pomiarowych wysokiego rzędu. Jak wcześniej wspomniano, każdy nowy rząd danych jest związany z niezależnym czynnikiem wywołującym ich zróżnicowanie. W przypadku czujnikowych pomiarów gazów takich czynników jest bardzo wiele. Poniżej wymieniono najistotniejsze z podaniem rzędu wielkości dla liczby poziomów, które te czynniki mogą przyjąć [33, 36]:

- materiały chemoczułe, np. polimery, półprzewodniki, kompozyty – 10^8,
- przetworniki, np. chemirezystory, mikrowagi, włókna optyczne – 10^1,
- geometria przetworników, np. kształt, rozmieszczenie elektrod – 10^2,
- parametry zewnętrzne, np. użycie filtrów, katalizatorów, przepływ gazu i parametry wewnętrzne, np. temperatura pracy, napięcie na elektrodach, częstotliwość światła – 10^2,
- sposoby modulacji parametrów wewnętrznych oraz zewnętrznych, np. według funkcji sinusoidalnej, skokowej – 10^2,
- sposoby modulacji równoczesnej, np. symultaniczna zmiana temperatury i przepływu gazu – 10^6.

Daje to łącznie niebagatelną liczbę 10^{21}, która oczywiście ma charakter hipotetyczny. Realistyczne oszacowanie jest wielokrotnie mniejsze. Jednak praktyka badawcza pokazuje, że do dyspozycji pozostaje nadal bardzo duża liczba różnorodnych czynników, które można wykorzystać w pomiarach czujnikowych w celu poprawy

identyfikacji i ilościowego określania badanych gazów. Wymienione czynniki stanowiące potencjalne źródła kolejnych rzędów w danych czujnikowych można uznać za przyczynek do tzw. sensoryki chemicznej wyższego rzędu (ang. *higher-order chemical sensing*) [3]. Kierunek badań związany z otrzymywaniem i analizą danych pomiarowych rzędu wyższego niż dwa jest obecnie coraz intensywniej rozwijany. Jest to efekt poszukiwania sposobów na zwiększenie zawartości informacji w danych pomiarowych [36, 60].

W praktyce spotyka się czujnikowe dane pomiarowe co najwyżej drugiego rzędu. Za każdy z rzędów odpowiada nie pojedynczy czynnik, lecz raczej ich grupy. Zazwyczaj jedna grupa związana jest z czynnikami materiałowymi, technologicznymi i technicznymi, będącymi przyczyną zróżnicowania czujników stanowiących elementy matrycy. Druga grupa natomiast obejmuje czynniki odpowiadające za szeroko rozumiany tryb pracy czujników oraz kształtowanie warunków ekspozycji na badane gazy. Znane się metody analizy czujnikowych danych pomiarowych rzędu drugiego i mniejszych.

W związku z dążeniem do pozyskiwania danych czujnikowych wyższego rzędu zachodzi potrzeba opracowywania i rozwoju metod analizy danych adekwatnych do ich złożoności. Duże nadzieje wiąże się z metodami analizy wielokierunkowej. Inspiracją do rozwoju tego typu metod analizy danych czujnikowych są metody chemometryczne, opracowywane na potrzeby instrumentów analitycznych wyższego rzędu, np. realizujących spektroskopię w podczerwieni czy spektroskopię Ramana. Należy jednak podkreślić, że metody poprawy selektywności sygnału pomiarowego względem badanych gazów są zupełnie inne w przypadku klasycznych metod analitycznych i w przypadku metod czujnikowych. Nie ma zatem możliwości posługiwania się metodami analizy wielokierunkowej w obu przypadkach w sposób identyczny.

Określanie gazów, a w szczególności zanieczyszczeń gazowych, na podstawie złożonych czujnikowych danych pomiarowych, można przedstawić jako problem rozpoznawania wzorców. Jak dowodzą wyniki licznych prac badawczych, zastosowanie strategii wypracowanej dla systemów rozpoznawania wzorców prowadzi do uzyskania zadowalających rozwiązań tego problemu. Należy jednak jeszcze raz podkreślić, że przedstawiona w rozdz. 2 metodologia rozpoznawania wzorców określa tylko ogólne zasady i ramy postępowania z danymi. W systemach rozpoznawania wzorców konstruowanych na potrzeby czujnikowych pomiarów gazów rozłożenie akcentów między poszczególnymi elementami systemu, dobór metod stosowanych w poszczególnych etapach jego budowy i struktura powiązań są dość charakterystyczne i wynikają ze specyficznego charakteru czujnikowych danych pomiarowych oraz funkcji tych systemów. Dlatego mówimy raczej o systemach analizy danych w pomiarach czujnikowych. Specyfika ta pogłębia się jeszcze, jeżeli ograniczy się zainteresowania do systemów opracowywanych do szczególnych zadań, np. pomiarów zanieczyszczeń powietrza.

5. Reprezentacja gazu w danych czujnikowych

Złożoność danych pomiarowych jest podstawową przyczyną, dla której pierwszym istotnym etapem ich analizy jest skonstruowanie możliwie najprostszej reprezentacji badanego gazu. Ważne jest przy tym, by proces ten przebiegał z zapewnieniem minimalnej straty informacji o gazie. Dobrze przeprowadzony proces budowy takiej reprezentacji jest w stanie doprowadzić do sytuacji, w której zadanie dalszych etapów analizy, polegające na odczycie informacji, będzie relatywnie mało skomplikowane [64].

5.1. Generowanie cech

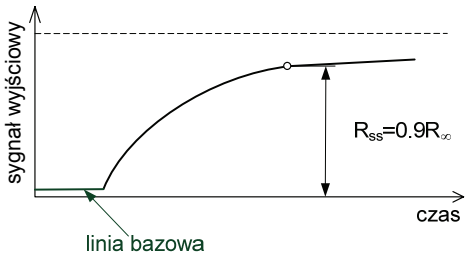
Cecha odgrywa zasadniczą rolę w konstruowaniu reprezentacji badanego gazu na podstawie czujnikowych danych pomiarowych. Jest to kluczowe pojęcie w dziedzinie rozpoznawania wzorców. Cecha jest zmienną. W analizie danych z pomiarów czujnikowych za cechę jawną uważa się parametr sygnału wyjściowego czujnika. Wyróżnia się też cechy niejawne (ang. *latent*), które powstają w wyniku przekształcenia zbioru cech jawnych w procesie określanym jako ekstrakcja cech. Opracowanie systemu analizy danych dla pomiarów czujnikowych wymaga najpierw wyłonienia tzw. cech kandydujących. Etap ten, określany z perspektywy systemu analizy wzorców jako generowanie cech, w dziedzinie pomiarów czujnikowych jest na ogół rozumiany jako etap wstępnego przetwarzania danych (ang. *data preprocessing*). Biorąc pod uwagę specyfikę etapu generowania cech w kontekście pomiarów czujnikowych, należy wyróżnić jego następujące elementy:

- korektę linii bazowej sygnału czujnikowego,
- wyłonienie parametrów sygnału czujnikowego.

5.1.1. Korekta linii bazowej sygnału czujnikowego

Zabiegi wykonywane na linii bazowej sygnału mają na celu przede wszystkim zredukowanie skutków dryfu czujników. Czujnik jest stabilny, jeżeli na pobudzenie

o określonej, stałej wartości odpowiada w taki sam sposób przez pewien czas (rys. 5.1). Terminem dryf (inna stosowana nazwa to dryft) określa się sytuację, w której zachodzi zmiana odpowiedzi czujnika w takich okolicznościach. Problem dotyczy zarówno zmiany położenia linii bazowej, jak i zmiany czułości czujnika [23].



Rys. 5.1. Odpowiedź czujnika zdefiniowana jako sygnał czujnika w stanie ustalonym

W przypadku czujników półprzewodnikowych, najpowszechniej stosowanych w systemach czujnikowych do pomiaru gazów, dryf najczęściej jest związany ze starzeniem czujnika. Starzenie jest wywołane zmianami mikrostruktury materiału, które następują z upływem czasu w wyniku pracy czujnika, jak również ze względu na to, że niektóre reakcje zachodzące na powierzchni materiału są nieodwracalne [65]. Dryf może mieć charakter krótkookresowy i wynikać ze zmiany zachowania materiału chemoczułego ze względu na cząstki zaadsorbowane na nim podczas kolejnych pomiarów. Efekt ten jest odwracalny, jeśli elementy zaadsorbowane podlegają desorpcji w krótkim czasie. Taką sytuację określa się często jako efekt pamięci (ang. *memory effect*). Pod względem rozpoznawania wzorców skutkiem dryfu jest zmiana położenia wzorców gazów w przestrzeni cech, a zatem zmiana położenia klas wzorców, jak również deformacja zależności ilościowych. Może to skutkować utratą adekwatności klasyfikatora lub modelu regresji [66]. Podstawowe techniki korekty linii bazowej przedstawiono w tabeli 5.1 [22, 23, 67].

W praktyce pomiarowej stosowane są również rozwiązania polegające na odnoszeniu sygnału pomiarowego do sygnału wzorcowego uzyskanego podczas ekspozycji na tzw. próbę zerową, tj. gaz o ściśle określonych właściwościach jakościowych i ilościowych [38, 68]. Metody korekty są podobne jak w przypadku linii bazowej, ale działania matematyczne wykonuje się na wartościach sygnałów mierzonego i referencyjnego uzyskanych w odpowiadającym sobie czasie ekspozycji.

Efekty zastosowania określonej techniki korekty linii bazowej w znacznym stopniu zależą od takich czynników, jak rodzaj czujników zastosowanych w systemie pomiarowym, rodzaj wielkości mierzonej, badane gazy oraz rodzaj informacji o gazach, którą system ma pozyskać. Dokonując wyboru techniki, warto kierować się własnym doświadczeniem w pracy z danym rodzajem czujników, dodatkowo uwzględniając kontekst rozwiązań zastosowanych w układzie pomiarowym. Dla czujników półprze-

wodnikowych podaje się na przykład teoretyczne uzasadnienie, z którego wynika, że najlepsze efekty powinna przynieść ilorazowa korekta linii bazowej [3]. Uzyskano jednak dobre rezultaty rozpoznawania gazów, stosując również korektę różnicową [69]. Gdy wielkością mierzoną dla tego typu czujników była konduktancja, wskazywano przewagę podejścia frakcyjnego [70] lub w ogóle braku korekty [71]. Z kolei dla napięcia odkładanego na czujniku referencyjnym jako sygnału czujnikowego [72] najlepsze rezultaty uzyskano nie wykonując żadnej korekty lub stosując podejście różnicowe.

Tabela 5.1. Podstawowe techniki korekty linii bazowej[a]

Metoda	Równanie	Uwagi
Różnicowa	$r_b(t) = r(t) - r_0$	redukcja dryfu o charakterze addytywnym
Ilorazowa	$r_b(t) = \frac{r(t)}{r_0}$	redukcja dryfu o charakterze multiplikatywnym
Frakcyjna	$r_b(t) = \frac{r(t) - r_0}{r_0}$	redukcja dryfu o charakterze addytywnym i multiplikatywnym
Logarytmiczna	$r_b(t) = \ln\left(\frac{r(t)}{r_0}\right)$	redukcja korzystna w wypadku dużego zakresu stężeń [67]

[a]Oznaczenia: $r_b(t)$ – wartość sygnału w chwili t po korekcie linii bazowej; $r(t)$ – wartość sygnału w chwili t przed korektą linii bazowej; r_0 – wartość sygnału odpowiadająca linii bazowej.

Korekta linii bazowej jest najprostszą metodą korekty dryfu. W literaturze przedmiotu dyskutuje się wiele bardziej złożonych rozwiązań, zintegrowanych z dalszymi stadiami analizy danych w ramach systemu rozpoznawania wzorców [73–76]. Mimo rozwoju tych technik obliczeniowych do korekty dryfu zaleca się rekalibrację systemu czujnikowego [68, 77].

5.1.2. Parametry sygnału czujnikowego

Cechy jawne są parametrami sygnału czujnikowego. W ujęciu konwencjonalnym cechą jest odpowiedź czujnika. Za odpowiedź uważa się najczęściej sygnał wyjściowy czujnika w tzw. stanie ustalonym (rys. 5.1) [78]. W stanie ustalonym czujnik znajduje się w równowadze termodynamicznej z badanym gazem. Za odpowiedź czujnika w stanie ustalonym przyjęto uważać wartość sygnału równą 90% maksymalnej wartości sygnału teoretycznie dostępnej po nieskończonym czasie ekspozycji [37].

W praktyce definicja ta nie jest ściśle przestrzegana i za odpowiedź czujnika przyjmuje się na ogół wartość sygnału zarejestrowaną po określonym czasie ekspozy-

cji, uznanym za wystarczający do osiągnięcia stanu ustalonego. W rzeczywistości jest to zwykle stan kwaziustalony. Tak zdefiniowana odpowiedź jest najpowszechniej stosowanym parametrem sygnału czujnikowego. Na podstawie mechanizmu działania czujnika można przyjąć, że parametr ten jest w sposób jednoznaczny związany ze stężeniem badanego gazu. Wyniki badań pokazują, że odpowiedź czujnika w stanie ustalonym przenosi również informację o rodzaju gazu. Prostota pomiaru, którą zyskuje się, wykorzystując omawiany parametr jako podstawę oznaczeń gazów, jest jednak osiągana za cenę ignorowania zawartości informacyjnej innych fragmentów sygnału.

Przesłanki teoretyczne oraz rezultaty eksperymentów wskazują, że również sygnał czujnika w stanach nieustalonych jest interesujący pod względem zawartych w nim informacji o badanym gazie [3, 60, 79]. Parametryzacja sygnału czujnika w stanach nieustalonych jest bardzo obiecującym kierunkiem poszukiwań sposobów rozbudowy przestrzeni cech. Ma on istotny związek z pracami nad możliwością wykonywania pomiarów gazów z wykorzystaniem pojedynczego czujnika [80–82].

Stany nieustalone czujnika są na tyle interesujące, że prowadzi się intensywne badania nad różnorodnymi metodami ich wywoływania i kształtowania. Do najbardziej znanych należy modulacja termiczna [83–89]. Jest to przykład metody opartej na zmianie parametrów wewnętrznych. Polega ona na ingerencji w temperaturę pracy czujnika podczas ekspozycji na badany gaz. Bardzo obiecujący kierunek badań koncentruje się na wywoływaniu stanów nieustalonych przez zmianę parametrów zewnętrznych. Przykładem są tu prace poświęcone zastosowaniu kontrolowanej zmiany stężenia badanych gazów [90–92], modulacji przepływu gazów [71] czy też wywołanie szeroko rozumianej zmienności warunków ekspozycji czujnika [45].

Uważa się, że za kształt sygnału w stanach nieustalonych wywołanych nagłą zmianą stężenia badanej substancji odpowiadają dwa różne mechanizmy. Pierwszy stanowi dyfuzja gazu w warstwie chemoczułej. Drugim jest kinetyka reakcji przebiegających na powierzchni i w objętości materiału chemoczułego. Dla danej pary gaz –czujnik charakter tych interakcji jest typowy, informacja o nich jest zatem potencjalnie użyteczna do celów analitycznych. Na podstawie teoretycznych modeli odpowiedzi opracowanych dla czujników półprzewodnikowych i polimerowych stwierdzono, że czas odpowiedzi jest kontrolowany przez procesy dyfuzji, nie zaś przez reakcje chemiczne [93]. Jest to teoretyczne uzasadnienie możliwości wpływania na odpowiedzi czujników przez ingerencję w warunki ekspozycji [45].

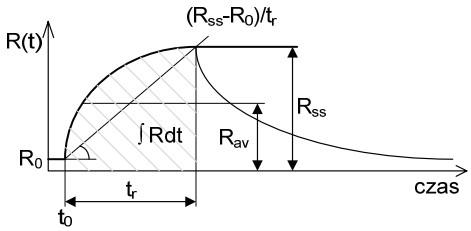
Według niektórych autorów odpowiedź czujnika konstruowana na podstawie sygnału w stanie nieustalonym jest bardziej odporna na zakłócenia pomiaru i dryf [91]. Stanowi ona bardziej stabilne źródło informacji o badanych gazach. Ze względów praktycznych posługiwanie się odpowiedzią czujnika w stanach innych niż ustalony, poprzedzających jego osiągnięcie, umożliwia skrócenie czasu akwizycji danych pomiarowych oraz zmniejszenie ich liczby. Wskutek krótszej ekspozycji na badany gaz również proces regeneracji czujnika przebiega sprawniej. W wyniku tego linia bazowa

czujnika jest odtwarzana szybciej, a cała procedura pomiarowa może zostać skrócona. W niektórych rozwiązaniach pomiarowych, np. podczas pomiaru ciągłego z dużą rozdzielczością czasową, tj. odstępem między kolejnymi pomiarami krótszym niż czas osiągnięcia stanu ustalonego przez czujnik, wykorzystanie stanów nieustalonych czujnika w celach pomiarowych jest nieuniknione. Nasuwa się tu kontekst zastosowań czujników w monitoringu zanieczyszczenia środowiska w czasie rzeczywistym [15]. Parametry sygnału związane ze stanami nieustalonymi czujnika określa się często jako parametry dynamiczne.

Wyróżnia się trzy grupy metod obliczeniowych, za pomocą których wyłaniane są parametry sygnałów czujników. Są to: i) wyznaczenie parametrów *ad-hoc*, ii) modelowanie matematyczne sygnału wyjściowego czujnika oraz iii) próbkowanie (ang. *subsampling*).

Parametry *ad hoc*

Parametry typu *ad hoc* opisują w sposób syntetyczny różne aspekty zmienności sygnału czujnika [94]. Podstawowe założenie dla ich konstrukcji jest takie, że informacja o badanych gazach jest przenoszona w różnym zakresie przez różne rodzaje zmienności sygnału. Ponieważ te rodzaje zmienności można na ogół związać z fragmentami sygnału, parametry typu *ad hoc* z reguły odnoszą się do fragmentów sygnału.



Rys. 5.2. Najczęściej stosowane parametry sygnału czujnika typu *ad-hoc*

Do najczęściej stosowanych parametrów sygnału czujnika typu *ad hoc* należą:

- wartość sygnału w określonym czasie, np. w stanie ustalonym (R_{ss} na rys. 5.2) [63, 95] lub w momentach specyficznych dla danej procedury pomiarowej (R_{av} na rys. 5.2) [95, 96],
- maksymalna lub minimalna wartość sygnału w określonym przedziale czasu (R_t na rys. 5.2) [97, 98],
- średnia z wartości sygnału w wybranym przedziale czasu (R_{av} dla okresu t_r na rys. 5.2) [97],

- całka z sygnału czujnika w wybranym przedziale czasu, np. dla okresu narostu sygnału ($\int Rdt$ dla okresu t_r na rys. 5.2) lub dla okresu zaniku sygnału [99];
- iloraz różnicowy sygnału czujnika w wybranym przedziale czasu [95], np. dla okresu narostu sygnału ($(R_{ss} - R_0)/t_r$ dla okresu t_r na rys. 5.2) dla okresu zaniku sygnału, pochodna chwilowa sygnału [15, 83],
- różnica pochodnej sygnału w czasie ekspozycji czujnika na badany gaz i gaz odniesienia [100];
- stała czasowa odpowiedzi sensora, czas do osiągnięcia maksimum/minimum sygnału [97, 98], czas do osiągnięcia zadanej wartości sygnału (t_r na rys. 5.2).

Tak rozumiane cechy określa się niekiedy jako cechy geometryczne [101].

Wspólną cechą parametrów należących do grupy *ad hoc* jest stosunkowa łatwość i szybkość ich wyznaczenia bez potrzeby stosowania złożonego aparatu matematycznego. Podstawowa wada parametrów *ad hoc* dotyczy fragmentarycznego traktowania sygnału czujnika powstałego w wyniku ekspozycji na badany gaz.

Model matematyczny sygnału

Metody wyłaniania parametrów na podstawie modelu matematycznego sygnału czujnika traktują ten sygnał całościowo. Zakłada się, że informacja o badanym gazie jest wówczas najlepiej widoczna. Wyłonienie cech na podstawie parametrów pochodzących z modelu matematycznego umożliwia najbardziej syntetyczną reprezentację informacji zawartej w całym sygnale wyjściowym czujnika. Zależnie od stosowanych metod modelowania matematycznego wyznacza się parametry:

- funkcji analitycznych aproksymujących sygnał w dziedzinie czasu, np. funkcji linowych, wielomianów, funkcji potęgowych, wykładniczych [102, 103], logarytmicznych, Lorentza [103], podwójnie sigmoidalnych [103], jak również parametry złożeń tych funkcji;
- modeli statystycznych zastosowanych do opisu sygnału czujnika w czasie [104], np. parametry funkcji autoregresji, maksimum sygnału po zastosowaniu wygładzania wykładniczego [92], parametry modelu autoregresji i średniej ruchomej (ARMA) [105];
- transformat uzyskanych w wyniku poddania sygnału odpowiedzi czujnika różnego rodzaju transformacjom, np. transformaty Fouriera [88, 106], transformaty falkowej [67, 71, 84, 88, 106];
- parametry uzyskane w wyniku przedstawienia sygnału odpowiedzi czujnika w przestrzeni stanów [107] konstruowane często na wzór parametrów sygnału czujnika w dziedzinie czasu, np. maksimum, całka, różniczka sygnału w przestrzeni stanów [99, 108, 109] oraz momenty dynamiczne [110].

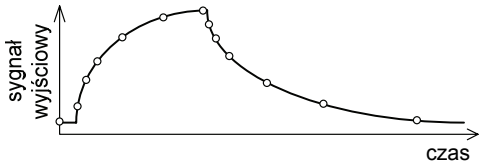
Choć atrakcyjne koncepcyjnie, pozyskiwanie parametrów na podstawie modelu matematycznego sygnału wyjściowego czujnika nastręcza wiele trudności oblicze-

niowych, zwłaszcza w wypadku stosowania funkcji analitycznych. Niejednokrotnie jest to jednak konieczne, np. dla sygnałów o charakterze cyklicznym, będących wynikiem zastosowania modulacji sygnału w pomiarach czujnikowych.

Próbkowanie

W ramach próbkowania za parametry sygnału przyjmuje się dyskretne wartości sygnału wyjściowego czujnika w wybranych chwilach s, $s \in \{1, \ldots, N_s\}$ [3]. Maksymalna liczba tych wartości wynosi $N_s = m$ i jest określona przez częstość próbkowania sygnału podczas jego rejestracji. Próbkowanie służące uzyskaniu parametrów sygnału jest jednak czymś innym niż próbkowanie na potrzeby jego rejestracji. W pierwszym przypadku próbkowany jest sygnał dyskretny, w drugim – sygnał ciągły. W wyniku procedury powstaje od razu cały wektor cech $\mathbf{r}_s$, który stanowi formę skompresowanej informacji zawartej w całym sygnale.

Sygnał czujnika może się charakteryzować zróżnicowaną dynamiką. W dużej mierze zależy ona od sposobu wywołania stanów nieustalonych czujnika. Jeżeli założy się, że zachowanie dynamiki oryginalnego sygnału w sygnale po spróbkowaniu jest kluczowe dla uniknięcia istotnych strat informacji podczas kompresji, zachodzi potrzeba dostosowania częstości próbkowania (rys. 5.3).



Rys. 5.3. Próbkowanie sygnału czujnika. Fragmenty sygnału o większej dynamice zostały spróbkowane gęściej. Zabieg pokazano również w odniesieniu do fragmentu sygnału uzyskanego podczas regeneracji czujnika (sygnał opadający)

Powszechnie przyjmuje się, że w przeciwieństwie do szybkozmiennych fragmentów sygnału stany kwaziustalone wymagają najmniejszej częstości próbkowania. W pracy [111] pokazano jednak przewagę próbkowania równomiernego nad próbkowaniem dostosowanym do dynamiki sygnału czujnikowego. Jeżeli za parametry sygnału przyjąć wyniki pomiarów dyskretnych w czasie, to reprezentacja dynamiki sygnału w uzyskanym zbiorze cech jest ukryta we wzajemnym skorelowaniu parametrów.

Rozwiązaniem pokrewnym dla próbkowania jest tzw. *windowing*. Polega on na przesuwaniu „okna” po osi czasu i pobieraniu wartości lub wybranego parametru geometrycznego sygnału z przedziału czasu ograniczonego oknem [87, 81]. W wyniku

zastosowania tej procedury powstaje cały wektor cech podobnie jak podczas próbkowania.

Podział parametrów sygnałów czujników według metody ich pozyskania nie jest jedynym możliwym. Istnieje też rzadziej spotykany podział ze względu na dziedzinę, w jakiej są określone parametry sygnału czujnikowego. Wyróżnia on parametry określone w dziedzinie czasu, częstotliwości i w dziedzinie stanów [112]. Kryteria tego podziału wydają się dość oczywiste, a przypisanie paramentów sygnału do poszczególnych kategorii we wspomnianym systemie nie powinno nastręczać trudności.

Brakuje jednoznacznych reguł wskazujących na ewidentną przewagę jednego typu parametrów sygnału nad innymi. Dużą rolę w ich wyborze odgrywa intuicja i doświadczenie. Opublikowano wiele analiz porównawczych różnych typów cech. Wnioski z nich płynące mogą się niekiedy wydawać sprzeczne. Należy jednak pamiętać, że sformułowano je dla bardzo konkretnych przypadków czujników, rozwiązań układu pomiarowego, procedur pomiarowych, wreszcie badanych gazów. Ich proste uogólnianie nie jest uzasadnione. Dla przykładu w [99] wykazano przewagę cech uzyskanych w wyniku przedstawienia sygnału czujnika w przestrzeni stanów nad uzyskanymi w dziedzinie czasu. Rezultaty przedstawione w pracy [88] wskazywały na większą zawartość informacyjną cech geometrycznych niż parametrów transformat sygnału – Fouriera i falkowej, w pracy [67] stwierdzono przewagę transformaty falkowej nad siedmioma innymi rodzajami cech.

Nie bez znaczenia dla wyboru rodzaju cech jest aparat matematyczny, którym dysponuje badacz, oraz zadania systemu analizy danych. W wielu wypadkach konieczne jest uciekanie się do rozwiązań najprostszych, co jednak niekoniecznie wiąże się ze zmniejszeniem możliwości pomiarowych systemu czujnikowego.

5.2. Redukcja wymiaru przestrzeni cech

Przyjmuje się, że cechy zarówno jawne (parametry sygnału czujnika), jak i niejawne (matematyczne przekształcenia cech jawnych) są nośnikami informacji o badanych gazach [113]. Poszczególne cechy różnią się jednak pod względem przydatności do pozyskania informacji. W przypadku czujników częściowo selektywnych nie można mówić o selektywności parametrów sygnałów wyjściowych czujników względem badanych gazów. Wprowadza się natomiast pojęcie powiązania (ang. *relevance*), które dotyczy wyrazistości odzwierciedlenia właściwości gazów za pomocą cech. Wśród parametrów wyłonionych z sygnałów wyjściowych czujników mogą się znaleźć cechy niepowiązane (ang. *irrelevant*), które nie zawierają żadnej przydatnej informacji. Z kolei grupa cech powiązanych również nie jest homogeniczna. Zazwyczaj obserwuje się w niej tzw. nadmiarowość (ang. *redundancy*) [22]. Polega ona przede wszystkim

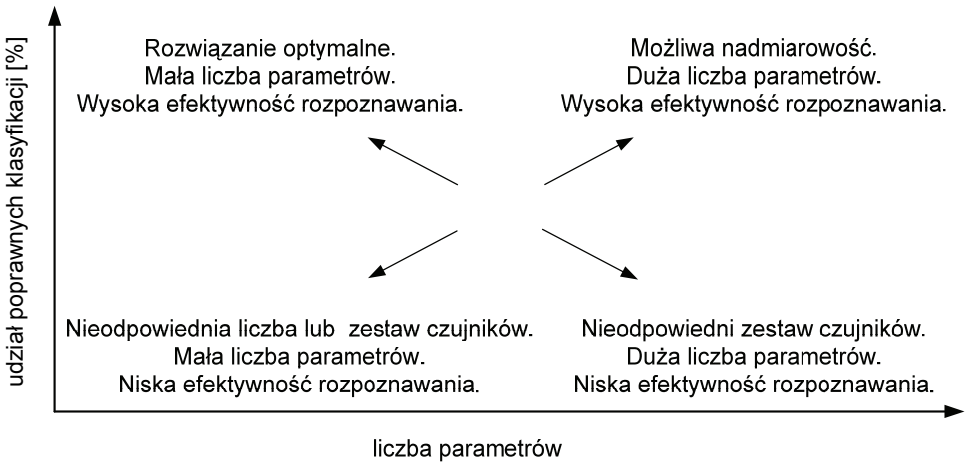
na dublowaniu informacji przez różne cechy. Najczęściej jest to widoczne jako wzajemne skorelowanie cech. Ponadto stopień przydatności poszczególnych cech do określania gazów jest przeważnie niejednakowy. Problem nadmiarowości w zbiorze cech wygenerowanych analizuje się w kategoriach proporcji między liczbą dostępnych nośników informacji a liczbą nośników niezbędnych [114, 115].

Liczba potencjalnych cech, które można wyłonić z sygnału czujnika jest bardzo duża. Jeśli posługujemy się matrycą czujników, to zwiększa się ona proporcjonalnie do liczby czujników w matrycy. Biorąc pod uwagę zróżnicowaną przydatność cech, wyselekcjonowanie cech przenoszących pożądaną informację i wyeliminowanie cech niepowiązanych oraz nadmiarowych jest warunkiem koniecznym efektywnego określania gazów.

Istnieją też inne powody, dla których nadmiarowość jest zjawiskiem niekorzystnym. Najistotniejszy z nich jest związany z tzw. przekleństwem wymiarowości (ang. *dimensionality curse*). Zagadnienie to jest dobrze znane w dziedzinie analizy danych. Problem został sformułowany w 1961 r. przez Bellmana i dotyczy wykładniczego wzrostu objętości, związanego z dodaniem kolejnych wymiarów w przestrzeni matematycznej [37, 114]. Wiąże się z nim szereg utrudnień w procesie dopasowywania modeli, szacowania ich parametrów czy też optymalizacji funkcji w przestrzeniach wielowymiarowych. Wraz ze wzrostem wymiaru przestrzeni cech trudność w znalezieniu optimum globalnego w przestrzeni parametrów wzrasta w sposób wykładniczy. Problem klasyfikacji wzorców (np. wzorców gazów) w przestrzeni cech wiąże się z koniecznością odniesienia się przez klasyfikator do każdej części tej przestrzeni, tak aby było wiadomo, która klasa wzorców ją zajmuje. Aby klasyfikator nie przeznaczał swoich zasobów do mapowania nieistotnych dla problemu klasyfikacji fragmentów przestrzeni cech, potrzebna jest odpowiednia liczba danych. Powinny one poprawnie wskazywać, które fragmenty tej przestrzeni są mniej, a które bardziej istotne. Im większa wymiarowość przestrzeni cech, tym więcej potrzeba takich danych. Wykazano, że liczba wyników pomiarów potrzebnych do sparametryzowania modelu rozpoznawania wzorców w systemach czujnikowych zwiększa się wykładniczo z liczbą cech uwzględnionych w modelu [116].

Dostarczenie dużej liczby pomiarów kalibracyjnych może być niewykonalne ze względu na potrzebny czas, nakład pracy oraz ponoszone koszty. Z drugiej strony pokazano, że gdy korzysta się z określonej liczby wyników pomiaru, istnieje maksymalna liczba cech, której dalsze zwiększanie powoduje pogorszenie właściwości modelu obliczeniowego. Wynika to na ogół ze zbyt dobrego dopasowania modelu do danych uczących, co powoduje utratę zdolności uogólniania. Posługiwanie się dużą liczbą cech wydłuża ponadto czas obliczeń i zwiększa wymagania co do mocy obliczeniowej oraz zasobów pamięci. Wybór najlepszego zbioru cech stwarza możliwość opracowania szybko działających i obliczeniowo efektywnych modeli do określania gazów na podstawie pomiarów czujnikowych z lepszą dokładnością niż bez tej operacji (rys. 5.4) [117]. Pardo i Sberveglieri [99] pokazali, że najlepszy podzbiór

30-elementowego zbioru cech prowadził zawsze do lepszych rezultatów w określaniu gazów niż cały zbiór.



Rys. 5.4. Powiązanie konfiguracji matrycy czujnikowej,
liczby składowych wektora cech i efektywności rozpoznawania [118]

Z przesłanek teoretycznych i z praktyki analizy danych wynika, że bardzo pożądane jest ograniczenie liczby cech do niezbędnego minimum [119] przez wybór najlepszego zbioru cech. Wszakże przestrzeń cech wyłoniona ostatecznie w wyniku redukcji wymiarowości musi zapewniać jak najlepszą efektywność rozwiązania problemu określania gazów. Stąd wśród kryteriów oceny metod kompresji warto brać pod uwagę liczbę wymiarów przestrzeni cech, efektywność jakościowego/ilościowego określania badanych gazów w tej przestrzeni oraz czas, który jest potrzebny na jej znalezienie [98].

Podstawowe strategie redukcji wymiarowości przestrzeni cech to selekcja oraz ekstrakcja cech, nazywana również mapowaniem.

5.2.1. Selekcja cech

Problem selekcji cech można zdefiniować jako wybór zbioru cech $\widehat{F}\left(\widehat{F} \subseteq X\right)$, będącego podzbiorem zbioru cech wygenerowanych X w celu osiągnięcia minimalizacji kryterium wyboru $J(F)$, zgodnie ze wzorem [23]:

$$\widehat{F} = \arg\min_{F \subseteq X} J\left(F\right) \tag{5.1}$$

Kryterium selekcji jest najczęściej błąd odczytu informacji na podstawie zestawu cech, np. błąd klasyfikacji.

Operacja selekcji cech polega na przeszukaniu pełnego zbioru cech i ocenie poszczególnych ich kombinacji (włączając cechy jednoelementowe) w świetle zadanego kryterium [115]. Atrakcyjność selekcji jako metody redukcji początkowego zbioru cech jest w dużej mierze związana z faktem, że cechy wyselekcjonowane zachowują interpretację fizykochemiczną cech początkowych. W pomiarach czujnikowych jest to bardzo istotna właściwość. Rozumienie natury cech przyczynia się do do rozumienia procesów, które odpowiadają za przeniesienie informacji o badanych gazach do danych pomiarowych.

W dziedzinie analizy danych z czujnikowych pomiarów gazów można spotkać przykłady rozwiązań problemu selekcji cech wykorzystujących kombinacje różnych strategii selekcji i podejść do oceny podzbiorów cech. Należy je rozumieć jako sposoby poszukiwania najistotniejszych źródeł informacji o badanych gazach. Niejednokrotnie skład optymalnego zbioru cech jest dodatkowo wskazówką, którą można wykorzystać do znalezienia najlepszego zestawu czujników w matrycy czy optymalnych parametrów ich pracy [29, 114, 118, 120–124].

Przegląd przykładów zastosowania selekcji cech w analizie danych z czujnikowych pomiarów gazów pozwala zauważyć, że jest ona stosowana głównie w kontekście problemów jakościowych. Przypadki włączania selekcji cech w poszukiwanie wzorca optymalnego w aspekcie ilościowego określania gazów nie są natomiast liczne [17, 92, 117, 125].

Metody oceny kombinacji cech

W literaturze dotyczącej selekcji cech wymienia się trzy główne podejścia do oceny podzbioru cech, określane jako [22, 114]:

- opakowane (ang. *wrapper approach*);
- filtr (ang. *filter approach*);
- wbudowane (ang. *embedded approach*).

Podejście typu opakowane jest najchętniej stosowanym sposobem oceny kombinacji cech w czujnikowych pomiarach gazów. Charakteryzuje się ono tym, że podzbiór cech jest oceniany w kontekście konkretnej metody obliczeniowej, np. danego rodzaju klasyfikatora lub modelu regresji. Poszczególne przykłady podejścia typu wrapper approach różnią się głównie pod względem metody rozwiązywania problemu klasyfikacji lub regresji. Ze względu na wymagany nakład obliczeniowy preferowane są modele proste, tzn. takie, których uczenie, walidacja i testowanie trwa relatywnie krótko z zachowaniem adekwatności do skali trudności rozwiązywanego problemu. Stosowane są tu np. analiza skupień [123], modele neuronowe jak perceptrony wielowarstwowe [96, 119] czy sieci probabilistyczne [96, 126], maszyny wektorów podpie-

rających [113], modele rozmyte [98], analiza dyskryminacyjna [17, 125, 127] oraz analiza regresji [17, 127]. Kryterium oceny kombinacji cech zależy od rodzaju rozwiązywanego problemu. W przypadku zagadnień jakościowych, np. określania rodzaju gazu lub innych jego właściwości jakościowych, jest to zwykle różnie definiowana efektywność klasyfikacji [96, 98, 113, 119, 126]. Z kolei do problemów ilościowych, np. do określania stężenia gazu, najczęściej stosowany jest błąd wyznaczenia wartości zmiennej wyjściowej [17, 127]. Im większa efektywność rozwiązania problemu, tym wyższa ocena podzbioru cech. Kombinacja cech o najwyższej ocenie jest uznawana za najlepszą. Podejście typu opakowane jest relatywnie wymagające obliczeniowo i podatne na zbytnie dopasowanie modelu obliczeniowego do zbioru danych (ang. *overfitting*).

Popularność podejścia opakowanego w czujnikowych pomiarach zanieczyszczeń uzasadniają względy praktyczne. Zestaw cech uznany za najlepszy staje się podstawą określania gazów w systemie rozpoznawania wzorców. Fakt, że model zaimplementowany w tym systemie został wcześniej wykorzystany do wyselekcjonowania wzorca, pozwala się spodziewać bardzo dobrych rezultatów określania gazów. Jest to zasadniczy cel do osiągnięcia przez system analizy danych.

Wyróżnikiem podejścia typu filtr jest ocenianie podzbioru cech bez kontekstu metody obliczeniowej. Skupia się ono na określeniu możliwości różnych podzbiorów cech niezależnie od klasyfikatora, tj. w kontekście tak zwanego klasyfikatora uniwersalnego. W rezultacie znajdowane cechy powinny dobrze współpracować z różnymi klasyfikatorami. Nie daje to jednak gwarancji uzyskania najlepszego wyniku. Podejście typu filtr ma więc mniejszy walor praktyczny niż podejście typu opakowane. Jest to podstawowy powód, dla którego jest rzadziej spotykane w publikacjach dotyczących pomiarów czujnikowych. Wśród kryteriów oceny należy wymienić:

- miary zdolności dyskryminacyjnych cech, np. odległości między klasami wzorców w przestrzeni cech (np. odległość euklidesowa, odległość Mahalanobisa) [97], stopień zróżnicowania wartości cechy w obrębie klas na tle zróżnicowania między klasami (np. stosunek międzygrupowej i wewnątrzgrupowej wariancji cechy, statystyka *t* [87], zdolność rozdzielcza cechy (ang. *resolution factor*) [128]), współczynnik zmienności [97];
- miary spójności cechy, np. powtarzalność wartości cechy (w odniesieniu do wyników pomiaru), czy nadmiarowości (stopień skorelowania z innymi cechami) [97];
- czułość i selektywność cechy [122];
- miary zawartości informacji w cechach, rozumiane jako redukcja niepewności określenia gazu dzięki uwzględnieniu cechy w podzbiorze (np. miary entropii) [115];
- miary współzależności, np. współczynnik korelacji Pearsona między cechą a miarą ilościowo opisującą badane gazy [115, 129] lub między różnymi cechami [63].

Filtrowanie jest najczęściej stosowane w odniesieniu do pojedynczych cech (jednoelementowych podzbiorów cech), prowadząc do powstania ich rankingu. Filtracja wielowymiarowa nie jest spotykana często, choć rzadziej prowadzi do uzyskania

kombinacji cech wywołujących zbytnie dopasowanie modeli do danych niż podejście typu opakowane. Wiadomo jednak, że kombinacja cech optymalna pod względem kryteriów filtra jest często suboptymalna w warunkach współpracy z konkretnym modelem obliczeniowym. Ponadto przydatność cech jest zazwyczaj różna, gdy są rozważane pojedynczo i gdy występują w kontekście innych cech [115]. Ze względu na ten fakt, filtracja z reguły nie jest stosowana jako ostateczna instancja, wyrokująca o jakości podzbiorów cech na potrzeby określania gazów na podstawie pomiarów czujnikowych. Ze względu na dużą efektywność obliczeniową podejście to jest natomiast często używane do wykonania preselekcji cech, poprzedzającej dalszą redukcję ich liczby już z wykorzystaniem podejścia typu opakowane [130]. W czystej formie podejście typu filtr ma ogromne znaczenie poznawcze ze względu na możliwość porównania różnych typów cech pod względem np. zawartości informacji o badanych gazach [115]. W pracy [87] wykazano na podstawie statystyki *t*, że cechy, które dobrze różnicowały grupy wzorców były zlokalizowane w określonych fragmentach sygnału czujników. Filtracja jest sporadycznie stosowana jako metoda wielowymiarowej selekcji cech. Pionierską pracę na ten temat przedstawił Eklov [119]. Inny przykład filtracji wielowymiarowej można znaleźć w [118], gdzie wykazano możliwość zredukowania tym sposobem liczby cech o 50–70% z jednoczesną poprawą wyniku określania gazów. Idące kolejny krok dalej rozwiązanie, polegające na zastosowaniu filtracji do wielokryterialnej optymalizacji matrycy czujników, przedstawiono w [128].

Interesującą propozycją jest połączenie podejścia typu *wrapper* i typu *filter*. Kombinację taką określa się ogólnie terminem *fraper*. W praktyce czujnikowych pomiarów zanieczyszczeń gazowych szczególnie przydatny jest pewien rodzaj frapera dwuczłonowego. Jego pierwszy człon realizuje podejście typu *wrapper* z zastosowaniem modelu linowego. Działa ono jak filtr dokonujący preselekcji podzbiorów cech, które w drugim członie są ponownie przeglądane, również w ramach podejścia opakowanego, lecz w sprzężeniu z modelem nieliniowym [114, 131]. Choć problemy oznaczania gazów są z reguły nieliniowe, za pomocą metod liniowych można szybko uzyskać zgrubną ocenę możliwości różnych parametrów sygnału czujnika w tym zakresie. Oceny szczegółowej dokonuje się za pomocą bardziej adekwatnych, lecz zarazem złożonych modeli. Etap ten jest jednak mniej uciążliwy ze względu na ograniczenie zbioru cech przez preselekcję.

Podejście typu wbudowane polega na zintegrowaniu sposobu przeglądania podzbiorów cech i ich oceny z procesem budowy modelu matematycznego służącego do rozwiązania określonego problemu. Jest ono wydajne obliczeniowo, a wybór cech jest ściśle związany ze strukturą modelu matematycznego. Podejście wbudowane realizują pewne szczególne rodzaje klasyfikatorów, np. drzewa klasyfikacji i/lub regresji (CART, C4.5) [114]. Choć bardzo popularne w dziedzinie rozpoznawania wzorców, podejście wbudowane jest relatywnie rzadko spotykane w publikacjach dotyczących selekcji cech na potrzeby określania gazów na podstawie pomiarów czujnikowych.

Pojedyncze prace przedstawiają przykłady zastosowania algorytmów CART [133, 134] oraz lasów losowych [132]. Można się jednak spodziewać zmiany tej sytuacji, gdyż omawiana metoda oferuje wgląd w znaczenie poszczególnych cech dla oznaczania gazów. Jest to bardzo przydatna właściwość. Ze względu na nią lasy losowe mają szansę zostać jednym z podstawowych narzędzi interpretacji danych z pomiarów czujnikowych [132].

Algorytmy przeszukiwania zbioru cech

Zbiór cech może zostać przeszukany w sposób zupełny bądź niezupełny. Przegląd zupełny obejmuje wszystkie podzbiory cech i pozwala porównać ich możliwości. Sposób ten daje gwarancję znalezienia podzbioru cech optymalnego globalnie. Jest również bardzo cenny ze względów poznawczych, gdyż pozwala uzyskać kompletny obraz możliwości danego zbioru cech. Istnieją przykłady zupełnego przeszukiwania zbioru kombinacji cech ocenianych w podejściu opakowanym na potrzeby analizy danych z czujnikowych pomiarów gazów [17, 99, 135]. W tych wypadkach zazwyczaj celem jest nie tylko znalezienie kombinacji cech, która pozwala najlepiej określić badany gaz, lecz przede wszystkim porównanie możliwości różnych kombinacji cech.

Niestety, najczęściej wymiar przestrzeni cech jest zbyt duży, by pozwolić na jej zupełne przeszukanie w akceptowalnym czasie za pomocą realnie dostępnych zasobów obliczeniowych [115]. W przypadku N cech istnieje 2^N kombinacji cech o liczbie elementów od 1 do N. Obecnie uważa się, że $N = 20$ stanowi górną granicę zasadności przeszukiwania zupełnego. W celu ominięcia problemu wykładniczej eksplozji liczby kombinacji cech wymagających przejrzenia opracowano szereg metod umożliwiających przegląd przestrzeni cech w sposób bardziej wydajny.

Przeszukiwanie niezupełne zajmuje się fragmentami pełnej przestrzeni cech. Jest znacznie szybsze, lecz nie gwarantuje uzyskania rozwiązania optymalnego w sensie globalnym. Jego zastosowanie wymaga zaakceptowania kompromisu między jakością rozwiązania a kosztem jego uzyskania.

Wyróżnia się dwie główne grupy strategii niezupełnego przeglądania przestrzeni cech [98, 119]:

- strategie deterministyczne,
- strategie stochastyczne.

Optymalną w sensie globalnym, deterministyczną strategią przeszukiwania niezupełnego jest algorytm *branch and bound* [23]. Algorytm unika przeszukiwania zupełnego przez odrzucanie suboptymalnych podzbiorów cech z pominięciem bezpośredniego ich oceniania i gwarantuje, że wybrany podzbiór jest najlepszy globalnie w sensie kryterium, które spełnia warunek monotoniczności (np. miary odległości, funkcje dyskryminacyjne). W praktyce jednak występują trudności z zapewnieniem

monotoniczności kryterium w przypadku dużego udziału składnika losowego w danych. Rozwiązanie ma postać drzewa, którego węzły są oznaczane numerem cechy odrzuconej w danym węźle. Przykłady zastosowania tego algorytmu w analizie danych z czujnikowych pomiarów gazów można znaleźć w [65, 136].

Z powodzeniem stosowane jest również prostsze i mniej wymagające obliczeniowo podejście sekwencyjne. Taki charakter mają podstawowe strategie deterministyczne o charakterze suboptymalnym. W szczególności należą tu [22, 114]:

- Selekcja postępująca (ang. *forward selection*) – procedura zaczyna od podzbioru cech, który może być pusty. W każdym kroku procedury do podzbioru jest dołączana jedna cecha – ta, która najbardziej poprawia jego ocenę. Procedura kończy się, gdy włączenie cechy spoza podzbioru nie poprawia jego oceny [98, 119, 126].
- Selekcja wsteczna (ang. *backward selection*) – jest odwróceniem selekcji postępującej. Procedura zaczyna od całego zbioru cech. W każdym kroku ze zbioru jest eliminowana jedna cecha – ta, która najbardziej pogarsza jego ocenę. Procedura kończy się, gdy eliminacja jakiejś cechy z pozostałych w podzbiorze nie poprawia jego oceny [98, 119, 126].
- Selekcja krokowa (ang. *stepwise selection*) – polega na naprzemiennym stosowaniu selekcji postępującej i wstecznej w kolejnych krokach algorytmu. Z dużym prawdopodobieństwem algorytm prowadzi do uzyskania optymalnego lub suboptymalnego zbioru cech, zużytkowując na ten cel mniej czasu niż algorytm *branch and bound* [98].
- Selekcja plus-*l*-minus-*r* – liczbę cech dołączanych do podzbioru w danym kroku oznacza *l*, a liczbę cech wyłączanych z podzbioru oznacza *r*. Jeśli $r > l$, to mamy do czynienia z wersją selekcji postępującej, a jeśli $r < l$, to otrzymujemy wersję selekcji wstecznej.
- Selekcja płynna (ang. *floating selection*) – liczba cech dołączanych jak też wyłączanych ze zbioru cech zmienia się w każdym kroku.
- Przeszukiwanie tabu (ang. *taboo-search*) – algorytm pozwalający uzyskać rozwiązanie optymalne lub niewiele różniące się od niego. W sekwencji ruchów istnieją ruchy niedozwolone, tzw. ruchy tabu. Algorytm unika oscylacji wokół optimum lokalnego dzięki przechowywaniu informacji o sprawdzonych już rozwiązaniach w postaci listy tabu [124].

Wymienione algorytmy uważa się za algorytmy optymalizacji lokalnej. Przeglądają zazwyczaj niewielką część wszystkich możliwych kombinacji cech. Według niektórych autorów jest to około 30%.

W metodach deterministycznych występuje tzw. problem zagnieżdżania. Wiadomo, że w ogólnym wypadku optymalny zbiór cech o rozmiarze $n + 1$ nie zawiera wszystkich cech należących do optymalnego zbioru o rozmiarze n, lecz mogą to być zupełnie różne zbiory. Wymienione powyżej strategie w dużym stopniu „ciągną" skład podzbioru cech przez kolejne kroki procedury selekcji. Algorytmy przedstawiono w kolejności malejącej roli problemu zagnieżdżania. Z wyjątkiem dwóch ostatnich metod znajdują one rozwiązania stosunkowo szybko. Wśród strategii stochastycznych należy wymienić przede wszystkim:

• Algorytmy genetyczne (ang. *genetic algorithms*) [98, 137, 138]. Algorytm genetyczny przeszukuje kombinacje cech w sposób losowy. W kolejnych krokach algorytmu oceniane są zbiory możliwych rozwiązań (zestawy podzbiorów cech), tzw. populacje. Pojedynczy osobnik populacji to odpowiednio zakodowany jeden z możliwych podzbiorów cech. Nowa populacja, tzw. potomstwo, powstaje w wyniku operacji określanych jako mutacja, krzyżowanie i selekcja, którym poddawana jest populacja poprzedzająca, tzw. rodzice. Dzięki zasadom rządzącym tymi operacjami zapewnia się losowość powstania nowej populacji. Mimo to w kolejnych krokach algorytmu uzyskiwane są coraz lepsze osobniki, tj. zestawy cech o coraz wyższej ocenie. Algorytmy genetyczne cieszą się dużą popularnością wśród strategii przeszukiwania zbioru kombinacji cech w analizie danych z czujnikowych pomiarów zanieczyszczeń [96, 113, 123, 126];

• Symulowane wyżarzanie (ang. *simulated annealing*) [98]. Algorytm porusza się krokowo wśród wszystkich możliwych kombinacji cech. Ocena podzbioru cech po wykonaniu kroku, tj. po wyeliminowaniu jednej cechy jest porównywana z oceną sprzed kroku. Zachodzi pewne prawdopodobieństwo, że cecha zostanie usunięta z podzbioru, mimo że ocena powstałego podzbioru jest gorsza. Prawdopodobieństwo to zależy od tzw. temperatury wyżarzania, T_i, oraz różnicy między rzeczywistą a pożądaną wartością oceny podzbioru cech, ΔE, w następujący sposób:

$$p = \exp\left(\frac{-\Delta E}{T_i}\right) \tag{5.2}$$

Możliwość usunięcia cechy mimo pogorszenia oceny danego zbioru stanowi zabezpieczenie algorytmu, dzięki któremu nie grzęźnie on w optimach lokalnych. Początkowa temperatura wyżarzania jest ustalana arbitralnie, a algorytm jest powtarzany ustaloną liczbę razy, podczas gdy temperatura się zmniejsza monotonicznie. Po każdej zmianie temperatury wyżarzania algorytm rozpoczyna działanie od pełnego zbioru cech. Pojedyncze prace dotyczą zastosowania tego interesującego algorytmu stochastycznego w analizie danych z czujnikowych pomiarów gazów [45, 98].

Wymienione algorytmy są uważane za algorytmy optymalizacji globalnej. Znajdują one rozwiązanie optymalne lub bliskie optymalnemu. W przedstawionych strategiach występuje element losowości w wyborze podzbioru cech w kolejnych krokach przeszukiwania. Z tego tytułu są one mniej podatne na utknięcie w minimum lokalnym niż strategie sekwencyjne. Ceną są większe wymagania pod względem obliczeniowym, a znalezienie rozwiązania może być czasochłonne.

Na koniec warto wspomnieć o rozwiązaniach proponujących wieloetapową (kaskadową) selekcję cech. Wydaje się, że jest to metoda umożliwiająca uzyskanie większej redukcji przestrzeni cech [137] oraz zbieżności między efektami różnych strategii selekcji zastosowanych w ramach etapów [98]. Możliwość wyłonienia obiektywnie

najlepszego, a zarazem najmniejszego zestawu cech jest istotną zaletą tego rodzaju selekcji.

5.2.2. Ekstrakcja cech

Inne podejście do problemu redukcji wymiarowości początkowego wektora cech określa się jako ekstrakcję lub mapowanie. Mapowanie polega na utworzeniu nowego zbioru cech w wyniku transformacji wektora cech wyjściowych. Zadaniem tej operacji jest znalezienie niskowymiarowego wektora **z**, który zachowuje większość informacji zawartych w wektorze początkowym **x**, zgodnie ze wzorem [23]:

$$f : \mathrm{x} \in R^d \rightarrow \mathrm{z} \in R^q \left(q < d\right) \tag{5.3}$$

W celu przedstawienia danych z wysokowymiarowej przestrzeni w przestrzeni niskowymiarowej potrzebna jest projekcja, optymalna ze względu na zadane kryterium *f*. Przy redukcji wymiaru przestrzeni cech techniką liniową mapowanie jest realizowane zgodnie ze wzorem:

$$\mathbf{z} = \mathbf{T}\mathbf{x} \tag{5.4}$$

gdzie **T** jest macierzą przekształcenia.

W czujnikowych pomiarach gazów najczęściej stosowane są liniowe techniki ekstrakcji cech. Podstawowa jest analiza składowych głównych. Inne stosowane techniki to np. liniowa analiza dyskryminacyjna Fishera, analiza składowych niezależnych, analiza składowych głównych z funkcją rdzeniową, skalowanie neuronowe.

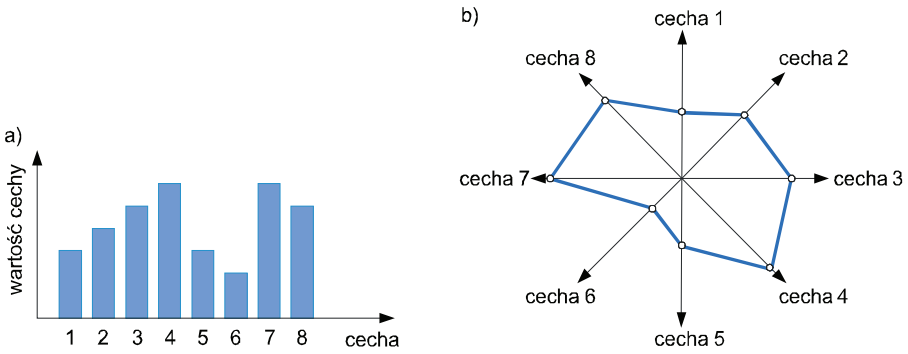
Redukcja wymiarowości początkowej przestrzeni cech metodami ekstrakcji jest często stosowana nie tylko jako etap wstępnego przetwarzania danych, poprzedzający dalszą analizę, lecz w celu ujawnienia struktury danych. Na ogół w wyniku graficznego przedstawienia danych w zredukowanej przestrzeni cech obserwuje się zasadnicze kierunki zmienności danych, co daje wgląd w rodzaje przenoszonej informacji. Metody ekstrakcji cech bardzo dobrze służą eksploracji danych. Omówiono je w p. 6.1, poświęconym eksploracyjnej analizie danych.

5.3. Wektor cech a wzorzec

W wyniku selekcji i/lub ekstrakcji cech ze zbioru cech kandydujących powstaje uporządkowany zestaw cech określany jako wektor cech $\mathbf{x} = [x_1, x_2, \ldots, x_k]^T$, gdzie x_i

jest pojedynczą cechą, $i = 1, 2, \ldots, k$. Należy podkreślić, że jego składowymi są zmienne. Wynika stąd wprost geometryczna interpretacja wektora cech. Wektor cech o liczbie składowych k ustanawia przestrzeń o k wymiarach. Z każdą spośród k zmiennych wiąże się jeden wymiar tej przestrzeni. Liczba wymiarów przestrzeni cech jest równa liczbie składowych wektora cech, ta z kolei jest równa liczbie cech wyselekcjonowanych/wyekstrahowanych. W ogólnym przypadku przestrzeń cech nie jest ortogonalna, ze względu na możliwe skorelowanie cech.

Na podstawie danych pomiarowych dla badanego gazu można otrzymać wektor danych będący liczbową realizacją wektora cech. Jest to tzw. wzorzec badanego gazu. Ze względu na wielowymiarowy charakter przestrzeni cech w analizie danych z pomiarów czujnikowych wzorce gazów przyjęto przedstawiać graficznie, najczęściej za pomocą wykresów kolumnowych lub radarowych (rys. 5.5) [12, 29, 37, 139]. Poszczególne kolumny lub osie (w przypadku wykresów radarowych) odpowiadają składowym wektora cech.

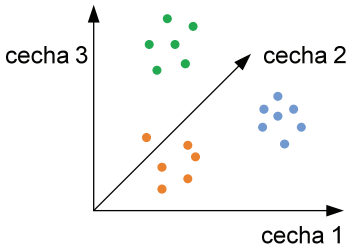


Rys. 5.5. Graficzne interpretacje wzorca gazu: a) wykres kolumnowy, b) wykres radarowy

Metody graficzne (rys. 5.5) pozwalają zaobserwować, w jakim stopniu dany wektor cech odzwierciedla różnice między wzorcami rozważanych gazów, a zatem pośrednio między gazami (zanieczyszczeniami). W zastosowaniach technicznych wizualna rozpoznawalność gazów na podstawie wzorca jest jednak niewystarczająca. Potrzebna jest raczej obiektywna miara wskazująca na przydatność wzorca do określania gazów. Z tego względu przydatniejsza jest geometryczna interpretacja wzorca jako punktu w przestrzeni określonej przez składowe wektora cech (rys. 5.6). Układ punktów w tej przestrzeni wskazuje na jej przydatność do pozyskania informacji jakościowej i ilościowej o gazach reprezentowanych przez wzorce.

Ze względu na informację jakościową w przestrzeni cech definiowane są miary podobieństwa wzorców, pozwalające wyróżnić kategorie wzorców, a zatem i gazów podobnych. W alternatywnym rozwiązaniu opracowuje się zasady podziału przestrzeni cech prowadzące do wydzielenia jej fragmentów, w których położone są punkty (wzorce) reprezentujące poszczególne kategorie zanieczyszczeń [22]. W ten sposób

powstają obiektywne miary/zasady pozwalające określić przynależność wzorców do odpowiednich kategorii. Są to zarazem miary/zasady przypisania reprezentowanych przez te wzorce gazów do odpowiednich klas. Zabieg ten jest równoznaczny z opracowaniem metody jakościowego określania zanieczyszczenia, polegającej na wyznaczaniu klasy gazów, do której należy badane zanieczyszczenie na podstawie pomiaru czujnikowego.



Rys. 5.6. Wzorzec gazu jako punkt w przestrzeni cech; kolorami zaznaczono różne kategorie wzorców

Ze względu na informację ilościową definiowane jest odwzorowanie przestrzeni cech w zmienną określającą gazy, reprezentowane przez te wzorce pod względem ilościowym. Parametryzacja tego odwzorowania jest dokonywana z wykorzystaniem zestawu wzorców i odpowiadających im wartości miary ilościowej. W ten sposób powstaje narzędzie pozwalające określić wartość tej miary dla dowolnego wzorca, którego elementy mieszczą się w zakresie wartości elementów wzorców uczących. Jest to równoznaczne z opracowaniem metody ilościowego określania zanieczyszczenia, polegającej na wyznaczaniu wartości zmiennej charakteryzującej je pod względem ilościowym na podstawie pomiaru czujnikowego.

W ogólnym wypadku wektory cech najlepsze do określania danego gazu pod względem jakościowym i ilościowym mogą być różne. Pozyskanie tego rodzaju informacji wiąże się bowiem z wykonaniem zupełnie różnego rodzaju operacji na przestrzeni cech. Stąd tak ogromną rolę spełnia selekcja/ekstrakcja cech prowadzona w kontekście konkretnego problemu oznaczania gazów.

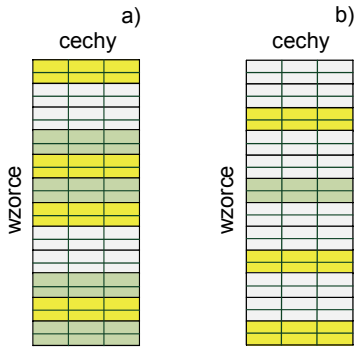
Koncepcja wektora cech oraz wzorca jest uniwersalna. Pozwala rozwiązywać problemy dotyczące określania gazów bez względu na złożoność ich składu. Fakt ten jest kluczowy dla pozyskiwania informacji o zanieczyszczeniu na podstawie pomiarów czujnikowych.

5.4. Dane wielowymiarowe

Zgodnie z ogólną definicją dane wielowymiarowe składają się z wyników obserwacji wielu różnych zmiennych dla licznych obiektów. Każda zmienna może być

uważana za ustanawiającą inny wymiar w ten sposób, że jeżeli występuje *n* zmiennych, to każdy obiekt można uważać za zajmujący określone miejsce w abstrakcie określanym jako *n*-wymiarowa przestrzeń [140]. W świetle wcześniejszych rozważań łatwo zauważyć, że szczególnym przypadkiem danych wielowymiarowych jest zestaw wzorców. W tym wypadku rolę podanych w definicji zmiennych pełnią cechy.

Klasycznie dane wielowymiarowe mają postać macierzy (rys. 5.7). Poszczególne kolumny odnoszą się do cech, wiersze zaś zawierają wzorce.



Rys. 5.7. Dane wielowymiarowe zbudowane na podstawie trójelementowego wektora cech. Kategorie wzorców zaznaczono kolorami: a) odpowiednie proporcje liczby wzorców różnych kategorii, b) nieodpowiednie proporcje liczby wzorców różnych kategorii. Stosunek wektorów danych do liczby cech wynosi osiem

Dane wielowymiarowe w takim rozumieniu są niezbędne do opracowania modeli klasyfikacji czy regresji w systemie rozpoznawania wzorców na potrzeby oznaczania gazów w pomiarach czujnikowych. Na tej zasadzie konstruuje się dane uczące, walidujące oraz testujące. Poszczególne wzorce zawarte w macierzy dotyczą różnych prób gazów. Zastosowanie metod uczenia nadzorowanego wymaga włączenia do analizy, równolegle z danymi wielowymiarowymi, wektora etykiet jakościowych, mówiących o przynależności poszczególnych wzorców do odpowiednich kategorii wzorców. Taka konfiguracja danych służy opracowaniu modeli klasyfikacji na potrzeby jakościowego oznaczania gazów. Do konstrukcji odwzorowań ilościowych na potrzeby oznaczania gazów o takim charakterze konieczny jest natomiast wektor wartości zmiennej charakteryzującej ilościowo poszczególne gazy, reprezentowane przez wzorce zawarte w macierzy danych wielowymiarowych. Eksploracja danych nie wymaga żadnych dodatkowych wektorów, jej przedmiotem zaś są wyłącznie dane wielowymiarowe.

Należy jeszcze podkreślić podstawowe znaczenie doboru wzorców uczących dla rezultatów pracy systemu rozpoznawania wzorców. Jednym z istotnych zagadnień jest stosunek liczby wektorów danych do liczby elementów wektora cech. Im mniejsza jest liczba danych w grupie, tym gorszy wynik klasyfikacji. Większe próby umożliwiają minimalizację skutków rozrzutu występującego w puli danych, wywołanego czynnikiem

losowym, i umożliwiają ujawnienie się rzeczywistych różnic między klasami. Jednocześnie uprzywilejowana jest jak najmniejsza liczba cech. W pracy [141] wykazano na podstawie badań z wykorzystaniem danych wygenerowanych o kontrolowanym poziomie zaszumienia, że stosunek liczby wektorów danych do liczby cech powinien wynosić co najmniej sześć. Pokazano, że taka wartość zapewniała zabezpieczenie przed zbytnim dopasowaniem klasyfikatora do danych, jeżeli jego struktura była opracowywana w trybie walidacji. W myśl powszechnie przyjmowanej zasady w dziedzinie rozpoznawania wzorców stosunek liczby wektorów danych do liczby cech powinien wynosić 5–10 [139]. Przykład danych wielowymiarowych spełniających te kryteria oraz niepozostających w zgodzie z nimi przedstawiano na rys. 5.7.

Dane wielowymiarowe poddaje się niekiedy zabiegom mającym poprawić ich jakość. Do podstawowych należą skalowanie i linearyzacja. Warto o nich wspomnieć, choć ich stosowanie nie jest konieczne.

5.4.1. Skalowanie

Wyróżnia się globalne i lokalne techniki skalowania [29, 38, 72]. Każda grupa zajmuje się innym aspektem danych, który może mieć negatywny wpływ na jakość rozpoznawania wzorców.

Skalowanie globalne koncentruje się na poszczególnych cechach. Jego użyteczność jest widoczna w warunkach istotnego zróżnicowania zakresów wartości poszczególnych parametrów sygnału czujników wyłonionych podczas kompresji. Dominujący wpływ na działanie klasyfikatora mają parametry o największych wartościach i zakresie zmienności. Jest to niekorzystne, jeżeli najbardziej użyteczna informacja jest przenoszona przez cechy o relatywnie niewielkich wartościach i zakresie zmienności. Zadaniem skalowania jest wyrównanie szans poszczególnych cech. Dwa podstawowe sposoby skalowania cech to normalizacja i autoskalowanie [142]. Normalizacja cechy przeskalowuje ją do zakresu [0, 1]. Następuje to w wyniku odjęcia od każdej wartości cechy jej wartości minimalnej i podzielenia tak otrzymanej różnicy przez zakres wartości parametru. Normalizacja prowadzi do ujednolicenia zakresu zmienności wszystkich cech [68]. Wykorzystuje cały zakres dynamiczny cechy. Niestety w związku z tym nie jest odporna na wartości skrajne. Autoskalowanie, inaczej standaryzacja cechy, jest wolne od tej wady. Polega ono na odjęciu od każdej wartości cechy jej średniej wartości i podzieleniu otrzymanej różnicy przez odchylenie standardowe parametru. Niestety odporność na wyniki skrajne uzyskuje się kosztem nieidentycznych zakresów zmienności wyskalowanych cech. Wadą technik globalnych jest możliwość wzmocnienia składnika losowego w danych.

Skalowanie lokalne dotyczy wektorów danych. Jego zadaniem jest zmniejszenie różnorodności wektorów należących do jednej kategorii. W szczególności powodem jej występowania jest zróżnicowanie stężeń poszczególnych prób lub dryf czujników.

Podstawową techniką globalną jest normowanie. Polega ono na podzieleniu każdego elementu wektora danych przez normę tego wektora. Jako normę stosuje się pierwiastek sumy kwadratów wszystkich elementów wektora, sumę wartości bezwzględnych wszystkich elementów wektora lub wybrany element wektora związany z czujnikiem przyjętym za referencyjny. Usunięcie efektu stężenia jest skuteczne, jeżeli wszystkie cechy wykazują taki sam rodzaj związku ze stężeniem gazu. Wówczas normalizacja może poprawić efektywność klasyfikacji. Nie powinna być natomiast stosowana w rozwiązywaniu problemów ilościowych, gdyż w takich sytuacjach norma wektora cech przenosi istotną informację. Sugeruje się, aby w przypadku matryc składających się z różnych czujników normalizację przeprowadzać w obrębie tych grup i ewentualnie w drugim kroku na całym wektorze cech. Inny rodzaj skalowania lokalnego to standaryzacja wektora cech. Polega ona na odjęciu od każdego elementu wektora średniej ze wszystkich jego elementów i podzieleniu przez ich odchylenie standardowe.

Stosowanie technik lokalnych wymaga ostrożności w związku z tym, że powodują one „zamknięcie" danych. Standaryzacja wymusza np. sumę cech bliską zeru, a normowanie wymusza sumę kwadratów cech równą jeden. Może to wywoływać wzmocnienie wzajemnego skorelowania cech o dużych lub małych wartościach. Stąd często wskazuje się na potrzebę zastosowania metod globalnych, wyrównujących zakres wartości cech przed skalowaniem czy normowaniem cech.

Zdania na temat przydatności technik skalowania są podzielone. Wyniki prac porównawczych wykazują np. znikome różnice między uzyskanymi efektami ich zastosowania [72].

5.4.2. Linearyzacja

Zastosowanie technik linearyzacji wiąże się z przygotowaniem danych do rozwiązywania problemów ilościowych. Zadaniem tych technik jest redukcja nieliniowości związku między wartościami cech a miarą ilościową opisującą próby gazu. Dotyczy to zwłaszcza redukcji efektu nasycenia, który występuje w zakresie większych stężeń gazów. Do najprostszych rozwiązań należą logarytmizacja i pierwiastkowanie [3]. Przykłady bardziej złożonych technik to przekształcenie Box–Coxa czy Horner–Hierolda [3]. Ostatnie wykazuje dobre właściwości w odniesieniu do danych czujników półprzewodnikowych. Doniesienia literaturowe o stosowaniu linearyzacji do czujnikowych pomiarów gazów są jednak sporadyczne.

6. Odczyt informacji na podstawie reprezentacji gazu

Podstawowym czynnikiem decydującym o możliwości odczytu informacji o badanych gazach z danych pomiarowych jest obecność tej informacji w danych. Można ją wówczas pozyskać, dobierając odpowiednią metodę. Spełnienie warunku występowania informacji w danych nie należy jednak do zakresu zadań analizy danych, lecz w dużej mierze jest kwestią odpowiedniego doboru czujników (por. rozdz. 3).

Jednym z najważniejszych czynników decydujących o wyborze metody odczytu informacji jest sposób jej występowania w danych. Wpływa on istotnie na stopnień skomplikowania zastosowanej metody. Szczególnie ważną rolę w zapewnieniu możliwości korzystania z prostszych modeli obliczeniowych odgrywa dobrze przeprowadzona selekcja lub ekstrakcja cech. Zmiana przestrzeni cech, w jakiej są na przykład rozwiązywane problemy klasyfikacji, może wywołać zmianę stopnia rozdzielenia kategorii wzorców (lepiej–gorzej), jak również zmianę jego charakteru (np. liniowo separowane – nieliniowo separowane). W zakresie problemów ilościowych ten sam zabieg może wpłynąć na postać zależności i jej wyrazistość.

Wybór metody odczytu informacji zależy również od liczby składowych wektora cech i wielkości zbioru danych uczących. Wynika to z konieczności zapewnienia, że model jakościowy/ilościowy odwzorowuje rzeczywistą strukturę danych, nie zaś przypadkowe zależności występujące w ich zbiorze. W razie niekorzystnej proporcji liczby składowych wektora cech do liczby wzorców zachodzi znaczne ryzyko zbytniego dopasowania modelu obliczeniowego do danych uczących i pozbawienia go możliwości uogólniania. Zdolność poprawnego rozpoznawania wzorców innych niż uczące jest zaś podstawowym wymaganiem stawianym tym modelom. Wskazówek teoretycznych do zapewnienia odpowiedniego zbioru danych uczących należy szukać w dziedzinie programowania eksperymentu.

Nie bez znaczenia są też względy praktyczne, takie jak znajomość poszczególnych metod przez badacza, dostępność odpowiedniego oprogramowania oraz przeznaczenie systemu analizy danych. Ze względu na ostatnie kryterium bardzo istotne są wymagania poszczególnych modeli pod względem zasobów pamięci, mocy obliczeniowej i czasu obliczeń.

6.1. Metody odczytu niezdefiniowanej informacji

W analizie danych z pomiarów czujnikowych istotną rolę odgrywają metody umożliwiające rozpoznanie wewnętrznej struktury danych wielowymiarowych bez narzucania zewnętrznych kryteriów porządkujących ten zbiór. Znakomicie nadają się do tego celu metody eksploracji danych. Należą do metod statystycznych, jednak nie analizują danych pod kątem hipotez statystycznych, które są potwierdzane lub odrzucane. Realizują natomiast podejście, które można określić, jako otwarte, tj. zorientowane na przedstawienie w sposób jak najbardziej czytelny prawidłowości, jakie ujawniają się w danych. W tym celu większość metod wykorzystuje narzędzia graficzne, umożliwiające wizualizację danych oraz transformacje, głównie zmierzające do zredukowania wymiarowości danych, a przez to ułatwienie ich rozumienia.

Trzy zasadnicze aspekty analizy eksploracyjnej wielowymiarowych danych czujnikowych dotyczą: i) samoorganizacji wektorów danych reprezentujących badane gazy, ii) samoorganizacji cech wyłonionych z sygnałów czujników w zestawy cech podobnych oraz iii) roli poszczególnych cech w grupowaniu wektorów danych. W pierwszym przypadku interesująca jest przede wszystkim zasada samoorganizacji, tj. właściwości gazów, które stanowią podstawę podobieństwa wektorów danych i skutkują ich przynależnością do określonego zgrupowania. Klasycznie rozważa się czynniki związane z właściwościami chemicznymi substancji oraz ich stężenie. Nierzadko analiza eksploracyjna ujawnia dodatkowo inne możliwości organizacji wektorów danych, które nie są oczywiste i wymagają refleksji nad ich przyczyną, leżącą po stronie innych właściwości badanych gazów. Innym interesującym zagadnieniem jest wyrazistość struktury danych. Wyraża się ona stopniem zwartości zgrupowań punktów danych, które są związane z gazami tej samej kategorii oraz stopniem rozłączności zgrupowań punktów reprezentujących różne kategorie gazów. Dzięki temu ujawnia się siła czynników wywołujących zróżnicowanie danych. Z ujawnionej struktury danych wynikają przesłanki dotyczące efektywnej klasyfikacji gazów.

Analiza eksploracyjna ze względu na strukturę zbioru cech ma związek z zagadnieniem nadmiarowości cech. Cechy wyłaniane z sygnałów czujników charakteryzują się różnym stopniem podobieństwa pod względem przenoszonej informacji. W toku analizy eksploracyjnej zachodzi możliwość wyłonienia grup cech podobnych. Otwiera to drogę do heurystycznej redukcji wymiarowości wektora cech z nieistotną stratą informacji. Efektem dodanym analizy struktury zbioru cech jest wgląd w czynniki wywołujące podobieństwo cech. Wiedza taka jest szczególnie przydatna do projektowania źródeł danych, np. doboru czujników.

Analiza eksploracyjna pozwala również wnioskować o roli poszczególnych cech w grupowaniu danych według kryteriów ujawniających się w ramach tej analizy. Ostatecznie warto wspomnieć, że interesujące jest zestawienie wyników analizy eksploracyjnej wykonanej dla danych wielowymiarowych dotyczących gazów o ściśle

zdefiniowanym składzie, lecz zbudowanych na podstawie różnych zestawów cech. Pozwala ono przeanalizować rolę czynników wpływających na zawartość informacji w sygnale czujnika przez porównanie zawartości informacyjnej różnych zestawów cech.

Metody uczenia bez nadzoru są w zasadzie metodami rozpoznawania wzorców *sensu stricte*, gdyż czynią to bez żadnej podpowiedzi. Niektóre z nich określa się jako metody eksploracji danych, podkreślając ich rolę w zrozumieniu rodzaju informacji i sposobu jej przenoszenia przez zbiór danych. Wiele z nich jest wymienianych jako metody redukcji wymiarowości czy graficznej prezentacji struktury danych. Funkcje te są w pewnym stopniu powiązane. Zmniejszenie liczby wymiarów wektora cech do trzech, a częściej nawet do dwóch pozwala zilustrować strukturę oryginalnie wielowymiarowego zbioru danych w przestrzeni trójwymiarowej lub na płaszczyźnie. Możliwość wizualizacji danych jest niezwykle istotna dla systemów czujnikowych ze względów aplikacyjnych. Redukcja wymiarowości może mieć też czysto instrumentalny wymiar, tj. służyć do przygotowania zredukowanego zestawu cech do dalszej analizy metodami z nadzorem. Stąd niektóre z metod eksploracji danych są również uważane za metody kompresji danych. Ten rodzaj kompresji określa się jako ekstrakcję cech.

6.1.1. Analiza składowych głównych

Metodą uczenia bez nadzoru najpowszechniej stosowaną w analizie danych czujnikowych jest analiza składowych głównych (ang. *principal component analysis*, PCA). Polega ona na dekompozycji macierzy wzorców **X**, w wyniku której powstają: macierz składowych głównych **A**, macierz ładunków **C** i macierz składnika losowego **E**:

$$\mathbf{X} = \mathbf{A}\mathbf{C}^{\mathrm{T}} + \mathbf{E} \tag{6.1}$$

Analiza przekształca przestrzeń oryginalnych zmiennych w przestrzeń innych zmiennych, określanych jako składowe główne. Wyznaczają one kierunki największej zmienności w danych. Na ogół dwie lub trzy pierwsze składowe wyjaśniają praktycznie całą zmienność zmiennych oryginalnych. W wyniku PCA zachodzi więc redukcja przestrzeni *n*-wymiarowej do 2–3-wymiarowej z niewielką stratą informacji. Nowe zmienne są ortogonalne względem siebie. Redukcja wymiarów i ich ortogonalność umożliwia wizualizację wzorców badanych gazów w postaci punktów w układzie kartezjańskim, którego osiami są składowe główne. Punkty blisko położone tworzą zgrupowania wskazujące na podobieństwo wzorców i pośrednio gazów, które te wzorce reprezentują. Liczba zgrupowań, ich rozległość i wzajemne ułożenie sygnali-

zują strukturę klas wzorców, jaka wyłania się samoczynnie z danych pomiarowych. Analiza pochodzenia poszczególnych wzorców pozwala wykryć źródło zróżnicowania gazów, które odzwierciedla struktura klas wzorców.

Interesująca jest również macierz ładunków. Zawiera ona informację o tym, które cechy najbardziej przyczyniają się do wyróżnienia grup wzorców. Najlepszym sposobem oceny tego jest sporządzenie wykresu ładunków cech we współrzędnych składowych głównych i odczytanie go razem z wykresem wzorców. Punkty ulokowane blisko początku układu współrzędnych wskazują na cechy relatywnie nieaktywne. Punkty odległe reprezentują cechy czynnie uczestniczące w wyróżnieniu grup wzorców za pomocą tej składowej, która decyduje o ich dużej odległości od początku układu współrzędnych. Odniesienie do wykresu wyników analizy PCA pozwala określić klasy gazów, w których wyróżnieniu uczestniczy dana cecha.

Ortogonalność nowych zmiennych uzyskanych w wyniku analizy ma też inną zaletę. Ze względu na często występujące skorelowanie cech oryginalnych składowe główne są atrakcyjnym zestawem zmiennych wejściowych dla modeli wykorzystujących metody klasyfikacji z nadzorem lub regresji. Dlatego PCA jest także uważane za metodę ekstrakcji cech.

Analiza składowych głównych jest w zasadzie podstawową metodą analizy danych czujnikowych. Przykłady jej zastosowania są bardzo liczne. Przedstawiony dalej wybór ma zilustrować zasadnicze kategorie problemów, gdzie analiza ta jest przydatna.

PCA pozwala określić właściwości gazów, które są przyczyną samoczynnego grupowania wzorców gazów na podstawie danych z pomiarów czujnikowych. Należą tu np. właściwości utleniające/redukujące gazu [95, 96], właściwości zapachowe gazu [12, 48], tożsamość chemiczna substancji [10, 54, 120, 143], skład mieszanin gazów [13, 20, 70, 144], stężenie gazu [10, 95]. Dzięki temu metoda pozwala również analizować wpływ interferencji na możliwości pomiarowe czujników. W pracach [145, 146] pokazano przykłady zastosowania PCA do analizy wpływu wilgotności na możliwości rozpoznawania gazów za pomocą matrycy czujników.

Niewyraźna struktura grup wzorców ujawniona w wyniku analizy składowych głównych nie zawsze jest miarodajnym wskaźnikiem trudności w rozróżnianiu klas wzorców, a zatem i gazów [13, 147]. Jeżeli natomiast w przestrzeni składowych głównych są wyraźnie rozdzielone grupy wzorców, to uzyskuje się zazwyczaj satysfakcjonujące wyniki klasyfikacji metodami pod nadzorem [10, 63].

Składowa główna stanowi syntetyczną reprezentację czynników wpływających w podobny sposób na wszystkie cechy wyłonione z sygnałów czujników. Można ten fakt wykorzystać do monitorowania zmian takiego czynnika, często w prosty sposób za pomocą jednej składowej głównej [49, 148]. Praca [148] ilustruje zastosowania tego rozwiązania do monitorowania procesu chemicznego – fermentacji wina.

PCA daje możliwość określenia roli poszczególnych cech w grupowaniu wzorców. Jest to problem o charakterze ogólniejszym, bardzo istotny dla optymalizacji matrycy czujników i z tego powodu podejmowany przez wielu badaczy. W razie ko-

rzystania z PCA podstawą oceny istotności cech jest analiza ładunków cech na poszczególne składowe główne w kontekście struktury wzorców [97].

6.1.2. Analiza składowych niezależnych

Zadanie rozdzielania niezależnych źródeł informacji w problemach wielowymiarowych można zrealizować z zastosowaniem analizy składowych niezależnych (ang. *independent component analysis*, ICA). Podstawowe równanie metody jest następujące:

$$\mathbf{X} = \mathbf{\Delta S} \tag{6.2}$$

gdzie: $\mathbf{X}$ jest macierzą wzorców, $\mathbf{\Delta}$ jest ortogonalną macierzą współczynników, a $\mathbf{S}$ jest macierzą składowych niezależnych. Znalezienie $\mathbf{\Delta}$, a w zasadzie macierzy odwrotnej do niej $\mathbf{\Delta}^{-1}$, jest wymagające obliczeniowo. Z tego względu rozwój metody ICA nastąpił dopiero obecnie, w dużej mierze dzięki zwiększeniu mocy obliczeniowej komputerów. Z metodologicznego punktu widzenia analiza składowych niezależnych stanowi szczególny przypadek analizy czynnikowej. Na rozwiązanie problemu dekompozycji macierzy danych nałożone jest w tym wypadku dodatkowe wymaganie wzajemnej niezależności wyłanianych składowych. Jest to warunek silniejszy niż wymaganie wzajemnego nieskorelowania składowych obowiązujący w przypadku PCA. Ze względu na potrzebę skorzystania ze statystyk wyższego rzędu uzyskanie rozwiązania wymaga niegaussowskiego rozkładu składowych niezależnych $\mathbf{S}$. Oczekiwanie to pozostaje w opozycji do zwykle milcząco przyjmowanego założenia, że dane mają rozkład normalny. W rzeczywistości ściśle gaussowskie zmienne są bardzo rzadko spotykane. Jeżeli odstępstwo od normalności jest znaczne, dostęp do części informacji za pomocą PCA nie jest możliwy i ICA staje się bardziej odpowiednią metodą analizy.

W przeciwieństwie do słabszego wymagania braku skorelowania, który obowiązuje w analizie składowych głównych, statystyczna niezależność składowych oznacza, że przenoszą one niezależną informację. Idąc tym tropem, można myśleć o informacji pochodzącej z niezależnych źródeł. Duże zainteresowanie ICA w dziedzinie analizy danych z pomiarów czujnikowych wynika stąd, że metoda ta wydaje się idealnie nadawać do zastosowania, gdy matryca czujników jest eksponowana na działające wspólnie bodźce różnego pochodzenia, z których każdy jest przedmiotem osobnego zainteresowania. Dzięki identyfikacji i wyodrębnieniu składowych niezależnych w sygnale można rozdzielić źródła informacji. Nie jest przy tym potrzebny dostęp do sygnałów odzwierciedlających poszczególne źródła oddziałujące osobno. ICA realizuje tzw. poszukiwanie źródeł „w ciemno", (ang. *blind source selection*, BSS) [149]. Najprostszym przykładem praktycznego wykorzystania takiej właściwości metody jest rozdzielenie na poziomie danych informacji pochodzącej od badanych zanieczyszczeń

i związków interferujących. Bardziej złożonym, a jeszcze bardziej interesującym zadaniem jest rozdzielenie informacji pochodzących od różnych substancji. Problemu takiego nie uda się rozwiązać np. za pomocą PCA lub metody cząstkowych najmniejszych kwadratów, jeżeli kierunki zmienności sygnału uwarunkowane przez różne bodźce są w dużej mierze zbieżne.

W analizie składowych niezależnych trudność sprawia wybór liczby składowych, na które będzie rozkładana macierz danych. Z tego względu ICA jest zwykle poprzedzana PCA w celu określenia proponowanej liczby składowych. Osobną kwestią jest ocena przydatności poszczególnych składowych. W przeciwieństwie do PCA nie ma tu możliwości określenia udziału poszczególnych składowych w ogólnej zmienności danych poddanych dekompozycji. Nie można więc zastosować tego kryterium wyboru najważniejszych składowych. W praktyce wybór jest przeprowadzany w systemie pod nadzorem. Składowe, które wykazują skorelowanie z badaną wielkością, przechodzą do dalszej analizy, a cechy nieskorelowane są z niej wyłączane jako reprezentujące składnik losowy. Taki zabieg ma na celu poprawę rezultatów klasyfikacji lub określania ilościowego badanych związków na podstawie pomiarów czujnikowych.

Di Natale i inni [150] zastosowali ICA do eliminacji zakłóceń wywołanych zmienną wilgotnością i temperaturą powietrza podczas pomiarów czujnikowych. Usunięcie składowej niezależnej, skorelowanej z tymi parametrami powietrza, z zestawu zmiennych wejściowych umożliwiło poprawę stopnia rozdzielenia dwóch klas badanych gazów na podstawie uzyskanego wektora cech. Kermit i Tomic [151] pokazali, że ICA jest skuteczną metodą korekty dryfu czujników. W tej samej pracy potwierdzono możliwość rozdzielenia informacji jakościowej i ilościowej o badanych gazach tak, że każda jest reprezentowana przez inną składową niezależną. Również Meng [152] z sukcesem oddzielił informację ilościową o badanej mieszaninie gazów od zakłóceń. Inspiracji dla nowych zastosowań metody w dziedzinie pomiarów czujnikowych można poszukiwać w obszarze pomiarów analitycznych. W szczególności należy mieć tu na uwadze analizę sygnału wyjściowego czujnika uzyskanego w odpowiedzi na złożoną mieszaninę gazów [153].

6.1.3. Analiza skupień

Analiza skupień poszukuje przestrzennych związków lub podobieństw między wektorami danych i grupuje je na tej podstawie. Pojęcie skupienia nie jest oczywiste, stąd próby jego zdefiniowania w różny sposób, np. i) wzorce w obrębie skupienia są bardziej podobne do siebie niż wzorce należące do różnych skupień, ii) w skupieniu gęstość punktów jest relatywnie duża, a między skupieniami istotnie mniejsza. Skupienia mogą być różnej wielkości, mieć różny kształt i mogą być różnie rozmieszczone względem siebie w przestrzeni cech. Wyłonienie skupień z zachowaniem ich charakteru i z uwzględnieniem kontekstu innych skupień jest zadaniem trudnym. Stosując

analizę skupień, należy pamiętać o kilku zasadach: Po pierwsze, metody z tej grupy wymagają dużej liczby wzorców uczących. Wzorce przynależne do różnych kategorii wypełnią wówczas regiony przestrzeni cech odpowiadające różnym typom obiektów. Po drugie, każdy algorytm analizy skupień wykazuje tendencję do wyróżnienia skupień w analizowanym zbiorze danych bez względu na to, czy one tam występują, czy nie. Istnieje więc konieczność weryfikacji otrzymanego rozwiązania z udziałem eksperta. Po trzecie, otrzymany rezultat jest subiektywny, tj. zależny od zastosowanej metody. Zarazem jednak trudno wskazać najlepszy algorytm analizy skupień. Ocena za pomocą metod obiektywnych stosowanych do rozwiązań pod nadzorem jest tu niedostępna. Istnieje natomiast zespół narzędzi określanych mianem technik walidacji skupień (ang. *cluster validation techniques*), które pozwalają ilościowo ocenić ich nieprzypadkowość [154]. Ostatnio zaproponowano bardzo interesującą metrykę rozdzielczości skupień [130], która umożliwia obiektywne i automatyczne określenie stopnia rozłączności skupień w sposób, który uwzględnia ich wielkość, położenie oraz wzajemne ułożenie.

Wyróżnia się następujące wspólne elementy analizy skupień: i) wybór miary niepodobieństwa wektorów, najczęściej jest wybierana odległość euklidesowa, inna opcja to np. odległość Mahalanobisa; ii) zdefiniowanie kryterium, które podlega optymalizacji, na ogół dotyczy ono struktury wewnętrznej klas oraz wzajemnego układu klas, iii) określenie algorytmu przeszukiwania niezupełnego, który zapewni dobre przypisanie wzorców do skupień; przeszukanie zupełne wszystkich możliwości jest najczęściej niewykonalne.

Większość algorytmów analizy skupień stosuje jedną z dwóch technik:

- iteracyjny podział w świetle wybranego kryterium, np. minimalizacji rozrzutu wewnątrz skupienia,
- aglomerację hierarchiczną.

Przykładem algorytmu stosującego pierwszą technikę jest analiza skupień metodą *k*-średnich. Hierarchiczna analiza skupień wykorzystuje drugą z wymienionych technik.

Metoda *k*-średnich

W wyniku zastosowania metody *k*-średnich dane podlegają podziałowi horyzontalnemu na *k* rozłącznych skupień [122]. Miarą rozrzutu w obrębie skupienia jest suma kwadratów odległości poszczególnych wektorów danych od centrum skupienia, którym jest wektor średni. Suma tych miar po wszystkich skupieniach jest określana jako średni błąd kwadratowy dla układu *k* skupień. Algorytm startuje od losowego podziału zbioru danych na skupienia. W kolejnych iteracjach wektory są przesuwane między skupieniami, a centra skupień przeliczane. Zadaniem algorytmu jest doprowadzenie do takiego podziału na *k* skupień, który minimalizuje błąd dla całego układu. Objawia się to brakiem dalszych przesunięć wektorów. W podstawowej wersji algorytm wy-

maga podania wartości *k*. Dostępne są procedury heurystyczne umożliwiające automatyczne określenie liczby skupień. Wprowadzają one do analizy takie elementy, jak łączenie i podział istniejących skupień, eliminacja małych bądź skrajnych skupień.

Podział metodą *k*-średnich jest efektywny obliczeniowo i daje bardzo dobre rezultaty w przypadku zwartych, hipersferycznych skupień, które są dobrze rozsunięte w przestrzeni cech. Po zastosowaniu odległości Mahalanobisa w miejsce odległości Euklidesa algorytm radzi sobie również ze skupieniami hiperelipsoidalnymi.

W podstawowej wersji metoda *k*-średnich jest rzadko stosowana w analizie danych z pomiarów czujnikowych. Powodem jest głównie trudność wizualizacji rezultatów. Największym zainteresowaniem cieszy się wzmocniona wersja metody z zastosowaniem logiki rozmytej [155]. Inspiracje pochodzą tu z obszaru modelowania przetwarzania bodźców zapachowych przez człowieka. W klasycznej metodzie *k*-średnich funkcja przynależności przyjmuje tylko dwie wartości – zero określa brak przynależności wektora do skupienia, jeden wskazuje na jego przynależność. Ideą główną metody *fuzzy C-means* jest zastosowanie funkcji przynależności, której wartości są ciągłe i iteracyjnie zmieniane, aż stopień przynależności wzorców do poszczególnych skupień zostanie określony najlepiej. Metoda *k*-średnich jest zwykle stosowana równolegle z innymi, np. z PCA, sieciami Kohonena [156, 157], często instrumentalnie, na jednym z etapów rozpoznawania wzorców, w szczególności jako metoda ekstrakcji cech [50] lub prototypów klas [158]. Xu i inni [123] pokazali przykład sprzężenia analizy skupień metodą *k*-średnich z algorytmem genetycznym do optymalizacji matrycy czujników.

Na koniec warto wspomnieć, że uogólniony algorytm kwantyzacji wektorowej (ang. *vector quantization*) jest równoważny algorytmowi *k*-średnich.

Hierarchiczna analiza skupień

Analiza hierarchiczna prowadzi do uzyskania sekwencji zagnieżdżonych skupień, które można przedstawić w postaci dendrogramu. U podstawy dendrogramu znajdują się wszystkie elementy zbioru danych. Oś pionowa podaje stopień odmienności skupień. Wielkość ta jest na ogół unormowana w przedziale [0, 1]. Poziome linie łączące skupienia określają poziom odmienności, dla którego nastąpiła aglomeracja skupień składowych. Istnieją dwa podstawowe typy algorytmów uzyskania tej sekwencji – aglomeracji i dzielenia. Algorytmy aglomerujące konstruują dendrogram począwszy od liści (pojedynczych elementów zbioru) w kierunku korzenia (całe dane). W każdym kroku łączone są dwa najbliższe skupienia. Podstawą ich wyboru jest kryterium podobieństwa skupień. Zazwyczaj stosowane jest kryterium minimalnej, średniej lub maksymalnej odległości między elementami dwóch skupień. Kształt skupień w bardzo dużym stopniu zależy od przyjętego kryterium. Gdy stosuje się minimalną odległość, skupienia są wydłużone, gdy maksymalną – bardziej zwarte. Algorytmy wykorzystu-

jące podział postępują w kierunku przeciwnym. Są jednak wymagające obliczeniowo i dlatego rzadziej stosowane, choć dla małej liczby skupień dają większą szansę uzyskania interpretowalnych rezultatów.

Literatura przedmiotu zawiera wiele przykładów zastosowania hierarchicznej analizy skupień w analizie danych z czujnikowych pomiarów gazów. Można wyróżnić dwa zasadnicze cele zastosowania tego podejścia. Jeden jest związany z określeniem wyrazistości różnic między wzorcami obiektów poddawanych pomiarom czujnikowym. Przykład klasycznej analizy tego rodzaju, zastosowanej dla nieprzekształconych danych z pomiarów czujkowych można znaleźć w pracy [159]. Szczególnie często hierarchiczna analiza skupień jest stosowana w czujnikowych badaniach obiektów biologicznych. W wielu pracach wykazano tą metodą ewidentną bliskość wzorców dotyczących obiektów zdrowych lub chorych oraz wyraźny dystans między tymi dwoma grupami wzorców. Jako przykład można tu przywołać badania cebuli [159], jagód [160] i krwi ludzkiej [161]. Dzięki zastosowaniu hierarchicznej analizy skupień czytelne dla czujników były różnice między gatunkami, np. dermatofitów [162]. Istnieją też przykłady analizy skupień wzorców pojedynczych substancji chemicznych, np. alkoholi [70].

Drugi główny cel analizy skupień w pomiarach czujnikowych jest związany raczej z samymi sensorami. Metodę tę stosowano do wyróżnienia grup warstw chemoczułych [120, 143] lub czujników [163] odpowiadających podobnie na substancje badane. Przez wybór reprezentantów grup uzyskano selekcję warstw chemoczułych/czujników ze względu na ich konkretne właściwości [143] lub zamierzone zastosowanie [163].

6.2. Metody odczytu zdefiniowanej informacji jakościowej

Istotą pozyskania informacji jakościowej o badanym gazie za pomocą systemu rozpoznawania wzorców jest zaklasyfikowanie wzorca gazu do właściwej kategorii, obejmującej wzorce gazów o właściwościach jakościowych, takich jak badany gaz. Wymaga to skonstruowania reguły podziału przestrzeni cech na określoną liczbę regionów w taki sposób, że w każdym z nich znajdują się wektory danych należące do jednej klasy i nie ma tam wektorów należących do innych klas. Granice między regionami są określane jako powierzchnie rozdzielające lub granice decyzyjne. Określenie zasady podziału przebiega na podstawie zestawu danych uczących, tj. dotyczących gazów o znanych właściwościach jakościowych. Proces ten określa się jako klasyfikację pod nadzorem (z nauczycielem, z próbą uczącą).

Gotowy klasyfikator jest w stanie przypisać wektorowi cech o nieznanej przynależności klasowej etykietę odpowiedniej dla niego klasy. Pochodzi ona ze zbioru etykiet określających klasy uw,zględnione w procesie uczenia. Termin klasyfikacja ma

zatem dwa znaczenia: i) grupowanie obiektów na podstawie pewnych cech, ii) analiza nowych obiektów i przypisywanie ich do już zdefiniowanych klas.

Klasyfikacja z nadzorem jest powszechnie stosowana w systemach rozpoznawania wzorców współpracujących z matrycami czujnikowymi jako metoda pozyskiwania zdefiniowanej informacji jakościowej o badanych gazach.

Problem klasyfikacji wymagający podziału zestawu wektorów danych na k klas można zdefiniować na trzy różne sposoby, tj. jako problem typu:

- każdy przeciw każdemu (ang. *one against one, all against all*); jest to jedno zadanie klasyfikacyjne polegające na podziale zbioru danych na k podzbiorów, odpowiadających poszczególnym klasom;
- jeden przeciw wszystkim (ang. *one against all*); składa się na niego $k - 1$ równoległych zadań klasyfikacji, polegających na podziale zbioru danych na dwie części, tak że do jednej należą elementy zbioru przynależne do jednej klasy, a do drugiego elementy przynależne do wszystkich pozostałych klas;
- hierarchiczny; składa się na niego więcej niż jedno, a mniej niż $k - 1$ zadań klasyfikacji, w których zbiór danych podlega sukcesywnie coraz większemu podziałowi.

Podejście pierwsze jest bardzo wymagające zarówno dla przestrzeni cech, jak i dla klasyfikatora. Aby zapewnić dobre rozwiązanie problemu klasyfikacji, jeden zestaw cech musi zawierać informację umożliwiającą wyróżnienie k klas. Założenie takie może się okazać zbyt optymistyczne, gdy liczba klas jest duża, nie są one zwarte, a odległości między centrami klas są małe. W dziedzinie rozpoznawania gazów na podstawie pomiarów czujnikami problem ten występuje często, zwłaszcza gdy do poszczególnych klas należą mieszaniny gazów o ustalonym składzie jakościowym, lecz składzie ilościowym mieszczącym się w dość szerokim zakresie. Podejście typu każdy przeciw każdemu wymaga też na ogół klasyfikatora o złożonej strukturze, pracującego na wielu zmiennych wejściowych, a więc kosztownego obliczeniowo.

Istotna przewaga podejścia jeden przeciw wszystkim polega na tym, że daje ono możliwość zróżnicowania zbiorów cech dla poszczególnych zadań klasyfikacji. Sprofilowanie zbiorów cech powiększa szanse powodzenia klasyfikacji. Każde zadanie jest rozwiązywane osobnym klasyfikatorem i może być rozwiązywane w innej przestrzeni cech. Ponadto ze względu na binarny charakter podziału można się posłużyć prostszymi klasyfikatorami niż w przypadku jednorazowego podziału na k klas.

Najbardziej złożone, lecz równocześnie najbardziej elastyczne jest rozwiązanie hierarchiczne [87, 155, 164–166]. Jego podstawową zaletą jest możliwość wykorzystania zróżnicowanego stopnia podobieństwa danych w zbiorze i odwzorowania go w strukturze stopniowego podziału tego zbioru. Zadania klasyfikacji w poszczególnych węzłach struktury są z natury mniej złożone i mogą być realizowane różnymi klasyfikatorami oraz na podstawie najbardziej odpowiednich zbiorów cech. W rezultacie dodatkowo pozyskuje się wiedzę o tym, na czym polega główna trudność całego problemu klasyfikacji. Dzi,ęki wymienionym właściwościom rozwiązania hierarchicznego daje ono duże szanse rozwiązania nawet bardzo złożonych problemów kla-

syfikacji. Warto też wspomnieć o możliwości opracowywania struktur hierarchicznych, które służą do rozwiązywania problemów klasyfikacji oraz regresji [18].

Dostępne dziś klasyfikatory w większości mieszczą się w ramach klasyfikacji statystycznej, nawet jeśli są bardziej znane pod innym szyldem, np. jako metody sztucznej inteligencji. Najpowszechniej stosowane w klasyfikacji podejście wykorzystuje koncepcję prawdopodobieństwa *a posteriori* przynależności wektora **x** do klasy *k*:

$$p(k|\mathbf{x}) = \frac{\pi_k p(\mathbf{x}|k)}{\sum_{r=1}^{g} \pi_r p(\mathbf{x}|r)} \tag{6.3}$$

gdzie: prawdopodobieństwo *a priori*, że wektor **x** pochodzi z klasy *k* wynosi π_k, prawdopodobieństwo *a priori*, że wektor **x** pochodzi z klasy $r = 1, ..., g$ wynosi π_r oraz dany jest dyskretny rozkład prawdopodobieństwa $p(\mathbf{x}|k)$ lub funkcja gęstości prawdopodobieństwa $f_k(x)$, opisująca rozkład obserwacji **x** w klasie k, $k = 1, ..., g$. Podobnie interpretuje się $p(\mathbf{x}|r)$.

Aby zminimalizować prawdopodobieństwo błędnej klasyfikacji, wektor obserwacji **x** powinien zostać zaliczony do klasy, którą charakteryzuje największe prawdopodobieństwo *a posteriori* przynależności, $p(k|\mathbf{x})$. Równoważny temu jest wybór klasy, dla której największą wartość przyjmuje iloczyn $\pi_k p(\mathbf{x}|k)$. Jest to tzw. reguła Bayesa.

Dla znanych rozkładów prawdopodobieństwa warunkowego reguła decyzyjna Bayesa pozwala uzyskać klasyfikator optymalny. Oznacza to, że dla określonego prawdopodobieństwa *a priori*, prawdopodobieństwa warunkowego i zadanej funkcji strat żadna inna reguła nie zapewni mniejszego ryzyka, tj. wartości oczekiwanej funkcji strat, np. prawdopodobieństwa popełnienia błędu.

Uzyskanie klasyfikatora optymalnego jest z reguły niemożliwe ze względu na nieznajomość prawdopodobieństw π_k oraz rozkładów prawdopodobieństwa $p(\mathbf{x}|k)$. W praktyce stosowana jest tzw. empiryczna reguła Bayesa, która wykorzystuje oszacowania tych wielkości na podstawie danych. W rozwiązaniu tym na podstawie danych uczących wyznacza się warunkową gęstość prawdopodobieństwa wystąpienia obserwacji dla każdej klasy po kolei. Metody wyznaczania rozkładów można podzielić na parametryczne i nieparametryczne. W pierwszym wypadku zakładana jest postać funkcji gęstości prawdopodobieństwa, np. funkcja opisująca rozkład normalny, której parametry są wyznaczane na podstawie danych. Niestety, podejście to może nie pozwolić na dostatecznie wierne odwzorowanie rzeczywistego rozkładu prawdopodobieństwa w klasach. Ze względu na ten niedostatek stosuje się podejście nieparametryczne lub dużo bardziej złożone rozwiązania mieszane. Do najistotniejszych metod klasyfikacji realizujących wymienione sposoby estymacji prawdopodobieństwa w klasach należą:

- klasyfikatory parametryczne – liniowa i kwadratowa analiza dyskryminacyjna,
- klasyfikatory nieparametryczne – metoda *k*-najbliższych sąsiadów, Parzena, analiza skupień,
- klasyfikatory mieszane – modele mieszanin rozkładów Gaussa (ang. *gaussian mixture models*), generatywne mapowanie topograficzne (ang. *genarative topographic mapping*), modele mieszanin probabilistycznych składowych głównych (ang. *probabilistic PCA mixture*).

Ze względów praktycznych bardzo dużą popularnością cieszą się metody klasyfikacji, w których konstrukcja reguł decyzyjnych ma za podstawę optymalizację ze względu na wybrane kryt,erium, np. błąd klasyfikacji. Wśród nich należy wymienić: klasyfikatory neuronowe, maszyny wektorów podpierających oraz rozwiązania, które dopiero torują sobie drogę w obszarze analizy danych z pomiarów czujnikowych – drzewa decyzyjne, rodziny drzew klasyfikacji i regresji, zwłaszcza lasy losowe.

Przedstawiony w dalszej części rozdziału wybór metod klasyfikacji determinuje tematyka monografii. Wybrano grupy metod przydatne w analizie danych z czujnikowych pomiarów gazów. Dokonując wyboru, brano pod uwagę przede wszystkim częstość stosowania metody, różnorodność kontekstów jej zastosowania oraz perspektywiczność.

6.2.1. Analiza dyskryminacyjna

Wśród klasyfikatorów realizujących parametryczną estymację rozkładów prawdopodobieństwa obserwacji w klasach najprostszym przykładem jest klasyfikator bazujący na założeniu, że obserwacje w klasach pochodzą z wielowymiarowego rozkładu normalnego. Występuje on w dwóch wersjach: liniowej i kwadratowej. Różnica między nimi wynika z przyjętego założenia dotyczącego macierzy kowariancji w klasach.

W przypadku liniowej analizy dyskryminacyjnej (ang. *linear discriminant analysis*, LDA) obowiązuje założenie, że macierze kowariancji w klasach są identyczne. W takiej sytuacji, hiperpowierzchnia rozdzielająca klasy jest hiperpłaszczyzną. Reguła decyzyjna rozstrzygająca o podziale przestrzeni cech korzysta z wartości funkcji dyskryminacyjnej $\delta_k(\mathbf{x})$ w klasach. Wektor $\mathbf{x}$ jest klasyfikowany jako należący do tej klasy, dla której wartość funkcji jest największa. Funkcje dyskryminacyjne są liniowe względem $\mathbf{x}$, stąd nazwa klasyfikatora. LDA jest bardzo atrakcyjna obliczeniowo. Dodatkowo daje możliwość graficznej prezentacji podziału przestrzeni cech, co umożliwia wizualną ocenę uzyskanych rezultatów [72]. Dzięki założeniu o jednakowych macierzach kowariancji w klasach oblicza się kowariancję na podstawie całego zestawu danych. W ten sposób zostają znacznie zmniejszone wymagania względem odpowiedniej liczebności danych w klasach. Przekłada się to wprost na liczbę pomiarów, które muszą zostać wykonane, aby liczba danych była wystarczająca do zbudowania klasyfikatora.

Gdy obserwacje w klasach pochodzą z wielowymiarowego rozkładu normalnego, lecz macierze kowariancji dla klas są różne, funkcja dyskryminacyjna ma postać nieliniową. Wówczas dwie klasy są rozdzielone hiperpowierzchnią kwadratową, i stąd pochodzi nazwa – kwadratowa analiza dyskryminacyjna (ang. *quadratic discriminant analysis*, QDA). Klasyfikator taki umożliwia rozdzielenie klas, których nie da się rozdzielić liniowo. Jest więc przeznaczony do rozwiązywania trudniejszych problemów klasyfikacyjnych niż klasyfikator liniowy. Jego zdefiniowanie wymaga jednak odpowiedniej liczby wektorów cech reprezentujących poszczególne klasy. W literaturze przedmiotu można znaleźć propozycje tzw. *shrinkage methods*, które są rozwiązaniami pośrednimi pomiędzy klasyfikatorami liniowymi i nieliniowymi [167].

Analiza dyskryminacyjna jest powszechnie stosowana w analizie danych z czujnikowych pomiarów gazów. Pozwala uzyskać bardzo dobre rezultaty klasyfikacji. Na przykład w pracy [108] rozróżniono 11 lotnych związków organicznych ze 100% efektywnością na podstawie parametrów sygnału czujnika uzyskanych w przestrzeni stanów. Dzięki dużej efektywności klasyfikacji w połączeniu z prostotą metody analiza dyskryminacyjna bywa stosowana jako metoda odniesienia dla metod bardziej złożonych. Dla przykładu w pracy [88] zastosowano jednoetapową klasyfikację metodą LDA, którą porównano z metodą hierarchiczną. Wykorzystano w niej różne wektory cech w poszczególnych węzłach struktury hierarchicznej. Wykazano przewagę drugiego rozwiązania pod względem poprawności klasyfikacji, jednak kosztem znacznego wzrostu złożoności. Ze względu na szybkość prowadzenia obliczeń analiza dyskryminacyjna bardzo dobrze sprawdza się jako klasyfikator oceniający różne wektory cech w ramach strategii opakowanej selekcji cech [146]. W przypadku niezbyt licznych zbiorów cech wygenerowanych umożliwia ona przeszukanie przestrzeni cech nawet w sposób zupełny w akceptowalnym czasie. W pracy [168] zaproponowano zastosowanie LDA wprost jako metody selekcji cech. Cechom przypisano wagi określające ich udział w wygenerowaniu kierunku, który zapewnia najlepszy rozdział klas po zrzutowaniu na niego wektorów danych. Metoda była lepsza niż sekwencyjna selekcja w przód, a nawet algorytm genetyczny. Jednak autorzy zwrócili uwagę na jej nieprzydatność w przypadku danych wielowymiarowych i małej liczby prób. Idea analizy dyskryminacyjnej inspiruje prace nad metodami pochodnymi. Przykładem jest dyskryminacyjna analiza czynnikowa (ang. *discriminant factorial analysis*), którą zastosowano w pracy [95] do rozpoznawania złożonyc,h mieszanin zanieczyszczeń w powietrzu o silnie zróżnicowanej wilgotności na podstawie pomiarów czujnikowych.

6.2.2. Metoda *k*-najbliższych sąsiadów

Nieznajomość analitycznej postaci rozkładu prawdopodobieństwa w klasach wymaga skorzystania z podejścia nieparametrycznego lub skonstruowania granicy decyzyjnej na podstawie zbioru uczącego według innych zasad. Wśród metod klasyfikacji

opartych na nieparametrycznej estymacji rozkładów prawdopodobieństwa najbardziej znana jest metoda *k*-najbliższych sąsiadów (ang. *k-nearest neighbor*, *k*-NN). Predykcja przynależności wektora danych polega na znalezieniu jego najbliższych sąsiadów i określeniu dominującej wśród nich klasy [169]. Występuje tu jeden parametr – *k*, którego wartość należy zoptymalizować empirycznie, np. w toku procedury *leave-one-out* (patrz rozdz. 7.1) ze względu na kryterium minimalnego błędu klasyfikacji. W najprostszej wersji metody, tj. gdy $k = 1$ (1-*NN*) i gdy założy się, że liczba wektorów danych zmierza do nieskończoności, błąd klasyfikacji uzyskany tą metodą jest nie większy niż dwukrotny błąd klasyfikatora optymalnego (Bayesa). W tym wypadku granica decyzyjna jest liniowa. Dla $k > 1$ granice decyzyjne są bardziej złożone. Metoda *k*-NN ma bardzo dobre właściwości w zakresie klasyfikacji nieliniowej w warunkach ograniczonej liczby danych. Jest w stanie wygenerować silnie lokalne granice decyzyjne, gdy dobierze się odpowiednie wartości *k*. Jej podstawowe ograniczenia wynikają z: i) dużego zapotrzebowania na zasoby pamięci, gdyż wymaga przechowywania całego zbioru danych, ii) dużych nakładów obliczeniowych, gdyż dla każdego klasyfikowanego wektora należy obliczyć i posortować odległości od wszystkich wektorów uczących (złożoność czasowa metody wynosi $O(n^2)$), co oznacza, że po podwojeniu liczby obserwacji czas obliczeń wzrasta czterokrotnie), iii) wrażliwości na skalowanie cech. Najwygodniejszym sposobem redukcji dwóch pierwszych ograniczeń jest zastosowanie jednej z metod redukcji zbioru danych, np. przez wygenerowanie podzbioru prototypów klas.

W analizie danych z pomiarów czujnikowych metoda *k*-najbliższych sąsiadów najczęściej nie jest proponowana jako jedyny algorytm klasyfikacji. Powodem są istniejące ograniczenia, głównie duże wymagania pod względem pamięci i mocy obliczeniowej, które wpływają na możliwość jej implementacji w gotowych systemach czujnikowych. Metoda jest natomiast chętnie stosowana w analizach porównawczych jako odniesienie dla klasyfikatorów nieliniowych działających na innych zasadach, np. parametrycznych, jak mieszane modele gaussowskie [25] czy sieci neuronowych [170, 171] i innych metod. Z reguły wyniki uzyskiwane za pomocną *k*-NN są porównywalne. Ten rodzaj klasyfikatora bywa też proponowany jako niezależny element komitetu klasyfikatorów [172]. Prostota i szybkość algorytmu powoduje, że jest on konkurencyjny jako środowisko testowania różnych hipotez dotyczących konstrukcji optymalnego wektora cech. Metodą *k*-NN posłużono się na przykład do przeanalizowania wpływu różnych technik kompresji zbioru danych uczących na efektywność rozpoznawania wzorców [173]. Z jego zastosowaniem [99] porównano możliwości dyskryminacyjne różnego rodzaju cech. W kilku pracach [17, 125, 127, 146] oceniano możliwości dyskryminacyjne wektorów cech uzyskanych w różnych warunkach ekspozycji na badane gazy, posługując się metodą *k*-NN. Podejmowane są próby modyfikacji i rozwijania metody. Zaproponowano na przykład [174] adaptacyjną wersję metody *k*-NN i wykazano możliwość jej zastosowania w problemach rozróżniania klas związanych z różnymi zakresami stężeń gazów.

6.2.3. Modele mieszanin rozkładów Gaussa

W klasyfikatorach statystycznych stosuje się też bardziej zaawansowane metody estymacji rozkładów prawdopodobieństwa obserwacji w klasach wykorzystujące zalety obu podejść – parametrycznego i nieparametrycznego [29]. Uważa się, że do analizy danych z pomiarów czujnikowych najlepiej stosować modele mieszanin rozkładów Gaussa [175]. Metoda polega na zbudowaniu liniowego złożenia wielu rozkładów, których udział w mieszaninie określają tzw. współczynniki zmieszania (ang. *mixing coefficients*) [23]. Zabieg ten wywołuje miedzy innymi zwiększenie liczby parametrów rozkładu, które można dobierać adaptacyjnie. Możliwa jest przez to budowa bardziej elastycznych, lepiej dopasowanych modeli rzeczywistych rozkładów prawdopodobieństwa wektorów cech w klasach. Wykazano przewagę klasyfikacji gazów z zastosowaniem modeli mieszanin rozkładów Gaussa nad różnymi architekturami sieci neuronowych [176], a nawet maszynami wektorów podpierających [177].

6.2.4. Maszyny wektorów podpierających

Koncepcja tej nieparametrycznej metody klasyfikacji została opracowana przez Vapnika [27]. Inspiracją dla maszyn wektorów podpierających (ang. *support vector machines*, SVM) była potrzeba uzyskania nieliniowego klasyfikatora, który mimo uczenia na niewielkim zbiorze danych dobrze generalizuje w wielowymiarowych przestrzeniach cech. Dla liniowego rozdzielenia klas wzorców poszukiwana jest granica decyzyjna w postaci optymalnej hiperpłaszczyzny, taka że odległość między najbliższym punktem danych (wektorem uczącym) a tą hiperpłaszczyzną (tzw. margines) jest jak największa [178]. Punkty danych, przez które ostatecznie przechodzą granice marginesów, określa się jako wektory podpierające. Odporność klasyfikatora na zakłócenia jest uzyskiwana przez maksymalizację marginesu, co predestynuje do uzyskania dobrych właściwości uogólniających. Podstawowa idea klasyfikatora SVM jest jednak taka, by oryginalne, liniowo nieseparowane dane przedstawić w nowej, bardziej wielowymiarowej przestrzeni, w której uzyskuje się liniową separowalność. W nowej przestrzeni należy znaleźć optymalną hiperpłaszczyznę, która rozdziela klasy. Przekształcenie oryginalnej przestrzeni w nową jest realizowane za pomocą tzw. funkcji jądrowych (ang. *kernel function*). W pierwotnej wersji SVM stosowano liniową funkcję jądrową i klasyfikator miał charakter liniowy. Kluczowym dla rozwoju metody był pomysł zastosowania nieliniowych funkcji jądrowych, np. wielomianowej, radialnej czy sigmoidalnej. W wyniku tego zyskano możliwość przekształcenia problemu silnie nieliniowego w przestrzeni oryginalnej w problem liniowy w nowej przestrzeni pod warunkiem wyboru odpowiedniego odwzorowania. Z matematycznego punktu widzenia problem optymalizacji jest tutaj problemem z zakresu programowa-

nia kwadratowego z jednym rozwiązaniem. Liczba wektorów podpierających jest określana automatycznie w toku procesu optymalizacji. W oryginale metodę opracowano dla problemu dwóch klas. Obecnie istnieją jej wersje dla problemu *k* klas.

Maszyny wektorów podpierających mają wiele istotnych zalet. Pozwalają uzyskać nieliniowe klasyfikatory, dobrze generalizujące w wysoko wielowymiarowych przestrzeniach cech mimo wykorzystania ograniczonych zbiorów danych. Klasyfikatory te są bardzo stabilne, tzn. odporne na zakłócenia występujące w zbiorze danych uczących. Ze względu na posługiwanie się niewielkim podzbiorem danych pełniących rolę wektorów podpierających nakłady obliczeniowe w etapie testowania są małe, również w przypadku dużych zbiorów danych [179].

Ze względu na omówione zalety maszyny wektorów podpierających są ugruntowaną metodą analizy danych z czujnikowych pomiarów gazów. Podstawowa praca w tym obszarze [65] podaje podstawy teorii SVM oraz przykłady zastosowania. Często w badaniach stosowano SVM do rozwiązania problemów klasyfikacji zarówno z podziałem na dwie, jak i na wiele klas. Uzyskane rezultaty porównywano z wynikami sieci neuronowych i wykazywano zazwyczaj przewagę SVM [178, 180]. Podjęto problem wpływu różnych czynników, takich jak liczba zmiennych wejściowych, parametr funkcji jądrowej i parametr regularyzacji na możliwości uogólniające SVM [179], wskazując kluczową rolę ostatniego czynnika. Analizowano również wpływ sposobu skalowania danych na rezultaty rozpoznawania i obliczania stężeń LZO [71]. W kilku pracach przedstawiono zastosowanie modyfikacji SVM, określanej jako maszyna wektorów związanych (ang. *relevance vector machine*) do rozróżniania mieszanin gazów. Przykładowe porównania tej metody z SVM wskazały na podobne możliwości uogólniania z użyciem znacznie mniejszej liczby funkcji jądrowych, a zatem i przy mniejszych kosztach obliczeniowych [181] bądź na przewagę SVM [182].

6.2.5. Drzewa decyzyjne (klasyfikacyjne)

Analiza danych metodą drzewa decyzyjnego polega na sukcesywnym podziale całego zbioru danych wielowymiarowych. Rozwiązanie to przedstawili po raz pierwszy Breiman i in. w 1984 [233]. Początkiem drzewa jest korzeń, w którym znajduje się cały zbiór danych. W każdym kroku podziału (węźle drzewa) wyłaniane są podzbiory danych, które mogą być dalej dzielone. Ostatni podział prowadzi do uzyskania zbiorów wzorców należących do osobnych klas. Są to tzw. liście. Najczęściej stosowany jest podział w systemie binarnym, tzn. z każdego węzła, który nie jest liściem, odchodzą dwie gałęzie. Podział w węźle odbywa się na podstawie jednej zmiennej, wybranej ze wszystkich zmiennych wejściowych w ten sposób, że zapewniony jest najlepszy podział podzbioru danych, który trafił do węzła. Najlepszy oznacza zapewniający najmniejszą różnorodność w obrębie nowo powstałych podzbiorów danych. Stosuje się wiele miar różnorodności, a zbiór cech jest przeszukiwany w sposób zupełny ze

względu na wybrane kryterium podziału. Drzewa decyzyjne można z powodzeniem przedstawić w postaci zestawu reguł *jeżeli … to*, pozwalającego na wykonanie podziału zbioru danych. Algorytm budowy drzewa decyzyjnego jest nieparametryczny i mało wymagający pod względem obliczeniowym.

Drzewa decyzyjne są najpowszechniej stosowanym rodzajem klasyfikatorów w obszarze uczenia maszynowego. W dziedzinie pomiarów czujnikowych są jednak mało znane i nie należą jeszcze do klasycznych technik analizy danych. Być może, dlatego że mimo dużej uniwersalności drzewa decyzyjne wymagają zazwyczaj stabilizacji i poprawy właściwości predykcyjnych. Efekt taki zyskuje się np. przez zastosowanie rozmaitych technik, takich jak przycinanie, kroswalidacja czy *boosting* (patrz rozdz. 7). Wreszcie bardzo dobrym rozwiązaniem są zespoły klasyfikatorów, których pojedyncze elementy stanowią drzewa.

Podjęto pojedyncze próby zastosowania drzew klasyfikacyjnych, a ściślej drzew klasyfikacji i regresji (ang. *classification and regression trees*, CART), do analizy danych z czujnikowych pomiarów gazów [133, 134]. W cytowanych pracach uzyskano drzewo o bardzo prostej strukturze, które w dwóch krokach realizowało podział zbioru wzorców na trzy klasy z błędem rzędu kilku procent. Rozwiązania ujawniły też możliwość spojrzenia na algorytm jako bardzo efektywne narzędzie selekcji cech. Drzewa klasyfikacji z sukcesem realizowały zadanie po pięciokrotnym [134], a nawet dziesięciokrotnym [133] zredukowaniu liczby składowych wektora cech. Bardziej zawansowany przykład zastosowania drzew decyzyjnych w analizie danych czujnikowych, dotyczący możliwości tego algorytmu jako selekcjonera cech na potrzeby rozpoznawania LZO, przedstawiono w pracy [169]. Badania porównawcze drzew klasyfikacyjnych (C4.5 i CART) i innych klasyfikatorów (perceptron wielowarstwowy i fuzzy ARTMAP) w zakresie rozpoznawania gazów na podstawie pomiarów czujnikowych pokazały, że metoda ta jest atrakcyjna jako klasyfikator oraz jako metoda redukcji przestrzeni cech na potrzeby innych klasyfikatorów [131].

6.2.6. Zespoły klasyfikatorów

Ciesząca się coraz większym zainteresowaniem zaawansowana technika klasyfikacji polega na zastosowaniu nie jednego, lecz zespołu klasyfikatorów (ang. *committee machine*). Źródłem tego pomysłu było dostrzeżenie, że nie istnieje jedno, optymalne podejście do problemu klasyfikacji i w związku z tym warto połączyć zróżnicowane podejścia i metody [23, 24]. Wykazano, że dzięki zastosowaniu zespołu klasyfikatorów można znacznie zwiększyć efektywność klasyfikacji, zwłaszcza gdy poszczególne klasyfikatory są słabe, np. trafność niewiele powyżej 50%. Stosowanie zespołu klasyfikatorów jest szczególnie korzystne, jeżeli poszczególne klasyfikatory są od siebie w dużej mierze niezależne. Do istotnych źródeł niezależności klasyfikatorów należą [24]:

- zastosowanie klasyfikatorów opartych na informacji pochodzącej z różnych źródeł, np. pomiarów wykonanych czujnikami, które różnią się mechanizmem powstawania odpowiedzi,
- dysponowanie różnymi zbiorami danych uczących, np. zebranymi w pewnych odstępach czasu czy w innych warunkach, jak wyniki pomiarów gazów w powietrzu o różnej wilgotności czy temperaturze powietrza,
- opracowanie klasyfikatorów pracujących z innymi zestawami cech, które różnią się pod względem pojemności informacyjnej o poszczególnych klasach, np. cechy związane z odpowiedzią czujników w stanie ustalonym i nieustalonym, z różnymi czujnikami itp.,
- zastosowanie klasyfikatorów, opracowanych na podstawie tego samego zbioru danych, lecz wykorzystujących różne metody klasyfikacji, np. klasyfikator neuronowy, analiza dyskryminacyjna, klasyfikacja metodą *k*-najbliższych sąsiadów.

Istnieje wiele metod agregacji wyników poszczególnych klasyfikatorów w zbiorczą ocenę zespołu [24]. Zasadniczo wyróżnia się tu metody statyczne i z zastosowaniem uczenia. Do pierwszych należą przede wszystkim różne systemy głosowania i uśredniania [183]. Obecnie wiadomo, że dają one gorsze wyniki niż podejście pozwalające na wypracowanie sposobu podjęcia wspólnej decyzji w rezultacie procesu uczenia [184]. Zasadniczym instrumentem realizacji tego podejścia są wagi kształtowane podczas trenowania zespołu i przypisywane rezultatom poszczególnych klasyfikatorów. Innym ciekawym rozwiązaniem są np. komitety lokalnych ekspertów.

Nie bez znaczenia dla wyboru metody fuzji jest postać wyniku pracy pojedynczego klasyfikatora. Poczynając od najbardziej pojemnych informacyjnie, wyróżnia się następujące ich rodzaje: i) miara ufności (zestaw prawdopodobieństw przynależności wzorca do poszczególnych klas), ii) ranga (ranking klas pod względem szansy przynależności wzorca do nich), iii) etykieta klasy (wskazanie klasy, do której przynależy wzorzec) [24].

Należy też wspomnieć, że stosowane są różne architektury zespołu klasyfikatorów, z których podstawowe to architektura: i) równoległa, ii) kaskadowa, iii) hierarchiczna. W praktyce najczęściej spotykane jest pierwsze rozwiązanie, w którym wszystkie klasyfikatory pracują niezależnie od siebie, a uzyskane przez nie wyniki są uzgadniane w ocenę łączną. W przypadku architektury kaskadowej mamy do czynienia z sekwencją klasyfikatorów, poczynając od najmniej dokładnych, lecz zarazem najmniej wymagających, skończywszy na najbardziej złożonych, lecz kosztownych pod względem obliczeniowym, jak też zapotrzebowania na dane uczące. Wraz z uzyskiwaniem wyników z kolejnych stopni kaskady ulega redukcji liczba klas, do których może potencjalnie przynależeć analizowany wzorzec. Najbardziej zaawansowana jest architektura hierarchiczna. Poszczególne klasyfikatory są połączone w ramach struktury drzewiastej, zajmując miejsce w jej węzłach. Mogą się znacznie różnić zarówno pod względem metody klasyfikacji, jak i liczby oraz rodzaju cech, z których korzystają. Choć złożona i wymagająca zarówno pod względem obliczeniowym, jak i zapo-

trzebowania na dane uczące, struktura taka jest najbardziej efektywna ze względu na możliwości wykorzystania informacji zawartej w różnych rodzajach cech.

Lasy losowe

Szczególnym przykładem zespołu klasyfikatorów są lasy losowe. Koncepcyjnie wywodzą sie one z algorytmu *baggingu* (patrz rozdz. 7), lecz wprowadzono tu dodatkowy element losowości. Lasy losowe są rodzinami drzew losowych. Nie stosuje się w tym wypadku innego rodzaju klasyfikatorów. Każde drzewo jest konstruowane na podstawie innej pseudopróby losowej, pobieranej z danych wielowymiarowych, lecz dodatkowo różny jest sposób konstruowania poszczególnych drzew. O ile w przypadku klasycznych drzew klasyfikujących podział w węźle jest najlepszym podziałem dla zbioru wszystkich zmiennych wejściowych, o tyle w przypadku lasu podstawą podziału jest wylosowany podzbiór zmiennych. Wielkości tego podzbioru wraz z liczbą drzew w lesie to jedyne parametry wymagające zdefiniowania. Oba parametry można optymalizować ze względu na kryterium minimalnego prawdopodobieństwa błędu klasyfikacji. Lasy nie są jednak szczególnie czułe na wartości wymienionych parametrów. Klasyfikacja wektora danych odbywa się wyłącznie za pomocą tych klasyfikatorów, w których budowie nie brał on udziału (nie został wylosowany). Agregacja wyników uzyskanych przez poszczególne klasyfikatory przebiega na podstawie głosowania większościowego.

Z racji przyjętej w lasach losowych zasady podziału w węźle bardzo dobrze nadają się one do rozwiązywania problemów klasyfikacyjnych, gdy liczba zmiennych wejściowych jest bardzo duża, np. rzędu kilku tysięcy. Stąd algorytmy te są na szeroką skalę stosowane w badaniach genetycznych, w szczególności do analizy matrycy DNA. Zasada ta ma też inne, bardzo interesujące konsekwencje. Umożliwia mianowicie uzyskanie rankingu poszczególnych atrybutów wektora cech pod względem rozwiązywanego problemu klasyfikacji. Ocena jest wykonywana dla poszczególnych klasyfikatorów i dotyczy tych wektorów, które nie stanowiły elementów pseudopróby wylosowanej i zastosowanej do konstrukcji klasyfikatora. Automatyczne opracowanie rankingu atrybutów jest atutem przemawiającym za zastosowaniem lasów losowych do analizy danych z czujnikowych pomiarów gazów.

Mimo niewątpliwych zalet zespoły klasyfikatorów są nadal rzadko proponowane jako metoda rozpoznawania wzorców w czujnikowych pomiarach gazów. Wydaje się, że w zakresie rozważanych dotychczas problemów satysfakcjonujące wyniki klasyfikacji są uzyskiwane za pomocą pojedynczych, prostszych klasyfikatorów. Jest tak przede wszystkim dzięki poświęceniu odpowiedniej uwagi wyłonieniu odpowiednich parametrów sygnałów czujnikowych oraz wykonywanym następnie zabiegom selekcji i ekstrakcji cech. Budowa zespołu może się wiązać ze znacznie większym nakładem pracy niż wyuczenie jednego klasyfikatora, zwłaszcza w sytuacji, gdy zespół składa się z klasyfikatorów różnych typów (np. MLP, RBFNN, *k*-NN, GMM i probabili-

styczna PCA [172]) lub są one trenowane na różnych zestawach danych. Ze względu na dużą efektywność poszczególnych klasyfikatorów uzyskana w zamian poprawa efektywności rozpoznawania wzorców względem najlepszego klasyfikatora w zestawie bywa niewielka, np. kilka procent [116, 172] lub wręcz żadna [185]. Wartość dodana wynikająca z zastosowania zespołu klasyfikatorów uwidocznia się najbardziej w warunkach małej efektywności klasyfikatorów składowych. W rozwiązaniach takich jak rodziny drzew klasyfikacji lub lasy losowe [132] dodatkową zaletą jest wbudowana w algorytm konstrukcji klasyfikatora selekcja cech, dzięki czemu nie ma konieczności poświęcania temu zagadnieniu osobnej uwagi.

Można się spodziewać wzrostu zainteresowania zespołami klasyfikatorów w tych obszarach zastosowania, gdzie złożoność problemów klasyfikacji rozwiązywanych przez systemy rozpoznawania wzorców w systemach czujnikowych jest bardzo duża. Pokazano [132], że poprawa efektywności klasyfikacji (w tym przypadku las losowy vs. SVM) może sięgać kilkunastu procent dla tzw. trudnych danych, dotyczących złożonych mieszanin gazów nieznacznie różniących się od siebie. Interesujące rozwiązanie analizowano w pracy [49]. Zaproponowano zasadę organizacji i podejmowania decyzji przez zespół klasyfikatorów, które pozwalają wykonać jednocześnie jakościową i ilościową ocenę odorową zanieczyszczeń na podstawie pomiarów czujnikowych. Jeszcze innym istotnym czynnikiem stymulującym stosowanie zespołów klasyfikatorów jest przechodzenie w stadium realizacji komercyjnych systemów czujnikowych [186]. Udział czynnika losowego w pomiarach nielaboratoryjnych i jego wpływ na efektywność rozpoznawania gazów jest bardzo istotny. Oparcie predykcji dotyczącej właściwości badanego gazu na podstawie szerszej niż wynik jednego klasyfikatora zwiększa zatem poziom zaufania do uzyskanego wyniku. W pracy [76] zaproponowano technikę kompensacji dryfu czujników z zastosowaniem zespołu klasyfikatorów. Pokazano, że opracowane rozwiązanie jest skuteczniejsze niż inne. Nie wymaga ono żadnych założeń dotyczących charakteru dryfu i jest niezależne od rodzaju klasyfikatora podstawowego, stosowanego do rozpoznawania gazów.

6.2.7. Inne oryginalne metody pozyskiwania informacji jakościowej o badanych gazach

W literaturze przedmiotu można znaleźć rozwiązania problemów rozpoznawania gazów, które nie wykorzystują klasyfikatorów w sensie ścisłym. Rozwiązania te najczęściej mieszczą się w obszarze tzw. dopasowywania wzorców (ang. *pattern matching*) [24, 187]. Jego istotą jest porównanie wzorca badanego gazu ze wzorcami znanych gazów, które są przechowywane w bibliotece wzorców. Porównania dokonuje się według zadanego algorytmu i z wykorzystaniem określonych kryteriów dopasowania. Podejście takie stwarza odpowiednie ramy dla niestandardowego traktowa-

nia sygnału czujnika jako nośnika informacji o badanych gazach. Interesujące propozycje z tego zakresu to np. rozpoznawanie oparte na rzadkiej reprezentacji sygnału (ang. *sparse representation*) [101] bądź na porównywaniu pochodnej sygnału czujnika z modulacją termiczną w określonych przedziałach czasu [83], czy porównywanie zestawów parametrów modelu odpowiedzi czujnika w przestrzeni stanów [109]. W literaturze przedmiotu można znaleźć jeszcze inne rozwiązania problemu rozpoznawania gazów. Bardzo ciekawy przykład intuicyjnego i prostego pomysłu został zainspirowany algorytmem Hopfielda – propozycją modelu przetwarzania bodźców zapachowych przez człowieka. Pomysł polegał na określeniu ilościowym zanieczyszczenia na podstawie odpowiedzi wielu czujników według krzywych kalibracji opracowanych dla różnych gazów [188]. Badana próba była uważana za reprezentującą ten gaz, którego krzywe kalibracji pozwoliły uzyskać najbardziej zgodne wyniki oszacowania stężenia. Chemiczne inspiracje znalazły się u podstaw metody przedstawionej w pracy [87]. Zbudowano hierarchiczną strukturę klasyfikującą i w jej poszczególnych węzłach umieszczono różne profile temperaturowe sygnału czujnika. W zależności od stopnia zagłębienia w strukturę profile pozwalały rozróżnić: czyste powietrze, lotne związki organiczne i nieorganiczne, następnie kategorie związków organicznych, np. alkany, związków aromatyczne, alkohole czy ketony, wreszcie poszczególne związki. Proces klasyfikacji przebiegał przez porównanie profili temperaturowych otrzymanych dla badanej substancji z profilami umieszczonymi w strukturze hierarchicznej.

6.3. Metody odczytu zdefiniowanej informacji ilościowej

Można wyróżnić dwa podejścia do ilościowego określania gazów na podstawie pomiarów czujnikami nieselektywnymi. Ich podstawy zostały opracowane w ramach dziedziny określanej jako kalibracja wielowymiarowa (ang. *multivariate calibration*) [140]. Dziedzina ta należy do chemometrii. Jest to niezależna gałąź chemii, która wykrystalizowała się we wczesnych latach siedemdziesiątych XX wieku w wyniku potrzeby wprowadzenia do tej nauki wielowymiarowych metod obliczeniowych.

Podejście określane jako klasyczne charakteryzuje się tym, że wielkość mierzona jest wyrażana jako funkcja stężeń badanych substancji. Jego realizacja w dziedzinie pomiarów czujnikowych polega na modelowaniu wybranego parametru sygnału czujnika jako funkcji składu mieszaniny gazów. Podstawową zaletą tego podejścia jest dostępność interpretacji fizycznej. Dzięki zjawisku wykorzystywanemu w przyrządzie pomiarowym bodziec wywołuje odpowiedź, odpowiedź jest zatem modelowana jako funkcja bodźca. Podstawowa wada rozwiązania klasycznego polega na konieczności odwracania modelu w celu obliczenia stężenia substancji. Jest to niekorzystne pod

względem poprawności oszacowania stężenia, zwłaszcza gdy rozpatruje się przekształcenia nieliniowe. Czasami przekształcenia takie są wręcz niewykonalne.

Idea alternatywnego podejścia, nazywanego odwróconym, polega na wyrażeniu wielkości badanej jako funkcji wielkości mierzonych. W zastosowaniu do pomiarów czujnikowych podejście to realizuje odwzorowanie parametrów sygnałów wyjściowych czujników w stężenia składników mieszaniny. Propozycja ta jest bardziej pragmatyczna i praktyczna zarazem. Brak konieczności odwracania modelu w celu wyznaczenia wielkości badanej zwalnia z ograniczeń co do złożoności modeli i sankcjonuje korzystanie z wielu zmiennych wejściowych. Dzięki temu mogą powstawać odwzorowania pozwalające szacować stężenia składników mieszanin z niewielkim błędem. Podstawową wadą kalibracji odwróconej jest brak zależności przyczynowo-skutkowej między zmiennymi objaśniającymi i objaśnianymi. Uzyskane modele nie są jednak pozbawione walorów poznawczych. Na podstawie udziału zmiennych wejściowych w wyjaśnianiu zmiennej wyjściowej można na przykład wnioskować o ich zawartości informacyjnej. Rozwiązania typu odwróconego dominują w nowoczesnej chemii instrumentalnej (np. spektroskopia w bliskiej podczerwieni (NIR), spektroskopia Ramana), gdzie kluczowe dla uzyskania wyniku analizy jest posługiwanie się wielowymiarową reprezentacją badanych gazów. Ze względu na szereg analogii do dyspozycji jest obszerny zestaw metod, które można zaadaptować na potrzeby pomiarów czujnikowych [61].

Warto również wspomnieć, że z natury odwzorowanie cech w zmienne ilościowe określające gazy ma charakter ciągły. Proponowane są jednak również – interesujące ze względów praktycznych – rozwiązania z uwzględnieniem dyskretyzacji dziedziny miar ilościowych [189].

6.3.1. Podejście klasyczne

W najbardziej atrakcyjnej obliczeniowo koncepcji odpowiedzi czujnika nieselektywnego na mieszaninę substancji przyjmuje się, że sygnał czujnika zarejestrowany w wyniku ekspozycji na taką mieszaninę jest liniowym złożeniem odpowiedzi tego czujnika na poszczególne jej składniki [37]. Tak wyrażoną zasadę superpozycji liniowej przedstawia równanie:

$$r = a_1 c_1 + a_2 c_2 + ... + a_k c_k + a_0 + \varepsilon \tag{6.4}$$

gdzie: r jest parametrem sygnału wyjściowego czujnika, c_1, c_2, …, c_k są stężeniami poszczególnych składników mieszaniny k-składnikowej, a_1, a_2, …, a_k reprezentują czułość czujnika na poszczególne składniki mieszaniny, a_0 jest bazową wartością sy-

gnału czujnika, ε jest składnikiem losowym, o którym zakłada się, że pochodzi z rozkładu $N(0, 1)$.

Jeżeli przedstawione założenie jest spełnione, to określenie składu ilościowego mieszaniny k-składnikowej o znanym składzie jakościowym wymaga matrycy składającej się, z co najmniej n czujników (założywszy, że każdy czujnik dostarcza tylko jedną cechę). Dla n elementów matrycy można zbudować układ n równań:

$$\begin{aligned} r^{\mathrm{I}} &= a_1^{\mathrm{I}} c_1 + a_2^{\mathrm{I}} c_2 + \ldots + a_k^{\mathrm{I}} c_k + a_0^{\mathrm{I}} + \varepsilon \\ r^{\mathrm{II}} &= a_1^{\mathrm{II}} c_1 + a_2^{\mathrm{II}} c_2 + \ldots + a_k^{\mathrm{II}} c_k + a_0^{\mathrm{II}} + \varepsilon \\ &\vdots \\ r^{n} &= a_1^{n} c_1 + a_2^{n} c_2 + \ldots + a_k^{n} c_k + a_0^{n} + \varepsilon \end{aligned} \tag{6.5}$$

gdzie I, II, ..., n wskazuje na czujnik. Układ ten w uproszczonej formie ma postać:

$$\mathbf{r} = \mathbf{A}\mathbf{c} + \mathbf{e} \tag{6.6}$$

Współczynniki w równaniach są określane w wyniku aproksymacji zależności **r** od **c** na podstawie pomiarów wykonanych dla pojedynczych składników mieszaniny. Jeżeli $n = k$ i układ równań jest oznaczony, to jego rozwiązanie można łatwo znaleźć przez odwrócenie macierzy **A**. Prowadzi to do oszacowania stężeń k składników mieszaniny

$$\hat{\mathbf{c}} = \mathbf{A}^{-1}\mathbf{r} \tag{6.7}$$

Gdy liczba czujników jest większa, podejście algebraiczne jest niewystarczające i konieczne jest skorzystanie z metod regresji.

Niestety założenie o liniowości odpowiedzi czujnika na badany gaz oraz addytywności odpowiedzi na poszczególne składniki mieszaniny, choć często przyjmowane w pomiarach czujnikowych [41], jest uzasadnione wyłącznie w zakresie małych stężeń. W szerokim zakresie stężeń charakterystyki czujników są na ogół nieliniowe, a ich odpowiedzi na mieszaniny gazów nie są liniowym złożeniem odpowiedzi na składniki mieszanin. Do głównych strategii obliczeniowych stosowanych w takich okolicznościach należą:

- wymuszenie superpozycji w wyniku odpowiedniego przekształcenia danych, np. przez znalezienie cechy, tj. parametru sygnału czujnika, dla którego spełniona jest zasada superpozycji,

• przyjęcie nieliniowego modelu odpowiedzi czujnika na pojedynczy składnik mieszaniny i poszukiwanie superpozycji liniowej bądź nieliniowej takich modeli; do najczęściej rozważanych należą tu:

– funkcja potęgowa [37]:

$$r = a_1 c^{b_1} + a_0 + \varepsilon \tag{6.8}$$

– funkcja wykładnicza [146]:

$$r = a_1 b_1^c + a_0 + \varepsilon \tag{6.9}$$

– funkcja logarytmiczna [190]:

$$r = a_1 \ln(c) + \varepsilon \tag{6.10}$$

– model izotermy Langmuira [37]:

$$r = \frac{b_1 a_1 c}{1 + a_1 c} + a_0 + \varepsilon \tag{6.11}$$

gdzie r jest odpowiedzią czujnika, c jest stężeniem badanego gazu, a_0, a_1 i b_1 są współczynnikami w modelu;

• poszukiwanie modelu odpowiedzi czujnika na mieszaninę, który nie jest superpozycją modeli odpowiedzi na składniki mieszaniny.

6.3.2. Podejście odwrócone

Alternatywą dla modelowania odpowiedzi czujnika jako funkcji składu badanej mieszaniny jest wyrażenie stężeń składników lub ilościowego parametru mieszaniny jako funkcji odpowiedzi czujników lub innego rodzaju cech. Łatwo zauważyć, że podejście odwrócone mieści się w koncepcji analizy wzorców o charakterze ilościowym. Rolę zestawu zmiennych wejściowych z powodzeniem może pełnić wektor cech. W zastosowaniu do ilościowego określania gazów na podstawie pomiarów czujnikowych sparametryzowany model ilościowy pozwala obliczyć wartości miary ilościowej gazu reprezentowanego konkretnym wzorcem.

Model odwrócony w najprostszej wersji określa stężenie jednego składnika mieszaniny jako złożenie liniowe odpowiedzi czujników i ma postać:

$$c = b_1 r_1 + b_2 r_2 + \ldots + b_n r_n + b_0 \tag{6.12}$$

gdzie: c jest stężeniem substancji, r_1, r_2, …, r_n są odpowiedziami poszczególnych czujników w matrycy n-elementowej, b_1, b_2, …, b_n określają udział poszczególnych czujników w określeniu stężenia substancji. Zależność wyrażona modelem nie odpowiada kierunkowi bodziec–skutek, to bowiem stężenie gazu wywołuje odpowiedź czujnika, a nie odwrotnie. Niemniej jednak wartości współczynników pozwalają ocenić, który czujnik niesie najwięcej informacji ilościowej o badanym gazie.

Opis stężeń składników k-składnikowej mieszaniny, a tym bardziej opis k różnych parametrów ilościowych mieszaniny na ogół wiąże się z budową osobnego modelu dla każdego składnika czy parametru:

$$\begin{aligned} c^{\mathrm{I}} &= b_1^{\mathrm{I}} r_1 + b_2^{\mathrm{I}} r_2 + \ldots + b_n^{\mathrm{I}} r_n + b_0^{\mathrm{I}} + e \\ c^{\mathrm{II}} &= b_1^{\mathrm{II}} r_1 + b_2^{\mathrm{II}} r_2 + \ldots + b_n^{\mathrm{II}} r_n + b_0^{\mathrm{II}} + e \\ &\vdots \\ c^{k} &= b_1^{k} r_1 + b_2^{k} r_2 + \ldots + b_n^{k} r_n + b_0^{k} + e \end{aligned} \tag{6.13}$$

gdzie I, II, ... i, k wskazuje na składnik mieszaniny lub parametr ilościowy. Modele te nie funkcjonują jednak jako układ równań. Oszacowanie współczynników dokonywane jest metodami regresyjnymi na podstawie wyników pomiarów mieszanin o ustalonym składzie jakościowym, lecz różnych składach ilościowych.

Z równania (6.12) wynika, że liczba prób gazu, których pomiary należy wykonać w celu sparametryzowania modelu ilościowego zależy od liczby zmiennych, tj. od liczby cech, z których składa się wzorzec ilościowy. Ze względu zatem na koszt i czas potrzebny na wykonanie pomiarów korzystne jest posługiwanie się modelami z jak najmniejszą liczbą zmiennych wejściowych. To spostrzeżenie ma charakter ogólny i dotyczy wszystkich rodzajów modeli realizujących koncepcję odwróconej kalibracji wielowymiarowej w pomiarach czujnikowych. Jest ono jedną z głównych przyczyn poszukiwania jak najmniejszego zbioru cech, który przenosi informację zawartą w sygnałach wyjściowych czujników z jak najmniejszą stratą.

Istnieją metody pozwalające na opis stężeń więcej niż jednego składnika za pomocą jednego modelu matematycznego, np. sztuczne sieci neuronowe, regresja metodą cząstkowych najmniejszych kwadratów. Z praktyki badawczej wynika jednak, że uzyskiwane w taki sposób oszacowania ilościowego składu mieszaniny są obciążone większym błędem niż przy opisie każdego składnika osobnym modelem. Dlatego też modele z więcej niż jedną zmienną wyjściową są stosowane rzadko.

6.3.3. Analiza regresji

Zespół metod służących do ilościowego odwzorowania statystycznych zależności między zmiennymi określa się jako metody analizy regresji. W klasycznej wersji metody te pozwalają na uzyskanie parametrycznych modeli zależności zmiennej objaśnianej (zależnej) od zmiennych objaśniających (niezależnych). Szereg metod regresji znalazło zastosowanie w ilościowym określaniu gazów na podstawie pomiarów czujnikowych.

Regresja liniowa

Ogólny model liniowy (ang. *general linear model*) ma postać:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6.14}$$

gdzie: $\mathbf{Y}$ jest macierzą, której kolumny stanowią zmienne zależne, $\mathbf{X}$ jest macierzą danych, $\boldsymbol{\beta}$ jest wektorem parametrów modelu, a $\boldsymbol{\varepsilon}$ reprezentuje składnik losowy o rozkładzie normalnym $N(1, 0)$. Najczęściej stosowaną w praktyce odmianą tego modelu jest tzw. model regresji liniowej wielokrotnej (*multiple linear regression*, MLR):

$$y = \alpha_0 + \sum_{k=1}^{K} \alpha_k x_k + \varepsilon \tag{6.15}$$

gdzie: x_k jest zmienną objaśniającą, y jest zmienną objaśnianą, K jest liczbą zmiennych objaśniających, α_0, α_k są współczynnikami modelu, ε reprezentuje czynnik losowy.

Model regresji wielokrotnej służy do odwzorowania zależności liniowej między wieloma zmiennymi objaśniającymi i jedną zmienną objaśnianą. Parametry modelu są szacowane na podstawie danych eksperymentalnych. Istnieje kilka metod szacowania parametrów równania regresji. Najważniejsze z nich to metody: i) najmniejszych kwadratów, ii) największego prawdopodobieństwa, iii) Bayesa. Opracowano też szereg metod i narzędzi oceny modeli regresji. Podstawowe mierniki jakości modelu to statystyka F i współczynnik determinacji, R^2. Pierwsza z nich jest miarą adekwatności modelu, druga zaś służy do oceny stopnia jego dopasowania do danych. Inna miara, statystyka t, jest podstawą oceny istotności parametrów modelu. Umożliwia eliminację nieistotnych zmiennych i uproszczenie struktury modelu. W rezultacie uzyskuje się poprawę jego właściwości uogólniających. Zwraca się uwagę, że model regresji wielokrotnej skutecznie odwzorowuje zależność ilościową między zmiennymi, jeżeli: i) liczba istotnych zmiennych niezależnych jest niewielka, rzędu kilku, ii) nie są one silnie współliniowe, iii) występuje ewidentna zależność między zmiennymi niezależnymi a zmienną zależną. W przypadku nadmiaru

liczby zmiennych niezależnych w stosunku do liczby obserwacji występuje ryzyko zbytniego dopasowania modelu do danych.

Metoda regresji wielokrotnej jest atrakcyjna ze względu na prostotę obliczeniową i interpretacyjną. Niemniej jednak w praktyce analizy danych z pomiarów czujnikowych model regresji wielokrotnej w czystej formie nie jest spotykany często. Podstawową przyczyną jest występujący zazwyczaj problem współliniowości parametrów sygnałów czujnikowych. Jak wcześniej wspomniano, wynika on z zachodzących zakresów selektywności czujników. Na przykład Boeker [191] zastosował MLR do obliczania stężeń amoniaku w wilgotnym powietrzu na podstawie odpowiedzi dwóch czujników. Uzyskane wyniki były zadowalające ze względu na różnice we właściwościach czujników i ich zbliżoną do liniowej charakterystykę. W pracy [14] zastosowano MLR do określania stężeń w dwuskładnikowych mieszaninach LZO w powietrzu. Dla poszczególnych składników zbudowano osobne modele na podstawie odpowiedzi matrycy 15 czujników półprzewodnikowych. Liczba zastosowanych wzorców gazów znacznie przewyższała liczbę cech. Uzyskano błędy względne predykcji stężeń mniejsze niż 5%. Niekonwencjonalny przykład zastosowania MLR podano w pracy [148]. Za pomocą modelu pokazano, że pierwszą składową główną wielowymiarowej odpowiedzi czujnika można bardzo dobrze wytłumaczyć składem chemicznym badanej próby.

Możliwości metody regresji liniowej wielokrotnej w zakresie pozyskania z danych czujnikowych użytecznej informacji ilościowej o badanych gazach są naturalnie ograniczone przez rzeczywisty charakter zależności między opisywaną miarą ilościową a parametrami wyłonionymi z sygnałów czujników lub odpowiednio skonstruowanymi cechami niejawnymi. Informacja zakodowana w zależnościach o charakterze nieliniowym jest tylko częściowo dostępna dla liniowej regresji wielokrotnej. Przykładem ilustrującym ten problem jest poprawa jakości modelu obliczającego wskaźnik jakości owoców, uzyskana przez Zhanga [192] w wyniku przełączenia z regresji liniowej ($R^2 = 0{,}87$) na nieliniową – wielomian drugiego stopnia ($R^2 = 0{,}92$).

Regresja nieparametryczna

W licznych przypadkach nieliniowej charakterystyki czujników wymagane jest stosowanie rozwiązań, które umożliwiają modelowanie zależności nieliniowych. W zakresie analizy regresji interesującą propozycją odpowiadającą na takie zapotrzebowanie jest regresja nieparametryczna. Korzystanie z tych metod ułatwia brak założeń dotyczących parametrycznej postaci funkcji. Konieczne jest jedynie założenie o jej ciągłości, a sprzyjające warunki zachodzą, gdy funkcja jest dość gładka. Model jest złożeniem liniowym wybranych funkcji, tzw. bazowych, za pomocą których można przybliżyć daną funkcję, uwypuklając bardziej lokalne lub raczej globalne cechy zależności. W celu estymacji funkcji jej dziedzina jest dzielona na fragmenty, a cza-

sami prowadzi się ją w otoczeniu pojedynczych punktów dziedziny. Do tej grupy metod należy np. regresja z zastosowaniem funkcji jądrowych i odwróconych funkcji jądrowych. Można znaleźć pojedyncze przykłady zastosowania tych metod w dziedzinie analizy czujnikowych danych pomiarowych. Autorzy donoszą o bardzo dobrym dopasowaniu modelu nawet dla niewielkiej liczby danych uczących [193], jak również o możliwości ekstrapolacji, tj. predykcji stężeń poza zakresem wartości ze zbioru danych uczących dzięki zastosowaniu tej metody [194, 195]. Warto tu wspomnieć pracę [68], w której porównano możliwości innych metod regresji nieparametrycznej, takich jak: i) regresja lokalnie ważona (ang. *locally weighted regession*), oparta na idei lokalnej liniowości i bliska koncepcyjnie regresji składowych głównych, ii) regresja z poszukiwaniem interesujących kierunków w danych, będąca *de facto* adaptacyjną metodą estymacji, iii) regresja wykorzystujaca warunkowo zmienne wartości oczekiwane, która uwypukla cechy globalne estymowanej funkcji z innymi, częściej stosowanymi metodami. Najlepsze rezultaty określania dwu- i trójskładnikowych mieszanin gazów uzyskano metodami ACE i PP.

Regresja odporna

Jakość modeli ilościowych uzyskanych z zastosowaniem metod klasycznych zarówno parametrycznych, jak i nieparametrycznych, jest niska, gdy dane wykorzystane do ich opracowania są obciążone z powodu występowania w nich tzw. wyników skrajnych (ang. *outliers*). Zgodnie z koncepcją wyników skrajnych niekoniecznie są to dane błędne, lecz tylko wyraźnie różniące się od większości. Ich pojawienie się można wytłumaczyć wystąpieniem podczas pomiaru zjawisk niespodziewanych. Zastosowanie nieodpornych technik regresji do analizy danych zawierających wyniki skrajne powoduje pogorszenie zdolności uzyskanych modeli do odtworzenia zależności obowiązującej dla tych danych. Detekcja wyników skrajnych jest trudna. W odpowiedzi na ten problem powstało szereg metod regresji odpornej (ang. *robust regression*), która zapewnia prawidłowy opis danych mimo występowania w nich wyników skrajnych. Do najbardziej znanych metod należą: estymatory największej wiarygodności, tzw. estymatory typu M, estymatory skali, tzw. estymatory typu S, metoda uciętych najmniejszych kwadratów (ang. *least trimmed squares*) oraz metoda median najmniejszych kwadratów (ang. *least median squares*).

W pracy [196] Czytelnik znajdzie wyczerpujący materiał przeglądowy dotyczący roli metod odpornych w ilościowym oznaczaniu substancji, w okolicznościach występowania wyników skrajnych. Regresja odporna jest chętnie stosowana w analizie danych z pomiarów czujnikowych do ilościowych oznaczeń gazów [17, 127]. Pomiary czujnikowe ze względu na ich charakter są podatne na dostarczanie wyników obarczonych rozrzutem. Duży rozrzut ma bardzo niekorzystny wpływ na możliwości po-

miarowe przyrządów czujnikowych, jeżeli nie zostaną zastosowane odporne techniki przetwarzania i analizy danych.

Regresja składowych głównych

Jednym ze sposobów na dotrzymanie wymagań dotyczących niezależności liniowej zmiennych objaśniających w modelu regresji jest obsadzenie w tej roli składowych głównych, wyłonionych w wyniku analizy składowych głównych danych wielowymiarowych. Jak wiadomo, składowe takie są ortogonalne. Uzyskane rozwiązanie jest określane jako regresja składowych głównych (ang. *principal components regression*, PCR). Jest to metoda liniowa. Liczba składowych głównych, które należy uwzględnić w modelu nie przekracza zazwyczaj kilku, lecz powinna być określana w toku kroswalidacji. Z podanych powodów regresja składowych głównych nadaje się do zastosowania w ilościowym określaniu badanych gazów na podstawie pomiarów czujnikowych. W naturalny sposób nasuwa się jej porównanie z innymi metodami [68]. Pokazano [153], że PCR jest równoważna z regresją składowych niezależnych pod względem wyników oznaczeń ilościowych gazów, choć druga z metod umożliwia lepszą interpretację chemiczną modelu. Dla małej liczby próbek uczących w pomiarach LZO PCR lokowało się między regresją procesów Gaussa (ang. *Gaussian process regression*), dającą lepsze wyniki, a siecią neuronową, która była ewidentnie niekorzystnym rozwiązaniem [197]. W zastosowaniu do tego rodzaju oznaczeń wykazano przewagę rozważanej metody względem klasycznej metody regresji najmniejszych kwadratów, jak również regresji odwróconych najmniejszych kwadratów, a nawet cząstkowych najmniejszych kwadratów [198].

Aby uzyskać zadowalający wynik modelowania metodą PCR składowe główne powinny być istotnie skorelowane ze zmienną objaśnianą, co nie zawsze zachodzi. Jest to podstawowy czynnik, który może wpływać niekorzystnie na wyniki ilościowych oznaczeń gazów z zastosowaniem metody PCR. Głównie z tej przyczyny lepsze rezultaty uzyskuje się zazwyczaj z zastosowaniem regresji metodą cząstkowych najmniejszych kwadratów. Niemniej jednak istnieją doniesienia o dobrych rezultatach ilościowego oznaczania trójskładnikowej mieszaniny gazów (CO, CH_4 i para wodna) metodą PCR [199]. Inni autorzy pokazali, że regresja składowych głównych pozwala uzyskać dobre wyniki obliczania stężeń składników dwuskładnikowej mieszaniny gazów po zaimplementowaniu w gotowym przyrządzie czujnikowym [200].

Regresja metodą cząstkowych najmniejszych kwadratów

Jeżeli założenia umożliwiające rozwiązanie problemu odwzorowania zmiennych niezależnych w zależne za pomocą regresji wielokrotnej nie są spełnione, to dobrą

alternatywę stanowi regresja metodą cząstkowych najmniejszych kwadratów (ang. *partial least squares regression, PLS regression*). Metoda została opracowana dla przypadków, gdy liczba zmiennych niezależnych jest duża, również względem liczby dostępnych wyników pomiarów, i zmienne te są silnie skorelowane [187]. Należy więc do tzw. metod miękkich. Ideą metody jest znalezienie kierunku w wielowymiarowej przestrzeni cech **X**, który odpowiada kierunkowi maksymalnej zmienności w przestrzeni zmiennych objaśnianych **Y**. Analiza polega na symultanicznej dekompozycji macierzy zmiennych niezależnych **X** i zależnych **Y** według równań:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{6.16}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \mathbf{F} \tag{6.17}$$

gdzie **T** jest macierzą wyników, **P** i **Q** – macierzami ładunków, a **E** i **F** – macierzami błędu [140]. Macierz **T** uzyskana w metodzie cząstkowych najmniejszych kwadratów jest wspólna dla zmiennych niezależnych i zależnych. Reprezentuje ona zmienność zmiennych zależnych, która jest związana ze zmiennością zmiennych niezależnych. Analizę i interpretację wzajemnych powiązań zmiennych niezależnych można przeprowadzić na podstawie macierzy **P**, zmiennych zależnych zaś na podstawie macierzy **Q**. Przede wszystkim jednak regresja PLS może być użyta do najbardziej kompletnego wyjaśnienia zmienności zmiennej zależnej za pomocą zmienności zmiennych niezależnych. Metoda umożliwia budowę jednego modelu ilościowego dla wielu zamiennych objaśnianych. Mimo to najczęściej jest stosowana w wersji z jedną zmienną zależną. Opracowano statystyki służące ocenie modelu. Statystyka R^2 służy do określenia stopnia dopasowania modelu do danych, natomiast statystyka Q^2 jest podstawą oceny zdolności predykcyjnych modelu. PLS jest uważany za tzw. złoty standard wśród metod ilościowych w chemii analitycznej. Nieliniowa wersja metody – PLS z funkcją rdzeniową (*kernel* PLS) jest obecnie bardzo popularną metodą chemometryczną.

Regresja PLS jest często stosowana w analizie ilościowej gazów na podstawie pomiarów czujnikowych. Jedną z głównych cech danych tego typu jest znaczne skorelowanie zmiennych i niekorzystny stosunek liczby zmiennych do liczby obserwacji. Zwłaszcza dotyczy to pomiarów wykonywanych w warunkach pozalaboratoryjnych. Jedna z pierwszych prac dotyczących zastosowania PLS w pomiarach czujnikowych [201] udowadniała, że w warunkach nadmiarowości cech i mimo nieliniowości czujników można uzyskać zadowalającą dokładność predykcji stężeń par metanolu i acetonu tą liniową metodą. Then [202] pokazał wyniki określania stężeń składników mieszaniny trójskładnikowej (1-butanol, toluen, *n*-oktan) w zakresie 40–1400 ppm. Zastosowano regresję PLS z jedną oraz trzema zmiennymi wyjściowymi i wykazano przewagę pierwszego rozwiązania. Metoda ta bardzo dobrze sprawdziła się w zastosowaniu do obliczania stężeń składników dwuskładnikowych rozpuszczalników. Uzyskano błędy predykcji stężeń mniejsze od kilku procent [165]. Regresja PLS

może też być z powodzeniem zastosowana do zagadnień jakościowych, jeśli zostaną przedstawione jako ilościowe [150]. Ze względu na genealogię metody cząstkowych najmniejszych kwadratów powstało wiele prac porównujących jakość oznaczeń gazów na podstawie pomiarów zaawansowanymi metodami analitycznymi, np. metodą GC-MS i na podstawie pomiarów czujnikowych, z jej zastosowaniem. Oceny wystawiane tym dwom sposobom pozyskiwania danych pomiarowych są często rozbieżne. Jedne rezultaty wskazują na przewagę metod czujnikowych [203, 204], inne wręcz przeciwnie [205]. Warto podkreślić, że na uzyskane oceny składa się szereg czynników, do których należą np. wybór czujników, wybór badanych gazów, rodzaj oznaczanego parametru ilościowego. Z aplikacyjnego punktu widzenia istotne są zachęcające rezultaty pomiaru intensywności odoru metodami czujnikowymi z zastosowaniem regresji metodą cząstkowych najmniejszych kwadratów [51, 206, 207]. Z tych samych względów warto wspomnieć algorytm hybrydowy złożony z klasycznej metody najmniejszych kwadratów i PLS [208]. Wykazano jego bardzo dobre właściwości w zastosowaniu do uaktualniania modelu kalibracji na podstawie jednego standardu rekalibrującego. Chodziło o uzyskanie kompensacji interferencji oraz dryfu czujników, których to czynników nie uwzględniono w modelu oryginalnym. Rezultaty porównywalne z modelem hybrydowym uzyskiwano tylko z zastosowaniem PLS, lecz pod warunkiem rekalibracji za pomocą kilku standardów.

6.4. Sztuczne sieci neuronowe

Sieci neuronowe (*artificial neural networks*, ANN) są przykładem podejścia koneksjonistycznego. Jego zasadniczą cechą jest niejawna struktura modelu matematycznego podlegająca *dostrajaniu* do danych w toku kolejnych iteracji procesu uczenia. Strategia ta bardzo dobrze sprawdza się w modelowaniu, którego zadaniem jest wygenerowanie poprawnych predykcji bez przedstawienia analitycznej zależności między zmiennymi, które są podstawą predykcji, a zmienną zależną. Sieci neuronowe są silnym narzędziem mapowania nieliniowego. Przez dopasowywanie się sieci do danych uczących jest modelowana funkcja, która przekształca zmienne wejściowe w wyjściowe (metody pod nadzorem), ewentualnie w strukturze sieci odzwierciedlana jest struktura danych (metody bez nadzoru). Mimo podkreślanych różnic istnieją istotne podobieństwa między sieciami neuronowymi a klasyfikacją statystyczną [28].

Ze względu na to, że w analizie danych z pomiarów czujnikowych często występują problemy nieliniowe, sieci neuronowe są powszechnie stosowane w tej dziedzinie. Problemy nieliniowe dotyczą zarówno odwzorowań przestrzeni cech w zmienną ciągłą (nieliniowe charakterystyki czujników), jak i w zmienne dyskretne (nieliniowy podział na klasy). W związku z tym omawiane metody są w równym stopniu przydatne do rozwiązywania problemów o charakterze zarówno jakościowym, jak i ilościowym.

Sieć neuronową charakteryzują trzy podstawowe elementy: i) funkcja aktywacji neuronów, ii) topologia sieci, iii) algorytm uczenia. Stanowi ona strukturę elementów obliczeniowych (neuronów), które są zorganizowane w warstwy. Liczba warstw i liczby neuronów w warstwach mogą być bardzo różne, zależnie od potrzeb. Wejścia neuronów są przeliczane w wyjścia za pomocą funkcji aktywacji. Nieliniowy charakter funkcji przejścia, np. sigmoidalne, tangensoidalne, logistyczne, gaussowskie sprawia, że sieci neuronowe realizują przekształcenia nieliniowe. W niektórych sytuacjach stosowana jest liniowa funkcja przejścia (np. w warstwie wyjściowej). Neurony poszczególnych warstw są połączone ze sobą połączeniami z przypisanymi im wagami. Wagi połączeń są zmieniane w toku procesu uczenia. Klasycznym algorytmem uczenia nadzorowanego sieci neuronowej jest algorytm propagacji wstecznej. Błąd odtworzenia zależności wejście–wyjście jest zdefiniowany jako funkcja wag neuronów sieci. Algorytm ma za zadanie iteracyjnie korygować wagi połączeń w sieci, tak aby przesuwać ją w kierunku minimalnej wartości błędu odwzorowania wejść sieci w jej wyjście. W podstawowej wersji algorytm jest wolnozbieżny. Obecnie funkcjonuje kilkadziesiąt wersji tego algorytmu, opracowanych z myślą o poprawie tempa oraz jakości uczenia sieci [29]. W podobnym celu rozwijane są również inne algorytmy [171]. Wysiłek ten jest bardzo istotny dla pomiarów czujnikowych, gdzie dąży się do wbudowania algorytmów obliczeniowych w przyrząd pomiarowy z zapewnieniem jego kompaktowości i szybkiego działania.

Zdolności predykcyjne sieci są zależne od ich możliwości uogólniających, a na te w decydujący sposób wpływa wielkość sieci. Rozmiary sieci są kontrolowane przez liczbę neuronów w warstwach ukrytych. Optymalna wielkość sieci zależy od wielu czynników, np. złożoności problemu klasyfikacji/regresji, liczby wejść i wyjść sieci, doboru funkcji aktywacji, algorytmu uczącego, wielkości zbioru uczącego, poziomu szumu w danych.

Istnieją dwa podstawowe sposoby poszukiwania optymalnej liczby i rozmieszczenia neuronów ukrytych. Pierwszy z nich polega na wyuczeniu kilku, kilkunastu, lub więcej sieci na tym samym zbiorze uczącym i porównaniu ich możliwości predykcyjnych dla zbioru testowego. Drugi ucieka się do różnych technik modyfikacji konkretnej sieci. Należą tu techniki inkrementacyjne, gdzie algorytm startuje z małą liczbą neuronów i stopniowo ją zwiększa, oraz techniki przycinające, które wygaszają wagi neuronów nadmiarowych w dużej puli początkowej. Za zdolności uogólniające sieci o ustalonej strukturze odpowiada układ wag w sieci. Ograniczenie stopnia dopasowania do zbioru uczącego osiąga się różnymi metodami. Do najbardziej efektywnych należą: wygaszanie wag (ang. *weight decay*), wczesne zatrzymanie (ang. *early stopping*) czy uczenie z szumem (ang. *training with noise*) [22].

6.4.1. Perceptron wielowarstwowy

Architektura sieci decyduje o liczbie parametrów (wag), których wartości należy określić w procesie uczenia. Ta zaś ustanawia wymagania względem wielkości zbioru

danych. Dla perceptronu wielowarstwowego (ang. *multilayer perceptron*, MLP), który jest najpowszechniej stosowanym typem sieci, liczbę parametrów określa wzór [29]:

$$P = H + O + (I \times O) + (H \times O) \tag{6.18}$$

gdzie: I, H, O są liczbą neuronów w warstwie wejściowej, ukrytej i wyjściowej. Zalecane jest, by liczba wzorców uczących była co najmniej dwukrotnie większa od liczby parametrów [38]. Wśród opinii krytycznych dotyczących sieci neuronowych często wysuwany jest argument dużej chłonności danych. Brak wystarczającej liczby próbek, na których można wyuczyć sieć, powoduje zbytnie dostosowanie wag do zbioru uczącego (przeuczenie) i osłabienie właściwości uogólniających sieci. Zapewnienie zbioru danych o odpowiedniej wielkości jest zaś w wielu przypadkach mało realne. Dotyczy to również zastosowania sieci neuronowych w analizie danych z pomiarów czujnikowych. Podstawowym sposobem redukcji wpływu tego problemu na jakość uzyskanego modelu neuronowego jest odpowiednie prowadzenie uczenia, walidacji i testowania, a zwłaszcza wybór zbiorów danych: uczącego, walidującego i testującego.

Osobną kwestią jest zależność liczby klas, które można wyróżnić od struktury sieci. Uważa się, że decydujące znaczenia ma tutaj liczba neuronów w warstwie ukrytej, a orientacyjne proporcje przedstawia wzór [209]:

$$L \le \frac{H(H+1)}{2} + 1 \tag{6.19}$$

gdzie: L jest liczbą klas, a H liczbą neuronów w warstwie ukrytej. Wynika z niego, że możliwości perceptronu są bardzo duże.

MLP jest często traktowany jako algorytm odniesienia, do którego porównywane są wyniki uzyskiwane za pomocą innych rozwiązań neuronowych. Perceptrony wykonują aproksymację globalną, dzieląc przestrzeń cech za pomocą hiperpłaszczyzn. Na ogół sieci te potrzebują mniejszej liczby parametrów do rozwiązania określonego zadania, lecz mogą mieć więcej warstw i wymagają relatywnie dłuższego czasu uczenia. Perceptrony wielowarstwowe są obok analizy składowych głównych najpopularniejszą metodą analizy danych z pomiarów czujnikowych. Z równym powodzeniem są stosowane do rozwiązywania problemów klasyfikacji oraz regresji.

W pracy [144] czytelnik znajdzie empiryczne potwierdzenie tezy, że algorytmy liniowe (np. liniowa dyskryminacja Fishera) i MLP dają podobne rezultaty w zastosowaniu do prostych zadań klasyfikacji. Natomiast tam, gdzie granice decyzyjne są nieliniowe, MLP pokonuje prostsze metody rozpoznawania gazów. Przykład rozwiązania problemu ilościowego (wyznaczania stężenia odoru gazu wysypiskowego) z wykorzystaniem MLP na podstawie pomiarów czujnikowych podano w pracy [210]. Konieczność dostosowania architektury MLP do potrzeb konkretnego problemu określania gazów zilustrowano w pracy [164], w [94] zaś przedstawiono praktyczną implementa-

cję optymalizacji architektury MLP w gotowym przyrządzie czujnikowym. Nierzadką praktyką jest opracowywanie rozwiązań modułowych z udziałem MLP i innej sieci. Perceptron występuje na ogół jako jednostka przetwarzająca wyniki działania innej sieci, która go poprzedza, najczęściej przeprowadzając wstępną organizację przestrzeni cech. Przykład takiego rozwiązania w zastosowaniu do określania stężeń CO, metanu, metanolu i propan-butanu podano w pracy [211]. Jako pierwszy moduł zastosowano sieć Kohonena. Inną hybrydą z udziałem warstwy z radialną funkcją bazową rozwiązywano w pracy [212] problemy klasyfikacji. W obu przykładach wykazano poprawę jakości predykcji względem rozwiązania opartego wyłącznie na MLP. Jeszcze inny rodzaj kombinacji z modułowym perceptronem wielowarstwowym zaproponowano w pracy [213]. Perceptrony wielowarstwowe są też najchętniej rozważane jako narzędzia rozpoznawania gazów do bezpośredniej implementacji w układach cyfrowych zintegrowanych z czujnikowymi systemami pomiarowymi [214].

6.4.2. Sieci z radialnymi funkcjami bazowymi

Innym interesującym typem rozwiązania neuronowego, również relatywnie często stosowanym w analizie danych z pomiarów czujnikowych, jest sieć z radialną funkcją bazową (ang. *radial basis function*, RBF). Sieć ta składa się z trzech warstw. Neurony jednej warstwy (ukrytej) są wyposażone w radialne funkcje przejścia. Funkcje te mierzą odległości między wektorami danych oraz wektorami wag neuronów i generują lokalne wzmocnienia w przypadku ich wzajemnej bliskości. Kryterium odległości, tzw. szerokość jądra, jest dodatkowym parametrem sieci. Na ogół liczba neuronów w warstwie ukrytej nie jest definiowana i może wzrastać w trakcie uczenia. Pozwala to uzyskać odpowiednie mapowanie danych wejściowych. Funkcje aktywacji neuronów warstwy wyjściowej są liniowe. Dzięki temu na wyjście sieci podawana jest wartość liniowej kombinacji funkcji radialnych z warstwy ukrytej. Ze względu na zróżnicowanie funkcji neuronów poszczególnych warstw stosuje się różne techniki optymalizacji. Zadaniem warstwy ukrytej jest znalezienie struktury klas i w tym celu jest ona uczona bez nadzoru. Natomiast wagi warstwy wyjściowej są obliczane metodą najmniejszych kwadratów. Strategia ta jest bardzo wydajna. Jednak ze względu na jej wrażliwość na szum opracowano też bardziej odporne algorytmy ucznia [22]. Sieci neuronowe z radialną funkcją bazową bardzo dobrze nadają się do rozwiązywania problemów klasyfikacji nieliniowej i z tego wynika zainteresowanie nimi na potrzeby analizy danych w pomiarach czujnikowych.

W pracy [9] porównano rezultaty rozpoznawania benzenu, toluenu i ksylenu w powietrzu o różnej wilgotności na podstawie pomiarów czujnikowych z zastosowaniem klasyfikatora liniowego i RBF, wskazując na przewagę ostatniego. Sieć ta okazała się lepsza niż MLP i PNN, również gdy rozróżniano sześć szczepów bakterii na podstawie pomiarów komercyjnym elektronicznym nosem Cyranose 320 [155]. Inte-

resujący przykład zastosowania RBF do problemu klasyfikacji, który jest *de facto* dyskretną reprezentacją problemu o charakterze ciągłym, podano w pracy [215]. Za pomocą sieci z RBF określano na podstawie wyniku pomiaru czujnikowego świeżość ryb przez podanie liczby dni, które upłynęły od połowu (1–15). Uzyskano błąd na poziomie ułamka procenta. Eksperymentalne potwierdzenie, że sieci tego typu mogą efektywnie rozwiązywać złożone problemy klasyfikacji, charakteryzujące się dużą liczbą klas, dużą wymiarowością przestrzeni cech i dużą licznością prób podano w pracy [212] na przykładzie rozróżniania elektronicznym nosem 21 rodzajów prostych i złożonych kompozycji zapachowych.

6.4.3. Probabilistyczne sieci neuronowe

Probabilistyczne sieci neuronowe (PNN) bywają przedstawiane jako neuronowa realizacja analizy dyskryminacyjnej z funkcją jądrową [216]. W zasadzie sieć jest wersją perceptronu, lecz dzięki zastosowaniu gaussowskiej funkcji przejścia w warstwie ukrytej ma pewne charakterystyczne właściwości. Podstawą działania probabilistycznej sieci neuronowej jest określenie funkcji gęstości prawdopodobieństwa dla każdej z klas występujących w zbiorze danych uczących. Funkcje te definiują granice poszczególnych klas i podczas predykcji są wykorzystywane do określania prawdopodobieństwa przynależności nowych wzorców do poszczególnych klas. Sieć ma tylko jeden parametr. Jest nim szerokość jądra, która wyznacza zakres interpolacji dopuszczalny podczas określania funkcji gęstości prawdopodobieństwa. Za zastosowaniem PNN w rozpoznawaniu gazów na podstawie pomiarów czujnikowych przemawia kilka zalet.

Dla dobrze zdefiniowanych problemów klasyfikacji algorytmy neuronowe prowadzą do rezultatów podobnych jak metody liniowe, np. DFA. Sieci są natomiast skuteczniejsze, jeżeli wchodzą w grę nieliniowe granice decyzyjne lub nieliniowe funkcje odwzorowujące przestrzeń cech w zmienne objaśniane, to znaczy kiedy rozważane są złożone problemy klasyfikacji [13, 217]. Dzięki zastosowaniu PNN możliwe jest dostarczenie informacji o prawdopodobieństwie przynależności danych do poszczególnych klas. Właściwość ta jest bardzo korzystna, np. w warunkach zastosowania czujników w systemach ostrzegania [216], kiedy należy rozróżnić sytuacje zagrożenia i jego braku. Wreszcie na wybór sieci probabilistycznych może wpłynąć znacznie krótszy czas uczenia w porównaniu z najczęściej stosowanym perceptronem wielowarstwowym [96, 218], zwłaszcza że odpowiednia selekcja cech wpływa na zmniejszenie różnic między rezultatami oznaczania gazów tymi metodami [96].

6.4.4. Sieci Kohonena

Technikę mapowania *z zachowaniem topologii* realizują sieci Kohonena [219]. Pod względem algorytmu mapa samoorganizująca (ang. *selforganizing map*, SOM)

jest siecią neuronów rozmieszczonych w węzłach dwuwymiarowej siatki. Każdy neuron funkcjonuje w dwóch stanach: aktywny – nieaktywny. Uaktywnienie neuronu zachodzi, gdy wektor jego wag jest najbliższy określonemu wektorowi uczącemu. Wraz z neuronem wygrywającym uaktywniane są neurony sąsiednie (zwykle kilka węzłów wokół), po czym następuje adaptacja ich wag. Adaptacja prowadzi do zbliżenia wag neuronów i wektora, który je aktywował. W wyniku przedstawienia sieci całego zbioru wektorów uczących poszczególne neurony dostosowują swoje wagi w ten sposób, że każdy z nich reaguje tylko na niewielki region oryginalnej przestrzeni cech, tzw. pole recepcji neuronu. Zachodzi koordynacja budowy tych pól w taki sposób, że organizacja neuronów w siatce odzwierciedla układ ich pól recepcji w przestrzeni cech. Podobieństwa wzorców w wielowymiarowej przestrzeni cech są zatem przenoszone na sąsiedztwo neuronów. SOM jest atrakcyjną techniką wizualizacji danych [220, 221]. Metoda ma charakter adaptacyjny, jest prosta i odporna na zakłócenia w danych pomiarowych. Mapowanie danych tą techniką jest jednak problematyczne, jeżeli ich struktura jest z definicji wielowymiarowa.

Wyczerpującą, ilustrowaną przykładem dyskusję możliwości map samoorganizujących w analizie danych z pomiarów czujnikowych można znaleźć w pracy [222]. Zależnie od stopnia złożoności problemu rozpoznawania, skupienia neuronów aktywnych reprezentujących różne kategorie wzorców bywają wyraźnie oddzielone granicą nieaktywnych jednostek sieci [223] lub stopień ich rozdzielenia jest niepełny. Rezultat taki uzyskano np. dla pomiarów czujnikowych oleju silnikowego zanieczyszczonego olejem napędowym [220]. W pracy [224] pokazano sieć Kohonena wprost w roli klasyfikatora o dobrej zdolności rozróżniania powietrza zanieczyszczonego metanem i powietrza zanieczyszczonego tlenkiem węgla. Ze względu na adaptacyjny charakter SOM-y są proponowane jako rozwiązania pozwalające poprawić odporność klasyfikatorów na dryf czujników [225].

6.4.5. Inne sieci neuronowe

Spośród innych rodzajów sieci neuronowych na uwagę w obszarze analizy danych z pomiarów czujnikowych zasługują rozwiązania ARTMAP (ang. *adaptive resonance theory supervised predictive mapping*) oraz hybryda neuronowo rozmyta, fuzzy ARTMAP. Są to sieci o charakterze adaptacyjnym. Cechuje je samostabilizująca pamięć, która umożliwia dopasowanie zapisanej w sieci wiedzy do nowych wzorców, napotykanych podczas pracy sieci. Wśród innych rozwiązań adaptacyjnych wyróżnia je mechanizm pamięci krótko- i długookresowej. Dzięki temu sieć, poznając nowe wzorce, nie zapomina zupełnie wzorców wcześniej prezentowanych, podczas gdy większość algorytmów adaptacyjnych jest wyposażona wyłącznie w pamięć krótkookresową. Właściwość ta decyduje o przydatności ARTMAP i fuzzy ARTMAP w warunkach, gdy klasy wzorców wędrują w przestrzeni cech. Zjawisko takie dotyczy danych pochodzących

z pomiarów czujnikowych i jest związane z dryfem czujników. Omawiane sieci mają bardzo dobre właściwości jako klasyfikatory współpracujące z różnymi metodami selekcji cech [98, 137], częstokroć dając lepsze rezultaty niż MPL [226]. Stanowią one konkurencyjną metodę klasyfikacji, zdolną skompensować wpływ dryfu czujników na wynik określania gazów z zastosowaniem systemu czujnikowego [227, 228].

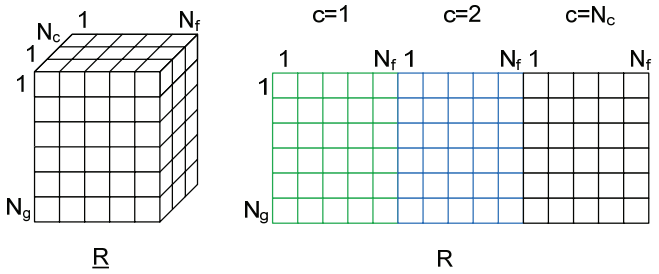
6.5. Metody wielokierunkowej analizy danych

Ostatnio w obszarze badań czujnikowych znajdują zastosowanie metody analizy wielokierunkowej (ang. *multiway analysis*). Jest to związane z rosnącymi możliwościami uzyskiwania danych czujnikowych coraz bardziej złożonych. Pierwotnie metody te były rozwijane na potrzeby kalibracji wielowymiarowej instrumentów analitycznych drugiego i wyższych rzędów, np. chromatografów gazowych ze spektrometrią mas (GC-MS), spektroskopów pracujących z wykorzystaniem bezdyspersyjnej spektroskopii w podczerwieni (ang. *non-dispersive*, IR). W związku z tym metody wielokierunkowe są przede wszystkim przeznaczone do takich zadań, jak analiza jakościowa i ilościowa wieloskładnikowych mieszanin gazów. Metody są interesujące ze względu na oferowane przez nie możliwości określania gazu w obecności nieznanych interferencji.

Specyfika pomiarów czujnikowych stanowi poważne wyzwanie dla technik analizy wielokierunkowej. Podstawowym powodem jest znacznie większe uwikłanie informacji o składnikach badanej mieszaniny w czujnikowych danych pomiarowych w porównaniu z danymi pochodzącymi z pomiarów technikami analitycznymi. Z drugiej strony praca nad technikami pomiaru pozwalającymi generować dane na potrzeby analizy wielokierunkowej oraz taka analiza jest jednym z możliwych kierunków rozwoju w pomiarach czujnikowych złożonych mieszanin gazów.

Dane wielokierunkowe można postrzegać jako rozszerzenie danych płaskich (dwukierunkowych) (patrz rozdz. 4). Kierunki są związane z niezależnymi od siebie czynnikami, które kształtują zawartość informacyjną danych pomiarowych. W związku z rozwojem techniki pomiarowej najczęściej przedmiotem uwagi są dane trójkierunkowe (rys. 6.1). Dane trójkierunkowe są zawarte w macierzy 3D, która bywa też określana jako kostka danych. Dane z pomiarów czujnikowych w naturalny sposób mogą zostać zorganizowanie w struktury 3D. Za poszczególne kierunki odpowiadają zazwyczaj: i) czujnik, ii) czas lub czynnik zmienny w czasie i modyfikujący odpowiedź czujnika, np. temperatura, przepływ, szeroko rozumiane warunki ekspozycji, iii) badana próba gazu. Zazwyczaj dane czujnikowe są analizowane w układzie dwukierunkowym w wyniku wyłonienia nośnych informacyjnie parametrów sygnału. Można jednak postawić tezę, że w ten sposób część istotnej informacji jest tracona. Uważa się, że analitycznie użyteczna informacja jest przede wszystkim związana

z kierunkiem czasu, który w tradycyjnym podejściu do danych nie jest traktowany z należytą uwagą. Metody wielokierunkowe są interesującą propozycją jego wykorzystania.



Rys. 6.1. Zasada rozwijania danych trójkierunkowych: N_c – liczba czujników ($c = 1, \ldots, N_c$), N_f – liczba poziomów czynnika wywołującego zmienność odpowiedzi czujnika ($f = 1, \ldots, N_f$), N_g – liczba zbadanych prób gazów ($g = 1, \ldots, N_g$) [231]

Ideą metod wielokierunkowych jest analiza danych w wielu kierunkach jednocześnie w celu wykorzystania informacji przenoszonej przez te kierunki do analizy jakościowej i/lub ilościowej gazów. W przypadku zastosowania tych metod wymagane jest bardzo uważne przetwarzanie wstępne danych. Nie może ono zaburzyć podstawowej właściwości zestawu danych, np. dwu- lub trójliniowości, której istnienie zakłada się, wykonując analizę. Omawiane metody są dobrze dostosowane do przetwarzania danych pomiarowych bez poddawania ich procesom selekcji lub ekstrakcji cech w tradycyjnym rozumieniu tych pojęć. Można je natomiast uznać za złożone, często bardzo złożone metody ekstrakcji cech.

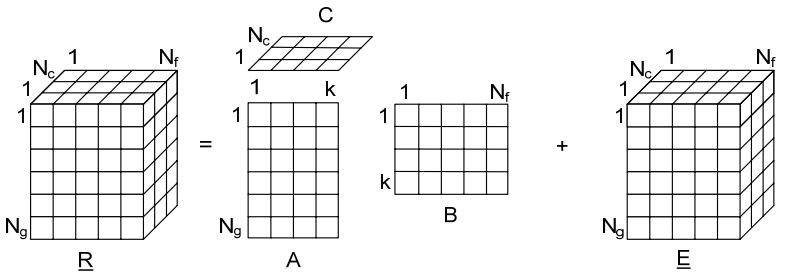
Podstawą analizy wielokierunkowej jest dekompozycja macierzy danych. Analizę danych wprost w postaci kostki określa się jako analizę danych nierozwiniętych. Istnieją również metody analizy danych tego typu w postaci rozwiniętej (ang. *unfolded data*), czyli w tradycyjnym układzie macierzy płaskiej, gdzie każdy wiersz dotyczy jednej próby, a każda kolumna jest jedną zmienną (cechą). Aby zilustrować istotne cechy metod wielokierunkowych, szerzej omówiono następujące metody dekompozycji: analizę składowych głównych z rozwinięciem danych (ang. *unfolded*-PCA), zrównoleglonąą analizę czynnikową (PARAFAC1) i jej wersję PARAFAC2 [229]. Można znaleźć przykłady ich zastosowania w dziedzinie czujnikowych pomiarów gazów, choć są one jeszcze nieliczne [230, 231].

Analiza składowych głównych z rozwinięciem danych jest klasyczną analizą składowych głównych, tyle że przeprowadzoną dla rozwiniętych danych trójkierunkowych. Zasadę rozwijania danych pokazano na rys. 6.1.

Stosowany jest model dekompozycji dwuliniowej. W przypadku modelowania danych z użyciem k składowych głównych macierz danych $\boldsymbol{R}(N_g \times N_c N_f)$ można przedstawić w postaci iloczynu:

$$\mathbf{R} = \mathbf{A}\mathbf{B}^T + \mathbf{E} \tag{6.20}$$

gdzie $\mathbf{A}(N_g \times k)$ jest macierzą wyników, która odnosi się do badanych próbek, a $\mathbf{B}(k \times N_c N_f)$ – macierzą ładunków, odnoszącą się do zmiennych. Jeżeli w jednym wierszu macierzy **R** znajdują się pomiary pochodzące z poszczególnych punktów czasu (rys. 6.1), to analiza jest próbą przedstawienia profilu czasowego czujnika (lub zestawu profili czasowych większej liczby czujników) za pomocą mniejszej liczby zmiennych nieskorelowanych.

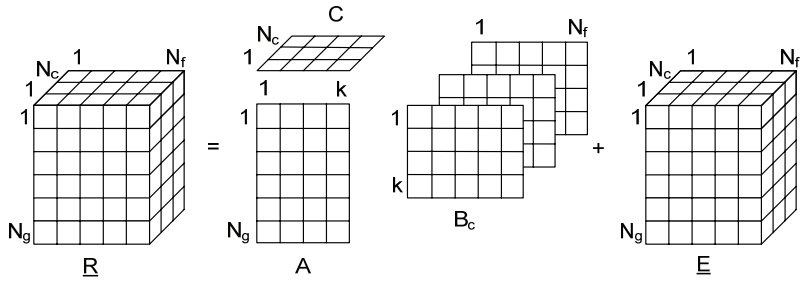


Rys. 6.2. Graficzna ilustracja metody PARAFAC1 [231]

Metoda PARAFAC1, której ideę przedstawiono na rys. 6.2, jest podobna do PCA, z tą różnicą, że pozwala wykonać dekompozycję trójliniową danych. Przebiega ona według równania:

$$\mathbf{R}_c = \mathbf{A}\mathbf{D}_c\mathbf{B}^T + \mathbf{E} \tag{6.21}$$

gdzie $\mathbf{R}_c(N_g \times N_f)$ jest dowolnym *c*-tym przekrojem podłużnym trójkierunkowej macierzy danych **R**, $\mathbf{D}_c$ jest *c*-tą macierzą diagonalną, której przekątna mieści *c*-ty wiersz macierzy **C**, **C** jest macierzą ładunków dotyczącą kierunku związanego z czujnikami. W przypadku zastosowania tej metody istnieje jedyne najlepsze rozwiązanie konkretnego problemu dekompozycji. Każda rotacja składowych głównych prowadzi do pogorszenia stopnia dopasowania do danych.



Rys. 6.3. Graficzna ilustracja metody PARAFAC2 [231]

Metoda PARAFAC2, którą zilustrowano graficznie na rys. 6.3, to bardziej złożona wersja metody PARAFAC1. Metoda pozwala przeanalizować dane, które nie są trójliniowe. Dekompozycja przebiega według równania:

$$\mathbf{R}_c = \mathbf{A}\mathbf{D}_c\mathbf{B}_c^T + \mathbf{E} \tag{6.22}$$

gdzie $\mathbf{B}_c$ jest macierzą ładunków dotyczącą kierunku związanego z czynnikiem wywołującym zmiany odpowiedzi czujnika, np. w czasie, dotyczącą c-tego czujnika. Jedyność rozwiązania jest gwarantowana przez zachowanie warunku: $\mathbf{B}_c^T\mathbf{B}_c = \mathbf{B}^T\mathbf{B}$, $c = 1, ..., N_c$.

Zaletą metody PARAFAC2 jest możliwość modelowania danych, gdy profile, np. sygnały czujników uzyskane dla różnych badanych prób gazu, wykazują cechy (np. lokalne maksima) przesunięte w czasie i o różnym kształcie. Wydobycie tych szczegółów nie jest możliwe z zastosowaniem PARAFAC1 ani PCA dla danych rozwiniętych. PARAFAC2 pozwala wprowadzić osobną macierz ładunków dla każdego czujnika ze względu na czynnik wywołujący zmiany jego sygnału, np. czas. Ta cecha metody jest bardzo interesująca ze względu na stosowanie różnych technik modulowania sygnału czujników właśnie w celu lepszego rozsunięcia informacji o badanych gazach w profilu czasowym odpowiedzi.

Zastosowanie metod analizy wielokierunkowej wymaga pewnego wysiłku podczas wizualizacji i interpretacji ich wyników. W przypadku analizy o charakterze eksploracyjnym rezultaty dekompozycji kostki danych są najczęściej przedstawiane w rozpisaniu na poszczególne kierunki. Kierunek związany z próbami najlepiej zilustrować wykresami wyników, kierunek związany z czujnikami – wykresami ładunków czujnikowych, a kierunek związany z czasem – przebiegami czasowymi odpowiednich ładunków w czasie.

Wszystkie trzy omawiane metody zapewniają prostą strukturę macierzy wyników (**A**) odnoszącą się do badanych prób gazu. Podstawowe różnice dotyczą macierzy ładunków dla kierunku czas (**B, B**$_c$). W przypadku PCA z rozwinięciem danych macierz ta dotyczy łącznie kierunku czas i kierunku czujniki. W pozostałych wypadkach dotyczy wyłącznie kierunku czas. PARAFAC2 daje możliwość przyjrzenia się roli poszczególnych czujników w każdym punkcie czasu. Rola poszczególnych czujników jest łatwiej widoczna, gdy zastosuje się metody PARAFAC1 i PARAFAC2 dzięki wyłonieniu osobnej macierzy ładunków czujnikowych (**C**).

Wykres wyników we współrzędnych składowych głównych jest traktowany jako wskaźnik możliwości zgrupowania badanych prób w różne kategorie. Oczywiście w tym wypadku należy pamiętać, że PARAFAC1, podobnie jak PCA, nie maksymalizuje odstępów między klasami, ujawnia tylko strukturę danych opierając się na kierunkach największej zmienności. Wykres ładunków czujników w zależności od poszczególnych składowych pozwala ocenić, które czujniki są przez te składowe reprezentowane. Z kolei wykres ładunków na kierunek czas daje podstawę do przyj-

rzenia się interesującym punktom czasowym w aspekcie rozróżniania gazów. Jeśli stosowano metodę PARAFAC2, to zyskuje się dodatkowo informację o przydatności odpowiedzi poszczególnych czujników w konkretnych punktach czasu.

Tabela 6.1. Charakterystyka wykresów przedstawiających rezultaty dekompozycji danych na pomocą różnych modeli. W nawiasach podano liczbę elementów zamieszczonych na każdym wykresie [231]

<table>
<tr><th>Model</th><th>Macierz danych</th><th>Kierunek A (próba gazu)</th><th>Kierunek B (czas)</th><th>Kierunek C (czujniki)</th></tr>
<tr><td>PCA</td><td>$N_g \times N_c$</td><td rowspan="4">wykres wyników (N_g)</td><td>wykres ładunków (N_c)</td><td rowspan="2">–</td></tr>
<tr><td>PCA z rozwinięciem danych</td><td rowspan="3">$N_g \times N_f \times N_c$</td><td>wykres ładunków ($N_f \times N_c$)</td></tr>
<tr><td>PARAFAC1</td><td>sygnał czujnika w funkcji czasu ($N_f \times k$)</td><td rowspan="2">wykres ładunków (N_c)</td></tr>
<tr><td>PARAFAC2</td><td>sygnał czujnika w funkcji czasu ($N_f \times N_c \times k$)</td></tr>
</table>

W tabeli 6.1 przedstawiono podstawowe informacje dotyczące charakterystyki wykresów przedstawiających wyniki dekompozycji trójwymiarowej.

Zazwyczaj jednak celem analizy wielokierunkowej jest zbudowanie modelu służącego do jakościowej i/lub ilościowej analizy wieloskładnikowych mieszanin gazów. Składowe główne, uzyskane w wyniku dekompozycji, traktowane są wówczas jako zmienne wejściowe modeli klasyfikujących dane [231, 232] lub odtwarzających zależności ilościowe [230].

7. Ocena systemu analizy danych w czujnikowych pomiarach gazów

Ocena systemu analizy danych w pomiarach czujnikowych to fragment szerszego zagadnienia, którym jest zdefiniowanie kryteriów jakości dla systemów czujnikowych w ogóle [37]. Podstawowy problem wynika tu z braku spójnej teorii, którą można by zastosować do opisu działania tego rodzaju systemów. Odróżnia to je od konkurujących metod pomiarowych, np. chromatograficznych, w których to przypadkach określenie ogólnie rozpoznawalnych miar, jak np. próg detekcji pojedynczych substancji lub ich kombinacji, nie nastręcza większych trudności.

Specyfika systemów czujnikowych wynikająca głównie z ich częściowej selektywności powoduje, że określenie właściwości pomiarowych ma w dużej mierze charakter empiryczny. Rozwiązania stosowane dla metod analitycznych, odpowiednie dla rozwiązań selektywnych, mają tu ograniczone zastosowanie. Tylko w pewnym zakresie stężeń możliwa jest ocena właściwości pomiarowych systemu czujnikowego w odniesieniu do wieloskładnikowych mieszanin gazów wyłącznie na podstawie pomiarów ich składników. W szerszym zakresie stężeń do wyuczenia systemu analizy danych potrzebne są również pomiary samych mieszanin z uwzględnieniem różnych konfiguracji ich składników. Wskaźniki jakości będą wówczas dotyczyły zakresu stężeń w zbadanych kombinacjach gazów, nie będą się zaś odnosiły do całości wyznaczonej przez przedziały stężeń pojedynczych składników. Uzyskane rezultaty nie będą też niezależne od zastosowanej strategii analizy danych, począwszy od metody ich wstępnego przetwarzania, a skończywszy na rodzaju klasyfikatora czy modelu regresji. Wręcz przeciwnie, będą od niej w znacznym stopniu uzależnione.

W pracach teoretycznych można znaleźć propozycje kryteriów oceny jakości systemów czujnikowych, takich jak np. próg detekcji bodźca odorowego, teoretyczna maksymalna liczba rozróżnialnych poziomów bodźca, rozdzielczość ze względu na bodźce sąsiadujące w przestrzeni cech [37]. Znajdują one jednak ograniczone zastosowanie w praktyce. Metody przedstawione w tym rozdziale odnoszą się raczej do oceny systemu analizy danych czujnikowych i w ograniczonym zakresie mogą być traktowane jako metody oceny całego systemu czujnikowego.

Wśród sposobów oceny systemu analizy danych w czujnikowych pomiarach gazów wyłaniają się dwa zasadnicze podejścia. Jedno ma walor diagnostyczny, a jego podstawą jest ocena zdolności systemu do poprawnego odczytania znanej informacji. Wyznacznikiem drugiego jest zaś ocena zdolności predykcyjnych systemu analizy danych. Podejścia te różnią się od siebie zarówno metodologicznie, jak i pod względem wartości (przydatności) ocen, które generują.

W podejściu diagnostycznym ocenia się efektywność działania systemu wyłącznie w odniesieniu do danych uczących. Można określić, na ile system jest w stanie wydostać określoną informację z tych danych. Występuje tu znaczne podobieństwo ze sposobem oceny w procesie filtracji cech. Uzyskana ocena systemu ma charakter ogólny i jest na ogół zbyt optymistyczna pod względem możliwości jego zastosowania w praktyce.

Drugi sposób uwzględnia perspektywę wykorzystania systemu do analizy danych innych niż uczące. Jest ona znacznie bardziej realna niż założenie, że nowe wzorce, z którymi system zostanie skonfrontowany podczas pracy w trybie rozpoznawania, będą takie same jak uczące. Z dużym prawdopodobieństwem wzorce, z którymi będzie miał do czynienia opracowany już system analizy danych, będą się różniły od wzorców stosowanych podczas jego opracowywania.

Osiągnięcie dużej rozpoznawalności nowych wzorców, niepoznanych w trakcie procesu uczenia, tj. nabycie dobrych zdolności uogólniających jest podstawowym celem projektowania systemu analizy danych. Dlatego procedura oceny takiego systemu ze względu na jego zdolności predykcyjne uwzględnia rozmaite techniki walidacji i testowania na danych spoza zakresu danych uczących. Operacja ta jest wyraźnie bardziej czasochłonna niż pierwsza, jednak pozwala uzyskać urealnioną ocenę systemu analizy danych. Jej istotnym aspektem jest czynny udział procesu oceny systemu analizy danych w konstruowaniu tego systemu.

7.1. Techniki walidacji i testowania

Zbytnie dopasowanie modelu klasyfikacji/regresji do danych uczących można ograniczyć przez zastosowanie procedury walidacji. Walidacja pozwala uzyskać ocenę zdolności uogólniających modelu o określonej architekturze. Ujawnia się w ten sposób rodzaj modelu, który wykazuje najlepsze zdolności uogólniające spośród wielu poddanych procesowi uczenia. Do najpopularniejszych technik walidacji należą:

- procedura *holdout*,
- *k*-krotna kroswalidacja,
- metoda *bootstrap*.

Podstawą procedury *holdout* jest podział zbioru danych na dwie części. Jedną część stanowi zbiór uczący, drugą – zbiór walidujący. Zbiór uczący służy do opracowania kilku klasyfikatorów, np. o różnych strukturach, innych parametrach i metaparametrach (np. czas uczenia). Za pomocą otrzymanych modeli wykonywana jest klasyfikacja wektorów danych należących do zbioru walidującego. Klasyfikator, z którym związane jest najmniejsze prawdopodobieństwo błędnej klasyfikacji na zbiorze walidującym, jest uznawany za najlepszy. Omawiane podejście sprawdza się w wielu sytuacjach. Ma jednak dwie podstawowe wady. Po pierwsze, nie można go zastosować do małych zbiorów danych, gdyż wówczas znacznie osłabia się zbiór uczący i zmniejsza możliwości klasyfikatora. Po drugie ocena klasyfikatora uzyskana w ten sposób może być niemiarodajna, jeżeli podział na dwa zbiory był obciążony.

Odpowiedzią na wady procedury *holdout* jest *k*-krotna kroswalidacja. Korzysta ona z podziału zbioru danych na *k* podzbiorów. Kolejne klasyfikatory o tej samej strukturze są opracowywane na zbiorze danych składającym się z $k - 1$ części początkowego zbioru. Ocena efektywności klasyfikatora jest przeprowadzana na pozostałej części danych, wyłączonej z udziału w konstruowaniu klasyfikatora. Te dwa kroki są powtarzane *k*-krotnie. Za każdym razem inna część oryginalnego zbioru jest wyłączana jako zbiór walidujący. Ocenę klasyfikatora stanowi uśredniona ocena uzyskana podczas *k* powtórzeń. Całą procedurę powtarza się dla klasyfikatorów o różnej architekturze, a za najlepszy uznaje się taki klasyfikator, który uzyskał najwyższą ocenę średnią w ramach *k* powtórzeń. Należy podkreślić, że efekty tej oceny zależą od wybranej wartości *k*. Dla dużego *k* średnie oceny klasyfikatora uzyskane w kolejnych powtórzeniach będzie charakteryzował duży rozrzut, natomiast obciążenie ostatecznej oceny będzie małe. Odwrotnie, w przypadku małych wartości *k* rozrzut średnich ocen będzie mały, lecz obciążenie ostatecznej oceny wzrośnie. Wybór wartości *k* jest uwarunkowany wielkością zbioru danych. Dla dużych zbiorów przyjmuje się za wartość wystarczającą $k \cong 3$. Im mniejszy zbiór, tym większe *k* jest bardziej odpowiednie. W granicznym przypadku *k* jest równe liczbie elementów w zbiorze danych i metoda przyjmuje nazwę *leave-one-out*. Z definicji stosuje się ją dla bardzo małych, rzadkich zbiorów, gdyż pozwala zachować jak najwięcej danych w zbiorze uczącym i przeanalizować jak najwięcej realizacji tego samego klasyfikatora. Wybierając wartość *k*, należy pamiętać, że wraz z jej zwiększeniem znacznie się zwiększają nakłady obliczeniowe.

W razie niedostatecznej liczby danych można skorzystać z technik pozwalających sztucznie zróżnicować zawartość zbioru danych uczących. Do najważniejszych należą *bagging* i *boosting*.

Bagging [23] jest przykładem agregacji *bootstrapowej* i opiera się na losowaniu ze zwracaniem z oryginalnego zbioru danych. Zbiór danych uzyskany w wyniku losowania ma taki sam rozmiar jak oryginalny. Losowanie jest wykonywane zgodnie z jednostajnym rozkładem prawdopodobieństwa. Przeciętnie trafia do niego 63% danych z początkowego zbioru. Reszta to powtórzenia. Poszczególne klasyfikatory są

budowane na różnych wylosowanych zbiorach z zastosowaniem tego samego algorytmu. Wynik zespołu klasyfikatorów jest uzgadniany na zasadzie głosowania większościowego. Wykazano, że dzięki zastosowaniu takiego rozwiązania można znacznie poprawić efektywność klasyfikatorów niestabilnych, tzn. takich, które są wrażliwe na niewielkie zakłócenia w zbiorze danych i wskutek tego mogą zmieniać ocenę przynależności wektora testowego. Procedura *baggingu* jest z powodzeniem stosowana do drzew decyzyjnych oraz sieci neuronowych jako elementów zespołu klasyfikatorów. Nie wnosi ona natomiast wiele, gdy stosuje się klasyfikatory odporne, jak *k*-NN czy SVM.

Modyfikacją *baggingu* jest *boosting* [23]. Metodą tą również losowany jest zbiór uczący o rozmiarze równym zbiorowi oryginalnemu. Powstaje klasyfikator, który jest testowany na oryginalnym zbiorze danych. W kolejnych losowaniach pseudozbioru uczącego zmienia się rozkład prawdopodobieństwa, według którego następuje losowanie. Wzrasta prawdopodobieństwo wylosowania tych wektorów cech, dla których w poprzednich krokach procedury uzyskano niepoprawny wynik klasyfikacji. Rezultaty uzyskane ze wszystkich klasyfikatorów w zespole są agregowane z zastosowaniem wag reprezentujących efektywność poszczególnych klasyfikatorów. Podstawowy algorytm *boostingu* nosi nazwę *AdaBoost*. Choć algorytm ma charakter ogólny, klasyfikatorami najczęściej opracowywanymi z jego udziałem są, podobnie jak w przypadku *baggingu*, drzewa decyzyjne. Jest to uwarunkowane możliwością ich szybkiej budowy z wykorzystaniem standardowych procedur, takich jak CART, C4.5, OC1 oraz interpretacji jako zestawu reguł. W praktyce najlepiej posługiwać się zespołami od kilkudziesięciu do kilkuset takich klasyfikatorów. Pokazano, że prawdopodobieństwo popełnienia błędu przez zespół klasyfikatorów jest gładką funkcją ich liczby. Niestety z rozważań teoretycznych wynika, że nie dąży ono do minimum, równego ryzyku bayesowskiemu w warunkach, gdy liczność próby uczącej nie jest nieskończona.

Obecnie algorytmy *boostingu* w zastosowaniu do drzew klasyfikacyjnych stanowią jeden z dwóch najlepszych algorytmów klasyfikacji o charakterze uniwersalnym [169]. Oznacza to, że można je stosować bez specjalnych zabiegów wstępnych, obejmujących np. analizę typu zmiennych wektora cech czy rozkładu prawdopodobieństwa w klasach.

Ocena efektywności klasyfikatora w stosunku do nieznanych danych jest zagadnieniem innym niż walidacja i nie powinna być z nim mylona. Użycie w tym miejscu oceny klasyfikatora uzyskanej w toku walidacji byłoby zbyt optymistyczne, gdyż wybór najlepszego klasyfikatora został dokonany właśnie na jej podstawie. Uzyskanie poszukiwanej oceny wymaga trzeciego zbioru danych, tzw. zbioru testowego, który jest niezależny od zbioru uczącego i od zbioru walidującego. Jest on potrzebny wyłącznie po to, by ocenić efektywność w pełni gotowego modelu, który został wybrany na podstawie efektywności klasyfikacji zbioru walidującego spośród wszystkich klasyfikatorów opracowanych dla określonych danych uczących. Gdy wielkość zbioru danych umożliwia realizację tego najbardziej poprawnego sposobu oceny metody obliczeniowej, przyjmuje się na ogół podział zbioru między uczący, walidujący i te-

stujący w proporcjach 1/2, 1/4 i 1/4. W innych sytuacjach zbiór jest dzielony na uczący i walidujący, gdy zachodzi potrzeba porównania kilku klasyfikatorów, lub uczący i testujący, jeśli takiej potrzeby nie ma. W skrajnych okolicznościach dysponujemy oceną efektywności klasyfikatora wyłącznie w odniesieniu do danych uczących (tzw. metoda resubstytucji.

W czujnikowych pomiarach gazów często występuje problem niewystarczająco licznych danych pomiarowych do budowy systemu rozpoznawania wzorców. Problem ten dotyczy przede wszystkimi opracowywania systemów na potrzeby oznaczeń ilościowych. Wynika on z dużej czasochłonności i pracochłonności procesu przygotowywania wzorcowych mieszanin gazów. Stąd stosowane są techniki zwiększania zbioru danych przez generowanie danych z wykorzystaniem rozkładów prawdopodobieństwa o założonych parametrach, najczęściej szacowanych z próby.

7.2. Miary efektywności systemu analizy danych

Ocena systemu analizy danych jest na ogół wykonywana z zastosowaniem miar reprezentujących błąd popełniany przez system. W wypadku klasyfikatorów za rozstrzygający dla oceny ich działania przyjmuje się błąd klasyfikacji. Inne miary efektywności dotyczą np. kosztu pozyskania cech będących podstawą klasyfikacji (również w kategoriach stopnia skomplikowania tej operacji), kosztu pozyskania odpowiedniej liczby danych, wymaganych zasobów obliczeniowych czy nakładu czasu koniecznych do opracowania klasyfikatora [24]. Są one stosowane jako miary uzupełniające.

W analizie danych z pomiarów czujnikowych podstawą oceny efektywności odczytu informacji o charakterze jakościowym jest najczęściej miara, określana jako udział błędnych klasyfikacji (ang. *misclassification rate*, MCR). Sposób obliczenia udziału błędnych klasyfikacji w przypadku zastosowania *k*-krotnej kroswalidacji określa wzór:

$$\mathrm{MCR} = \frac{1}{N_k} \sum_{k=1}^{N_k} \frac{n_{fk}}{n_k} \tag{7.1}$$

gdzie: n_{fk} jest liczbą wektorów danych sklasyfikowanych niepoprawnie w *k*-tym powtórzeniu, n_k jest liczbą wszystkich testowych wektorów danych w *k*-tym powtórzeniu, N_k jest liczbą powtórzeń. Im mniejsza wartość tej miary, tym większa efektywność rozpoznawania wzorców przez system. Warto nadmienić, że prawdopodobieństwo błędnej klasyfikacji jest równe wartości oczekiwanej udziału błędnych klasyfikacji. Ponadto, w przypadku zero-jedynkowej funkcji strat jest ono równe ryzyku całkowitemu.

W równaniu (7.1) wszystkie możliwe błędy klasyfikacji są traktowane w sposób symetryczny. Jest to bardzo wygodne, zwłaszcza w wypadku problemów klasyfikacji wieloklasowej, rozwiązywanych w systemie typu *all-against-all*. Istnieją jednak realne problemy klasyfikacji, gdy przydatne jest niesymetryczne traktowanie błędów określenia przynależności do różnych klas. Dla problemu dwóch klas stosuje się na przykład pojęcia czułości i specyficzności [233]. Definicję tych pojęć przedstawiono za pomocą równań (7.2) i (7.3), posługując się oznaczeniami jak w macierzy rezultatów klasyfikacji (tabela 7.1).

Tabela 7.1. Macierz rezultatów klasyfikacji

Klasa	Sklasyfikowane jako A	Sklasyfikowane jako B
A (stan akceptowalny)	n_{AA}	n_{AB}
B (stan niepożądany)	n_{BA}	n_{BB}

$$\text{czułość} = \frac{n_{BB}}{n_{BB} + n_{BA}} \tag{7.2}$$

$$\text{specyficzność} = \frac{n_{AA}}{n_{AA} + n_{AB}} \tag{7.3}$$

Czułość określa prawdopodobieństwo przewidzenia stanu niepożądanego pod warunkiem, że on taki rzeczywiście jest. Specyficzność natomiast informuje o prawdopodobieństwie zdiagnozowania stanu akceptowalnego dla takich samych założeń. Jest intuicyjnie czytelne, że bezbłędna detekcja stanu niepożądanego ma inną rangę niż stwierdzenie, że nic złego się nie dzieje, kiedy faktycznie brak zagrożenia. Kontekst środowiskowy takiego podziału na dwie klasy jest oczywisty. Dlatego w analizie danych z systemów czujnikowych, np. na potrzeby diagnostyki stanów zagrożeń środowiska, należy się liczyć z możliwością stosowania niesymetrycznej oceny błędów klasyfikacji.

W odniesieniu do problemów o charakterze ilościowym najczęściej stosuje się dwa rodzaje błędów – średni błąd kwadratowy predykcji, RMSE [68] i średni błąd względny predykcji, MRE). Sposób obliczenia średniego błędu kwadratowego predykcji w przypadku zastosowania *k*-krotnej kroswalidacji (RMSE) określa wzór:

$$\text{RMSE} = \frac{1}{n_k} \sum_{k=1}^{N_k} \text{RMSE}_k \tag{7.4}$$

$$\mathrm{RMSE}_k = \sqrt{\frac{\sum_{i=1}^{n_k}\left(c_i - \widehat{c}_i\right)^2}{n_k}} \tag{7.5}$$

gdzie: RMSE_k jest błędem obliczonym dla jednego powtórzenia, c_i jest pojedynczym stężeniem rzeczywistym, a $\widehat{c}_i$ jest odpowiadającym mu stężeniem obliczonym.

Sposób obliczenia średniego błędu względnego predykcji w przypadku zastosowania *k*-krotnej kroswalidacji (MRE) określa wzór:

$$\mathrm{MRE} = \frac{1}{N_k}\sum_{k=1}^{N_k}\mathrm{MRE}_k \tag{7.6}$$

$$\mathrm{MRE}_k = \frac{1}{n_k}\sum_{k=1}^{n_k}\left|\frac{\widehat{c}_i - c_i}{c_i}\right| \tag{7.7}$$

gdzie: MRE_k jest błędem obliczonym dla jednego powtórzenia.

Średni błąd kwadratowy predykcji jest miarą bezwzględną, podawaną w jednostkach wielkości określanej przez system analizy danych, np. w ppm. Średni błąd względny predykcji natomiast to miara względna, podawana na ogół jako ułamek jedności. Umożliwia on porównanie efektowności systemu w zastosowaniu do określania różnych miar ilościowych, jak też w zastosowaniu do różnych zakresów wartości tej samej miary.

8. Zakres analizy danych z czujnikowych pomiarów zanieczyszczeń

Przedstawiona w monografii analiza danych dotyczy sposobu pozyskiwania informacji o zanieczyszczeniach gazowych na podstawie czujnikowych danych pomiarowych. Zakres analizy określono, odnosząc się do następujących zagadnień:

- poszukiwane informacje o zanieczyszczeniach,
- zanieczyszczenia,
- matryca czujników,
- dane,
- reprezentacja informacji o zanieczyszczeniu,
- metody odczytu informacji,
- ocena metod pozyskiwania informacji.

Przyjęty zakres analizy odpowiada poczynionym przez autorkę założeniom dotyczącym efektywnego pozyskiwania informacji o zanieczyszczeniach na podstawie pomiarów czujnikowych. Najważniejsze z tych założeń są następujące:

- Na podstawie danych pomiarowych z pomiaru czujnikowego można uzyskać informacje dotyczące różnych właściwości zanieczyszczeń gazowych zarówno o charakterze jakościowym, jak i ilościowym.
- W tym celu uprzywilejowane jest korzystanie ze stosunkowo prostych metod analizy danych, niewymagających pod względem czasu i zasobów obliczeniowych oraz ilości danych koniecznych do parametryzacji modeli, lecz jednocześnie pozwalających uzyskać zadowalającą dokładność oznaczeń.
- Określony rodzaj informacji jest łatwiej dostępny, jeżeli dla jego pozyskania posłużyć się odpowiednią reprezentacją badanego gazu. Innymi słowy, zasadnicze znaczenie ma rozwiązywanie poszczególnych problemów klasyfikacji i regresji w przestrzeniach cech, gdzie można uzyskać najbardziej adekwatny rezultat.
- Wśród przestrzeni umożliwiających osiągnięcie zbliżonych rezultatów uprzywilejowane są przestrzenie niskowymiarowe.
- Najlepsze przestrzenie cech powinny być znajdowane w kontekście metod klasyfikacji lub regresji, stosowanych do odczytu konkretnego rodzaju informacji.

Analizę danych przeprowadzono w środowisku obliczeń naukowych Matlab. Autorka napisała w tym celu algorytmy realizujące różne konfiguracje systemów rozpoznawania wzorców o różnych konfiguracjach.

8.1. Poszukiwane informacje o zanieczyszczeniach

Analizując dane, poszukiwano informacji dotyczącej jakościowych i ilościowych właściwości zanieczyszczeń gazowych. Odniesiono się do nich w tradycyjnym rozumieniu, tj. do tożsamości chemicznej substancji zanieczyszczających oraz ich stężenia. Tyle samo uwagi poświęcono jednak pozyskiwaniu innej niż analityczna informacji o środowisku. Dane pochodzące z pomiarów czujnikowych bardzo dobrze nadają się do realizacji tego celu.

Poszukiwano następujących rodzajów informacji o charakterze jakościowym:

- rodzaj substancji zanieczyszczającej,
- przynależność do kategorii substancji zanieczyszczających,
- skład jakościowy mieszaniny substancji zanieczyszczających,
- przynależność do kategorii mieszanin substancji zanieczyszczających.

Poszukiwano następujących rodzajów informacji o charakterze ilościowym:

- stężenie substancji zanieczyszczającej,
- miara ilościowa inna niż stężenie substancji zanieczyszczającej; rozważono takie miary, jak suma stężeń substancji zanieczyszczających oraz stężenie atomów węgla pochodzących od substancji zanieczyszczających.

8.2. Zanieczyszczenia badane

Dogodnym obiektem badawczym pozwalającym przeanalizować problem pozyskiwania zarówno analitycznej, jak i innej niż analityczna informacji o środowisku jest grupa zanieczyszczeń określana jako lotne związki organiczne. Właściwości toksyczne i kancerogenne tych substancji mogą powodować zagrożenie dla zdrowia ludzi w wypadku narażenia na ich duże stężenia. Jednak również w małych stężeniach związki te nie są obojętne dla człowieka. Uważa się je za istotny czynnik wpływający na pogorszenie jakości powietrza wewnętrznego, współodpowiedzialny za syndrom chorego budynku. Z obecnością tych substancji w powietrzu związane jest wzmocnienie niektórych objawów, zwłaszcza astmatycznych i alergicznych, jak również ogólne zmniejszenie sprawności psychofizycznej użytkowników pomieszczeń [214]. LZO występują na ogół w środowisku w mieszaninach wieloskładnikowych, których oddziaływanie nie jest sumą oddziaływań poszczególnych składników. W związku z tym interesująca jest możliwość ich oceny jako zanieczyszczenia raczej w sposób łączny niż w podziale na poszczególne substancje.

Poddano analizie czujnikowe dane pomiarowe dotyczące mieszanin lotnych związków organicznych w powietrzu. Były to mieszaniny jednoskładnikowe (jeden LZO w powietrzu) i dwuskładnikowe (dwa LZO w powietrzu). Oznaczanie mieszanin

jednoskładnikowych jest uważane za najprostszy z możliwych problemów analizy danych w czujnikowych pomiarach zanieczyszczeń, choć w ogólnym wypadku przekonanie to nie odpowiada prawdzie. Trudność określenia zależy bowiem od rodzaju poszukiwanej informacji. Przypadek mieszanin dwuskładnikowych jest z kolei bliższy rzeczywistym problemom środowiskowym. W szczególności oznaczanie mieszanin charakteryzujących się ilościową dysproporcją składników dobrze ilustruje okoliczności związane z produkcją i stosowaniem rozpuszczalników organicznych. Według danych literaturowych jest to podstawowe źródło antropogenicznej emisji LZO.

Tabela 8.1. Jednoskładnikowe mieszaniny LZO

Substancja zanieczyszczająca	Zakres stężeń w mieszaninie z powietrzem [ppm]/[mmol C/m^3]	Liczba zestawów danych pomiarowych[a]
Heksan	17–204/5–55	4
Heptan	15–183/5–65	
Oktan	14–165/4–52	
Cykloheksan	21–249/6–67	
Benzen	25–302/8–94	
Toluen	21–255/8–91	
Ksylen	18–222/5–59	
Etylobenzen	18–220/6–79	

[a]Każdy zbiór danych (macierz danych pomiarowych) odnosi się do substancji zanieczyszczającej w innym stężeniu.

Tabela 8.2. Dwuskładnikowe mieszaniny LZO

Dominująca substancja zanieczyszczająca		Pozostałe substancje zanieczyszczające		Zakres stężeń mieszaniny zanieczyszczeń [mmol C/m^3]	Liczba zestawów danych pomiarowych[a]
Nazwa	Zakres stężeń [ppm]	Nazwa	Zakres stężeń [ppm]		
Heksan	128–817	heptan	8–183	46–231	25
		oktan	7–156	46–232	
		cykloheksan	10–249	46–240	
		benzen	13–302	47–254	
		toluen	11–255	47–253	
Toluen	159–1019	heksan	9–204	61–328	
		heptan	8–183	62–328	
		benzen	13–302	66–333	
		etylobenzen	8–183	66–332	
		ksylen	9–222	66–332	

[a]Każdy zbiór danych (macierz danych pomiarowych) odnosi się do innego zestawu stężeń dwóch substancji zanieczyszczających.

Informacje o składzie mieszanin LZO, których dotyczyły analizowane dane pomiarowe, podano w tabelach 8.1 i 8.2. Stężenia poszczególnych substancji podano w ppm oraz jako stężenia atomów węgla organicznego [mmol C/m^3].

8.3. Matryca czujników

Złożone dane pomiarowe użyto jako podstawę określania zanieczyszczeń. Badano możliwość uzyskania wielu różnych informacji o zanieczyszczeniu na podstawie pojedynczego pomiaru. Obiecującym źródłem złożonych danych pomiarowych dotyczących środowiska są matryce czujników. Do pomiaru gazowych zanieczyszczeń powietrza jako elementy takich matryc mogą być użyte rezystancyjne półprzewodnikowe czujniki gazów.

O konkurencyjności chemicznie czułych rezystorów w pomiarach środowiskowych decydują ich właściwości, takie jak bardzo niski koszt, duża czułość, krótki czas odpowiedzi, szybka regeneracja, prosty interfejs elektroniczny, łatwość obsługi, małe wymagania pod względem utrzymania oraz możliwość detekcji bardzo dużej liczby gazów [234, 235]. Czujniki oparte na tlenkach metali stanowią najlepszy wybór pod względem zastosowania w systemach monitoringu zanieczyszczeń pracujących w sposób ciągły [76, 236]. Korzystając z komercyjnie dostępnych czujników półprzewodnikowych, można wykonywać pomiary stężeń gazowych zanieczyszczeń powietrza, np. lotnych związków organicznych od kilku do kilku tysięcy ppm, co odpowiada stężeniom spotykanym na stanowiskach pracy. Wyniki licznych prac dowodzą możliwości obniżenia progu detekcji tego typu czujników do poziomu ppb [55, 57, 147, 237]. Dotychczas brakuje jednak zastosowania przedstawianych propozycji w rozwiązaniach komercyjnych.

Podczas ekspozycji na gazy redukujące półprzewodnikowe czujniki rezystancyjne wykazują zmiany rezystancji. Wynikają one z reakcji utleniania tych gazów, zachodzących z udziałem tlenu, chemicznie zaadsorbowanego na warstwie chemoczułej. Większość półprzewodzących tlenków metali (tlenek cynku, dwutlenek cyny, dwutlenek tytanu, tlenek żelaza(III)) wykazuje przewodnictwo typu n. Jego przyczyną jest proces uwalniania elektronów w wyniku powstawania wakancji tlenowych. Tego rodzaju tlenki są w przeważającej części stosowane w chemicznych czujnikach gazów. Mniej liczna jest grupa tlenków metali charakteryzujących się przewodnictwem typu p (tlenek niklu, tlenek kobaltu i niewiele innych). Tlenki te są również rzadziej wykorzystywane jako materiały chemoczułe. Szczegółowy opis mechanizmu działania rezystancyjnych półprzewodnikowych czujników gazów wykracza poza zakres tej pracy. Zainteresowany czytelnik znajdzie wyczerpujące omówienie zagadnienia w monografii [33].

Charakterystyka czujnika półprzewodnikowego, na którą składają się czas i dynamika odpowiedzi, czułość oraz selektywność, w znacznym stopniu zależy od powierzchni właściwej materiału chemoczułego, gęstości donorów ładunku, rodzaju i stopnia aglomeracji materiału, porowatości warstwy, równowagi kwasowo-zasadowej w materiale chemoczułym, występowania katalizatora, temperatury pracy i innych czynników. Odpowiednio kształtując te czynniki, można uzyskać czujniki o różnych charakterystykach. Matryca takich czujników jest atrakcyjnym źródłem złożonych danych pomiarowych.

Dane pomiarowe analizowane przez autorkę pochodziły z pomiarów wykonanych z użyciem matrycy czujników półprzewodnikowych. Składała się ona z piętnastu sensorów gazów Taguchi Gas Sensors (TGS) firmy Figaro. Nazwy katalogowe czujników oraz podstawowe informacje o ich przeznaczeniu zamieszczono w tabeli 8.3. Jeśli było to możliwe, dla poszczególnych czujników podawano zakres detekcji w odniesieniu do etanolu (C_2H_5OH). Dzięki temu można porównać zakres detekcji różnych sensorów matrycy dla przykładowego lotnego związku organicznego.

Tabela 8.3. Czujniki stanowiące elementy matrycy i obszar ich zastosowania

Nazwa katalogowa	Właściwości	Przykładowe zastosowanie	Zakres detekcji	Nr w matrycy
TGS 821	wysoka czułość na wodór	detekcja wodoru, konserwacja transformatorów, akumulatory, przemysł stalowniczy	1000–5000 ppm C_2H_5OH	1
TGS 822	wysoka czułość na pary rozpuszczalników organicznych (etanol)	alkomaty, detektory wycieku gazu, detektory rozpuszczalników dla przemysłu, pralni chemicznych i przemysłu półprzewodnikowego	50–5000 ppm C_2H_5OH	2
TGS 824	wysoka czułość na związki amonowe	detekcja wycieku amoniaku w systemach chłodzenia, kontrola wentylacji w przemyśle rolno-spożywczym i drobiarskim	30–300 ppm NH_3	3
TGS 825	wysoka czułość na pary H_2S	detektory H_2S	5–100 ppm H_2S	4
TGS 826	wysoka czułość na związki amonowe	detektory wycieku amoniaku z systemów chłodzenia, kontrola wentylacji w przemyśle rolno-spożywczym i drobiarskim	9–100 ppm NH_3	5
TGS 880	czułość na parę wodną i opary kuchenne (alkohol, związki zapachowe)	sterowanie procesem gotowania	30–3000 ppm C_2H_5OH	6

Nazwa katalogowa	Właściwości	Przykładowe zastosowanie	Zakres detekcji	Nr w matrycy
TGS 883	czułość na parę wodą	sterowanie procesem gotowania	1–100 g/m^3 H_2O	7
TGS 800	reaguje na zanieczyszczenia powietrza	sterowanie jakością powietrza, sterowanie wentylacją	1–100 ppm C_2H_5OH	8
TGS 2201[a]	wysoka czułość na spaliny z silników benzynowych	sterowanie wentylacją w pojazdach	10–100 ppm CH_3OH	9
TGS 2201[a]	wysoka czułość na spaliny z silników wysokoprężnych		0,1–1 ppm H_2S	10
TGS 2106			0,5–5 ppm NO_2	11
TGS 2104	wysoka czułość na gazy spalinowe z silników benzynowych		2–20 ppm $C_{10}H_{22}$	12
TGS 2602	wysoka czułość na LZO i gazy o właściwościach odorowych	urządzenia oczyszczające powietrze, sterowanie wentylacją, monitory jakości powietrza, monitory LZO, monitory gazu o właściwościach odorowych.	1–30 ppm C_2H_5OH	13
TGS 2620	wysoka czułość na alkohol i pary rozpuszczalników organicznych	alkomaty, detektory par związków organicznych, detektory rozpuszczalników dla przemysłu pralni chemicznych i przemysłu półprzewodnikowego	50–5000 ppm C_2H_5OH	14
TGS 2600	wysoka czułość na gazowe zanieczyszczenia powietrza	urządzenia oczyszczające powietrze, sterowanie wentylacją, pomiary jakości powietrza.	1–100 ppm C_2H_5OH	15

[a]Czujnik TGS 2201 ma dwa niezależne elementy chemoczułe na jednym podłożu i dostarcza dwa niezależne sygnały pomiarowe. W matrycy zastosowano dwa takie czujniki, za każdym razem wyprowadzając sygnał z innego elementu chemoczułego.

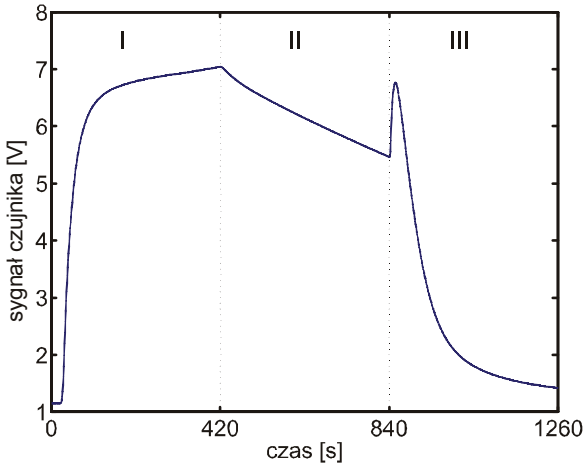
Czujniki typu TGS należą do rezystancyjnych, półprzewodnikowych czujników gazów. Ze względu na możliwość pozyskiwania różnorodnych informacji o zanieczyszczeniu powietrza istotne jest, że zasada działania tych sensorów wyklucza ich selektywność. Czujniki wymienione w tabeli 8.3 reagują na wiele gazów redukujących w dużym i zróżnicowanym zakresie stężeń, od kilku ppm do kilku procent. Z tego względu korzystne jest użycie wielu czujników jednocześnie. Nawet niewielkie różnice w czułości oraz częściowej selektywności poszczególnych sensorów poprawiają różnorodność informacji zawartej w danych pomiarowych. Matryca składająca się z piętnastu czujników stwarza duże możliwości w zakresie pozyskania takich informacji.

8.3.1. Tryb pracy czujników

Użycie matrycy różnych czujników jest podstawowym i dobrze znanym sposobem zwiększenia złożoności czujnikowych danych pomiarowych, a zarazem poszerzenia zakresu zawartej w nich informacji o badanych gazach. Obecnie uwaga badaczy skupia się w znacznym stopniu na innych metodach poprawy złożoności czujnikowych danych pomiarowych, które mogą, lecz nie muszą być stosowane równocześnie (patrz rozdz. 4). Do głównych kierunków poszukiwań należą ingerencja w parametry pracy czujnika lub w warunki jego ekspozycji. Ostatni z podanych sposobów wykorzystano w tej pracy.

Etap	I	II	III
Gaz	badany gaz	badany gaz	czyste powietrze
Natężenie przepływu	niezerowe stałe	zero	niezerowe stałe
Warunki ekspozycji	dynamiczne	statyczne	dynamiczne
Czas trwania etapu	określony	określony	określony

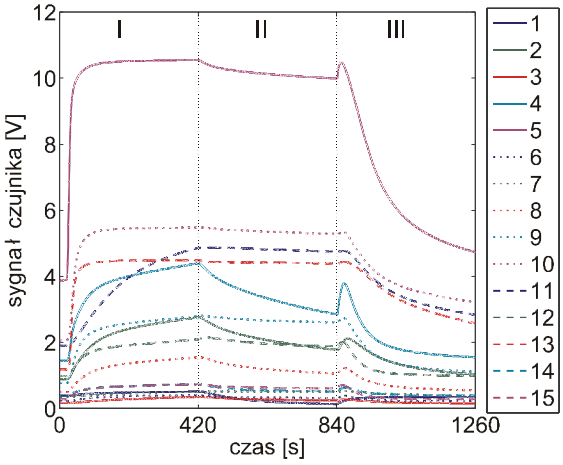
Rys. 8.1. Zasadnicze elementy trybu pracy czujników *stop flow*



Rys. 8.2. Kształt sygnału czujnikowego uzyskanego w trybie *stop flow*

Dane pochodziły z pomiarów czujnikowych wykonanych w trybie pracy *stop flow*. Polegał on na okresowym włączaniu i wyłączaniu przepływu badanego gazu oraz czystego powietrza przez komory czujników według określonego schematu (rys. 8.1). Powodowało to zmiany warunków ekspozycji czujników w czasie. Następowały zarówno zmiany otoczenia czujników (natężenia przepływu gazu, rodzaju

gazu, stężenia badanej substancji), jak i związane z tym zmiany parametrów samych czujników (temperatury czujnika, stopnia pokrycia warstwy chemoczułej gazami, kinetyki procesów adsorpcji/desorpcji, kinetyki reakcji redoks). Zmiany warunków ekspozycji zachodzące w czasie wpływały na sygnał wyjściowy czujnika i przejawiały się w jego charakterystycznym kształcie (rys. 8.2). Stwierdzono też ich zróżnicowany wpływ na sygnały poszczególnych czujników (rys. 8.3). Szczegółowy opis i omówienie trybu pracy *stop flow* znajdzie Czytelnik w pracach [17, 69, 125, 127, 135].



Rys. 8.3. Przykład odpowiedzi matrycy czujników uzyskanej podczas ekspozycji w trybie *stop flow*

Dane pomiarowe analizowane w pracy pochodziły z pomiarów, w których czas trwania wszystkich faz ekspozycji czujników (I–III) wynosił 7 min, natomiat natężenie przepływu gazu wynosiło 1 dm^3/min w fazie I, 0 dm^3/min w fazie II oraz 1 dm^3/min w fazie III. Sygnał uzyskany w trybie *stop flow* jest interesujący jako źródło parametrów sygnału o charakterze dynamicznym oraz statycznym. W pracy analizowano przydatność każdego rodzaju cech do określania zanieczyszczeń gazowych pod względem jakościowym oraz ilościowym.

8.4. Dane

8.4.1. Dane pomiarowe

Złożone dane pomiarowe analizowane w pracy stanowiły cyfrowy zapis sygnałów zarejestrowanych podczas ekspozycji czujników na badane gazy. Sygnał czujnika

składał się z wyników pomiarów wykonanych w kolejnych, dyskretnych momentach czasu ekspozycji. Sygnały wyjściowe czujników były rejestrowane z określoną rozdzielczością w czasie. Rozdzielczość ta określała długość dyskretnego momentu czasu, którego dotyczyła jedna zarejestrowana wartość sygnału czujnika. Moment taki jest określany jako *punkt czasowy*. Dla danych analizowanych w tej pracy jego czas trwania wynosił 1 s.

Do zapisu danych pomiarowych przyjęto notację macierzową (rys. 8.4). Kolumny macierzy danych pomiarowych odnosiły się do poszczególnych czujników. W pojedynczej kolumnie znajdowały się dyskretne wartości sygnału określonego czujnika, zarejestrowane podczas ekspozycji na badaną próbę gazu. Wiersze macierzy wskazywały na dyskretne momenty czasu w trakcie ekspozycji w trybie *stop flow*. W pojedynczym wierszu znajdowały się dyskretne wartości sygnałów wszystkich czujników związane z jednym punktem czasowym ekspozycji. W nomenklaturze tensorowej opisującej struktury danych są to dane rzędu II.

$$R = \begin{bmatrix} r_{11} & \cdots & r_{1j} & \cdots & r_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{k1} & \cdots & r_{kj} & \cdots & r_{kn} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{m1} & \cdots & r_{mj} & \cdots & r_{mn} \end{bmatrix} \begin{matrix} \\ \\ \text{czas} \\ \\ \\ \end{matrix}$$

czujnik

Rys. 8.4. Struktura czujnikowych danych pomiarowych

Macierz analizowanych danych pomiarowych miała wymiary $m \times n$, gdzie m jest liczbą dyskretnych pomiarów wykonanych podczas ekspozycji na badane gazy, n zaś jest liczbą czujników w zestawie. W notacji tej k, $k = 1, \ldots, m$ ($m = 1260$) oznacza kolejne dyskretne momenty czasu w trakcie ekspozycji, natomiast j, $j = 1, \ldots, n$ ($n = 15$) oznacza poszczególne czujniki. Pojedynczy element matrycy r_{kj} jest wynikiem pomiaru wykonanego j-tym czujnikiem, w k-tym punkcie czasowym ekspozycji.

8.4.2. Dane generowane

Uzyskanie poprawnie sparametryzowanych modeli klasyfikacji czy regresji wymaga odpowiednio dużego zbioru danych uczących. Właściwa liczba wektorów danych jest tym większa, im większą liczbę elementów ma wektor cech. W analizach, których wyniki przedstawiono w tej pracy, rozważano wektory cech o różnej liczbie elementów. Jeżeli liczba danych pomiarowych była niewystarczająca do spełnienia

tego warunku, to wdrażano procedurę generowania danych dodatkowych na podstawie rzeczywistych danych pomiarowych. Zaproponowano procedurę, która za punkt wyjścia przyjmowała macierz rzeczywistych danych pomiarowych $R(m \times n)$, w której pojedynczy punkt danych to r_{kj}. Procedura umożliwiała wygenerowanie macierzy danych $R'(m \times n)$, w której pojedynczy punkt danych oznaczono jako r'_{kj}. Pochodził on z rozkładu $t(r_{kj}, s_{kj}/n^{1/2})$. Wartość odchylenia standardowego s_{kj} określono na podstawie statystycznej analizy danych pochodzących z trzynastu powtórzeń pomiaru zanieczyszczenia odniesienia (etanolu) o stężeniu 769 ppm. Było to odchylenie standardowe wyniku pomiaru etanolu i-tym czujnikiem, w k-tym punkcie czasowym ekspozycji dla próby liczącej trzynaście powtórzeń pomiaru. Procedura pozwala wygenerować dowolną liczbę danych dodatkowych na podstawie jednej macierzy danych pomiarowych. Danym wygenerowanym przypisywano etykietę jakościową oraz parametry ilościowe takie same jak odpowiednim danym rzeczywistym. Struktura danych generowanych była taka sama jak struktura danych pomiarowych. Na ogół generowano nie więcej niż cztery macierze danych, uzupełniających na podstawie jednej macierzy danych pomiarowych.

8.4.3. Dane wielowymiarowe

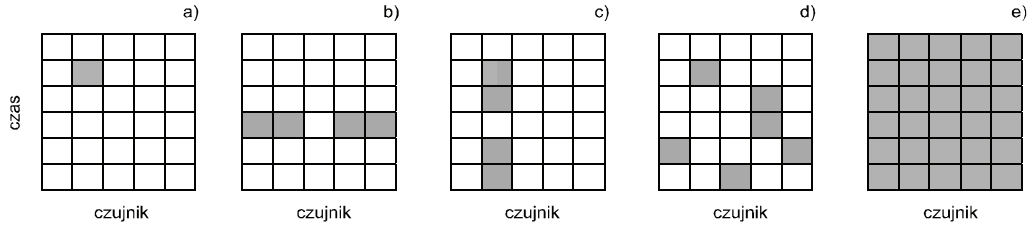
Do parametryzacji modeli klasyfikacji/regresji stosowane są dane wielowymiarowe (patrz rozdz. 4). Kolumny macierzy danych wielowymiarowych odnosiły się do poszczególnych elementów wektora cech. Cechy były zarazem zmiennymi wejściowymi modeli klasyfikacji/regresji, więc ich liczba odpowiadała liczbie wejść modeli. Każdy wiersz macierzy był liczbową realizacją wektora cech odnoszącą się do innej badanej próby gazu. Na potrzeby uczenia z nadzorem poszczególnym wierszom macierzy przypisywano etykietę jakościową lub wartość miary ilościowej charakteryzującej zanieczyszczenie, którego dotyczyły dane umieszczone w tym wierszu. Poszukiwanie przestrzeni cech, w której najlepiej można rozwiązać określony problem klasyfikacji/regresji, wiązało się z koniecznością rozważania danych wielowymiarowych zbudowanych na podstawie różnych wektorów cech.

8.5. Reprezentacja informacji o zanieczyszczeniach

Analizowano możliwość uzyskania wielu różnych informacji o zanieczyszczeniu na podstawie pojedynczego pomiaru czujnikowego. Istotą pomysłu prowadzącego do osiągnięcia tego celu było poszukiwanie rozwiązań poszczególnych problemów klasyfikacji i regresji w różnych przestrzeniach cech. W związku z tym przyjęto koncepcję

cechy, która pozwala uzyskać: i) jak największą liczbę potencjalnych przestrzeni cech, ii) jak największe zróżnicowanie wśród tych przestrzeni.

Cechę zdefiniowano jako wynik pomiaru czujnikowego w dyskretnym momencie ekspozycji. Nazwano ją cechą typu *P* (punktową). W związku z tak przyjętą definicją o macierzy danych pomiarowych można myśleć jako o liczbowej realizacji macierzy cech. Zaproponowano w ten sposób inne podejście do opracowania reprezentacji informacji o badanym gazie w postaci wektora cech. Kontekst sygnału pojedynczego czujnika, tradycyjnie uwzględniany podczas generowania cech zastąpiono kontekstem wszystkich dyskretnych odpowiedzi wszystkich czujników zarejestrowanych podczas pojedynczej ekspozycji matrycy czujników na badane gazy.



Rys. 8.5. Położenie elementów wektora cech w macierzy *R*. Przyjmując podaną interpretację cechy, macierz tę można traktować jako macierz cech

Różne możliwości konstrukcji wektora cech typu *P* pokazano na rys. 8.5. Dwa skrajne przypadki to: i) wektor jednoelementowy $\mathbf{r} = r_{jk}$, którego realizacją jest wartość sygnału jednego czujnika w jednej, dyskretnej chwili (rys. 8.5a) oraz wektor składający się ze wszystkich elementów macierzy *R* rozwiniętej do postaci wektora (rys. 8.5e). Dwa szczególne przypadki zachodzą, gdy wektor cech składa się z elementów r_{jk} takich, że k = const (rys. 8.5c) oraz gdy wektor cech składa się z elementów r_{jk}, gdzie j = const (rys. 8.5b). W pierwszym z nich elementy wektora pochodzą z sygnałów różnych czujników, lecz zostały zarejestrowane w tej samej dyskretnej chwili. Elementy wektora drugiego rodzaju pochodzą z sygnału jednego czujnika, lecz wystąpiły w różnych dyskretnych chwilach. Między wymienionymi skrajnymi i szczególnymi przypadkami mieszczą się wektory cech, których elementy pochodzą z różnych punktów czasowych sygnałów różnych czujników (rys. 8.5d). Liczba elementów w takich wektorach cech może wynosić od 2 do ($mn - 1$). Biorąc pod uwagę, że skład wektora o określonej liczbie elementów może być zróżnicowany, liczba potencjalnych wektorów cech jest bardzo duża. Określa ją wzór:

$$lwc = \sum_{q=1}^{l} \binom{l}{q} = \sum_{q=1}^{l} \frac{l!}{q!(l-q)!} \tag{8.1}$$

gdzie: $l = mn$ jest liczbą elementów macierzy cech (jest równa liczbie elementów macierzy danych); q jest liczbą elementów wektora cech.

Przyjęta formuła cechy umożliwia konstrukcję wektorów cech przenoszących informację o dowolnej skali reprezentacji w danych pomiarowych, od bardzo lokalnej (rys. 8.5a) po absolutnie globalną (rys. 8.5e). Rozwiązanie to jest od kilku lat badane przez autorkę. Dotychczas opublikowane prace [17, 69, 125, 135, 127] dotyczą wektora cech typu *P*, jaki pokazano na rys. 8.5b. Przedstawione w nich rezultaty dowodzą, że informacja o badanych gazach jest bardzo dobrze reprezentowana w wektorach cech zbudowanych z wartości sygnałów różnych czujników związanych z tym samym momentem ekspozycji. Jako punkt wyjścia do konstrukcji najlepszych wektorów cech przyjęto założenie, że ich elementy mogą mieć dowolne położenie w macierzy cech.

Aby przyspieszyć obliczenia, przeprowadzono wstępną redukcję wymiarów oryginalnej macierzy cech. Posłużono się procedurą *subsamplingu*. Zastosowano heurystyczny schemat próbkowania nawiązujący do zróżnicowanej dynamiki sygnału (rys. 8.2). Schemat *subsamplingu* sygnału przedstawiono w tabeli 8.4. Sygnał każdego z piętnastu czujników próbkowano według tego samego schematu. Uzyskano w ten sposób macierz cech 122 × 15. Mimo redukcji liczba wektorów cech, które można było zbudować na podstawie tej macierzy i przyjąć za podstawę rozwiązywania poszczególnych problemów klasyfikacji czy regresji, nadal była bardzo duża. Różniły się one pod względem przydatności. Przyjęto więc określone strategie ekstrakcji i selekcji, aby wyłonić zestawy cech najbardziej użytecznych do pozyskiwania konkretnych informacji o zanieczyszczeniach.

Tabela 8.4. Schemat heurystycznego *subsamplingu* sygnału czujnikowego zarejestrowanego podczas ekspozycji w trybie *stop flow* (por. rys. 8.2)

Etap	Przedział czasu [s]	Czas między próbkami [s]
I	20–59	3
	60–119	5
	120–239	10
	240–419	20
II	420–839	20
III	840–899	3
	900–959	5
	960–1079	10
	1080–1260	20

8.6. Wybór najlepszych przestrzeni cech

8.6.1. Selekcja cech

Ze względów praktycznych (prostota i krótki czas budowy modeli oraz rozsądne wymagania co do liczby wektorów uczących) atrakcyjne są wektory cech zbudowane

z jak najmniejszej liczby elementów. Innymi słowy, korzystne jest rozwiązywanie problemów klasyfikacji i/lub regresji w przestrzeniach jak najmniej wymiarowych. Przyjęta w pracy definicja cechy umożliwia konstrukcję olbrzymiej liczby wektorów cech. Wynosi ona $2^{330} - 1$ dla macierzy cech o wymiarach 122×15. Poszczególne wektory mogą mieć różną liczbę elementów, a ich elementy zróżnicowane rozmieszczenie w macierzy cech. Znalezienie najlepszych wektorów wymaga przeszukania zbioru wektorów możliwych. Za najlepsze uważane są te wektory, które pozwalają uzyskać poszukiwaną informację z jak największą dokładnością.

Biorąc pod uwagę preferencje w stosunku do małej liczby elementów wektorów cech, rozważono wektory 1-, 2-, 3-, 5- i 7-elementowe. Z wyjątkiem wektora jednoelementowego przegląd zupełny zbiorów wektorów kandydujących nie był możliwy w akceptowalnym czasie obliczeniowym (patrz rozdz. 5). Przyjęto zatem strategię selekcji cech z zastosowaniem przeszukiwania niezupełnego. Jako metodę przeszukiwania wybrano symulowane wyżarzanie [98]. Selekcjonowano 30 najlepszych wektorów cech o określonej liczbie elementów dla każdego rozważanego problemu klasyfikacji/regresji. Bazując na tej puli wektorów oceniano możliwości wektora cech typu *P* o określonej liczbie elementów. W sposób zupełny przejrzano wyłącznie zbiór jednoelementowych wektorów cech.

Poszukiwano wektorów cech jak najlepiej nadających się do rozwiązania poszczególnych problemów klasyfikacji czy regresji. Oceniając wektory cech, uwzględniano zawsze kontekst problemu. W różnym stopniu natomiast brano pod uwagę kontekst klasyfikatora/modelu regresji. Porównano efekty uzyskane w przypadku: i) braku odniesienia do niego (filtracja jednowymiarowa), ii) wykorzystania klasyfikatora jako narzędzia oceny wektora cech (podejście opakowane), iii) włączenia klasyfikatora w proces konstrukcji wektora cech (podejście wbudowane) (patrz rozdz. 5).

Jako narzędzie filtracji cech ze względu na pozyskiwanie z danych informacji jakościowej wybrano jednowymiarową analizę wariancji (ang. *analysis of variance*, ANOVA), jako narzędzie oceny cechy ze względu na pozyskiwanie informacji ilościowej przyjęto natomiast współczynnik korelacji między wartościami cechy a wartościami określonej miary ilościowej opisującej zanieczyszczenie. Selekcję cech w trybie opakowanym prowadzono w odniesieniu do metod klasyfikacji i metod regresji wybranych dla pozyskania informacji o zanieczyszczeniach. Selekcja cech ze względu na rozwiązywanie problemów ilościowych jest rzadko rozważana w literaturze czujnikowej, stąd przedstawione w pracy wyniki mogą wzbudzić szczególne zainteresowanie. Selekcję cech na sposób wbudowany zastosowano tylko w odniesieniu do problemu pozyskiwania informacji jakościowej. Zastosowano komitet drzew klasyfikacji.

8.6.2. Ekstrakcja cech

Równocześnie z selekcją cech typu *P* dla uzyskania wektorów cech o niewielkiej liczbie elementów zastosowano ekstrakcję cech. Operacja ta umożliwia szybką reduk-

cję wymiarowości wielowymiarowej przestrzeni cech, jeżeli cechy charakteryzuje współliniowość.

Czujniki, które stanowiły źródło analizowanych danych pomiarowych były częściowo selektywne i ich działanie opierało się na tym samym mechanizmie chemoczułym. W takich okolicznościach występuje zwykle wzajemne skorelowanie odpowiedzi sensorów. Z przyjętej koncepcji cechy wynikało wzajemne skorelowane składowych wyjściowej przestrzeni cech. Ponadto, przestrzeń ta miała kilkanaście tysięcy wymiarów.

Cechy wtórne otrzymane w wyniku ekstrakcji ujmują syntetycznie informację współdzieloną przez wszystkie cechy pierwotne. Wskutek wykorzystania wszystkich cech typu *P* wektor cech wyekstrahowanych zawiera skompresowaną informację występującą w skali całych danych pomiarowych, inaczej niż w przypadku wektora zbudowanego z wyselekcjonowanych cech typu *P*, który uwzględniał informację zawartą w pewnych tylko fragmentach macierzy danych, pomijając inne.

W celu ekstrakcji cech posłużono się analizą składowych głównych (patrz p. 6.1.1) oraz metodą cząstkowych najmniejszych kwadratów (patrz p. 6.3.3). Zasadnicza różnica między nimi polega na tym, że składowe główne wyłaniane są bez kontekstu poszukiwanej informacji. Ujawniają tylko hierarchię kierunków zmienności w danych. W wyniku ekstrakcji z zastosowaniem PCA powstaje jedna, wielowymiarowa przestrzeń cech, która jest obrazem oryginalnej przestrzeni cech. Tylko w tej jednej przestrzeni (ewentualnie w jej podprzestrzeniach) można poszukiwać rozwiązań różnorodnych problemów jakościowych/ilościowych. Natomiast metoda cząstkowych najmniejszych kwadratów wyłania składowe ze względu na ich udział w wyjaśnianiu zmian zmiennej objaśnianej. Zmiana zmiennej objaśnianej powoduje zmianę wyniku przekształcenia. Mamy tu do czynienia z rodzajem dostrajania przestrzeni cech do konkretnego problemu oznaczania gazów.

Cechy wyekstrahowanie w wyniku analizy składowych głównych zastosowano zarówno w rozwiązywaniu problemów jakościowych, jak i ilościowych. Cechy uzyskane metodą cząstkowych najmniejszych kwadratów rozważano tylko do pozyskiwania informacji ilościowej o zanieczyszczeniach powietrza. W obu przypadkach brano pod uwagę 1-, 2-, 3-, 5- i 7-elementowe wektory cech.

8.7. Metody odczytu informacji o zanieczyszczeniach

Przyjęto założenie, że określanie zanieczyszczeń na podstawie pomiarów czujnikowych powinno przebiegać na podstawie jak najbardziej oszczędnej, lecz zarazem efektywnej reprezentacji informacji o zanieczyszczeniu. W znalezienie tej reprezentacji powinien być włożony główny wysiłek obliczeniowy w procesie opracowywania systemu analizy danych (patrz rozdz. 2). Rozstrzygnięcie to pozwala zwrócić się ku modelom klasyfikacji i regresji o relatywnie prostej strukturze, efektywnie współpra-

cującym z niewieloma zmiennymi wejściowymi, mniej wymagającym pod względem zapotrzebowania na dane uczące oraz o krótkim czasie uczenia. Dzięki tym właściwościom modele mogą też uczestniczyć aktywnie w procesie konstrukcji najlepszych przestrzeni cech.

Z przedstawionych powodów w tej pracy nie zajmowano się w ogóle podejściem koneksjonistycznym, mimo jego niewątpliwej popularności w analizie danych z pomiarów czujnikowych (patrz p. 6.4). Zastosowano natomiast metody umożliwiające określenie struktury modelu i jego parametryzację na podstawie jednokrotnej prezentacji zbioru uczącego.

Analizę struktury zmienności danych oraz jej źródeł przeprowadzono z zastosowaniem analizy składowych głównych.

Do pozyskiwania informacji jakościowej zastosowano metody klasyfikacji reprezentujące główne podejścia do konstrukcji klasyfikatorów. Wybrano metody proste, lecz skuteczne w rozważanym zakresie zastosowania, takie jak:

- liniowa analiza dyskryminacyjna (patrz p. 6.2.1) będąca przedstawicielem klasyfikatorów działających na zasadzie horyzontalnego podziału przestrzeni cech,
- metoda *k*-najbliższych sąsiadów (patrz p. 6.2.2), która stanowi nieliniowy klasyfikator oparty na określaniu i porównywaniu odległości między wzorcami,
- komitet drzew klasyfikacji (patrz p. 6.2.5 i 6.2.6) reprezentujący klasyfikatory działające na zasadzie hierarchicznego podziału przestrzeni cech, a zarazem będący przykładem komitetu klasyfikatorów.

Każde zagadnienie pozyskiwania informacji jakościowej o zanieczyszczeniu przedstawiano i rozwiązywano jako problem klasyfikacji typu jeden przeciw wszystkim.

Do pozyskiwania informacji o charakterze ilościowym zastosowano analizę regresji. Wybrano następujące rodzaje modeli regresji:

- regresję liniową wielokrotną odporną (patrz p. 6.3),

– cech wyselekcjonowanych:

$$c = \alpha_0 + \sum_{k=1}^{K} \alpha_k X_{i,t} + \varepsilon \tag{8.2}$$

gdzie: X_{it} jest wartością sygnału czujnika *i*-tego w chwili *t* podczas ekspozycji matrycy czujników na badane gazy, związaną ze stężeniem lub miarą ilościową *c*; *K* jest liczbą cech uwzględnionych w modelu, α_0, α_k są współczynnikami modelu, ε zaś reprezentuje czynnik losowy;

– cech wyekstrahowanych:

$$c = \alpha_0 + \sum_{k=1}^{K} \alpha_k PC_k + \varepsilon \tag{8.3}$$

gdzie: PC_k jest k-tą składową główną wyłonioną w wyniku przekształcenia wektora cech, będącego rozwinięciem całej macierzy cech;

- jednowymiarową regresję nieliniową

– cechy wyselekcjonowanej:

$$c = \alpha_{i,t}\beta_{i,t}^{X_{i,t}} + \varepsilon \tag{8.4}$$

gdzie: α_{it}, β_{it} są współczynnikami modelu;

– cechy wyekstrahowanej:

$$c = \alpha\beta^{PC1} + \varepsilon \tag{8.5}$$

gdzie: $PC1$ jest pierwszą składową główną wyłonioną w wyniku przekształcenia wektora cech, będącego rozwinięciem całej macierzy cech, α i β są współczynnikami modelu.

- regresję metodą cząstkowych najmniejszych kwadratów (patrz p. 6.3).

Pod względem oznaczeń ilościowych zanieczyszczeń szczególnie istotne są rezultaty zastosowania regresji liniowej wielokrotnej dla wektorów cech wyselekcjonowanych oraz regresji nieliniowej dla jednoelementowych wektorów cech. Są to najprostsze i najmniej wymagające z możliwych rozwiązań obliczeniowych tego rodzaju problemów. Regresję metodą cząstkowych najmniejszych kwadratów zastosowano jako metodę odniesienia ze względu na bardzo dobre rezultaty, które pozwala uzyskać z definicji.

8.8. Ocena metod pozyskiwania informacji o zanieczyszczeniach

Zbudowano modele klasyfikacji i regresji i przeprowadzano ich walidację z zastosowaniem metody dziesięciokrotnej kroswalidacji (patrz p. 7.1). Jest to ugruntowana metoda podziału zbioru danych na cześć uczącą i walidującą, która gwarantuje nie nazbyt łagodną ocenę rezultatów analizy danych. Jedynie dla komitetów drzew klasyfikacyjnych ocena modeli dotyczyła zbioru wektorów wyłączonych z procesu uczenia w trybie *baggingu*.

Za miarę oceny metod pozyskiwania informacji o charakterze jakościowym przyjęto udział błędnych klasyfikacji obliczany według równaniem (7.1). Natomiast do informacji o charakterze ilościowym zastosowano średni błąd względny określony równaniem (7.6).

Miary te w sensie ścisłym odnoszą się do rezultatów klasyfikacji i regresji. Należy jednak pamiętać, że wejściami modeli klasyfikacji i regresji były wektory cech uzyskane w toku starannie przeprowadzonych procesów selekcji lub ekstrakcji. Uzyskane wartości miar odzwierciedlały zatem nie tylko jakość samych modeli, lecz zarazem jakość źródeł informacji, wykorzystywanych przez te modele. Umożliwiały one w rzeczywistości ocenę całego procesu pozyskiwania informacji z czujnikowych danych pomiarowych, począwszy od wyłonienia cech, a skończywszy na sparametryzowaniu modeli klasyfikacji czy regresji. Za ich pomocą oceniano cały system analizy danych, zorientowany na pozyskanie konkretnej informacji o zanieczyszczeniach gazowych na podstawie danych z pomiarów czujnikowych.

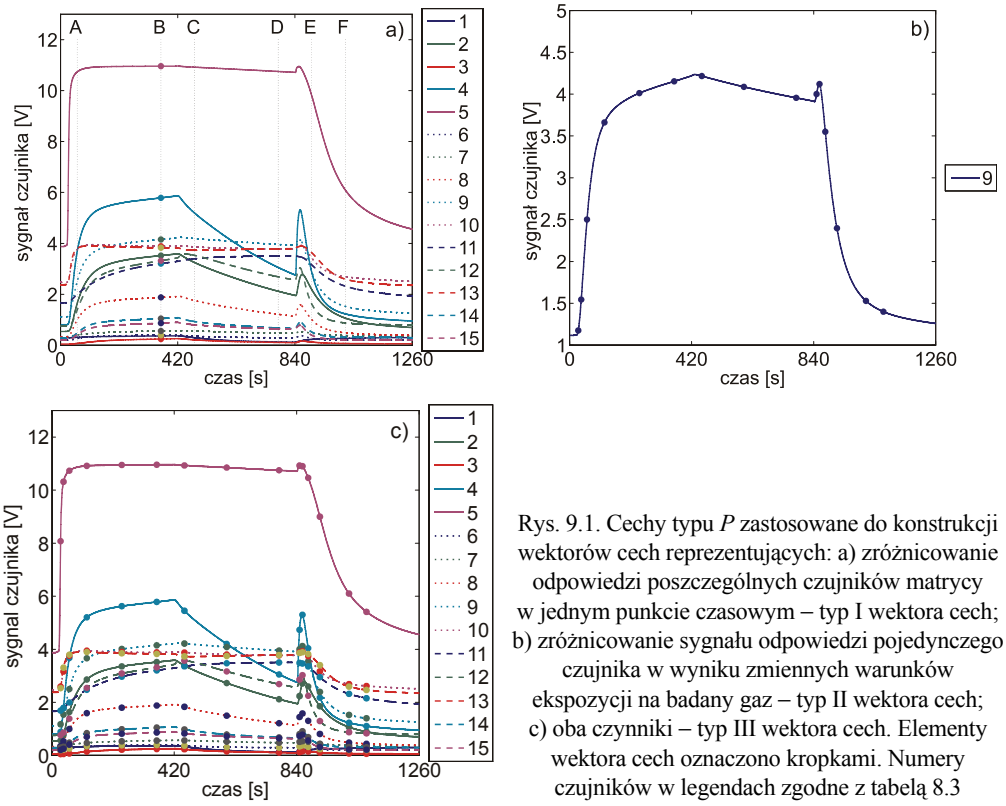
9. Eksploracja danych czujnikowych ze względu na informację o zanieczyszczeniach

Analizą eksploracyjną objęto dane wielowymiarowe II rzędu skonstruowane na podstawie czujnikowych danych pomiarowych. Dane dotyczyły ośmiu lotnych związków organicznych występujących w suchym powietrzu oraz powietrza wilgotnego o zróżnicowanej zawartości pary wodnej, niezawierającego LZO. Szczegóły dotyczące rodzajów i stężeń związków podano w tabeli 8.1. Jako metodę analizy eksploracyjnej zastosowano analizę składowych głównych ze względu na jej duże możliwości w połączeniu z prostotą (patrz rozdz. 6.1.1). Właściwości te decydują o powszechnym stosowaniu PCA w eksploracji danych. Analizie poddano dane wielowymiarowe zbudowane na podstawie trzech typów wektorów cech. Zasady ich konstrukcji przedstawiono na rys. 9.1.

Wektor cech typu I składał się z wartości sygnałów wszystkich czujników zarejestrowanych w ustalonym momencie ekspozycji (*punkcie czasowym* sygnału) na badany gaz (wektor piętnastoelementowy). Przeanalizowano sześć takich wektorów (A–F), każdy związany z innym punktem czasowym. Jak pokazano na rys. 9.1a, wybrane punkty czasowe reprezentują różne fragmenty sygnału czujnika. Fragmenty te są z kolei związane z różnymi warunkami ekspozycji czujników na badane zanieczyszczenia w trybie *stop flow* (patrz rys. 8.1 i 8.2). Typ I wektora cech umożliwiał przyjrzenie się zawartości informacyjnej przestrzeni cech, których składowe pochodziły z sygnałów różnych czujników. Dawał też możliwość stwierdzenia, czy reprezentacja informacji w przestrzeni cech tego rodzaju ewoluowała w wyniku zmiany warunków ekspozycji czujników.

Wektor cech typu II wyłoniono metodą *subsamplingu* sygnału jednego czujnika. Przyjęto, że wektor składa się z piętnastu wartości sygnału związanych z punktami czasowymi, których rozmieszczenie pokazano na rys. 9.1b. Wyboru punktów dokonano arbitralnie, kierując się zasadą, że strefy większej dynamiki sygnału są gęściej próbkowane. Rozważono piętnaście wektorów, każdy związany z innym czujnikiem. Typ II wektora cech umożliwiał poznanie zawartości informacji przestrzeni cech, której składowe pochodziły z sygnału jednego czujnika. Dawał też możliwość stwierdzenia, czy reprezentacja informacji w takiej przestrzeni zależy od czujnika.

Wektor cech typu III składał się z cech tworzących wektor typu II pobranych od wszystkich czujników, jak pokazano na rys. 9.1c. W wyniku uwzględnienia sygnałów piętnastu czujników wektor miał 15 × 15 = 225 elementów. Przestrzeń cech stanowiących elementy takiego wektora obejmowała możliwości informacyjne różnych czujników oraz różnych punktów czasowych ich sygnałów łącznie. Typ III wektora cech pozwalał porównać rolę takich czynników, jak czujnik oraz warunki ekspozycji w przenoszeniu informacji przez dane z pomiarów czujnikowych.

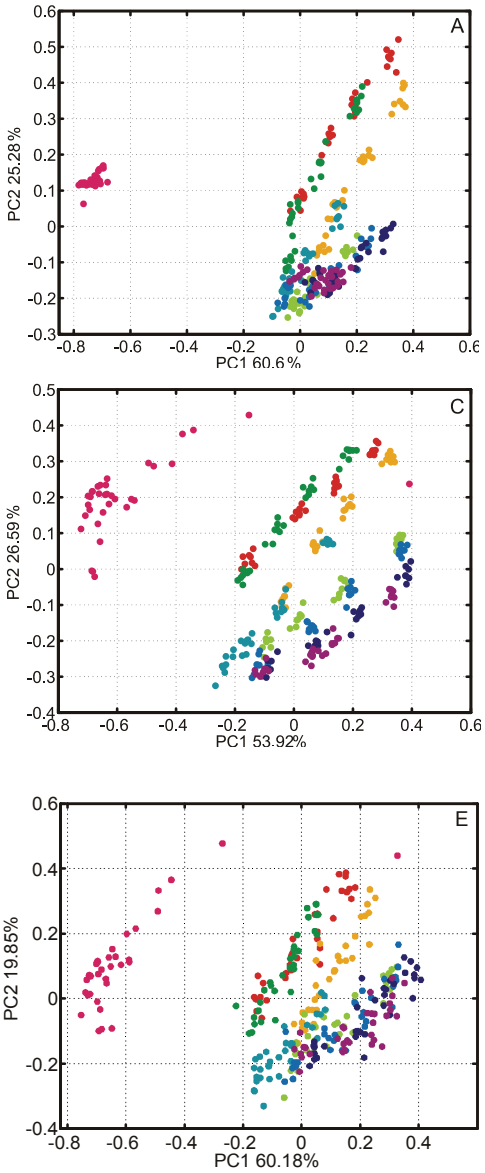


Rys. 9.1. Cechy typu *P* zastosowane do konstrukcji wektorów cech reprezentujących: a) zróżnicowanie odpowiedzi poszczególnych czujników matrycy w jednym punkcie czasowym – typ I wektora cech; b) zróżnicowanie sygnału odpowiedzi pojedynczego czujnika w wyniku zmiennych warunków ekspozycji na badany gaz – typ II wektora cech; c) oba czynniki – typ III wektora cech. Elementy wektora cech oznaczono kropkami. Numery czujników w legendach zgodne z tabelą 8.3

9.1. Eksploracja danych na podstawie wektora cech typu I

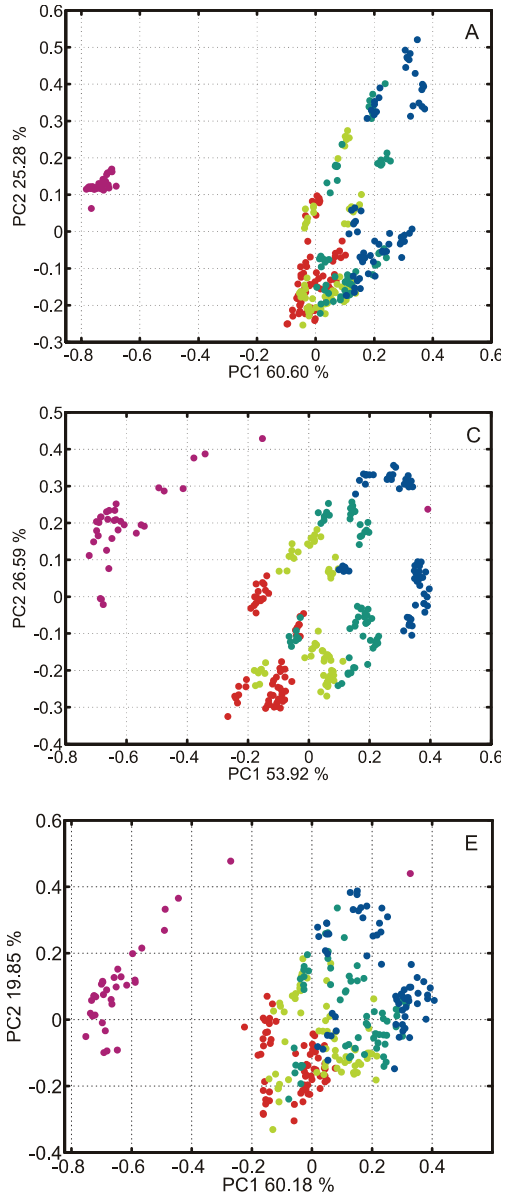
Na rysunkach 9.2 i 9.3 przedstawiono wyniki analizy składowych głównych danych wielowymiarowych dla wektora cech typu I. Dotyczą one wybranych momentów czasu ekspozycji czujników. Poszczególne wykresy, oznaczone literami A, C i E, odnoszą się do wektorów cech związanych z punktami czasowymi ekspozycji jak na

rys. 9.1a. Każdy wykres reprezentuje inny etap i inne warunki ekspozycji czujników w trybie *stop flow*.



Rys. 9.2. Rezultaty eksploracji danych we współrzędnych dwóch pierwszych składowych głównych. Dane zbudowano na podstawie wektora cech składającego się z wartości sygnałów różnych czujników zarejestrowanych w tym samym momencie ekspozycji na badane gazy. Przedstawiono rezultaty dla punktów czasowych A, C i E (patrz rys. 9.1a). Kolorami oznaczono tożsamość chemiczną badanych zanieczyszczeń: heksan – czerwony, heptan – pomarańczowy, oktan – pistacjowy, cykloheksan – zielony, benzen – morski, toluen – jasnoniebieski, ksylen – ciemnoniebieski, etylobenzen – fioletowy, para wodna – różowy

Wyniki analizy przedstawiono we współrzędnych dwóch pierwszych składowych głównych. Na rysunku 9.2 przyjęto oznaczenie punktów danych według rodzaju zanieczyszczeń, natomiast na rys. 9.3 oznaczenie tych samych punktów odnosi się do zakresów stężeń zanieczyszczeń.



Rys. 9.3. Rezultaty eksploracji danych jak na rys. 9.2, lecz kolorami oznaczono stężenia badanych zanieczyszczeń: 14–25 ppm – czerwony, 41–76 ppm – pistacjowy, 83–151 ppm – zielony, 165–302 ppm – niebieski, 493 ppm–2.65% – fioletowy

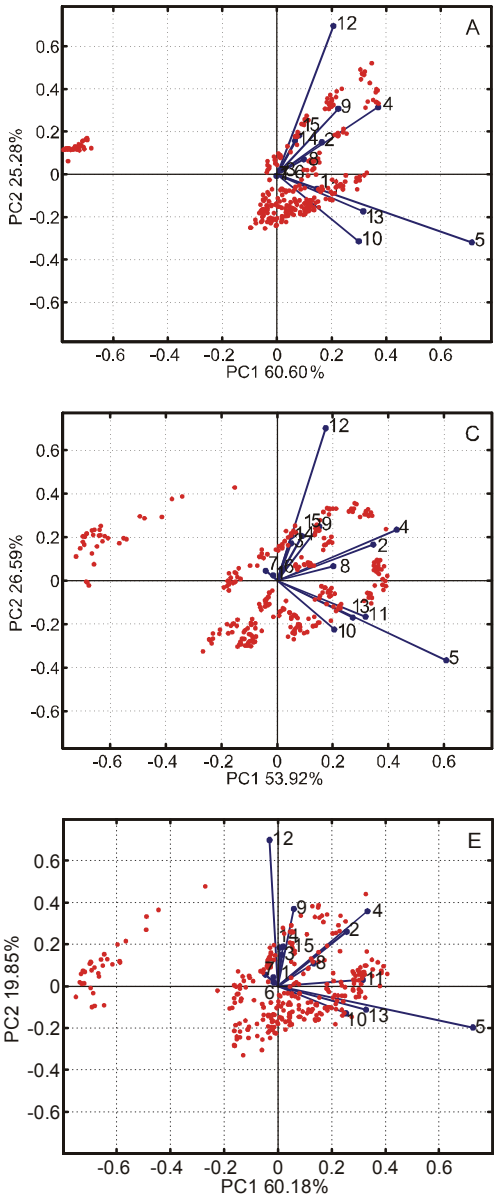
Jak wynika z rysunków 9.2 oraz 9.3, w czujnikowych danych pomiarowych zawarta jest informacja dotycząca jakościowych oraz ilościowych właściwości badanych zanieczyszczeń. Łatwo zauważyć, że zasadnicze kierunki zmienności danych, wyznaczone przez pierwszą i drugą składową główną, są związane z tymi dwoma rodzajami informacji. Widoczny na rysunkach układ punktów danych dotyczących LZO w powietrzu dowodzi, że pomiary czujnikowe umożliwiały identyfikację tych zanieczysz-

czeń w szerokim zakresie ich rodzajów i stężeń. Wyniki pomiarów czujnikowych wyraźnie różniły się w zależności od tego, czy w powietrzu występowały lotne związki organiczne. Na rysunku 9.2 skupiska punktów reprezentujących powietrze wilgotne i powietrze zawierające lotne związki organiczne są wyraźnie rozdzielone. Z układu punktów wskazujących na poszczególne substancje na rys. 9.2 wynika, że za ich rozsunięcie odpowiada przede wszystkim druga składowa główna. Zbiory punktów nie są jednak w pełni rozdzielone. Można zauważyć, że grupują się one w trzy szersze kategorie, do których należą: i) heksan i cykloheksan; ii) oktan, toluen, ksylen i etylobenzen; iii) heptan i benzen. Źródło podobieństwa składników grup nie jest oczywiste. Z ich istnienia wynika natomiast możliwość wyodrębnienia w toku analizy składowych głównych zgrupowań danych, które nie nawiązują do intuicyjnych kryteriów klasyfikacji badanych zanieczyszczeń.

Skupienia punktów odpowiadających poszczególnym związkom na rys. 9.2 nie są zwarte. Każde składa się z czterech mniejszych grup, które, jak wynika z porównania rys. 9.2 i 9.3, są związane z zakresem stężeń badanych zanieczyszczeń. Można też zauważyć, że PCA ujawniło kategoryzację badanych gazów według stężenia, w dużej mierze niezależnie od rodzaju gazu (obraz ten nieco zaburza benzen i heptan). Układ zgrupowań punktów reprezentujących grupy zanieczyszczeń o różnych stężeniach na rys. 9.3 pokazuje, że za ich rozsunięcie odpowiadała przede wszystkim pierwsza składowa główna. Ze względu na dominujące znaczenie pierwszej składowej głównej w analizie zmienności rozważanych tu danych z pomiarów czujnikowych można uznać, że informacja o charakterze ilościowym dominowała nad informacją o charakterze jakościowym.

Z porównania rezultatów PCA dla danych związanych z różnymi wektorami cech (poszczególne wykresy na rys. 9.2 i 9.3) wynikało, że wyrazistość oraz stopień rozsunięcia zbiorów punktów odpowiadających różnym lotnym związkom organicznym, jak również poszczególnym zakresom stężeń zależały od wyboru przestrzeni cech. W analizowanym przypadku czynnikiem odpowiadającym za zróżnicowanie przestrzeni cech były warunki ekspozycji czujników gazu. Największe możliwości przenoszenia zróżnicowanej informacji o badanych gazach oferowały warunki ekspozycji związane z momentami B, C i D. Odpowiadały one wolnozmiennym fragmentom sygnałów czujnikowych uzyskanych zarówno w warunkach przepływu (B i D), jak i braku przepływu (C) gazu nad czujnikiem podczas pracy w trybie *stop flow*.

Okoliczności te sprzyjały ujawnieniu informacji zarówno jakościowej, jak i ilościowej o badanych gazach. Nieco mniejsze możliwości pod tym względem oferowały warunki ekspozycji A, związane z fazą narostu sygnału czujnikowego tuż po rozpoczęciu doprowadzania gazu do komórek czujnikowych. Z analizy składowych głównych wynika, że bardzo interesujące pod względem przenoszonej informacji były warunki ekspozycji E, które odpowiadały fazie regeneracji czujników. Okazało się, że zanik informacji jakościowej o badanych gazach w danych pomiarowych następował szybciej niż zanik informacji ilościowej (por. rys. 9.2 E i 9.3 E).



Rys. 9.4. Rezultaty eksploracji danych przedstawionych na rys. 9.2 i 9.3 z pokazaniem roli odpowiedzi poszczególnych czujników w wyjaśnianiu zmian danych wielowymiarowych przez dwie pierwsze składowe główne. Numery podane w polu wykresu odnoszą się do poszczególnych czujników zgodnie z tabelą 8.3

Na rysunku 9.4 przedstawiono dane pomiarowe oraz ładunki poszczególnych cech (tu jednoznacznie związane z czujnikami) we współrzędnych dwóch pierwszych składowych głównych. Poszerzenie analizy PCA o rozważenie ładunków poszczególnych cech na składowe główne pozwala uzyskać informację, w jakim stopniu poszczególne cechy decydują o ujawnionej strukturze danych wielowymiarowych.

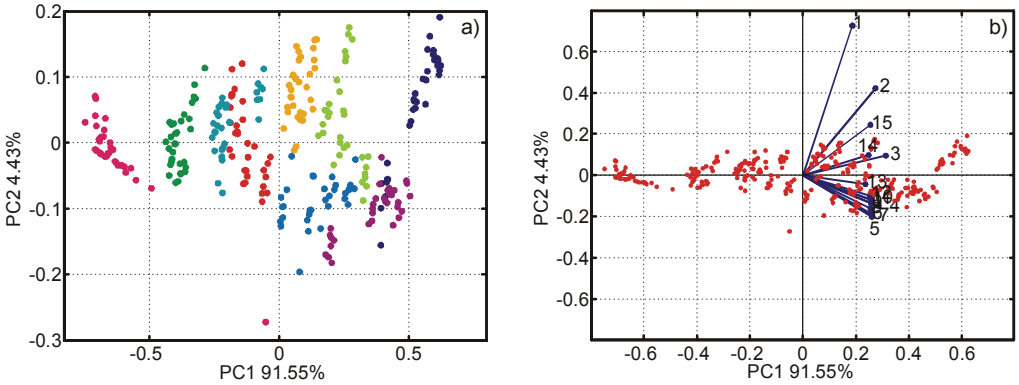
Jak pokazano na rysunku 9.4, za informację ilościową o badanych gazach odpowiadały przede wszystkim cechy pochodzące z sygnałów czujników 5, 10, 11 i 13 oraz 2, 4 i 8 (por. tabela 8.3). Cechy te miały znaczny udział w pierwszej składowej głównej. Jak wcześniej zauważono (rys. 9.3), składowa ta reprezentowała kierunek zmienności danych spowodowanej zróżnicowaniem stężeń badanych gazów. Najistotniejszy udział w składowej drugiej, odpowiedzialnej głównie za informację jakościową (rys. 9.2), miały zaś cechy pochodzące z sygnału czujnika 12. Znaczące były również ładunki czujników 5, 9, 4 i 10. Cechy, które niewiele wnosiły w wyjaśnienie zmienności danych pochodziły z sygnałów czujników 1, 3, 6, 7. Do grupy tej można zaliczyć także nieco lepsze czujniki 14, 15 i 8.

9.2. Eksploracja danych na podstawie wektora cech typu II

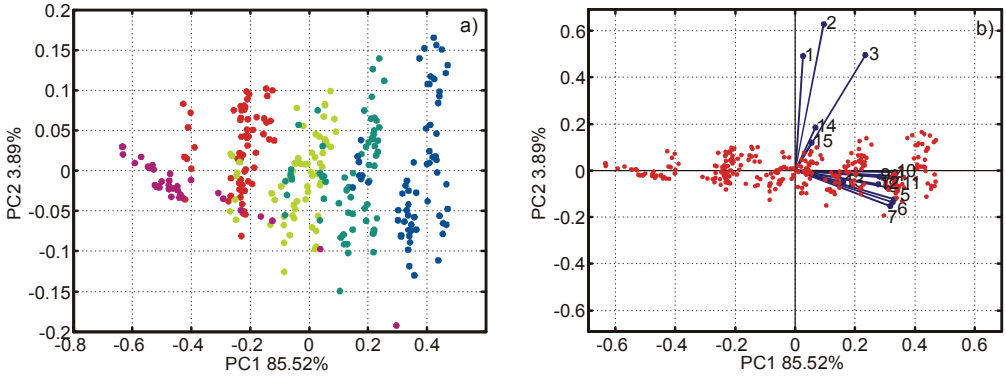
Analizie składowych głównych poddano również dane wielowymiarowe zbudowane na podstawie wektorów cech, które wyłoniono metodą *subsamplingu*. Konkretny wektor cech był związany z sygnałem jednego czujnika i składał się z jego wartości odpowiadających różnym punktom czasu ekspozycji w trybie *stop flow* (patrz rys. 9.1b). Analizę wykonano dla wszystkich czujników (tabela 8.3), a wybrane rezultaty przedstawiono na rys. 9.5 i 9.6.

Na podstawie uzyskanych wyników stwierdzono, że sygnały poszczególnych czujników wykazywały zróżnicowane pod względem rodzaju przenoszonej informacji o badanych gazach oraz jej zawartości. Na podstawie sygnałów większości czujników można było na przykład wyraźnie wyróżnić dwie kategorie wektorów danych, z których jedna odpowiadała powietrzu wilgotnemu, a druga powietrzu zanieczyszczonemu lotnymi związkami organicznymi. Stwierdzono, że sygnały niektórych czujników dobrze różnicowały wszystkie badane zanieczyszczania pod względem chemicznym. Należał do nich TGS 2201 (olej napędowy). Na wykresie rezultatów PCA dla tego czujnika (rys. 9.5a) widać dobrze wykształcone i w znacznej mierze rozdzielone grupy punktów reprezentujących różne LZO. Zróżnicowanie danych obserwowano wzdłuż pierwszej składowej głównej, odpowiedzialnej za ponad 90% ich zmienności. Zasadnicza część całkowitej zmienności sygnału czujnika TGS 2201 wynikała z rodzaju badanych substancji. Dla kilku sensorów wektor cech typu II okazał się bardzo dobrym nośnikiem informacji ilościowej. Przykładem jest czujnik TGS 825. Na rysunku 9.6a przedstawiono rezultaty analizy składowych głównych dla wektora cech zbudowanego na podstawie sygnału tego czujnika. Na rysunku widoczne są dobrze wykształcone grupy punktów odpowiadające poszczególnym zakresom stężeń badanych związków. Grupy te są rozmieszczone wzdłuż pierwszej składowej głównej, która jest odpowiedzialna za niemal 86% zmienności danych. Oznacza to, że zasadnicza część zmienności danych pochodzi-

ła od stężenia zanieczyszczeń. Można uznać, że sygnał czujnika TGS 825 przenosił przede wszystkim informację ilościową.



Rys. 9.5. Rezultaty eksploracji danych przedstawione we współrzędnych dwóch pierwszych składowych głównych. Analiza dotyczy wektora cech składającego się z piętnastu wartości sygnału czujnika TGS 2201 (olej napędowy) związanych z różnymi etapami ekspozycji na badane gazy (patrz rys. 9.1b): a) kolorami oznaczono rodzaj badanych zanieczyszczeń: heksan – czerwony, heptan – pomarańczowy, oktan – pistacjowy, cykloheksan – zielony, benzen – morski, toluen – jasnoniebieski, ksylen – ciemnoniebieski, etylobenzen – fioletowy, para wodna – różowy, b) numery w polu wykresu odnoszą się do punktów czasowych ekspozycji czujnika



Rys. 9.6. Rezultaty eksploracji danych przedstawione we współrzędnych dwóch pierwszych składowych głównych. Analiza dotyczy wektora cech składającego się z piętnastu wartości sygnału czujnika TGS 825 związanych z różnymi etapami ekspozycji na badane gazy (patrz rys. 9.1b): a) kolorami oznaczono stężenia badanych zanieczyszczeń: 14–25 ppm – czerwony, 41–76 ppm – pistacjowy, 83–151 ppm – zielony, 165–302 ppm – niebieski, 493 ppm – 2.65 % – fioletowy; b) numery w polu wykresu odnoszą się do punktów czasowych ekspozycji czujnika

Wśród rozważanych czujników były i takie, że dane zbudowane na podstawie ich sygnału nie wykazywały czytelnej struktury. Analiza składowych głównych nie ujaw-

niała uprzywilejowanych kierunków zmienności. Nie oznacza to jednak nieprzydatności danych pomiarowych pochodzących z takich czujników do oznaczania gazów. O ich faktycznej użyteczności można wnioskować dopiero na podstawie wyników analizy danych z zastosowaniem tzw. „metod z nadzorem".

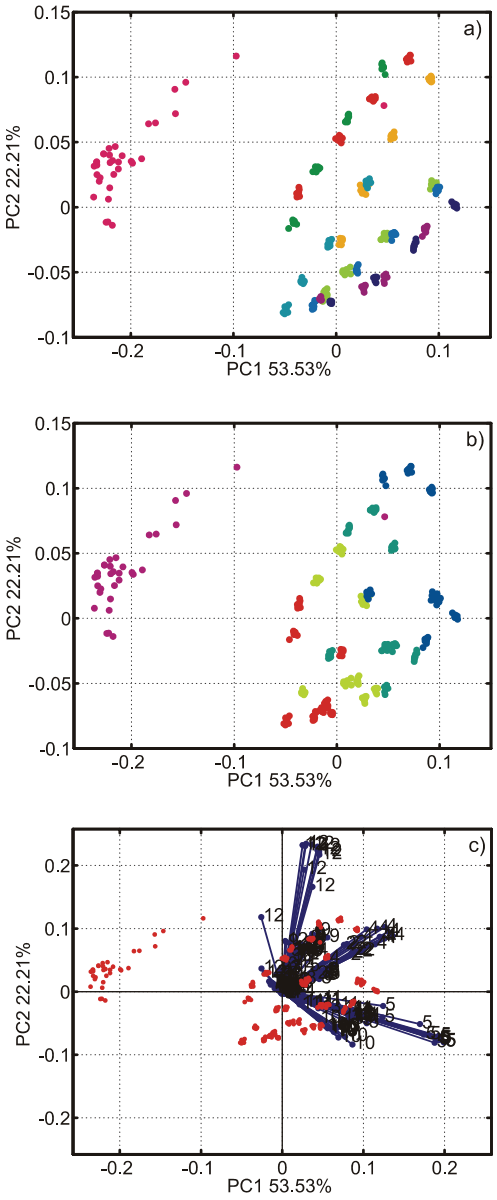
Z analizy składowych głównych wynika, że sygnały poszczególnych czujników dobrze przenosiły informację ilościową albo jakościową, albo źle przenosiły każdy z tych rodzajów informacji o zanieczyszczeniach. Nie stwierdzono, by wektor cech zbudowany na podstawie sygnału któregoś z sensorów dobrze przenosił oba rodzaje informacji. Zjawisko takie obserwowano natomiast dla wektora cech typu I. Nie zmienia to faktu, że zaobserwowany rodzaj selektywności czujników, polegający na przydatności do przenoszenia informacji jakościowej lub ilościowej, jest bardzo interesujący pod względem możliwości zastosowania w czujnikowych pomiarach zanieczyszczeń.

Ważne wnioski wynikały też z analizy ładunków cech. Na podstawie jej rezultatów zidentyfikowano warunki ekspozycji czujników odpowiadające za powstanie obserwowanej struktury danych. Przykłady przedstawione na rys. 9.5b i 9.6b dobrze ilustrują ogólne prawidłowości występujące w szerszej grupie czujników. Najbardziej nośnymi informacyjnie okazały się fragmenty sygnałów, na które przypadały punkty czasowe od 4 do 13 (por. rys. 9.1b). Były one związane ze zróżnicowanymi warunkami ekspozycji sensorów, występującymi w trakcie pomiaru prowadzonego w trybie *stop flow*. Punkty czasowe o małej wadze przypadały na początek pomiaru, kiedy badany gaz był doprowadzany do komory czujnikowej oraz na stadium zaawansowanej regeneracji czujników, gdy większość badanych substancji została już stamtąd usunięta.

9.3. Eksploracja danych na podstawie wektora cech typu III

Na rysunku 9.7 przedstawiono wyniki analizy składowych głównych dla danych wielowymiarowych skonstruowanych na podstawie wektora cech typu III (patrz rys. 9.1c). Widoczne jest podobieństwo obrazu struktury danych wielowymiarowych uzyskanych na podstawie wektora cech typu I (rys. 9.2 i 9.3) oraz typu III (rys. 9.7a, b). Dla obu kierunek zmienności wyznaczony pierwszą składową główną był związany przede wszystkim z informacją ilościową, zaś kierunek wyznaczony drugą składową dotyczył raczej właściwości jakościowych badanych gazów. Układ zgrupowań punktów na rys. 9.7a, b i rys. 9.2 lub 9.3 jest również bardzo podobny. Jednak stosując wektor cech typu III, uzyskano większą zwartość i rozdzielenie grup punktów. Wyraźnie widoczne są zestawy punktów dotyczące gazu określonego rodzaju i o określonym stężeniu. Wynika stąd, że dane oparte na wektorze cech typu III przenosiły informację dokładniej niż dane oparte na wektorze cech typu I lub II. Interesujące

wnioski wynikały z analizy wykresu ładunków cech będących elementami wektora typu III (patrz rys. 9.7c).



Rys. 9.7. Rezultaty eksploracji danych wielowymiarowych skonstruowanych na podstawie wektora cech typu III (patrz rys. 9.1c): a) kolorami oznaczono tożsamość chemiczną badanych zanieczyszczeń: heksan – czerwony, heptan – pomarańczowy, oktan – pistacjowy, cykloheksan – zielony, benzen – morski, toluen – jasnoniebieski, ksylen – ciemnoniebieski, etylobenzen – fioletowy, para wodna – różowy, b) kolorami oznaczono stężenia badanych zanieczyszczeń: 14–25 ppm – czerwony, 41–76 ppm – pistacjowy, 83–151 ppm – zielony, 165–302 ppm – niebieski, 493 ppm–2.65% – fioletowy, c) numery w polu wykresu odnoszą się do czujników. Wszystkie punkty czasowe z sygnału jednego czujnika oznaczono tym samym numerem

Cechy pochodzące z sygnału indywidualnego czujnika miały bardzo zbliżone udział w dwóch pierwszych składowych głównych. Układ ładunków cech związanych z różnymi czujnikami naśladował natomiast układ pokazany na rys. 9.2, dotyczący

danych związanych z wektorem cech typu I. Obserwacje te dowodzą, że informacja dostępna dzięki zróżnicowaniu czujników dominuje nad informacją przenoszoną dzięki zmienności sygnału czujnikowego, jeżeli wektor cech zawiera reprezentację sygnałów wielu czujników.

9.4. Wnioski z eksploracyjnej analizy danych czujnikowych ze względu na określanie zanieczyszczeń

Z przeprowadzonej eksploracji danych wynika, że analizowane czujnikowe dane pomiarowe przenosiły informację jakościową i ilościową o badanych zanieczyszczeniach. Właściwości zanieczyszczeń pod względem jakościowym i ilościowym były istotnymi, jeżeli nie głównymi źródłami zmienności danych. Ich wpływ na sygnał czujników był widoczny bardzo wyraźnie w strukturze danych wielowymiarowych ujawniającej się w toku analizy eksploracyjnej, np. metodą analizy składowych głównych.

Wyniki analizy pokazały, że wyrazistość konkretnego rodzaju informacji w danych zależała od przestrzeni cech wybranej jako przestrzeń analizy danych. Możliwości poszczególnych przestrzeni cech typu *P* dostępnych w skali macierzy cech były w znacznym stopniu kształtowane przez takie czynniki, jak rodzaj czujnika oraz tryb jego pracy.

Zasadniczą rolę w pozyskaniu określonej informacji o badanych gazach odgrywa zatem praca z odpowiednią przestrzenią cech. Poszukując różnych rodzajów informacji, warto pracować z różnymi przestrzeniami cech, wybierając takie, które są najlepsze w poszczególnych przypadkach. Aby zwiększyć niezawodność pozyskania konkretnej informacji, można skorzystać z więcej niż jednej przestrzeni cech, w których ta informacja jest dobrze reprezentowana.

10. Analiza danych czujnikowych pod względem informacji jakościowej o zanieczyszczeniach

Celem analizy danych czujnikowych pod względem informacji jakościowych o zanieczyszczeniach była ocena różnych konfiguracji systemu analizy danych pozwalającego ustalić:

- tożsamość chemiczną substancji zanieczyszczających,
- skład jakościowy mieszanin substancji zanieczyszczających,
- przynależność do kategorii substancji zanieczyszczających,
- przynależność do kategorii mieszanin substancji zanieczyszczających.

Ocena przynależności substancji do określonej kategorii jest przykładem innego podejścia do problemu zanieczyszczenia bez poznawania jego tożsamości chemicznej.

Do wykonania analizy danych ze względu na pozyskanie informacji jakościowej o zanieczyszczeniach zastosowano podejście oparte na koncepcji rozpoznawania wzorców. Podstawowe znaczenie ma w tym przypadku wybór odpowiedniej reprezentacji badanego gazu oraz właściwego rodzaju klasyfikatora. Rozważono następujące konfiguracje tych dwóch elementów systemu analizy danych:

- wektor cech wyselekcjonowanych i klasyfikator liniowy,
- wektor cech wyselekcjonowanych i klasyfikator nieliniowy,
- wektor cech wyekstrahowanych i klasyfikator liniowy,
- wektor cech wyekstrahowanych i klasyfikator nieliniowy,
- komitet drzew klasyfikacyjnych.

10.1. Określanie rodzaju substancji zanieczyszczającej

Rozpoznawanie tożsamości chemicznej substancji zanieczyszczających rozważono na przykładzie następujących lotnych związków organicznych: heksan (He), heptan (Hp), oktan (Ok), cykloheksan (Cy), benzen (Be), toluen (To), ksylen (Ks) oraz etylobenzen (Eb). Rodzaj substancji rozpoznawano niezależnie od stężenia w zakresie stężeń podanym w tabeli 8.1.

Klasyfikator liniowy i nieliniowy oraz wektor cech wyselekcjonowanych

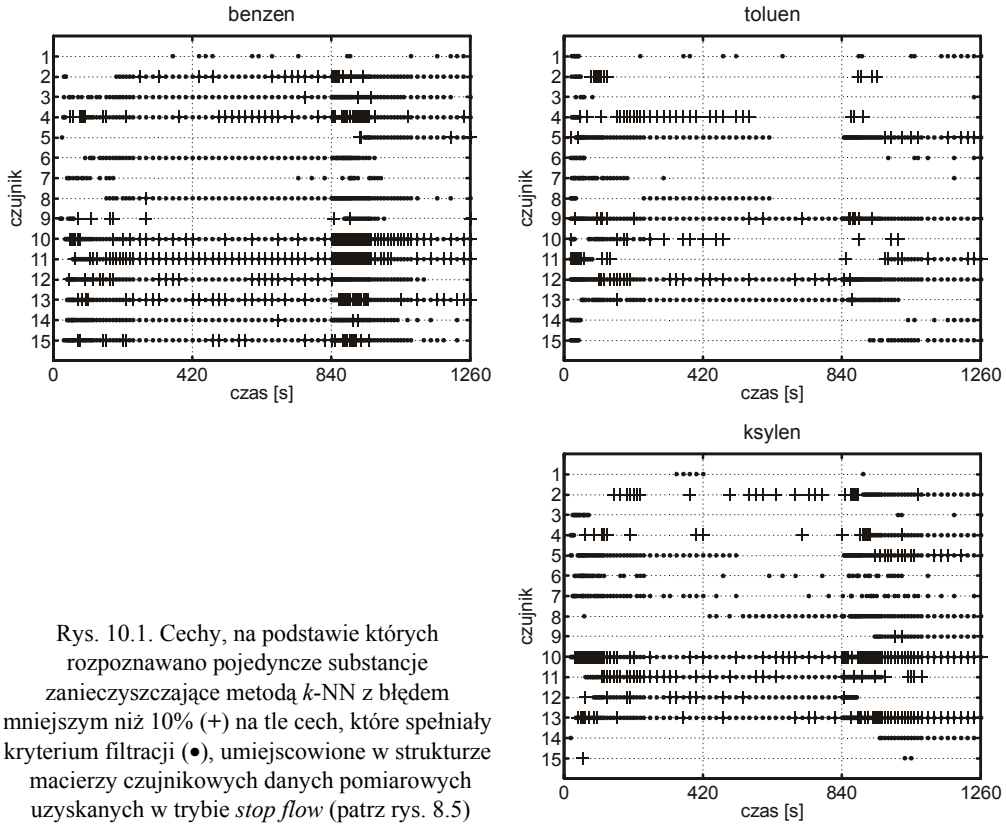
Analizowano możliwość uzyskania informacji o tożsamości chemicznej pojedynczych substancji zanieczyszczających na podstawie ich reprezentacji w postaci wektora cech wyselekcjonowanych. Rozważono wektory składające się z 1, 2, 3, 5 lub 7 cech. Jako liniową metodę klasyfikacji zastosowano liniową analizę dyskryminacyjną (LDA), jako nieliniową metodę klasyfikacji wybrano metodę *k* najbliższych sąsiadów (*k*-NN). Metody te służyły do oceny wektorów cech w ramach selekcji w systemie opakowanym. Przestrzeń cech przeszukiwano metodą symulowanego wyżarzania. Dla porównania przeprowadzono jednowymiarową filtrację metodą jednowymiarowej analizy wariancji (ANOVA) z przeglądem zupełnym przestrzeni cech.

Tabela 10.1. Liczba jednoelementowych wektorów cech umożliwiających rozpoznawanie tożsamości chemicznej pojedynczych substancji zanieczyszczających[a]

Tożsamość chemiczna substancji zanieczyszczającej	LDA	*k*-NN	ANOVA
Heksan	25	342	997
Heptan	6	280	675
Oktan	2	152	586
Cykloheksan	9	362	1146
Benzen	0	326	1218
Toluen	0	126	668
Ksylen	18	237	812
Etylobenzen	3	125	1029

[a]Rozpoznawanie metodami LDA i *k*-NN, tak aby udział błędnych klasyfikacji był mniejszy niż 10%. Liczba cech istotnie różnicujących substancje na podstawie ANOVA ($\alpha = 0{,}01$). Ogólna liczba cech – 1830.

W tabeli 10.1 podano liczbę jednoelementowych wektorów cech typu *P*, które na podstawie jednowymiarowej analizy wariancji ($\alpha = 0{,}01$) wyraźnie różnicowały rozważane substancje zanieczyszczające ze względu na ich tożsamość chemiczną. Zamieszczono tam również liczby jednoelementowych wektorów cech, które pozwalały rozpoznawać pojedyncze substancje zanieczyszczające metodami LDA i *k*-NN z rezultatem lepszym niż 10% błędnych klasyfikacji. Położenie elementów tych wektoró,w cech w macierzy cech pokazano na rys. 10.1. Przedstawiono wspólnie wektory spełniające kryterium analizy wariancji oraz dobrze współpracujące z klasyfikatorem *k*-NN. Dla przykładu zilustrowano rezultaty uzyskane dla takich zanieczyszczeń, jak benzen, toluen i ksylen.



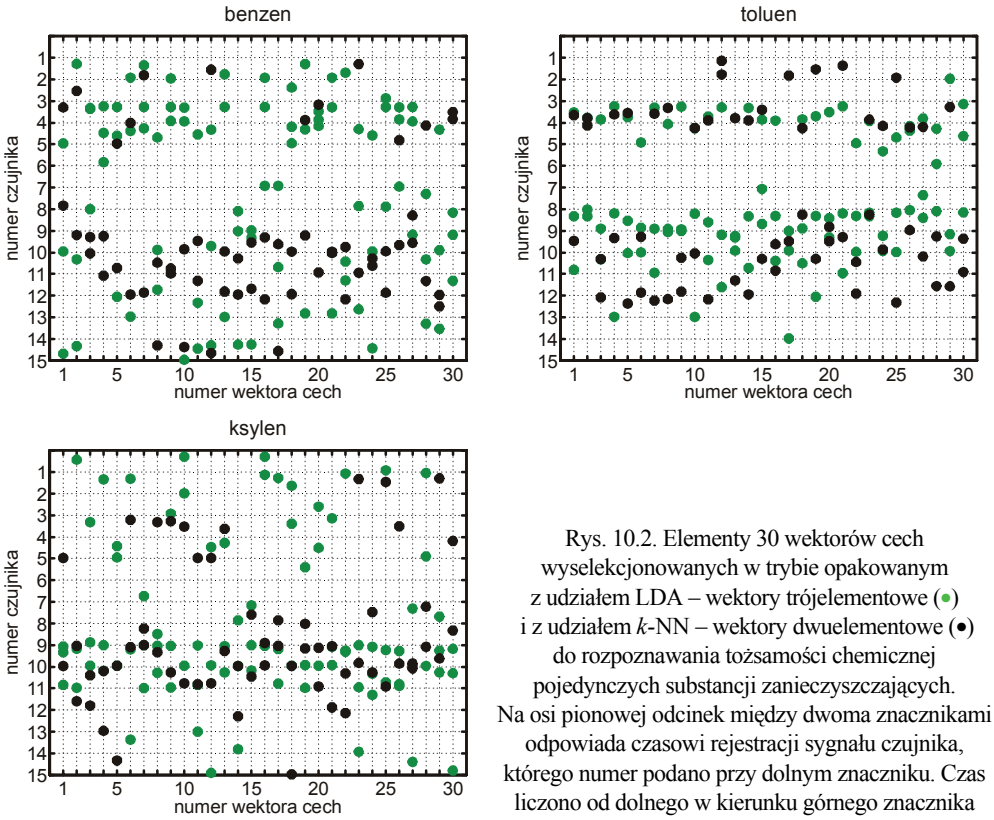
Rys. 10.1. Cechy, na podstawie których rozpoznawano pojedyncze substancje zanieczyszczające metodą *k*-NN z błędem mniejszym niż 10% (+) na tle cech, które spełniały kryterium filtracji (•), umiejscowione w strukturze macierzy czujnikowych danych pomiarowych uzyskanych w trybie *stop flow* (patrz rys. 8.5)

Tabela 10.2. Udział błędnych klasyfikacji (MCR) w rozpoznawaniu pojedynczych zanieczyszczeń[a]

Substancja zanieczyszczająca	LDA					*k*-NN				
	Liczba elementów wektora cech									
	1	2	3	5	7	1	2	3	5	7
Heksan	3–10	0–8	0–6	0–1	0–0	0–4	0–0			
Heptan	6–18	2–10	0–5			1–4				
Oktan	9–24	1–6	0–4			3–6				
Cykloheksan	8–13	0–0	0–0	0–0		0–0				
Benzen	13–16	0–3				0–3				
Toluen	24–28	2–8	0–3			1–6				
Ksylen	3–13	0–3	0–0	0–1		0–3				
Etylobenzen	3–18	3–6	0–3	0–0		1–6				

[a]Podano wartości min(MCR)–max(MCR) [%], dla 30 najlepszych wektorów cech o następującej liczbie elementów: 1, 2, 3, 5 i 7 uzyskanych w wyniku selekcji.

W tabeli 10.2 przedstawiono błąd oznaczania tożsamości chemicznej substancji zanieczyszczających metodą liniowej analizy dyskryminacyjnej i *k*-najbliższych sąsiadów, w zależności od liczby elementów w wektorze cech.



Rys. 10.2. Elementy 30 wektorów cech wyselekcjonowanych w trybie opakowanym z udziałem LDA – wektory trójelementowe (•) i z udziałem *k*-NN – wektory dwuelementowe (•) do rozpoznawania tożsamości chemicznej pojedynczych substancji zanieczyszczających. Na osi pionowej odcinek między dwoma znacznikami odpowiada czasowi rejestracji sygnału czujnika, którego numer podano przy dolnym znaczniku. Czas liczono od dolnego w kierunku górnego znacznika

Na rysunku 10.2 pokazano położenie w strukturze macierzy danych elementów trzydziestu wektorów cech wyselekcjonowanych ze względu na rozpoznawanie wybranych pojedynczych substancji zanieczyszczających z jak najmniejszym błędem. Liczbę 30 przyjęto arbitralnie. Dla metody liniowej przedstawiono wektory o trzech elementach, dla metody nieliniowej – wektory o dwóch elementach. Wektory o tej liczbie cech umożliwiały bezbłędne rozpoznawanie wszystkich substancji zanieczyszczających (tabela 10.2).

Algorytm selekcji wybierał najczęściej cechy pochodzące z sygnałów czujników wymienionych w tabeli 10.3 jako elementy najlepszych wektorów cech trójelementowych według oceny klasyfikatora LDA i dwuelementowych zgodnie z oceną klasyfikatora *k*-NN.

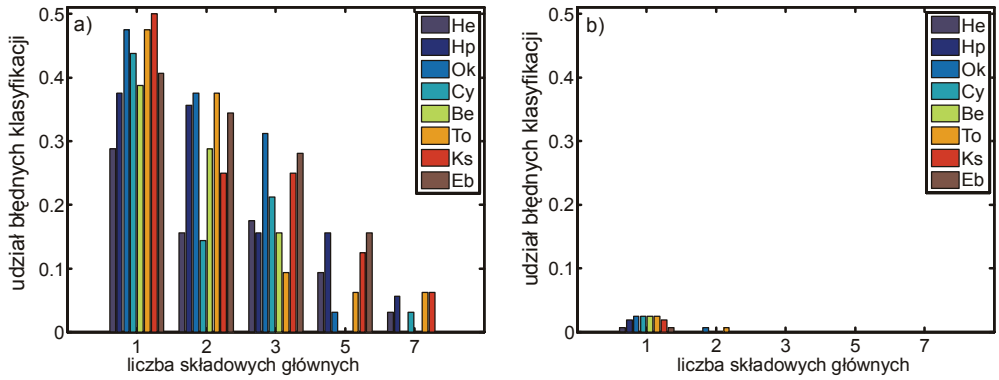
Tabela 10.3. Czujniki, które najczęściej stanowiły źródło elementów trzydziestu wektorów cech wyselekcjonowanych do rozpoznawania tożsamości chemicznej pojedynczych substancji zanieczyszczających[a]

Tożsamość chemiczna substancji zanieczyszczającej	LDA	*k*-NN	
	Liczba elementów wektora cech		
	3	1	2
Heksan	14	10, 12	5, 11
Heptan	12	10, 11, 12	12
Oktan	11, 9	10, 11, 12	11
Cykloheksan	10	10, 11, 12	10
Benzen	4, 5	10, 11	10
Toluen	4, 9	4	4, 10
Ksylen	10, 11	10, 13	10, 11
Etylobenzen	11	11	11

[a]Numery czujników według tabeli 8.3.

Klasyfikator liniowy i nieliniowy oraz wektor cech wyekstrahowanych

Analizowano możliwość uzyskania informacji o tożsamości chemicznej pojedynczych substancji zanieczyszczających na podstawie wektora cech uzyskanych w wyniku ekstrakcji. Jako metodę ekstrakcji cech zastosowano analizę składowych głównych. Cechami niejawnymi były składowe główne utworzone w wyniku zmapowania wektora cech powstałego po rozwinięciu całej macierzy cech. Rozważono wektory cech zbudowane z 1, 2, 3, 5 lub 7 pierwszych składowych głównych. Podobnie

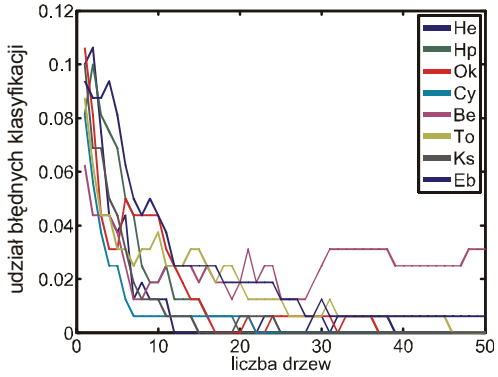


Rys. 10.3. Wyniki rozpoznawania tożsamości chemicznej substancji zanieczyszczających na podstawie wektorów cech, których elementy stanowiły odpowiednio 1, 2, 3, 5 lub 7 składowych głównych, wyłonionych za pomocą PCA. Klasyfikacji dokonano z zastosowaniem: a) LDA, b) *k*-NN

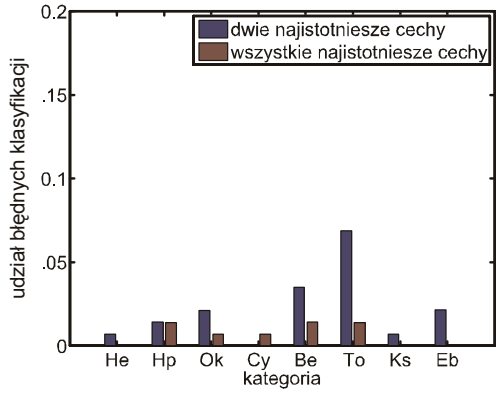
jak dla wektorów cech wyselekcjonowanych zastosowano liniową (LDA) i nieliniową (*k*-NN) metodę klasyfikacji. Rezultaty takiego rozpoznawania pojedynczych substancji przedstawiono na rys. 10.3.

Komitet drzew klasyfikacyjnych

Rozważono również zastosowanie rozwiązania wbudowanego w postaci komitetu drzew klasyfikacyjnych typu CART w celu pozyskania informacji dotyczących tożsamości chemicznej substancji zanieczyszczających. Wejściem klasyfikatora była cała macierz cech rozwinięta do postaci wektora. Na rysunku 10.4 pokazano zależność poprawności rozpoznawania substancji od liczby drzew w komitecie.



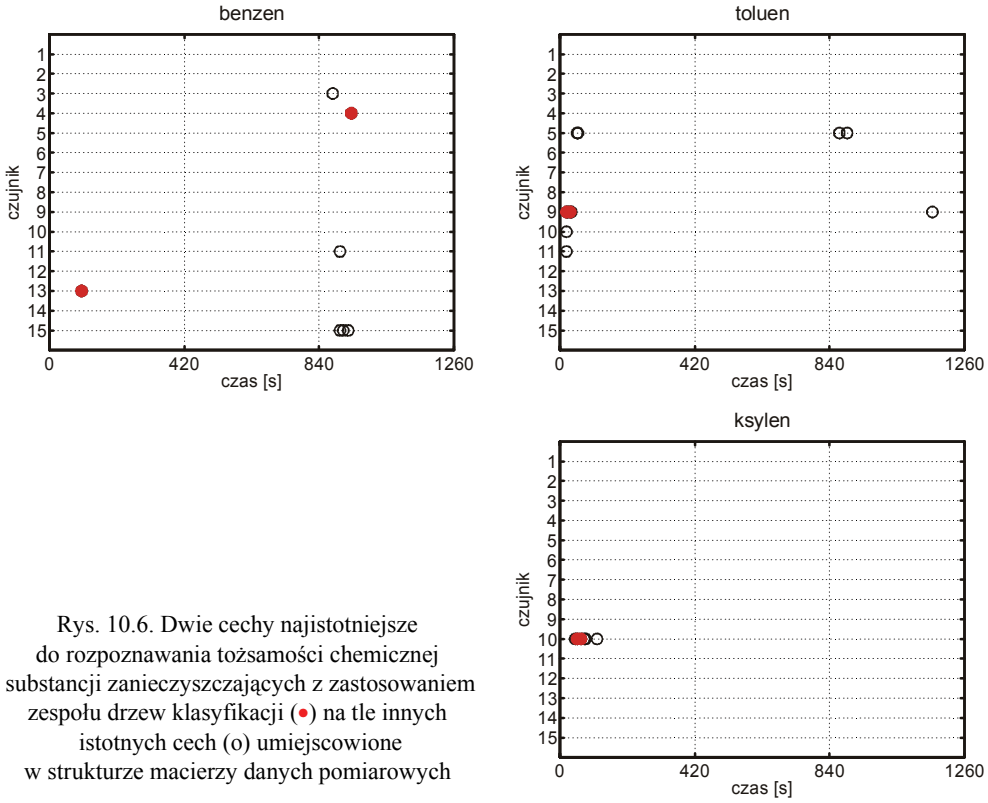
Rys. 10.4. Zależność efektywności rozpoznawania tożsamości chemicznej substancji zanieczyszczających od liczby drzew w komitecie klasyfikatorów



Rys. 10.5. Efektywność rozpoznawania pojedynczych substancji zanieczyszczających za pomocą komitetu 15 drzew klasyfikacyjnych. Zastosowano cechy istotne według klasyfikatora pracującego z wektorem cech zawierającym wszystkie elementy macierzy cech

Niezależnie od rodzaju substancji zanieczyszczających zastosowanie komitetów liczących więcej niż piętnaście drzew klasyfikacyjnych nie wiązało się z poprawą rezultatów rozpoznawania. Na rysunku 10.5 pokazano wyniki uzyskane na podstawie zredukowanego wektora cech po zastosowaniu komitetu 15 drzew klasyfikacyjnych.

Elementami wektora były cechy o wskaźniku istotności większym od 0,2. Wskaźnik ten porządkuje cechy według ich znaczenia dla poprawnej klasyfikacji wektorów danych wyłączanych z puli uczącej w trybie budowy komitetu.



Rys. 10.6. Dwie cechy najistotniejsze do rozpoznawania tożsamości chemicznej substancji zanieczyszczających z zastosowaniem zespołu drzew klasyfikacji (•) na tle innych istotnych cech (o) umiejscowione w strukturze macierzy danych pomiarowych

Dla porównania z metodami LDA i *k*-NN posłużono się również dwuelementowym wektorem cech najlepszych. Położenie cech stanowiących elementy obu typów wektorów w macierzy cech przedstawiono na rys. 10.6.

Dyskusja

Wykazano małą przydatność analizy wariancji jako metody selekcji cech na potrzeby rozpoznawanych tożsamości chemicznej substancji zanieczyszczających. Na podstawie jej wyników liczba pojedynczych cech różnicujących substancje w sposób statystycznie istotny była duża. Faktyczna przydatność jednoelementowych wektorów cech do rozpoznawania zanieczyszczeń była mniejsza i wyraźnie zależała od zastoso-

wanej metody klasyfikacji (patrz tabela 10.1). Potencjalne potraktowanie filtracji w tej wersji, jako metody preselekcji cech, również okazało się zawodne. Zasadniczo cechy pojedyncze najbardziej przydatne w klasyfikacji stanowiły podzbiór cech wyłonionych w toku filtracji, jak pokazano na rys. 10.1. Występowały również odstępstwa od tej reguły, np. w rozpoznawaniu heksanu, oktanu, heptanu czy toluenu. Jednak to oznaczenia na podstawie wielowymiarowej, nie jednowymiarowej reprezentacji badanych gazów dawały zadowalające rezultaty, jak pokazano w tabeli 10.2, na rys. 10.3 i 10.5.

Stwierdzono istotną przewagę nieliniowej metody klasyfikacji nad liniową w zastosowaniu do rozpoznawania tożsamości chemicznej substancji zanieczyszczających. Obserwacja ta dotyczyła zarówno reprezentacji zanieczyszczenia w postaci wektora cech wyselekcjonowanych, jak i wyekstrahowanych. Stosując metodę nieliniową, uzyskano bezbłędną klasyfikację substancji na podstawie dwuelementowych wektorów cech otrzymanych w wyniku selekcji (tabela 10.2) i trójelementowych wektorów cech uzyskanych na drodze ekstrakcji (rys. 10.3). Klasyfikator liniowy był w stanie zapewnić taki rezultat dopiero dzięki wykorzystaniu siedmioelementowych wektorów cech otrzymanych w wyniku selekcji. Zastosowanie wektora cech wyekstrahowanych o tej liczbie elementów nie zapewniało informacji wystarczającej do bezbłędnego rozpoznawania.

Analizując skład najlepszych wieloelementowych wektorów cech, stwierdzono, że bardzo rzadko pochodziły one z sygnałów czujników 1, 2, 3, 6, 7, 15, (patrz tabela 10.3). Dla sześciu spośród ośmiu rozważanych zanieczyszczeń obserwowano zgodność między oceną przydatności czujników ze strony algorytmów selekcji opakowanej współpracujących z liniową oraz nieliniową metodą klasyfikacji (tabela 10.3). W szczegółach skład wektorów cech wytypowanych tymi dwoma sposobami mógł jednak znacznie się różnić (rys. 10.2). Należy dodać, że co najmniej jeden z czujników stanowiących źródło najlepszych jednoelementowych wektorów cech był również uznawany za przydatny przez algorytmy selekcji wektorów wieloelementowych.

Komitet drzew klasyfikacyjnych trudno uznać za konkurencyjny w stosunku do metod *k*-NN lub LDA pod względem np. liczby cech potrzebnych do osiągnięcia bezbłędnego rozpoznawania substancji zanieczyszczających. Pokazano, że rozwiązanie takie można uzyskać, stosując od kilku do kilkunastu cech współpracujących z komitetem piętnastu klasyfikatorów (rys. 10.5 i rys. 10.6). Niemniej jednak uzyskane wyniki są zachęcające ze względu na prostotę algorytmu klasyfikującego, zwłaszcza pod względem aplikacji w przyrządzie pomiarowym. Interesującym jest też, że najistotniejsze cechy wybierane przez drzewa klasyfikacji pochodziły często z sygnału tego samego czujnika (patrz rys. 10.6). Dodatkowo, cechy istotne dla rozróżniania zanieczyszczeń były najczęściej umiejscowione w początkowym fragmencie odpowiedzi różnych czujników. Fakt ten potwierdził słuszność posługiwania się stanami nieustalonymi czujników jako źródłem informacji o gazowych zanieczyszczeniach powietrza.

10.2. Określanie składu jakościowego mieszanin substancji zanieczyszczających

Pozyskiwanie informacji o składzie jakościowym mieszanin substancji zanieczyszczających przeanalizowano na przykładzie następujących mieszanin: heksan i heptan (m1), heksan i oktan (m2), heksan i cykloheksan (m3), heksan i benzen (m4), heksan i toluen (m5), toluen i heksan (m6), toluen i heptan (m7), toluen i benzen (m8), toluen i ksylen (m9) oraz toluen i etylobenzen (m10). Przyjęto, że informacja ta powinna być dostępna bez względu na stężenia składników. Analizą objęto mieszaniny w zakresie stężeń podanym w tabeli 8.2.

Klasyfikator liniowy i nieliniowy oraz wektor cech wyselekcjonowanych

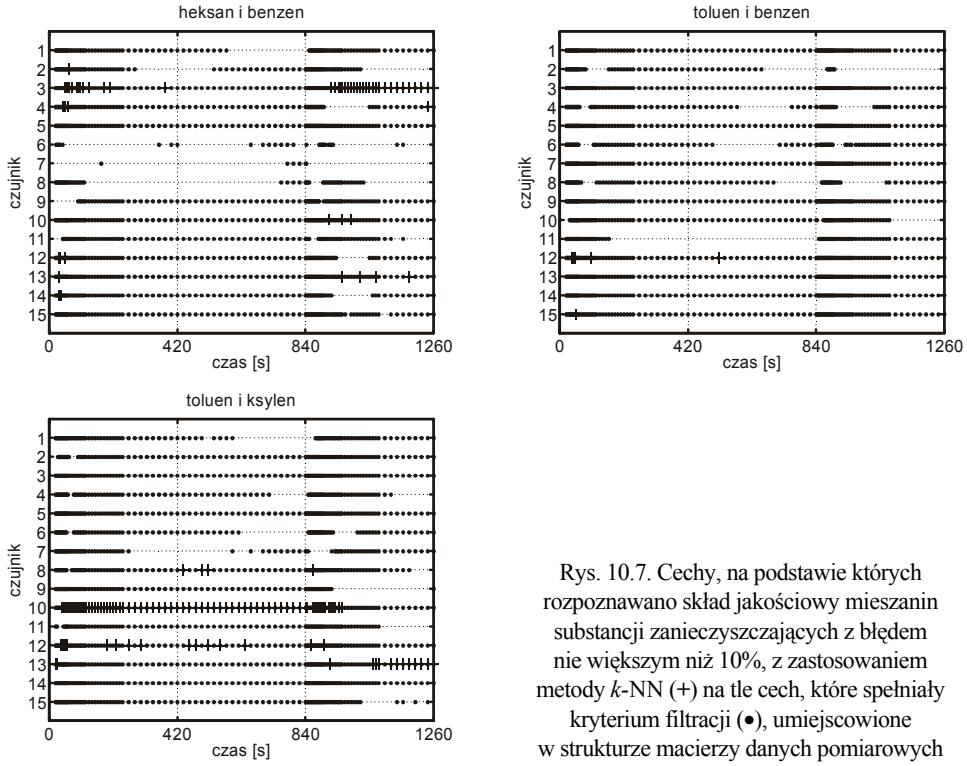
Analizowano rozpoznawanie składu mieszanin substancji zanieczyszczających na podstawie wektora cech składającego się cech wyselekcjonowanych. Rozważono wektory cech zbudowane z 1, 2, 3, 5 lub 7 cech. Zastosowano klasyfikację liniową (LDA) i nieliniową (*k*-NN). Metod tych użyto w ocenie wektorów cech w ramach selekcji w systemie opakowanym. Przestrzeń cech przeszukiwano metodą symulowanego wyżarzania. Przeprowadzono również filtrację cech metodą jednowymiarowej analizy wariancji, z przeglądem zupełnym przestrzeni cech.

Tabela 10.4. Liczba jednoelementowych wektorów cech umożliwiających rozpoznawanie składu jakościowego mieszanin substancji zanieczyszczających[a]

Skład jakościowy mieszaniny substancji zanieczyszczających	LDA	*k*-NN	ANOVA
Heksan i heptan	0	11	1534
Heksan i oktan	4	91	1409
Heksan i cykloheksan	25	282	1672
Heksan i benzen	8	59	1413
Heksan i toluen	181	439	1607
Toluen i heksan	0	44	1518
Toluen i heptan	0	3	1402
Toluen i benzen	0	8	1603
Toluen i ksylen	0	122	1650
Toluen i etylobenzen	0	82	1486

[a]Zastosowano metody klasyfikacji LDA i *k*-NN, tak aby udział błędnych klasyfikacji był mniejszy niż 10%. Liczba cech wyłonionych w toku filtracji z zastosowaniem ANOVA ($\alpha = 0{,}01$). Ogólna liczba cech – 1830.

W tabeli 10.4 podano liczbę jednoelementowych wektorów cech, które na podstawie jednowymiarowej analizy wariancji ($\alpha = 0{,}01$) wskazywały na statystycznie istotne różnice między składem jakościowym rozważanych mieszanin. W tabeli podano też liczbę jednoelementowych wektorów cech, które umożliwiały rozpoznawanie pojedynczych substancji zanieczyszczających metodami LDA i *k*-NN z rezultatem lepszym niż 10% błędnych klasyfikacji. Położenie tych wektorów w macierzy cech pokazano na rys. 10.7. Przedstawiono wspólnie wektory spełniające kryterium analizy wariancji oraz dobrze współpracujące z klasyfikatorem *k*-NN. Dla przykładu zilustrowano rezultaty otrzymane dla mieszanin: heksan i benzen, toluen i benzen oraz toluen i ksylen.



Rys. 10.7. Cechy, na podstawie których rozpoznawano skład jakościowy mieszanin substancji zanieczyszczających z błędem nie większym niż 10%, z zastosowaniem metody *k*-NN (+) na tle cech, które spełniały kryterium filtracji (•), umiejscowione w strukturze macierzy danych pomiarowych

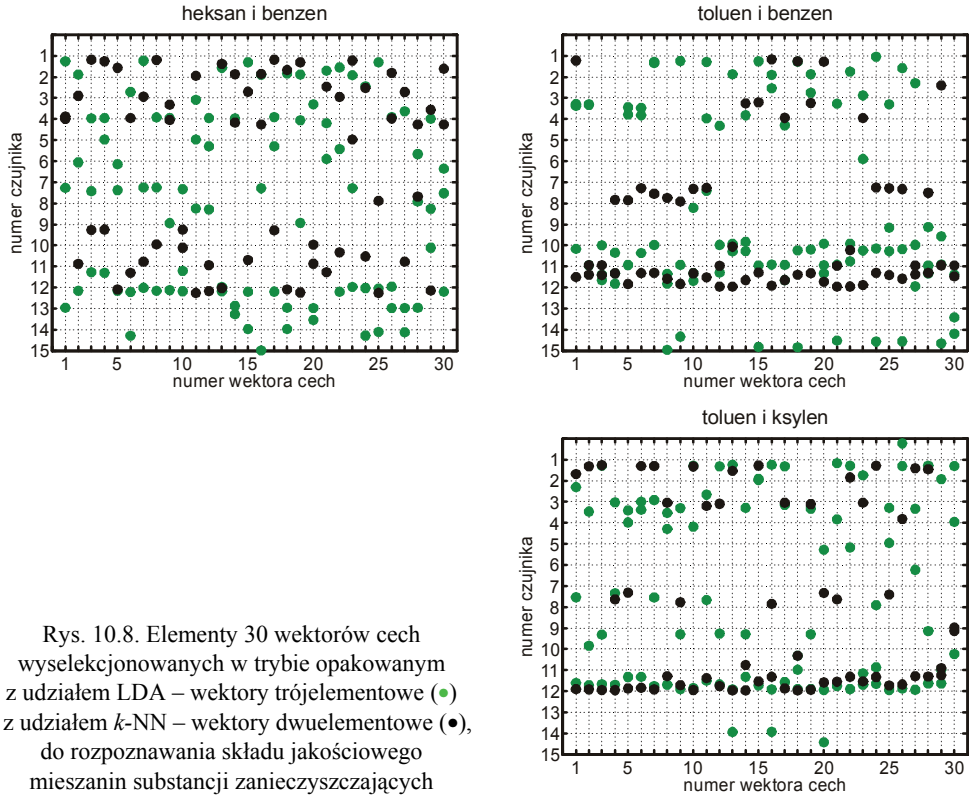
W tabeli 10.5 podano efektywność określania mieszanin substancji zanieczyszczających pod względem ich składu jakościowego różnymi metodami w zależności od liczby elementów w wektorze cech wyselekcjonowanych.

Wyłoniono trójelementowe wektory cech, które umożliwiały określenie składu jakościowego mieszanin substancji zanieczyszczających metodą LDA z jak najmniejszym błędem. Podobny zabieg wykonano względem wektorów dwuelementowych współpracujących z metodą *k*-NN. Położenie elementów trzydziestu wektorów z obu grup w macierzy cech pokazano na rys. 10.8 dla wybranych mieszanin zanieczyszczeń.

Tabela 10.5. Udział błędnych klasyfikacji (MCR) podczas rozpoznawania składu jakościowego mieszanin substancji zanieczyszczających[a]

Skład jakościowy mieszaniny substancji zanieczyszczających	LDA					*k*-NN				
	Liczba elementów wektora cech									
	1	2	3	5	7	1	2	3	5	7
Heksan i heptan	19–20	6–10	3–9	3–5	2–5	8–11	0–2	0–0		
Heksan i oktan	8–13	3–7	2–5	0–2	0–1	3–9	0–1			
Heksan i cykloheksan	5–10	1–2	0–1	0–1	0–1	1–3	0–1			
Heksan i benzen	9–15	2–7	1–4	0–0	0–1	7–10	0–1			
Heksan i toluen	3–4	0–1	0–0	0–0	0–0	1–2	0–0			
Toluen i heksan	19–22	0–4	0–1	2–4	0–0	8–10	0–0			
Toluen i heptan	28–32	5–9	4–7	3–6	0–3	10–12	0–1			
Toluen i benzen	18–24	8–9	4–8	0–1	3–4	9–11	0–1			
Toluen i ksylen	18–19	0–5	1–2	0–1	0–1	6–8	0–0			
Toluen i etylobenzen	12–16	1–3	0–2	0–0	0–0	5–8	0–0			

[a]W tabeli podano wartości min(MCR)–max(MCR) [%] dla 30 najlepszych wektorów cech o następującej liczbie elementów: 1, 2, 3, 5 i 7 uzyskanych w wyniku selekcji.



Rys. 10.8. Elementy 30 wektorów cech wyselekcjonowanych w trybie opakowanym z udziałem LDA – wektory trójelementowe (•) i z udziałem *k*-NN – wektory dwuelementowe (•), do rozpoznawania składu jakościowego mieszanin substancji zanieczyszczających

Tabela 10.6. Czujniki, które najczęściej stanowiły źródło elementów wyselekcjonowanych wektorów cech do określania składu jakościowego mieszanin substancji zanieczyszczających[a]

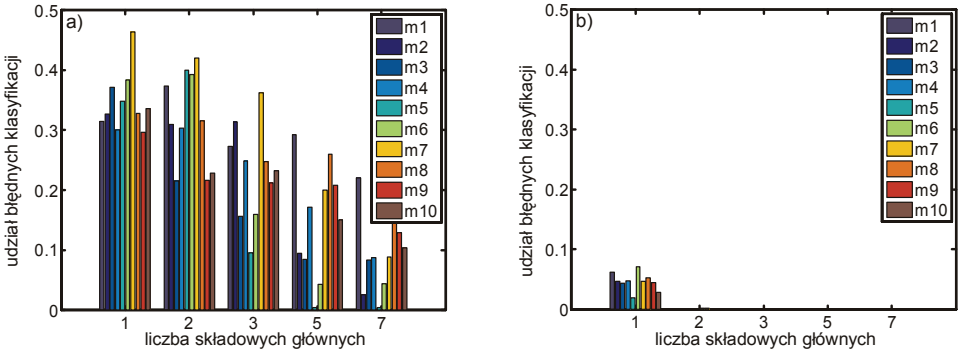
Skład jakościowy mieszaniny substancji zanieczyszczających	LDA	*k*-NN	
	Liczba elementów wektora cech		
	3	1	2
Heksan i heptan	10	–	3
Heksan i oktan	5, 3, 10	3	10, 3
Heksan i cykloheksan	10	10, 11	11, 10
Heksan i benzen	3, 12	3	12, 9, 2
Heksan i toluen	13	7, 9, 12, 14, 15	2
Toluen i heksan	11	4, 9	4, 11
Toluen i heptan	12, 5	–	12
Toluen i benzen	11	–	12, 8
Toluen i ksylen	12	10	12
Toluen i etylobenzen	4, 14	4, 11, 12	4, 12

[a]Numery czujników według tabeli 8.3.

Algorytm selekcji najczęściej wybierał cechy pochodzące z sygnałów czujników wymienionych w tabeli 10.6 jako elementy najlepszych wektorów cech trójelementowych w klasyfikacji z zastosowaniem LDA oraz dwuelementowych w klasyfikacji z zastosowaniem metody *k*-NN.

Klasyfikator liniowy i nieliniowy oraz wektor cech wyekstrahowanych

Skład jakościowy mieszanin substancji zanieczyszczających rozpoznawano na podstawie wektorów cech uzyskanych w wyniku ekstrakcji. Zastosowano składowe główne powstałe w rezultacie przekształcenia wektora cech, który był rozwinięciem

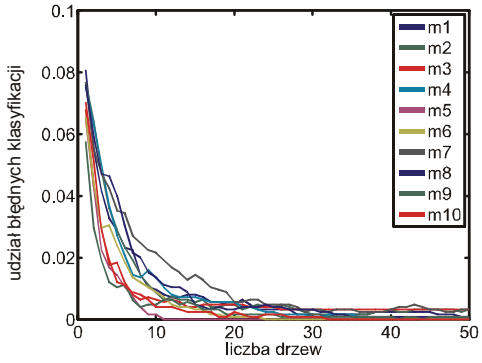


Rys. 10.9. Rezultaty rozpoznawania składu jakościowego mieszanin substancji zanieczyszczających na podstawie wektorów cech, których elementy stanowiło 1, 2, 3, 5 lub 7 składowych głównych wyłonionych za pomocą PCA. Klasyfikację wykonano z zastosowaniem: a) LDA, b) *k*-NN

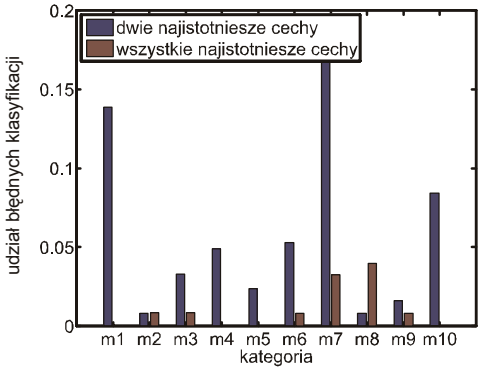
całej macierzy cech. Rozważono wektory cech zbudowane z 1, 2, 3, 5 lub 7 pierwszych składowych głównych. Rezultaty rozpoznawania z zastosowaniem metod klasyfikacji LDA i *k*-NN przedstawiono na rys. 10.9.

Komitet drzew klasyfikacyjnych

Analizowano możliwości rozwiązania wbudowanego w postaci komitetu drzew klasyfikacyjnych (CART) w zakresie określania składu jakościowego mieszanin substancji zanieczyszczających powietrze. Wejściem klasyfikatora była cała macierz cech, rozwinięta do postaci wektora. Na rysunku 10.10 pokazano zależność błędu rozpoznawania od liczby drzew w komitecie.

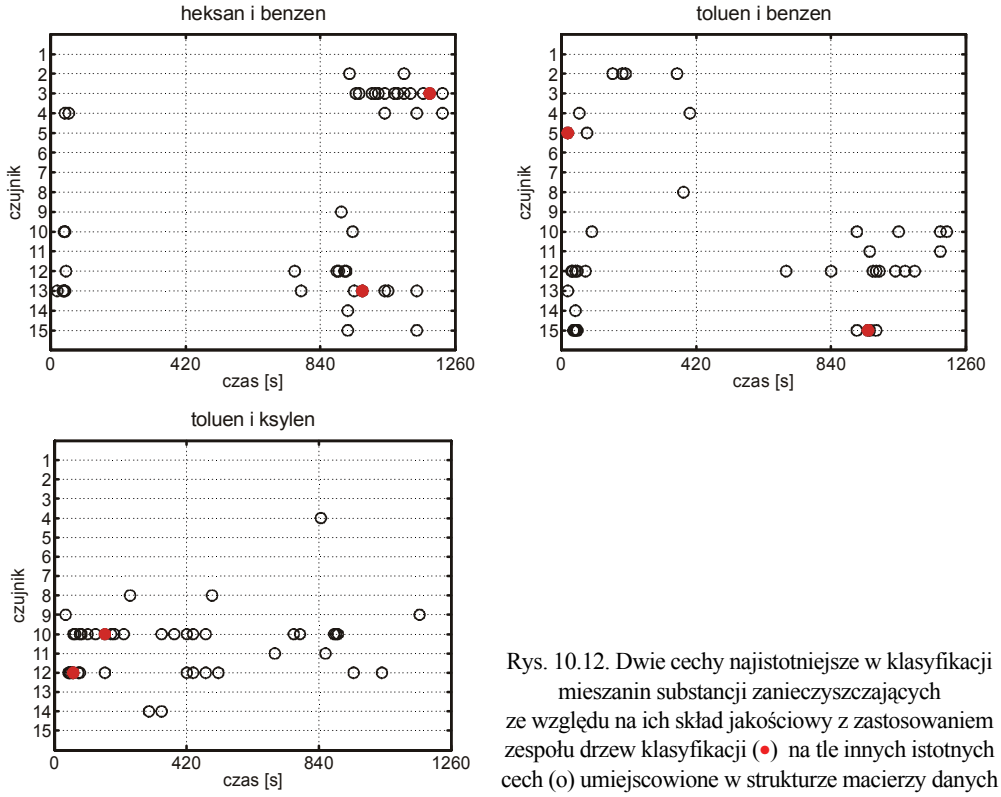


Rys. 10.10. Zależność efektywności rozpoznawania składu jakościowego mieszanin substancji zanieczyszczających od liczby drzew w komitecie klasyfikatorów



Rys. 10.11. Efektywność rozpoznawania składu jakościowego mieszanin substancji zanieczyszczających za pomocą komitetu 20 drzew klasyfikacyjnych. Zastosowano cechy istotne według klasyfikatora pracującego z wektorem zawierającym wszystkie elementy macierzy cech

Jak wynika z rysunku 10.10, zastosowanie komitetów liczących więcej niż dwadzieścia drzew klasyfikacyjnych nie wiązało się z poprawą rezultatów określania mieszanin substancji zanieczyszczających pod względem ich składu jakościowego. Prawidłowość ta dotyczyła wszystkich badanych mieszanin. Na rysunku 10.11 pokazano wyniki uzyskane z zastosowaniem komitetu 20 drzew klasyfikacyjnych i na podstawie zredukowanego wektora cech. Jego elementy stanowiły cechy o wskaźniku istotności powyżej 0.2. Dla porównania z metodami LDA i *k*-NN posłużono się wektorem wszystkich cech istotnych oraz dwuelementowym wektorem cech najlepszych. Położenie cech stanowiących elementy obu typów wektorów w macierzy cech przedstawiono na rys. 10.12.



Rys. 10.12. Dwie cechy najistotniejsze w klasyfikacji mieszanin substancji zanieczyszczających ze względu na ich skład jakościowy z zastosowaniem zespołu drzew klasyfikacji (•) na tle innych istotnych cech (o) umiejscowione w strukturze macierzy danych

Dyskusja

Na podstawie rezultatów analizy wariancji liczba pojedynczych cech różnicujących w sposób statystycznie istotny mieszaniny substancji zanieczyszczających była z reguły większa niż dla pojedynczych substancji zanieczyszczających (por. tabele 10.1 i 10.4). Przeciwnie, liczba cech realnie przydatnych do celów klasyfikacji jednowymiarowej była znacznie mniejsza. Mimo dysproporcji ilościowych między rezultatami jednowymiarowej selekcji i filtracji zachodziła natomiast między nimi zgodność w typowaniu cech istotnych. Cechy wyselekcjonowane z udziałem metody *k*-NN stanowiły podzbiór cech odfiltrowanych (rys. 10.7).

Wieloelementowe wektory cech wyselekcjonowanych zapewniały istotnie lepsze wyniki jakościowego określania składu mieszanin zanieczyszczeń niż wektory jednoelementowe (tabela 10.5). Niestety, uzyskanie rezultatów podobnych jak dla pojedynczych zanieczyszczeń wymagało zwiększenia wymiaru przestrzeni cech. Dla prawie wszystkich mieszanin błąd mniejszy od 5% uzyskano metodą liniową na podstawie reprezentacji złożonej z pięciu cech. Możliwości metody nieliniowej były większe, gdyż już trójelementowy wektor cech zapewniał błąd klasyfikacji poniżej ułamka procenta.

W określaniu składu jakościowego poszczególnych mieszanin substancji zanieczyszczających wiodącą role odgrywały jeden lub dwa czujniki. Były one najczęściej wybierane jako źródło elementów wektorów umożliwiających efektywne rozpoznawanie, jak pokazano na rys. 10.8 dla przykładowych mieszanin. Mimo znaczenia tych pojedynczych czujników duże błędy klasyfikacji jednowymiarowej wskazują na istotną rolę reprezentacji wielowymiarowej (patrz tabela 10.5). W praktyce oznacza to konieczność korzystania z danych pochodzących od innych czujników lub z innych fragmentów sygnału tego samego czujnika. Jak wynika z rys. 10.8, użyteczne dane mogą być zlokalizowane w bardzo różnych miejscach macierzy cech.

Wyniki analiz prowadzą do wniosku, że w porównaniu z selekcją cech ekstrakcja zwiększyła możliwości klasyfikacji nieliniowej i zmniejszyła możliwości metody liniowej (por. tabela 10.5 i rys. 10.9). Już dwuelementowy wektor cech wyekstrahowanych umożliwiał praktycznie bezbłędne rozpoznawanie składu jakościowego mieszanin gazów metodą nieliniową (rys. 10.9b). Nawet siedmioelementowy wektor tego typu w połączeniu z metodą liniową nie gwarantował natomiast błędu mniejszego niż 10%.

Analiza cech istotnych dla budowy komitetu drzew klasyfikacyjnych potwierdziła znaczenie sygnału pojedynczego czujnika w rozpoznawaniu mieszanin gazów. Jak wynika z rys. 10.12, wiele cech przydatnych w rozpoznawaniu poszczególnych mieszanin pochodziło z sygnałów pojedynczych czujników. Sensory wskazane w tej analizie jako istotne pokrywały się zasadniczo z czujnikami wskazanymi w ramach opakowanej selekcji cech (por. rys. 10.12 i tabela 10.6). Rezultaty rozpoznawania mieszanin substancji zanieczyszczających przez komitet drzew klasyfikacji na podstawie dwuelementowych wektorów cech (rys. 10.11) były gorsze niż przedstawione rozwiązania nieliniowe z zewnętrzną selekcją lub ekstrakcją cech (tabela 10.5). Zastosowanie wszystkich cech istotnych przyniosło natomiast rezultaty godne zainteresowania niemal dla wszystkich mieszanin.

10.3. Określanie przynależności do kategorii substancji zanieczyszczających

Analizowano również pozyskiwanie informacji o przynależności do kategorii substancji zanieczyszczających powietrze na podstawie pomiarów czujnikowych. Jako przykładowe kategorie wybrano: węglowodory alifatyczne i węglowodory aromatyczne. Kategorie te są związane ze strukturą chemiczną cząsteczki związku, lecz równocześnie mają związek z właściwościami toksycznymi zanieczyszczeń. W pierwszej kategorii znalazły się: heksan, heptan i oktan, do drugiej kategorii należały: benzen, toluen, ksylen i etylobenzen. Wymienione kategorie substancji zanieczyszczających były rozpoznawane niezależnie od stężeń poszczególnych substancji należących do tych kategorii w zakresie stężeń podanym w tabeli 8.1.

Klasyfikator liniowy i nieliniowy oraz wektor cech wyselekcjonowanych

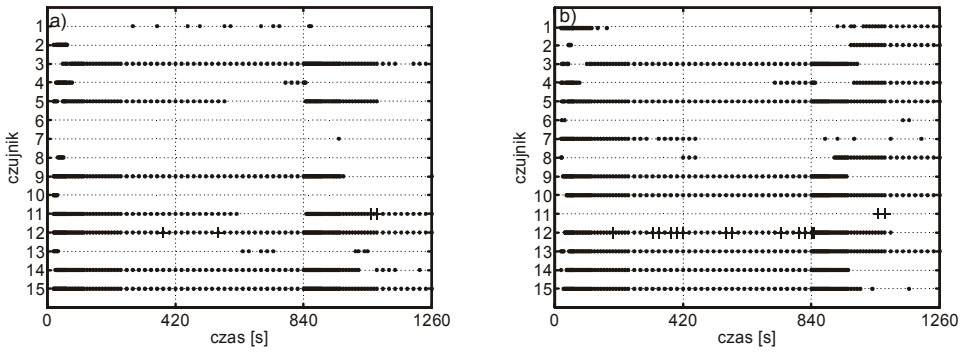
Jako podstawę rozpoznawania rozważono wektory składające się z 1, 2, 3, 5 lub 7 cech. Jako liniową metodę klasyfikacji zastosowano liniową analizę dyskryminacyjną, jako nieliniową – metodę *k*-najbliższych sąsiadów. Metody te służyły do oceny wektorów cech w ramach selekcji w systemie opakowanym. Przestrzeń cech przeszukiwano metodą symulowanego wyżarzania. Dla porównania przeprowadzono filtrację cech metodą jednowymiarowej analizy wariancji. W tym wypadku przestrzeń cech przejrzano zupełnie.

Tabela 10.7. Liczba jednoelementowych wektorów cech umożliwiających rozpoznawanie kategorii substancji zanieczyszczających[a]

Kategoria substancji zanieczyszczających	LDA	*k*-NN	ANOVA
LZO alifatyczne	0	4	839
LZO aromatyczne	0	15	1063

[a]Rozpoznawanie z zastosowaniem metody klasyfikacji LDA i *k*-NN, tak aby udział błędnych klasyfikacji był mniejszy niż 10%. Liczba cech istotnych wyłonionych w toku filtracji z zastosowaniem ANOVA ($\alpha = 0{,}01$). Ogólna liczba cech – 1830.

W tabeli 10.7 podano liczbę jednoelementowych wektorów cech, które wskazano jako istotnie różnicujące rozważane kategorie substancji zanieczyszczających na podstawie jednowymiarowej analizy wariancji dla przyjętego poziomu istotności $\alpha = 0{,}01$. Zestawiono je z liczbą jednoelementowych wektorów cech, które umożliwiały rozpoznawanie tych kategorii metodami LDA i *k*-NN z rezultatem lepszym niż 10% błędnych klasyfikacji.



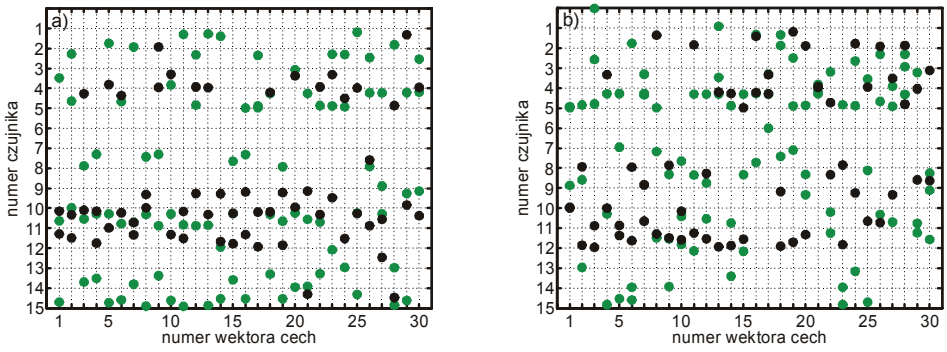
Rys. 10.13. Cechy, na podstawie których rozpoznawano kategorie substancji zanieczyszczających metodą *k*-NN z błędem nie większym niż 10% (+) na tle cech, które spełniały kryterium filtracji (•), umiejscowione w strukturze macierzy danych pomiarowych: a) rozpoznawanie węglowodorów alifatycznych, b) rozpoznawanie węglowodorów aromatycznych

Tabela 10.8. Udział błędnych klasyfikacji (MCR) w rozpoznawaniu kategorii substancji zanieczyszczających[a]

Kategoria substancji zanieczyszczających	LDA					*k*-NN				
	Liczba elementów wektora cech									
	1	2	3	5	7	1	2	3	5	7
LZO alifatyczne	20–23	8–16	2–11	0–4	0–2	6–13	0–0			
LZO aromatyczne	16–18	0–9	0–3	0–0	0–0	4–11				

[a]Podano wartości min(MCR)–max(MCR) [%] dla 30 najlepszych wektorów cech o następującej liczbie elementów: 1, 2, 3, 5 i 7 uzyskanych w wyniku selekcji.

Położenie wyselekcjonowanych wektorów cech w macierzy cech przedstawiono na rys. 10.13. Zaznaczono wektory spełniające kryterium analizy wariancji oraz cechy dobrze współpracujące z metodą *k*-NN. Błędy oznaczeń na podstawie wieloelementowych wektorów cech wyselekcjonowanych metodą LDA i *k*-NN podano w tabeli 10.8.



Rys. 10.14. Elementy 30 dwuelementowych wektorów cech wyselekcjonowanych w trybie opakowanym z udziałem LDA (•) i *k*-NN (•) do rozpoznawania kategorii pojedynczych substancji zanieczyszczających

Tabela 10.9. Wyniki rozpoznawania składu jakościowego mieszanin substancji zanieczyszczających na podstawie wektorów cech uzyskanych w wyniku ekstrakcji[a]

Kategoria substancji zanieczyszczających	LDA	*k*-NN	
	Liczba elementów wektora cech		
	3	1	2
LZO alifatyczne	11	11, 12	11
LZO aromatyczne	5	12	12

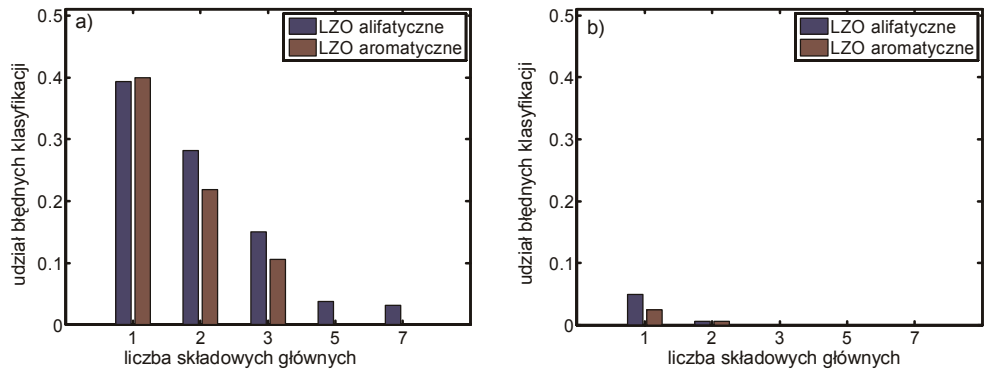
[a]Numery czujników zgodne z tabelą 8.3.

Wyłoniono trzydzieści trójelementowych wektorów cech, które umożliwiały rozpoznawanie kategorii substancji zanieczyszczających z jak najmniejszym błędem

z zastosowaniem metody LDA. Operację tę wykonano też dla wektorów dwuelementowych w kontekście metody *k*-NN. Położenie elementów tych wektorów w macierzy cech pokazano na rys. 10.14, czujniki zaś, które najczęściej ich dostarczały, zestawiono w tabeli 10.9.

Klasyfikator liniowy i nieliniowy oraz wektor cech wyekstrahowanych

Rozważono zastosowanie wektorów cech wyekstrahowanych jako podstawy określania przynależności do kategorii substancji zanieczyszczających. Ich elementy stanowiły składowe główne wyłonione z wektora cech będącego rozwiniętą macierzą cech. Wektory cech zbudowano z 1, 2, 3, 5 lub 7 pierwszych składowych głównych. Rezultaty rozpoznawania z zastosowaniem metod klasyfikacji LDA i *k*-NN przedstawiono na rys. 10.15.



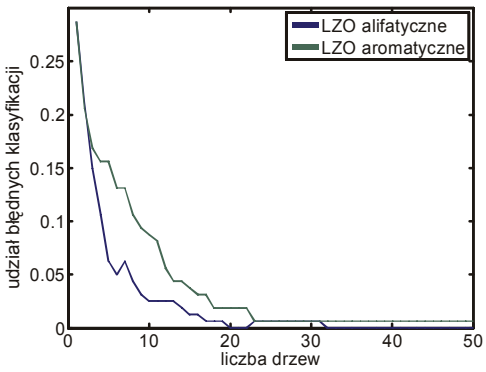
Rys. 10.15. Rezultaty rozpoznawania kategorii substancji zanieczyszczających na podstawie wektorów cech, których elementy stanowiły odpowiednio 1, 2, 3, 5 lub 7 składowych głównych, wyłonionych za pomocą PCA. Klasyfikację wykonano z zastosowaniem: (a) LDA, (b) *k*-NN

Komitet drzew klasyfikacyjnych

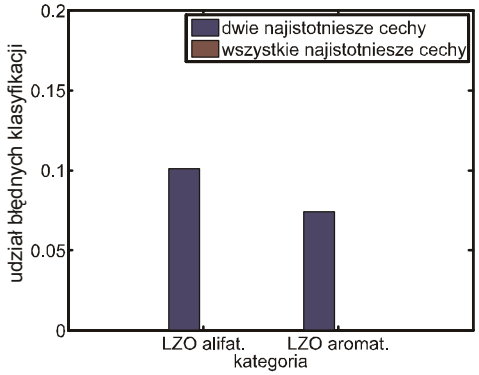
Zastosowano też rozwiązanie wbudowane w postaci komitetu drzew klasyfikacyjnych jako metodę określania przynależności do kategorii substancji zanieczyszczających. Wejściem klasyfikatora była cała macierz cech, rozwinięta do postaci wektora. Na rysunku 10.16 pokazano zależność poprawności rozpoznawania substancji od liczby drzew w komitecie.

Jak wynika z rysunku 10.16, zastosowanie komitetów liczących więcej niż dwadzieścia drzew klasyfikacyjnych nie przynosiło poprawy dokładności określenia przynależności do kategorii substancji zanieczyszczających. Na rysunku 10.17 pokazano

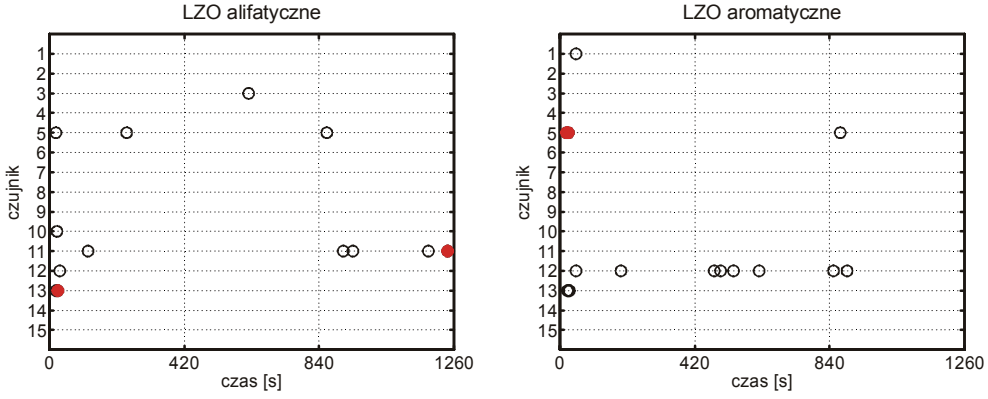
wyniki uzyskane z zastosowaniem komitetu 20 drzew klasyfikacyjnych i na podstawie zredukowanego wektora cech. Jego elementy stanowiły cechy o wskaźniku istotności powyżej 0,2. Posłużono się wektorem wszystkich takich cech oraz, dla porównania z wynikami innych klasyfikatorów, zastosowano dwuelementowy wektor cech najlepszych. Położenie cech stanowiących elementy obu typów wektorów w macierzy cech przedstawiono na rys. 10.18.



Rys. 10.16. Zależność efektywności rozpoznawania kategorii substancji zanieczyszczających od liczby drzew w komitecie klasyfikatorów



Rys. 10.17. Efektywność rozpoznawania kategorii substancji zanieczyszczających za pomocą komitetu 20 drzew klasyfikacyjnych. Zastosowano cechy istotne według klasyfikatora pracującego z wektorem zawierającym wszystkie elementy macierzy cech



Rys. 10.18. Dwie cechy najistotniejsze do określania przynależności do kategorii substancji zanieczyszczających z zastosowaniem zespołu drzew klasyfikacji (•) na tle innych istotnych cech (o) umiejscowione w strukturze macierzy cech

Dyskusja

Rozpoznawanie kategorii substancji zanieczyszczających na podstawie jednoelementowych wektorów cech wyselekcjonowanych było obarczone błędem od kilkunastu (metody nieliniowe) do kilkudziesięciu procent (metody liniowe) (tabela 10.8). Ten niesatysfakcjonujący rezultat cechowała w dodatku niewspółmierność względem wykazanej w toku analizy wariancji dużej liczby cech istotnie różnicujących kategorie substancji zanieczyszczających (tabela 10.7).

Jak wynika z tabeli 10.8, przełom w efektywności klasyfikacji nastąpił dzięki zastosowaniu wieloelementowych wektorów cech wyselekcjonowanych. Wraz z użyciem metody nieliniowej dwuelementowy wektor cech umożliwiał bezbłędną klasyfikację kategorii substancji zanieczyszczających. Z zastosowaniem metody liniowej rozpoznawanie kategorii związków aromatycznych z błędem mniejszym niż 3% wymagało trójelementowego wektora cech. Jednak dla związków alifatycznych ten sam poziom błędu osiągnięto dopiero dla pięcioelementowego wektora cech.

Ekstrakcja cech pogarszała rezultaty rozpoznawania kategorii substancji zanieczyszczających metodą LDA, jak również *k*-NN w przestrzeniach niskowymiarowych w stosunku do selekcji cech.

Analiza składu wektorów cech wyselekcjonowanych wskazywała na istotną rolę czujnika 12 w rozpoznawaniu węglowodorów aromatycznych oraz czujnika 11 w rozpoznawaniu węglowodorów alifatycznych metodą nieliniową (rys. 10.13, 10.14, tabela 10.9). W podejściu liniowym najczęściej były wybierane czujniki 5 i 11. Mimo ich znaczenia dla uzyskania informacji odpowiedniej jakości konieczny był udział dodatkowego źródła. W ramach podejścia opakowanego był to na ogół inny czujnik.

Selekcja cech w ramach podejścia wbudowanego ujawniła natomiast, że cechy najistotniejsze dla rozpoznawania kategorii mieszanin zanieczyszczeń mogły pochodzić z sygnału tego samego czujnika (rys. 10.18). Dla związków aromatycznych był to czujnik 12, a dla kategorii związki alifatyczne – czujnik 11. Rezultaty uzyskane tą metodą dla dwuelementowego wektora najistotniejszych cech były lepsze niż metodą LDA (rys. 10.18), lecz gorsze niż z zastosowaniem metody *k*-NN.

10.4. Określanie przynależności do kategorii mieszanin substancji zanieczyszczających

Analizowano czujnikowe dane pomiarowe pod względem pozyskania informacji o kategoriach mieszanin substancji zanieczyszczających. Przyjęto, że wyznacznikiem kategorii mieszanin jest wspólny składnik dominujący. W związku z tym rozważono kategorię mieszanin o składzie zdominowanym przez heksan i kategorię mieszanin,

których skład był zdominowany przez toluen. Do pierwszej kategorii należały mieszaniny, w których heksan występował razem z heptanem, oktanem, cykloheksanem, benzenem lub toluenem, do drugiej natomiast mieszaniny, w których toluen występował obok heksanu, heptanu, benzenu, ksylenu i etylobenzenu. Wymienione kategorie mieszanin substancji zanieczyszczających były rozpoznawane niezależnie od stężeń poszczególnych składników w zakresie stężeń podanym w tabeli 8.2.

Klasyfikator liniowy i nieliniowy oraz wektor cech wyselekcjonowanych

Rozpoznawano kategorie mieszanin substancji zanieczyszczających na podstawie wektora cech składającego się z wyselekcjonowanych cech. Rozważono wektory cech zbudowane z 1, 2, 3, 5 lub 7 cech. Zastosowano liniową analizę dyskryminacyjną oraz metodę *k*-najbliższych sąsiadów do oceny wektorów cech w ramach selekcji w systemie opakowanym. Przestrzeń cech przeszukiwano metodą symulowanego wyżarzania. Dla porównania przeprowadzono filtrację cech metodą jednowymiarowej analizy wariancji z przeglądem zupełnym przestrzeni cech.

Tabela 10.10. Liczba jednoelementowych wektorów cech umożliwiających rozpoznawanie mieszanin substancji zanieczyszczających[a]

Kategoria mieszanin substancji zanieczyszczających	LDA	*k*-NN	ANOVA
Mieszaniny o składzie zdominowanym przez heksan lub toluen	280	171	1760

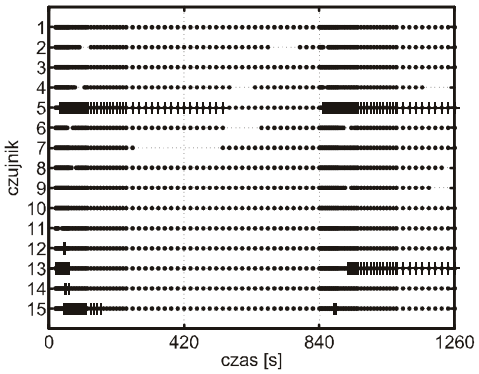
[a]Rozpoznawanie z zastosowaniem metody klasyfikacji LDA i *k*-NN, tak aby udział błędnych klasyfikacji był mniejszy niż 10%. Liczba cech wyłonionych w toku filtracji z zastosowaniem ANOVA. Ogólna liczba cech – 1830.

W tabeli 10.10 podano liczbę pojedynczych cech, które za pomocą analizy wariancji wyraźnie różnicowały rozważane kategorie mieszanin zanieczyszczeń na poziomie istotności $\alpha = 0{,}01$. Zestawiono je z liczbą jednoelementowych wektorów cech, które umożliwiały rozpoznawanie pojedynczych substancji zanieczyszczających metodami LDA i *k*-NN z rezultatem lepszym niż 10% błędnych klasyfikacji. Położenie tych wektorów cech w macierzy cech pokazano na rys. 10.19. Przedstawiono wspólnie wektory spełniające kryterium analizy wariancji oraz dobrze współpracujące z oboma typami klasyfikatorów.

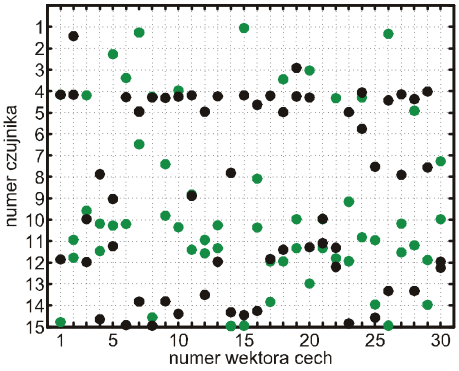
Z zastosowaniem *k*-NN, jak również LDA uzyskano bezbłędne rozpoznawanie kategorii mieszanin substancji zanieczyszczających na podstawie wektorów cech o liczbie elementów większej niż jeden (tabela 10.11).

Na rysunku 10.20 przedstawiono położenie elementów trzydziestu najlepszych dwuelementowych wektorów cech wyłonionych w trybie selekcji opakowanej z udzia-

łem LDA i *k*-NN, w strukturze macierzy cech. W tabeli 10.12 zestawiono numery czujników, z których sygnałów pochodziły najczęściej elementy owych wektorów cech.



Rys. 10.19. Cechy, dla których udział błędnych klasyfikacji kategorii mieszanin substancji zanieczyszczających z zastosowaniem metody *k*-NN nie przekraczał 10% (+) na tle cech, które spełniały kryterium filtracji (•) umiejscowione w strukturze macierzy danych pomiarowych



Rys. 10.20. Elementy 30 dwuelementowych wektorów cech wyselekcjonowanych w trybie opakowanym z udziałem LDA (•) i *k*-NN (•) na potrzeby rozpoznawania kategorii mieszanin substancji zanieczyszczających

Tabela 10.11. Udział błędnych klasyfikacji (MCR) w rozpoznawaniu kategorii mieszanin substancji zanieczyszczających[a]

Kategoria mieszanin substancji zanieczyszczających	LDA					*k*-NN				
	Liczba elementów wektora cech									
	1	2	3	5	7	1	2	3	5	7
Mieszaniny o składzie zdominowanym przez heksan lub toluen	1–3	0–0				1–3	0–0			

[a]W tabeli podano wartości min(MCR)–max(MCR) [%] dla 30 najlepszych wektorów cech o liczbie elementów 1, 2, 3, 5 lub 7 uzyskanych w wyniku selekcji.

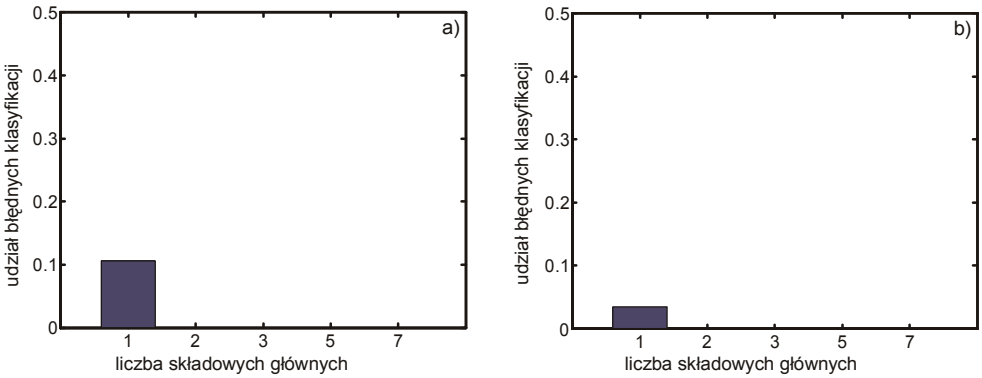
Tabela 10.12. Czujniki, dla których algorytm selekcji najczęściej wybierał cechy pochodzące z ich sygnałów jako elementy najlepszych wektorów cech dwuelementowych podczas klasyfikacji kategorii mieszanin substancji zanieczyszczających z zastosowaniem LDA oraz metody *k*-NN[a]

Kategoria mieszanin substancji zanieczyszczających	LDA	*k*-NN	
	Liczba elementów wektora cech		
	3	1	2
Mieszaniny o składzie zdominowanym przez heksan lub toluen	5, 14	5, 13	5

[a]Numery czujników podano zgodne z tabelą 8.3.

Klasyfikator liniowy i nieliniowy oraz wektor cech wyekstrahowanych

Przeprowadzono rozpoznawanie kategorii mieszanin substancji zanieczyszczających na podstawie wektorów cech wyekstrahowanych. Jako metodę ekstrakcji zastosowano PCA. Ich elementy stanowiły składowe główne wyłonione z wektora cech obejmującego wszystkie elementy macierzy cech. Rozważono wektory cech zbudowane z 1, 2, 3, 5 lub 7 pierwszych składowych głównych. Rezultaty rozpoznawania z zastosowaniem metod klasyfikacji LDA i *k*-NN przedstawiono na rys. 10.21.

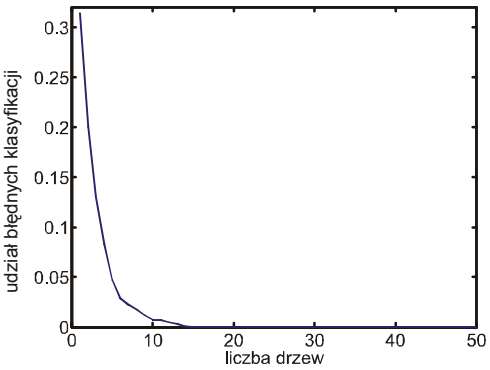


Rys. 10.21. Rezultaty rozpoznawania pojedynczych substancji zanieczyszczających na podstawie wektorów cech, których elementy stanowiły 1, 2, 3, 5 lub 7 składowych głównych wyłonionych za pomocą PCA. Wyniki uzyskano z zastosowaniem: a) LDA, b) *k*-NN

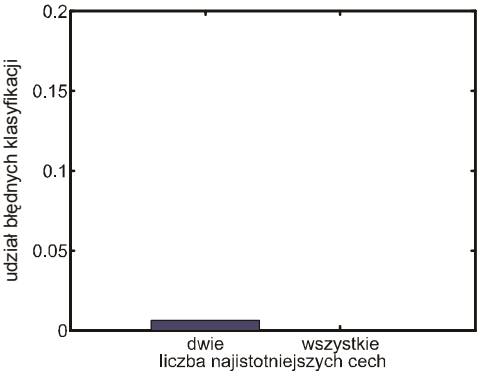
Komitet drzew klasyfikacyjnych

Analizowano możliwości rozwiązania wbudowanego w postaci komitetu drzew klasyfikacyjnych w zastosowaniu do określania przynależności gazu do kategorii mieszanin substancji zanieczyszczających. Wejściem klasyfikatora była cała macierz cech, rozwinięta do postaci wektora. Na rysunku 10.22 pokazano zależność poprawności rozpoznawania substancji od liczby drzew w komitecie.

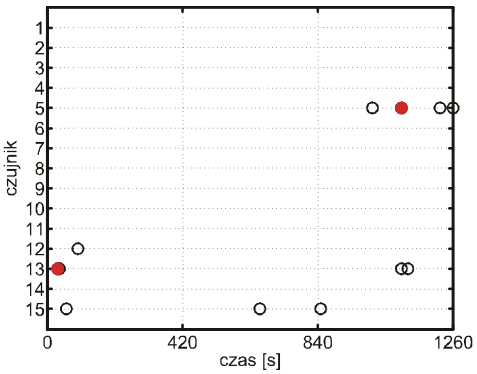
Jak wynika z rysunku 10.22, zastosowanie komitetów liczących więcej niż piętnaście drzew klasyfikacyjnych nie powodowało poprawy rezultatów rozpoznawania kategorii substancji zanieczyszczających. Na rysunku 10.23 pokazano wyniki uzyskane z zastosowaniem komitetu 15 drzew klasyfikacyjnych i na podstawie zredukowanego wektora cech. Jego elementy stanowiły cechy o wskaźniku istotności większym niż 0,2. Posłużono się wektorem wszystkich takich cech oraz dla porównania z innymi metodami klasyfikacji dwuelementowym wektorem cech najlepszych. Położenie cech stanowiących elementy obu typów wektorów w macierzy cech przedstawiono na rys. 10.24.



Rys. 10.22. Zależność efektywności rozpoznawania kategorii mieszanin substancji zanieczyszczających od liczby drzew w klasyfikatorze



Rys. 10.23. Efektywność rozpoznawania pojedynczych substancji zanieczyszczających za pomocą komitetu 15 drzew klasyfikacyjnych. Zastosowano cechy istotne według klasyfikatora pracującego z wektorem będącym rozwiniętą macierzą cech



Rys. 10.24. Dwie cechy najistotniejsze do określania przynależności do kategorii mieszanin substancji zanieczyszczających z zastosowaniem zespołu drzew klasyfikacji (•) na tle innych istotnych cech (o)

Dyskusja

Przykładem złożonych kategorii zanieczyszczeń są mieszaniny gazów mające w składzie wspólną substancję zanieczyszczającą. Kategorie takie rozpoznawano na podstawie danych z pomiarów czujnikowych z błędem mniejszym niż 10%, z zastosowaniem podejścia jednowymiarowego. Pozwalała na to stosunkowo liczna grupa cech (tabela 10.10).

Stwierdzono, że wystarczało zwiększyć liczbę elementów wektora cech do dwóch, by uzyskać bezbłędną klasyfikację kategorii zarówno metodą liniową, jak i nieliniową (tabela 10.11). Wiodącą rolę w rozpoznawaniu kategorii mieszanin odgrywał czuj-

nik 5, zarówno przy jednowymiarowym, jak i wielowymiarowym podejściu do problemu (tabela 10.12). Jednak pojedyncze cechy pochodzące z odpowiedzi tego czujnika były niewystarczające dla uzyskania najlepszych rezultatów. Algorytm selekcji opakowanej najczęściej wskazywał dodatkowo na czujnik 14, jeżeli klasyfikowano metodą LDA oraz czujnik 15, gdy metodą klasyfikacji była *k*-NN.

Jednowymiarowa klasyfikacja z zastosowaniem wektora cech wyekstrahowanych dawała lepsze rezultaty niż przy użyciu wektora cech wyselekcjonowanych.

Wyniki uzyskane z zastosowaniem komitetu drzew klasyfikacyjnych wskazywały na czujniki 5, 12, 13 i 15 jako źródła cech istotnych dla rozpoznawania (rys. 10.24). Wektor zbudowany z dwóch najlepszych cech pozwalał na prawie bezbłędne rozróżnianie kategorii mieszanin tą metodą (rys. 10.23). Taki rezultat uzyskano również, stosując inne klasyfikatory (tabela 10.11).

10.5. Wnioski z analizy danych czujnikowych ze względu na informację jakościową o zanieczyszczeniach

Efektywność odczytu informacji o zanieczyszczeniach na podstawie pomiarów czujnikowych zależy od doboru elementów systemu rozpoznawania wzorców do rodzaju poszukiwanej informacji. Wnioski z analizy danych czujnikowych ze względu na pozyskiwanie informacji jakościowej formułowano, mając na uwadze tę kwestię.

Rozważono szereg rodzajów informacji jakościowej. Dotyczyły one tożsamości chemicznej pojedynczych substancji zanieczyszczających, składu jakościowego ich mieszanin oraz przynależności do zdefiniowanej grupy substancji zanieczyszczających lub do grupy mieszanin takich substancji.

Tabela 10.13. Błąd uzyskiwania informacji o tożsamości chemicznej substancji zanieczyszczającej[a]

Liczba cech	Selekcja cech		Ekstrakcja cech		Komitet drzew klasyfikacyjnych
	LDA	*k*-NN	LDA	*k*-NN	
1	65	30	50	5	
2	10	0	40	1	10
3	6	0	35	0	
5	2	0	15	0	
7	0	0	10	0	
Wszystkie					5

[a]Podano maksymalny udział błędnych klasyfikacji dla najlepszych rozwiązań [%].

Tabela 10.14. Błąd uzyskiwania informacji o składzie jakościowym mieszaniny substancji zanieczyszczających[a]

Liczba cech	Selekcja cech		Ekstrakcja cech		Komitet drzew klasyfikacyjnych
	LDA	*k*-NN	LDA	*k*-NN	
1	70	20	40	5	
2	10	2	45	0	20
3	9	0.2	35	0	
5	6	0	30	0	
7	5	0	20	0	
Wszystkie					0

[a]Podano maksymalny udział błędnych klasyfikacji dla najlepszych rozwiązań [%].

Tabela 10.15. Błąd uzyskiwania informacji o kategorii substancji zanieczyszczających[a]

Liczba cech	Selekcja cech		Ekstrakcja cech		Komitet drzew klasyfikacyjnych
	LDA	*k*-NN	LDA	*k*-NN	
1	65	65	40	5	
2	17	9	30	1	4
3	10	4	15	0	
5	4	0	5	0	
7	2	0	5	0	
Wszystkie					1

[a]Podano maksymalny udział błędnych klasyfikacji dla najlepszych rozwiązań [%].

Tabela 10.16. Błąd uzyskiwania informacji o kategorii mieszanin substancji zanieczyszczających[a]

Liczba cech	Selekcja cech		Ekstrakcja cech		Komitet drzew klasyfikacyjnych
	LDA	*k*-NN	LDA	*k*-NN	
1	60	55	10	5	
2	0	0	0	0	1
3	0	0	0	0	
5	0	0	0	0	
7	0	0	0	0	
Wszystkie					0

[a]Podano maksymalny udział błędnych klasyfikacji dla najlepszych rozwiązań [%].

W tabelach 10.13–10.16 przedstawiono podsumowanie rezultatów odczytu poszczególnych rodzajów informacji jakościowej o zanieczyszczeniach wszystkimi rozważanymi metodami. Metody te stanowiły *de facto* propozycje różnych systemów

rozpoznawania wzorców. Podano górną granicę udziału błędnych klasyfikacji dla najlepszych rozwiązań w danej grupie metod.

Pod względem rodzaju informacji jakościowej najłatwiejszym problemem okazało się określenie kategorii mieszanin substancji zanieczyszczających (tabela 10.16). Nieco trudniejsze było rozróżnienie kategorii substancji zanieczyszczających (tabela 10.15). Pozyskanie pozostałych rodzajów informacji dotyczących tożsamości chemicznej substancji zanieczyszczających oraz składu jakościowego ich mieszanin odznaczało się większym stopniem trudności (tabela 10.13 i tabela 10.14).

Wyniki przeprowadzonych analiz wskazały na konieczność skorzystania z podejścia wielowymiarowego w celu zapewnienia rozwiązań obarczonych bardzo małym błędem lub bezbłędnych. Wykazano, że dokładność pozyskiwania informacji zależała zarówno od zastosowanej metody klasyfikacji, jak i rodzaju reprezentacji wielowymiarowej badanego gazu.

Rozwiązanie nieliniowe wymagało najmniejszej liczby elementów wektora cech do zapewnienia bezbłędnego określenia zanieczyszczeń pod względem jakościowym. Na ogół dwu- lub trójelementowy wektor cech wystarczał do uzyskania takiego rezultatu metodą *k*-NN. Stosując metodę liniową, najczęściej nie osiągano bezbłędnej klasyfikacji, nawet z użyciem siedmioelementowego wektora cech. Jest interesujące, że liczba elementów wektora cech zapewniającego bezbłędną klasyfikację określoną metodą była stosunkowo nieczuła na sposób wyłonienia wektora, tj. nie miało większego znaczenia, czy cechy były selekcjonowane czy ekstrahowane.

W zakresie wektorów cech niezapewniających bezbłędnej klasyfikacji metoda nieliniowa dawała lepsze wyniki w zastosowaniu z wektorami cech wyekstrahowanych, tzn. z reprezentacją informacji o charakterze globalnym. Dla takiej konfiguracji obserwowano mniejsze błędy określania zanieczyszczeń. Metoda liniowa dawała natomiast lepsze rezultaty dla wektora cech wyselekcjonowanych, tj. informacji rozproszonej w macierzy danych pomiarowych. W przypadku rozwiązań jednowymiarowych błędy uzyskane dla obu typów wektorów były z reguły porównywalne. Selekcjonując cechy dla wieloelementowych wektorów cech o tej samej liczbie elementów, uzyskiwano natomiast kilkakrotnie mniejsze błędy niż wykonując ekstrakcję.

Odnosząc się do reprezentacji badanych gazów uzyskanej w wyniku selekcji cech, można zauważyć, że niektóre czujniki są nośnikami konkretnej informacji jakościowej o zanieczyszczeniach. Stwierdzono, że czujniki te były różne dla różnych rodzajów informacji. Jest to skutek ograniczonej, lecz jednak pewnej selektywności półprzewodnikowych czujników gazów. Co interesujące, stwierdzono, że wybór czujnika kluczowego do pozyskania konkretnego rodzaju informacji o zanieczyszczeniach mógł być wrażliwy na zastosowaną metodę klasyfikacji.

Wykonane analizy prowadziły do wniosku, że jednowymiarowa reprezentacja gazów nie była wystarczająca do zapewnienia odpowiedniej jakości określenia zanieczyszczeń pod względem jakościowym. Dla każdego rozważanego rodzaju informacji konieczna była jeszcze co najmniej jedna dodatkowa cecha, która z reguły pochodziła

z sygnału innego czujnika. Dużą wartość miało zatem posługiwanie się złożonymi danymi pomiarowymi dostarczanymi przez matrycę czujnikową jako podstawą jakościowego określania zanieczyszczeń.

Z metodologicznego punktu widzenia warto podkreślić wyraźną przewagę podejścia opakowanego nad filtracją jako metodą wyboru najlepszej reprezentacji informacji jakościowej o zanieczyszczeniu, również względem reprezentacji jednowymiarowej. Stwierdzono, że analiza wariancji znacznie przeszacowuje liczbę cech przydatnych do pozyskania informacji jakościowej o badanych gazach.

Rezultaty klasyfikacji z opakowaną selekcją cech oraz ekstrakcją cech wskazywały na dominującą pozycję rozwiązania nieliniowego (*k*-NN) we współpracy z wektorem cech wyekstrahowanych. Na podstawie wyników analizy danych rezultatów należy je uznać za metodę preferowaną w poszukiwaniu informacji jakościowej o zanieczyszczeniach.

Rezultaty klasyfikacji z wbudowaną selekcją cech (komitet drzew klasyfikacyjnych) zwracały z kolei uwagę na możliwość rozpoznawania klas zanieczyszczeń na podstawie wielu cech pochodzących z sygnału jednego czujnika. Stwierdzono występowanie takich czujników dla różnych analizowanych klas zanieczyszczeń. Obserwacja ta wskazywała na możliwość zastosowania cech typu *P* pochodzących z sygnału odpowiedzi pojedynczego czujnika pracującego w trybie dynamicznym jako podstawy jakościowego określania zanieczyszczeń.

11. Analiza danych czujnikowych pod względem informacji ilościowej o zanieczyszczeniach

Dane czujnikowe analizowano pod względem możliwości oraz sposobów pozyskiwania informacji o właściwościach ilościowych zanieczyszczeń. W tej pracy skupiono się na dwóch rodzajach miar określających zanieczyszczenia ilościowo. Było to stężenie substancji zanieczyszczających oraz stężenie atomów węgla pochodzących od lotnych związków organicznych. Druga z miar jest przykładem podejścia do problemu zanieczyszczenia powietrza inaczej niż ze względu na tożsamość chemiczną substancji zanieczyszczającej.

Zastosowano podejście oparte na koncepcji rozpoznawania wzorców. Podstawowe znaczenie dla efektywności systemu analizy danych ma w tym wypadku wybór odpowiedniej reprezentacji badanego gazu oraz właściwego modelu ilościowego, które rozważono w następujących konfiguracjach:

- wektor cech wyselekcjonowanych i regresja liniowa wielokrotna, odporna,
- wektor cech wyselekcjonowanych i regresja nieliniowa,
- wektor cech wyekstrahowanych i regresja liniowa wielokrotna, odporna,
- wektor cech wyekstrahowanych i regresja nieliniowa,
- regresja metodą cząstkowych najmniejszych kwadratów.

11.1. Stężenia substancji zanieczyszczających

Analizowano metody wyznaczania stężeń substancji zanieczyszczających, gdy występują one w powietrzu jako:

- pojedyncza substancja zanieczyszczająca,
- substancja zanieczyszczająca ilościowo dominująca nad innymi,
- substancja zanieczyszczająca ilościowo zdominowana przez inne.

11.1.1. Zanieczyszczenie jako pojedyncza substancja zanieczyszczająca

Jako przykładowe pojedyncze substancje zanieczyszczające, których stężenia wyznaczano, wybrano benzen, toluen, ksylen, etylobenzen, heksan, heptan, oktan i cykloheksan. Zakres stężeń tych związków objęty analizą podano w tabeli 8.1.

Regresja liniowa i nieliniowa oraz wektor cech wyselekcjonowanych

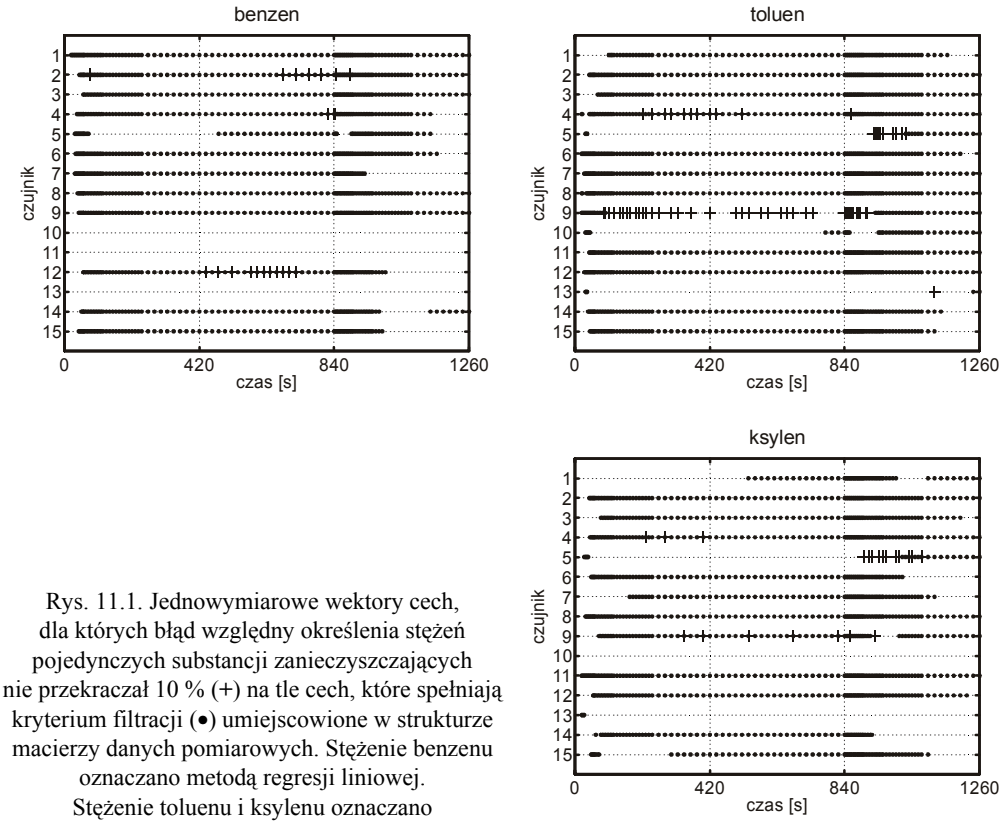
Wyznaczano stężenie pojedynczych substancji zanieczyszczających na podstawie ich reprezentacji w postaci wektorów cech wyselekcjonowanych. Rozważono wektory składające się z 1, 2, 3, 5 lub 7 cech. Selekcję cech przeprowadzono w trybie opakowanym. Dla porównania wykonano filtrację cech metodą analizy korelacji. Do oceny selekcjonowanych wektorów cech zastosowano metody regresji liniowej oraz nieliniowej (tylko dla jednoelementowych wektorów cech). Wielowymiarowe przestrzenie cech przeszukiwano metodą symulowanego wyżarzania. Dla jednowymiarowych przestrzeni cech wykonano przegląd zupełny.

Tabela 11.1. Liczba pojedynczych cech umożliwiających określenie stężenia pojedynczego zanieczyszczenia ze średnim błędem względnym MRE ≤ 10% oraz wskazana przez kryterium filtracji liczba cech istotnych[a]

Lotny związek organiczny	Selekcja cech		Filtracja cech
	Regresja liniowa	Regresja nieliniowa	
Heksan	4	127	1263
Heptan	13	34	1204
Oktan	101	1	1367
Cykloheksan	4	103	970
Benzen	21	0	1239
Toluen	8	63	1347
Ksylen	0	21	1194
Etylobenzen	1	18	856

[a]Współczynnik korelacji $R \geq 0{,}95$. Ogólna liczba cech – 1830.

W tabeli 11.1 podano liczbę jednoelementowych wektorów cech typu *P*, umożliwiających wyznaczenie stężeń pojedynczych substancji zanieczyszczających z błędem mniejszym niż 10%. Zamieszczono w niej również liczbę pojedynczych cech uznanych za istotne według kryterium filtracji (współczynnik korelacji $R \geq 0{,}95$). Na rysunku 11.1 pokazano położenie cech według tabeli 11.1 w strukturze macierzy danych pomiarowych. Dla porównania z wynikami analizy danych pod względem pozyskiwania informacji jakościowej jako przykładowe zanieczyszczenia wybrano benzen, toluen i ksylen.



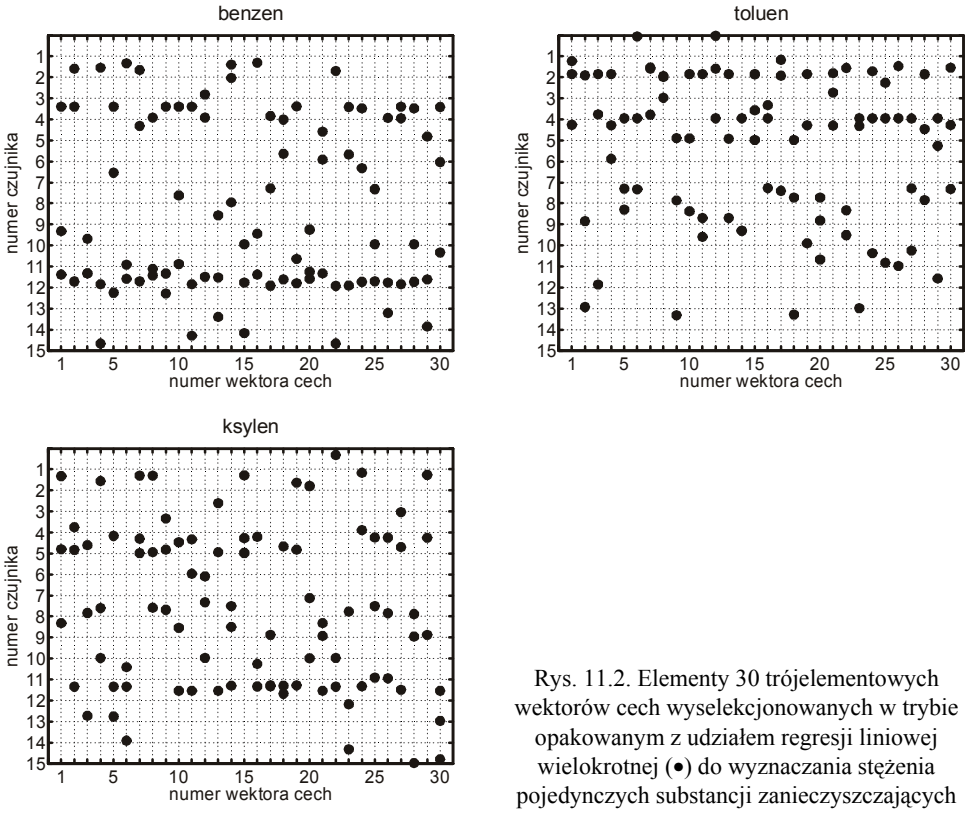
Rys. 11.1. Jednowymiarowe wektory cech, dla których błąd względny określenia stężeń pojedynczych substancji zanieczyszczających nie przekraczał 10 % (+) na tle cech, które spełniają kryterium filtracji (•) umiejscowione w strukturze macierzy danych pomiarowych. Stężenie benzenu oznaczano metodą regresji liniowej. Stężenie toluenu i ksylenu oznaczano metodą regresji nieliniowej

Tabela 11.2. Średni błąd względny (MRE) określania stężenia pojedynczego zanieczyszczenia[a]

Lotny związek organiczny	Regresja liniowa					Regresja nieliniowa
	Liczba elementów wektora cech					
	1	2	3	5	7	1
Heksan	7–16	4–6	4–6	3–5	3–5	3–5
Heptan	7–11	4–5	3–5	3–4	2–4	4–10
Oktan	5–8	4–5	4–5	3–4	3–4	9–16
Cykloheksan	8–14	5–7	5–6	4–5	3–5	4–6
Benzen	7–11	3–4	3–4	2–4	2–3	10–16
Toluen	7–13	5–7	4–6	4–5	3–5	3–8
Ksylen	11–19	5–8	5–7	5–7	5–6	5–10
Etylobenzen	10–15	4–8	4–6	3–5	3–5	5–11

[a]Podano wartości min(MRE)–max(MRE) [%] dla 30 najlepszych wektorów cech o liczbie elementów 1, 2, 3, 5 i 7, uzyskanych w wyniku selekcji.

W tabeli 11.2 zestawiono wartości średniego błędu względnego określania stężeń zanieczyszczeń na podstawie trzydziestu najlepszych wektorów cech wyłonionych w toku selekcji. Liczbę tę przyjęto arbitralnie, podobnie jak w rozdz. 10 poświęconym sposobom poszukiwania informacji o charakterze jakościowym. Dla wektorów wieloelementowych podane wyniki uzyskano z zastosowaniem regresji liniowej wielokrotnej odpornej. W odniesieniu do wektorów jednoelementowych podano wyniki uzyskane metodą regresji jednowymiarowej odpornej, liniowej i nieliniowej.



Rys. 11.2. Elementy 30 trójelementowych wektorów cech wyselekcjonowanych w trybie opakowanym z udziałem regresji liniowej wielokrotnej (•) do wyznaczania stężenia pojedynczych substancji zanieczyszczających

Na rysunku 11.2 pokazano przynależność elementów trzydziestu najlepszych wektorów cech trójelementowych (tabela 11.2) wyselekcjonowanych z udziałem regresji liniowej wielokrotnej odpornej do sygnałów czujników. W tabeli 11.3 podano listę czujników, takich że algorytm selekcji najczęściej wybierał cechy pochodzące z ich sygnałów jako elementy najlepszych wektorów cech. Wyniki dotyczą algorytmu współpracującego z regresją liniową. Informację tę zestawiono w odniesieniu do jednoelementowych i trójelementowych wektorów cech. Dla wektorów o większej liczbie elementów dominującą rolę odgrywały te same czujniki co dla wektorów trójelementowych.

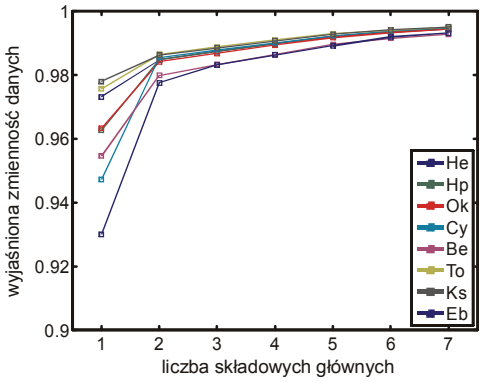
Tabela 11.3. Czujniki będące najczęściej źródłem elementów wektorów cech wyselekcjonowanych do obliczania stężenia pojedynczych substancji zanieczyszczających[a]

Lotny związek organiczny	Liczba elementów wektora cech	
	1	3
Heksan	4, 12	3
Heptan	9	3
Oktan	2, 8, 12	2, 8, 12
Cykloheksan	12	2, 3
Benzen	2, 12	4, 12
Toluen	9	2
Ksylen	5	5, 12
Etylobenzen	11	2, 11

[a]Algorytm selekcji opakowanej zastosowano do modelu regresji liniowej. Numery czujników według tabeli 8.3.

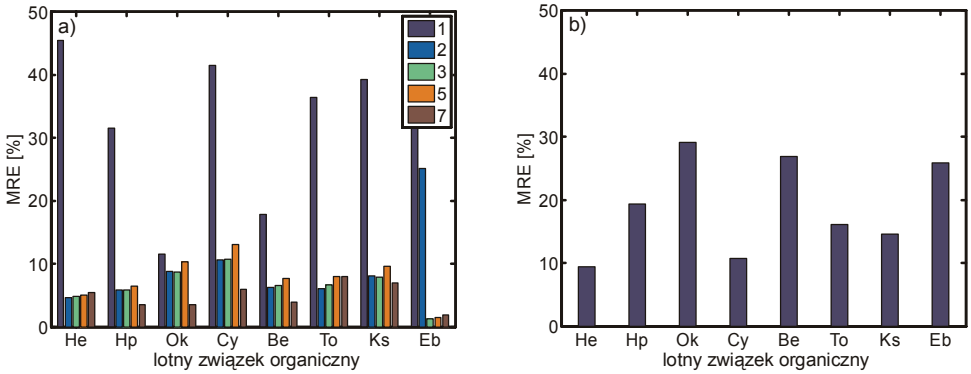
Regresja liniowa i nieliniowa oraz wektor cech wyekstrahowanych

Jako podstawę określania stężeń pojedynczych substancji zanieczyszczających zastosowano wektor cech wyekstrahowanych. Elementami wektora były składowe główne powstałe w wyniku przekształcenia wektora cech będącego rozwinięciem całej macierzy cech. Jako metodę ekstrakcji cech zastosowano analizę składowych głównych. Rozważono wektory cech zbudowane z 1, 2, 3, 5 lub 7 pierwszych składowych głównych. Podobnie jak dla wektorów cech wyselekcjonowanych do oceny reprezentacji gazów zastosowano metodę regresji liniowej oraz nieliniowej (tylko dla jednoelementowych wektorów cech).



Rys. 11.3. Łączny udział określonej liczby składowych głównych wyłonionych metodą PCA w wyjaśnianiu zmienności danych pomiarowych dotyczących poszczególnych zanieczyszczeń

Na rysunku 11.3 pokazano łączny udział określonej liczby składowych głównych w wyjaśnianiu zmienności danych wielowymiarowych dotyczących poszczególnych substancji zanieczyszczających.

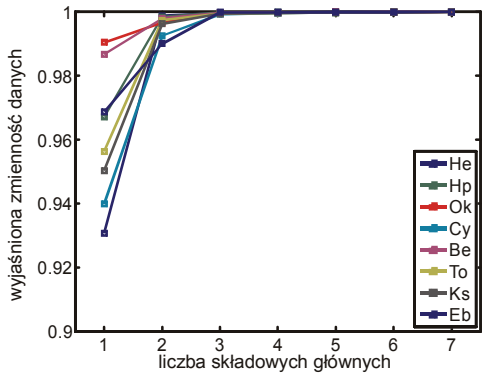


Rys. 11.4. Średni błąd względny określenia stężeń lotnych związków organicznych za pomocą: a) regresji liniowej 1, 2, 3, 5 i 7 składowych głównych, b) regresji nieliniowej pierwszej składowej głównej

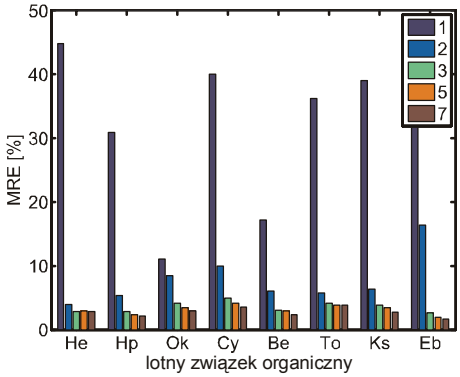
Na rysunku 11.4 przedstawiono średni błąd względny określenia stężeń poszczególnych LZO za pomocą modeli regresji liniowej składowych głównych z uwzględnieniem 1, 2, 3, 5 i 7 składowych (rys. 11.4a) oraz regresji nieliniowej pierwszej składowej głównej (rys. 11.4b).

Regresja metodą cząstkowych najmniejszych kwadratów

Do określania stężeń pojedynczych substancji zanieczyszczających zastosowano również regresję metodą cząstkowych najmniejszych kwadratów (PLS). Na rysunku 11.5 przedstawiono łączny udział określonej liczby składowych w opisie stężeń poszczególnych zanieczyszczeń.



Rys. 11.5. Łączny udział określonej liczby składowych wyłonionych metodą PLS w wyjaśnianiu zmienności stężeń poszczególnych substancji zanieczyszczających



Rys. 11.6. Średni błąd względny określenia stężeń lotnych związków organicznych za pomocą regresji metodą cząstkowych najmniejszych kwadratów z wykorzystaniem 1, 2, 3, 5 i 7 składowych

Na rysunku 11.6 pokazano średni błąd względny określenia stężeń substancji zanieczyszczających za pomocą modeli regresji PLS o różnej liczbie składowych.

Dyskusja

Na podstawie wykonanych analiz stwierdzono, że istnieje znaczna liczba cech typu *P* wyraźnie skorelowanych ze stężeniami określanych zanieczyszczeń. Udział cech, które spełniały założone kryterium filtracji jednowymiarowej (tabela 11.1) w zbiorze wszystkich rozważanych cech przekraczał na ogół 70%. Jednak przydatność tych cech do ilościowego określania zanieczyszczeń była zróżnicowana. W rzeczywistości znacznie mniejsza liczba jednoelementowych wektorów cech umożliwiała określenie stężeń zanieczyszczeń z błędem mniejszym niż 10%.

Jak pokazano na rysunku 11.1, typowanie cech za pomocą kryterium filtracji metodą regresji było zasadniczo zgodne z rozstrzygnięciami selekcji opakowanej, wykorzystującej ocenę liniową. Okazało się natomiast ryzykowną metodą preselekcji cech w wypadku zastosowania modeli nieliniowych do określania stężeń substancji zanieczyszczających.

Biorąc pod uwagę trzydzieści najlepszych jednoelementowych wektorów cech wyselekcjonowanych, należy stwierdzić, że występował duży rozrzut wartości błędu określania zanieczyszczenia na ich podstawie. Dla ośmiu rozważanych zanieczyszczeń pojedyncze cechy lepiej współpracowały z modelem nieliniowym, natomiast oktan i benzen, lepiej określano ilościowo za pomocą modeli liniowych.

Rezultaty uzyskane za pomocą jednowymiarowego podejścia liniowego znacznie poprawiono, posługując się zestawem cech wyłonionych w wyniku selekcji wielowymiarowej i stosując regresję wielokrotną odporną. Poprawa polegała na zmniejszeniu błędów określenia stężeń i zmniejszeniu ich rozrzutu. Jak wynika z tabeli 11.2, bardzo dobre rezultaty uzyskano już dla modeli z dwoma zmiennymi wejściowymi (MRE od 3% do 8%). Stwierdzono brak znacznej poprawy w wyniku dalszego zwiększenia liczby zmiennych w stosunku do wzrostu złożoności modelu. Na podstawie przeprowadzonych analiz uznano model liniowy z dwoma lub trzema zmiennymi objaśniającymi, którymi były wyselek-

cjonowane cechy typu *P*, za wystarczający do określania stężeń pojedynczych LZO na podstawie pomiarów czujnikowych.

Zauważono, że pewne czujniki były wskazywane częściej niż inne przez algorytm selekcji cech jako potrzebne do określania stężeń konkretnych substancji zanieczyszczających (tabela 11.3). Nie oznacza to, że były to te same czujniki w przypadku jedno- i wieloelementowych wektorów cech. Jak wynika z rysunku 11.2, cechy wchodzące w skład najlepszych wektorów wieloelementowych pochodziły na ogół z sygnałów różnych czujników. Dowodzi to, że poszczególne czujniki wnoszą istotną, komplementarną informację ilościową o badanych związkach. Posługiwanie się matrycą czujników zwiększa zestaw zanieczyszczeń, dla których istnieje możliwość ilościowego określania w porównaniu z możliwościami pojedynczego czujnika. Warto zwrócić uwagę, że informacja przydatna do określania stężenia konkretnego zanieczyszczenia była zawarta w danych zlokalizowanych w wielu różnych miejscach sygnału czujnikowego. Dla wyselekcjonowanych wektorów cech o większej liczbie elementów, tj. 5 i 7, stwierdzono istnienie wielu wektorów o więcej niż jednym elemencie pochodzącym z różnych części sygnału tego samego czujnika.

Konkurencyjną podstawą obliczania stężeń zanieczyszczeń w stosunku do wektora cech wyselekcjonowanych okazał się wektor cech wyekstrahowanych. Jak wynika z rysunku 11.3, już pierwsza składowa główna, wyłoniona w toku PCA na rozwiniętej macierzy cech, tłumaczyła 93–98% zmienności danych dotyczących rozważanych zanieczyszczeń. Mimo to określanie stężeń zanieczyszczeń z zastosowaniem regresji liniowej odpornej oraz regresji nieliniowej pierwszej składowej głównej nie dawało zadowalających rezultatów. Uzyskano natomiast bardzo dobre wyniki z zastosowaniem regresji liniowej dwóch składowych głównych (rys. 11.4a). Rozwiązanie to było wyraźnie lepsze niż regresja nieliniowa pierwszej składowej głównej (rys. 11.4b) Niestety jego skuteczność była mniejsza niż oznaczenia pojedynczych substancji zanieczyszczenia na podstawie wektora cech wyselekcjonowanych (tabela 11.2).

Inny możliwy wybór wiązał się z zastosowaniem metody cząstkowych najmniejszych kwadratów. Jak wynika z rys. 11.5, pierwsza składowa wyłoniona w ramach rozważanego podejścia tłumaczyła od 93% do ponad 99% zmienności stężeń w zależności od substancji zanieczyszczającej. Jednak model z dwoma składowymi dawał gorsze rezultaty niż regresja wielokrotna odporna, na wektorze dwóch cech wyselekcjonowanych (rys. 11.6, tabela 11.2). Począwszy od trzech elementów w wektorze cech, rezultaty obu metod były porównywalne.

11.1.2. Zanieczyszczenie jako dominująca substancja zanieczyszczająca

Ilościowe określanie substancji zanieczyszczających, gdy dominują one wśród innych pod względem ilościowym, rozważono na przykładzie heksanu i toluenu. Anali-

zowano przykłady występowania każdego z tych zanieczyszczeń w powietrzu wspólnie z inną substancją zanieczyszczającą: heksanu z heptanem (He1), oktanem (He2), cykloheksanem (He3), benzenem (He4) lub toluenem (He5) oraz toluenu z heksanem (To1), heptanem (To2), benzenem (To3), ksylenem (To4) lub etylobenzenem (To5). Udział związków dominujących w badanych mieszaninach zanieczyszczeń wynosił od 60% do 95% (tabela 8.2).

Regresja liniowa i nieliniowa oraz wektor cech wyselekcjonowanych

Obliczano stężenia dominujących substancji zanieczyszczających na podstawie ich reprezentacji w postaci wektorów cech wyselekcjonowanych. Rozważono wektory składające się z 1, 2, 3, 5 lub 7 cech. Zastosowano metodę regresji liniowej odpornej oraz nieliniowej (tylko dla jednoelementowych wektorów cech). Metody te zastosowano do oceny wektorów cech w ramach selekcji w systemie opakowanym. Przestrzeń cech przeszukiwano metodą symulowanego wyżarzania. Dla porównania przeprowadzono filtrację cech metodą analizy korelacji.

Tabela 11.4. Liczba cech umożliwiających określenie stężenia zanieczyszczenia dominującego ze średnim błędem względnym MRE ≤ 10% oraz liczba cech istotnych wskazana przez kryterium filtracji[a]

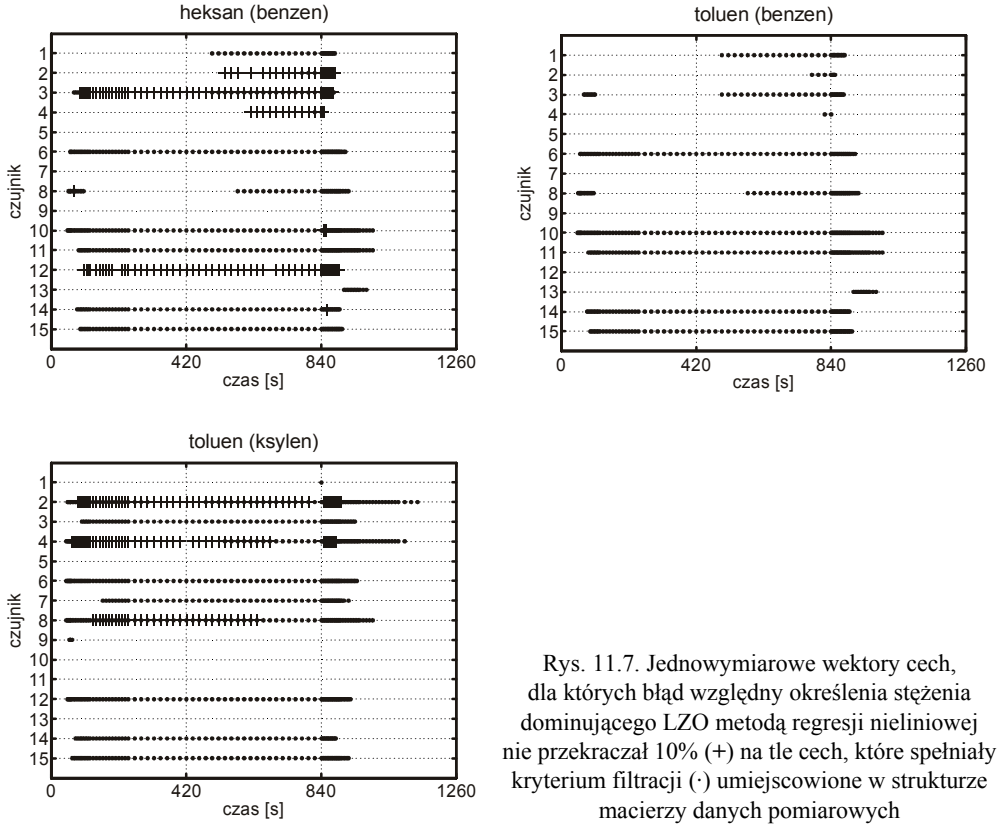
Lotny związek organiczny dominujący (zdominowany)	Selekcja cech		Filtracja cech
	Regresja liniowa	Regresja nieliniowa	
Heksan (heptan)	7	171	556
Heksan (oktan)	0	0	0
Heksan (cykloheksan)	47	180	335
Heksan (benzen)	0	166	545
Heksan (toluen)	0	3	46
Toluen (heksan)	6	263	396
Toluen (heptan)	3	238	278
Toluen (benzen)	18	241	572
Toluen (ksylen)	0	156	727
Toluen (etylobenzen)	8	207	641

[a]Współczynnik korelacji $R \geq 0{,}95$. Ogólna liczba cech – 1830.

W tabeli 11.4 podano liczbę jednoelementowych wektorów cech umożliwiających obliczenie stężeń zanieczyszczeń dominujących z błędem mniejszym niż 10%, w porównaniu z liczbą cech uznanych za istotne według kryterium filtracji ($R \geq 0{,}95$).

Na rysunku 11.7 pokazano położenie cech spełniających kryterium filtracji oraz kryterium jednowymiarowej, nieliniowej selekcji opakowanej w strukturze macierzy

danych pomiarowych. Do porównania z innymi analizami przedstawionymi wybrano konfigurację dominujący heksan z mniejszą ilością benzenu oraz według tego schematu toluen z benzenem i toluen z ksylenem.



Rys. 11.7. Jednowymiarowe wektory cech, dla których błąd względny określenia stężenia dominującego LZO metodą regresji nieliniowej nie przekraczał 10% (+) na tle cech, które spełniały kryterium filtracji (·) umiejscowione w strukturze macierzy danych pomiarowych

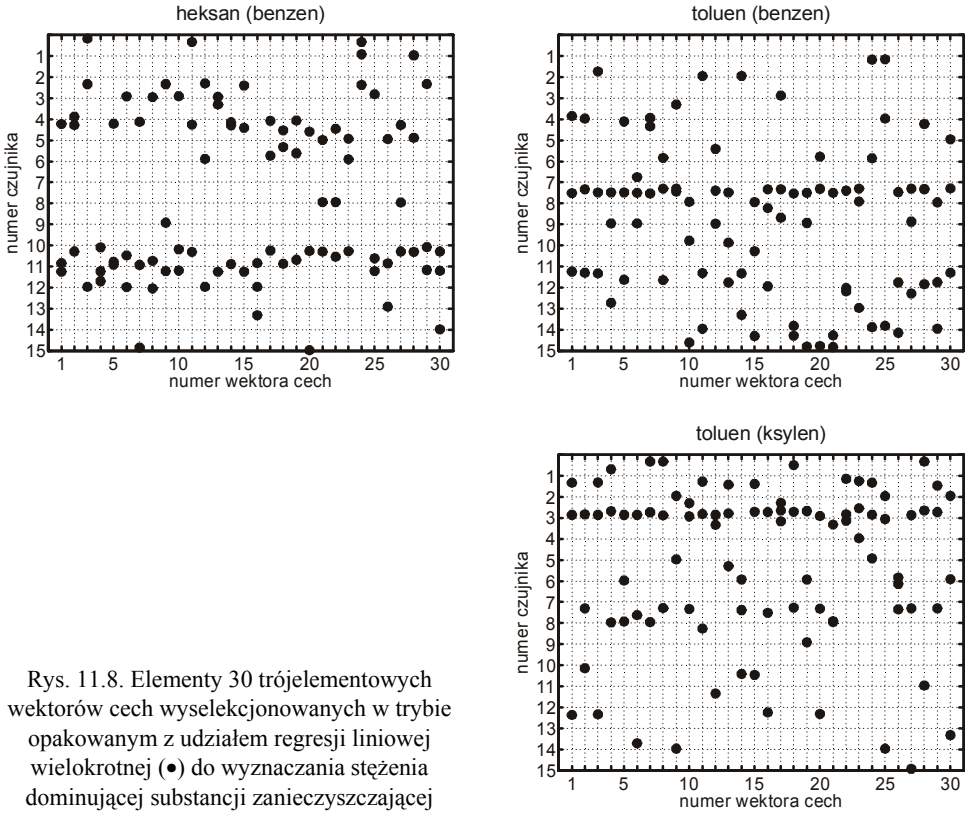
W tabeli 11.5 zestawiono wartości średniego błędu względnego określenia stężeń zanieczyszczeń na podstawie trzydziestu najlepszych wektorów cech wyłonionych w toku selekcji. Wyniki podane dla wektorów wielowymiarowych uzyskano z zastosowaniem regresji liniowej wielokrotnej odpornej. W odniesieniu do wektorów jednoelementowych przedstawiono rezultaty uzyskane metodą regresji jednowymiarowej odpornej, liniowej i nieliniowej.

Na rysunku 11.8 pokazano przynależność elementów trzydziestu wektorów cech trójelementowych (patrz, tabela 11.5) do sygnałów czujników. Taka liczba cech wektora zapewniała uzyskanie określenia zanieczyszczenia dominującego z błędem nieprzekraczającym 10%.

Tabela 11.5. Średni błąd względny (MRE) określania stężenia zanieczyszczenia dominującego[a]

Lotny związek organiczny dominujący (zdominowany)	Regresja liniowa					Regresja nieliniowa
	Liczba elementów wektora cech					
	1	2	3	5	7	1
Heksan (heptan)	9–11	8–12	6–8	6–7	6–7	8–9
Heksan (oktan)	13–15	11–17	8–10	7–10	5–8	14–15
Heksan (cykloheksan)	8–9	6–10	5–7	5–6	5–6	8–8
Heksan (benzen)	11–13	10–13	8–10	8–9	7–8	7–8
Heksan (toluen)	14–17	9–19	6–10	5–8	5–6	9–12
Toluen (heksan)	9–11	9–12	7–10	5–8	6–7	5–5
Toluen (heptan)	9–12	7–11	7–8	6–7	5–7	5–6
Toluen (benzen)	9–10	7–11	7–9	7–8	6–7	5–6
Toluen (ksylen)	10–12	8–12	8–10	8–10	7–9	5–6
Toluen (etylobenzen)	9–11	8–12	7–10	7–8	6–8	3–5

[a]Podano wartości min(MRE)–max(MRE) [%] dla 30 najlepszych wektorów cech o następującej liczbie elementów: 1, 2, 3, 5 i 7, uzyskanych w wyniku selekcji.



Rys. 11.8. Elementy 30 trójelementowych wektorów cech wyselekcjonowanych w trybie opakowanym z udziałem regresji liniowej wielokrotnej (•) do wyznaczania stężenia dominującej substancji zanieczyszczającej

Tabela 11.6. Czujniki będące najczęściej źródłem elementów wektorów cech wyselekcjonowanych do obliczania stężenia zanieczyszczenia dominującego

Lotny związek organiczny dominujący (zdominowany)	Liczba elementów wektora cech	
	1	3
Heksan (heptan)	2	11, 3
Heksan (oktan)	–	–
Heksan (cykloheksan)	2, 8	10, 3
Heksan (benzen)	3, 12	11
Heksan (toluen)	–	9
Toluen (heksan)	2, 4, 8, 11	2
Toluen (heptan)	2, 4, 8	8
Toluen (benzen)	2, 4, 8	8
Toluen (ksylen)	2, 4, 8	3
Toluen (etylobenzen)	2, 4, 8	8

Algorytm selekcji opakowanej zastosowano do modelu regresji liniowej. Numery czujników zgodne z tabelą 8.3.

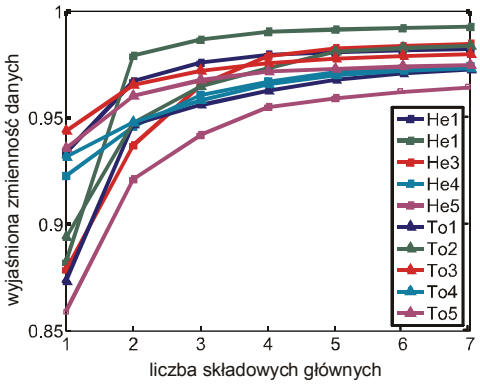
W tabeli 11.6 podano listę czujników takich, że algorytm selekcji najczęściej wybierał cechy pochodzące z ich sygnałów jako elementy najlepszych wektorów cech. Wyniki dotyczące algorytmu współpracującego z regresją liniową zestawiono dla jednoelementowych i trójelementowych wektorów cech.

Regresja liniowa i nieliniowa oraz wektor cech wyekstrahowanych

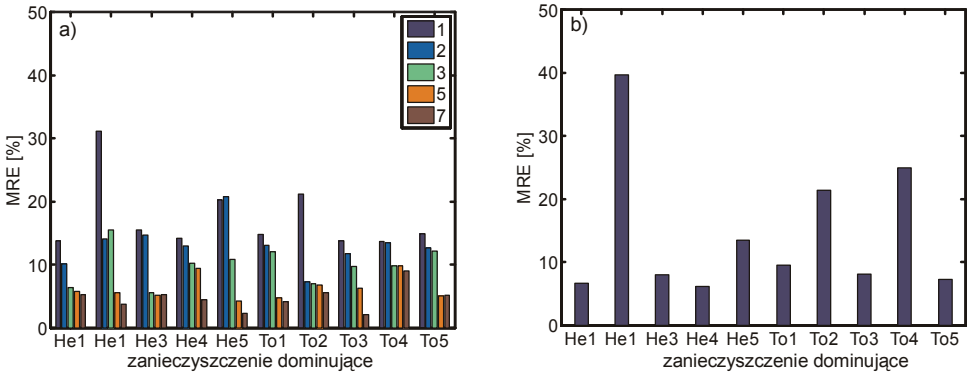
Określanie stężeń dominujących substancji zanieczyszczających przeprowadzono również na podstawie wektora cech wyekstrahowanych. Były to składowe główne, powstałe w wyniku przekształcenia wektora cech stanowiącego rozwinięcie całej macierzy cech. Jako metodę ekstrakcji cech zastosowano analizę składowych głównych. Rozważono wektory cech zbudowane z 1, 2, 3, 5 lub 7 pierwszych składowych głównych. Podobnie jak dla wektorów cech wyselekcjonowanych do obliczania stężeń zastosowano metodę regresji liniowej oraz nieliniowej (tylko dla jednoelementowych wektorów cech) odpornej.

Łączny udział składowych głównych w wyjaśnianiu zmienności danych wielowymiarowych dotyczących heksanu oraz toluenu występujących w powietrzu jako zanieczyszczenia dominujące w mieszaninie z różnymi substancjami zdominowanymi pokazano na rys. 11.9.

Na rysunku 11.10 przedstawiono wyniki określenia stężeń tych zanieczyszczeń z zastosowaniem regresji liniowej składowych głównych z uwzględnieniem 1, 2, 3, 5 i 7 składowych oraz za pomocą regresji nieliniowej pierwszej składowej głównej.



Rys. 11.9. Łączny udział składowych głównych w wyjaśnianiu zmienności danych wielowymiarowych dotyczących heksanu oraz toluenu jako zanieczyszczeń dominujących

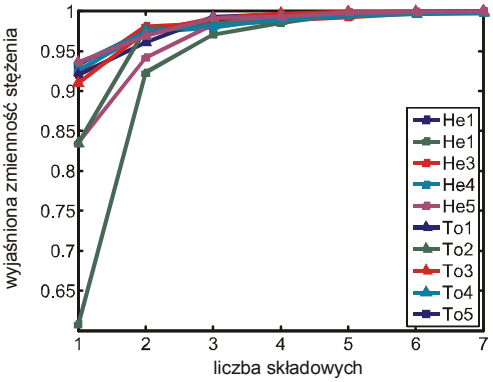


Rys. 11.10. Średni błąd względny (MRE) określenia stężeń dominujących LZO za pomocą modelu: a) regresji liniowej składowych głównych z następującą liczbą zmiennych wejściowych: 1, 2, 3, 5 i 7, b) regresji nieliniowej z jedną składową główną jako zmienną objaśniającą

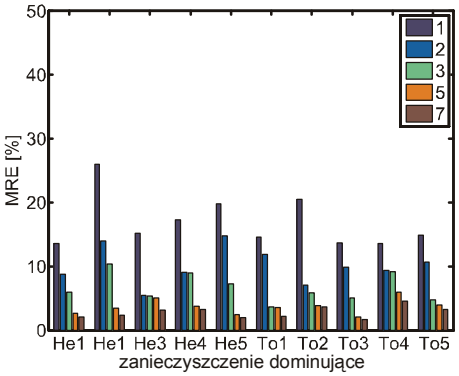
Regresja metodą cząstkowych najmniejszych kwadratów

Do określania stężeń dominujących substancji zanieczyszczających zastosowano też regresję metodą cząstkowych najmniejszych kwadratów. Na rysunku 11.11 przedstawiono łączny udział określonej liczby składowych w opisie stężeń heksanu oraz toluenu występujących jako zanieczyszczenia dominujące w mieszaninie z różnymi substancjami zdominowanymi.

Na rysunku 11.12 pokazano rezultaty określenia stężeń zanieczyszczeń dominujących z zastosowaniem regresji metodą cząstkowych najmniejszych kwadratów dla modeli o różnej liczbie składowych.



Rys. 11.11. Łączny udział określonej liczby składowych w wyjaśnianiu zmienności stężeń heksanu i toluenu jako zanieczyszczeń dominujących



Rys. 11.12. Średni błąd względny (MRE) określenia stężeń dominujących LZO za pomocą modelu regresji cząstkowych najmniejszych kwadratów z 1, 2 3, 5 i 7 zmiennymi objaśniającymi

Dyskusja

Na podstawie przeprowadzonej analizy danych czujnikowych pokazano, że liczba cech skorelowanych ze stężeniem zanieczyszczenia jest znacznie większa, gdy występuje ono w powietrzu pojedynczo, nie z innymi zanieczyszczeniami, nawet jeśli zanieczyszczenie to dominuje pod względem ilościowym (por. tabela 11.1 i tabela 11.4). W grupie wszystkich rozważanych par zanieczyszczeń zgodność metody selekcji i filtracji w typowaniu cech istotnych była umiarkowana, choć dla przykładów podanych na rys. 11.8 cechy wyselekcjonowane były podzbiorem zbioru cech odfiltrowanych.

Stwierdzono, że w podejściu jednowymiarowym określenie stężenia zanieczyszczenia dominującego z błędem mniejszym niż 10% wymagało zastosowania regresji nieliniowej (tabela 11.5). Znaczna liczba pojedynczych cech umożliwiała uzyskanie takiego rezultatu praktycznie bez względu na drugi składnik mieszaniny (tabela 11.4) z wyjątkiem oktanu oraz toluenu, których obecność w powietrzu pogarszała wynik oznaczenia heksanu.

W ramach podejścia wielowymiarowego z zastosowaniem wektorów cech wyselekcjonowanych dobre wyniki oceny stężenia zanieczyszczenia dominującego uzyskano za pomocą wielokrotnej regresji liniowej (tabela 11.5). Uznano, że zadowalająca liczba zmiennych objaśniających wynosiła trzy. Niemiej jednak uzyskanie rezultatów porównywalnych jak dla pojedynczych substancji wymagało posłużenia się wektorami cech o pięciu, a nawet siedmiu elementach. Z użyciem wektorów cech wieloelementowych wyniki określania stężenia zanieczyszczenia dominującego metodą liniową były porównywalne jak dla jednowymiarowej regresji nieliniowej, co

wskazuje na znaczenie metod nieliniowych w zastosowaniu do ilościowego oznaczania zanieczyszczeń dominujących w mieszaninach.

Jak wynika z tabeli 11.6, czujniki najczęściej wybierane jako źródła elementów najlepszych wektorów cech na ogół różniły się dla wektorów jedno- i wieloelementowych. Prawidłowość tę obserwowano również dla pojedynczych substancji zanieczyszczających (tabela 11.3). Czujniki wybierane najczęściej do obliczania stężeń substancji dominującej zależały od rodzaju substancji zdominowanej.

W ramach podejścia wielowymiarowego z wykorzystaniem cech wyekstrahowanych stężenie zanieczyszczenia dominującego obliczano z dobrym rezultatem za pomocą regresji liniowej składowych głównych oraz regresji PLS. W obu wypadkach zastosowanie pięciu składowych głównych w modelu pozwalało uzyskać błąd mniejszy niż 5%.

Poza przypadkiem zastosowania wyłącznie pierwszej składowej obserwowano przewagę regresji metodą cząstkowych najmniejszych kwadratów nad regresją składowych głównych. Rezultaty uzyskiwane metodą PCR były z kolei porównywalne z wynikami regresji wielokrotnej cech wyselekcjonowanych dla tej samej liczby zmiennych objaśniających.

11.1.3. Zanieczyszczenie jako zdominowana substancja zanieczyszczająca

Określanie stężeń substancji zanieczyszczających, które pod względem ilościowym zostały zdominowane przez inne zanieczyszczenia, rozważono na przykładzie heksanu, heptanu, oktanu, cykloheksanu, benzenu, toluenu, ksylenu i etylobenzenu. Analizowano przypadki, gdy związki te występowały w powietrzu wraz z heksanem lub toluenem. Udział związków zdominowanych w badanych mieszaninach zanieczyszczeń wynosił od 5% do 40% (tabela 8.2). Dla mieszanin substancji zdominowanych przez heksan przyjęto oznaczenia Hp1, Ok1, Cy1, Be1, To1, przez toluen – He2, Hp2, Be2, Ks2, Eb2.

Regresja liniowa i nieliniowa oraz wektor cech wyselekcjonowanych

Obliczano stężenia zdominowanych substancji zanieczyszczających na podstawie ich reprezentacji w postaci wektorów cech wyselekcjonowanych. Rozważono wektory składające się z 1, 2, 3, 5 lub 7 cech. Selekcję cech przeprowadzono w trybie opakowanym. Dla porównania wykonano filtrację cech metodą analizy korelacji. Do obliczania stężeń zastosowano metodę regresji liniowej oraz nieliniowej (tylko dla jednoelementowych wektorów cech) odpornej. Metody te służyły zarazem do oceny

wektorów cech w ramach selekcji w systemie opakowanym. Podczas selekcji wielowymiarowej przestrzeń cech przeszukiwano metodą symulowanego wyżarzania.

W wyniku przeprowadzonych analiz stwierdzono brak jednoelementowych wektorów cech umożliwiających obliczenie stężeń zanieczyszczeń zdominowanych z błędem mniejszym niż 10%. Podobnie liczba cech uznanych za istotne według kryterium filtracji (współczynnik korelacji $R \geq 0{,}95$) wynosiła zero.

Tabela 11.7. Średni błąd względny (MRE) określania stężenia zanieczyszczenia zdominowanego[a]

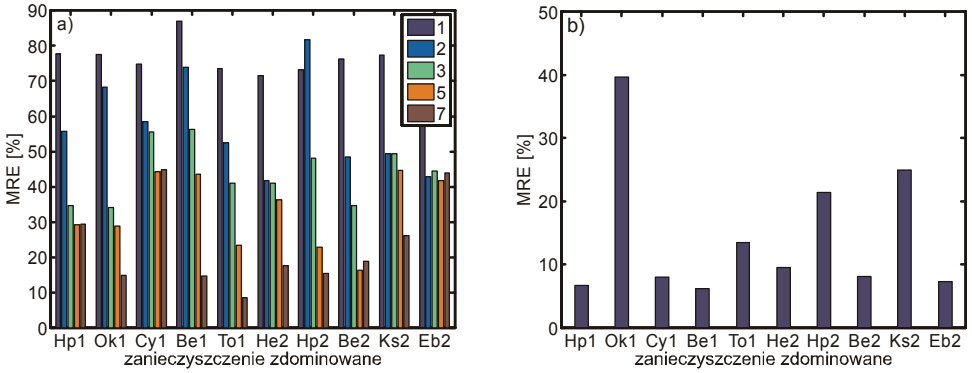
Lotny związek organiczny zdominowany (dominujący)	Regresja liniowa					Regresja nieliniowa
	Liczba elementów wektora cech					
	1	2	3	5	7	1
Heptan (heksan)	64–70	37–56	29–36	27–33	25–29	58–65
Oktan (heksan)	44–49	29–46	27–33	24–29	22–25	38–44
Cykloheksan (heksan)	53–69	36–49	34–42	35–39	32–37	60–66
Benzen (heksan)	52–67	41–58	36–45	34–39	30–36	50–56
Toluen (heksan)	28–51	32–54	21–33	18–29	18–25	13–28
Heksan (toluen)	31–44	26–38	26–34	25–31	23–29	17–23
Heptan (toluen)	53–57	34–51	27–36	28–33	22–31	31–36
Benzen (toluen)	52–58	38–52	30–40	29–35	27–32	44–47
Ksylen (toluen)	45–52	35–52	34–40	34–39	32–38	21–37
Etylobenzen (toluen)	44–55	40–51	35–43	37–41	33–40	25–40

[a]Podano wartości min(MRE)–max(MRE) [%] dla 30 najlepszych wektorów cech o następującej liczbie elementów: 1, 2, 3, 5 i 7 uzyskanych w wyniku selekcji.

W tabeli 11.7 porównano rezultaty określenia stężeń zanieczyszczeń zdominowanych na podstawie wektorów cech o liczbie elementów 1, 2, 3, 5 i 7, wyłonionych podczas selekcji opakowanej.

Regresja liniowa i wektor cech wyekstrahowanych

Jako podstawę określania stężeń zdominowanych substancji zanieczyszczających zastosowano również wektor cech wyekstrahowanych. Były nimi składowe główne powstałe w wyniku przekształcenia wektora cech stanowiącego rozwinięcie całej macierzy cech. Jako metodę ekstrakcji cech zastosowano analizę składowych głównych (PCA). Rozważono wektory cech zbudowane z 1, 2, 3, 5 lub 7 pierwszych składowych głównych. Podobnie jak dla wektorów cech wyselekcjonowanych stężenia zanieczyszczeń obliczano metodami regresji liniowej oraz nieliniowej (tylko dla jednoelementowych wektorów cech) odpornej. Łączny udział składowych głównych w opisie danych wielowymiarowych dotyczących substancji występujących w powietrzu jako zanieczyszczenia zdominowane przez heksan lub toluen pokazano na rys. 11.9.

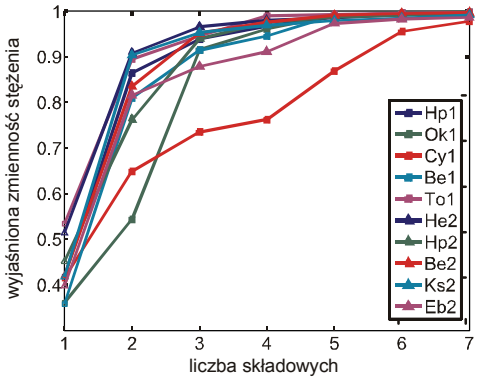


Rys. 11.13. Średni błąd względny (MRE) określenia stężeń zdominowanych lotnych związków organicznych za pomocą: a) modelu regresji liniowej składowych głównych z następującą liczbą zmiennych objaśniających: 1, 2, 3, 5 i 7, b) regresji nieliniowej z jedną składową główną jako zmienną objaśniającą

Na rysunku 11.13 pokazano wyniki określenia stężeń zdominowanych LZO za pomocą regresji liniowej składowych głównych z wykorzystaniem 1, 2, 3, 5 i 7 składowych oraz regresji nieliniowej z pierwszą składową główną jako zmienną objaśniającą.

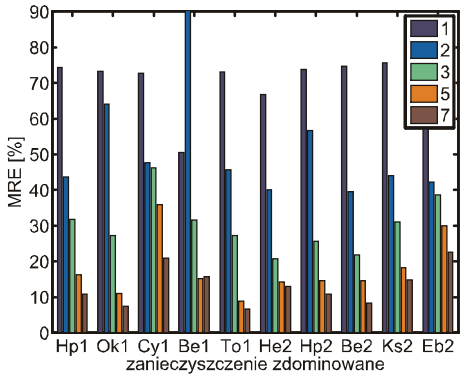
Regresja metodą cząstkowych najmniejszych kwadratów

Rozważono zastosowanie regresji metodą cząstkowych najmniejszych kwadratów do określania stężeń zdominowanych substancji zanieczyszczających.



Rys. 11.14. Łączny udział określonej liczby składowych w wyjaśnianiu zmienności stężeń zanieczyszczeń zdominowanych

Na rysunku 11.14 przedstawiono łączny udział określonej liczby składowych w wyjaśnianiu zmienności stężeń heptanu, oktanu, cykloheksanu, benzenu, toluenu, ksylenu, etylobenzenu, heksanu oraz toluenu występujących w powietrzu jako zanieczyszczenia zdominowane przez heksan lub toluen.



Rys. 11.15. Średni błąd względny (MRE) określenia stężeń zdominowanych LZO z zastosowaniem modelu regresji cząstkowych najmniejszych kwadratów o liczbie zmiennych objaśniających 1, 2, 3, 5 i 7

Na rysunku 11.15 pokazano rezultaty określania stężeń zdominowanych LZO z zastosowaniem regresji metodą cząstkowych najmniejszych kwadratów. Porównano modele o różnej liczbie zmiennych objaśniających.

Dyskusja

Jak wynika z tabeli 11.7, oznaczanie ilościowe zanieczyszczania zdominowanego na podstawie danych czujnikowych jest problemem trudnym. Jednowymiarowe podejście liniowe okazało się w tym przypadku zupełnie nieskuteczne (MRE od 30% do 70%). Dzięki zastosowaniu wieloelementowych wektorów cech wyselekcjonowanych i regresji wielokrotnej uzyskano błędy na poziomie 30–40% dla wszystkich rozważanych substancji. Wynik ten był jednak daleki od satysfakcjonującego. Metodą jednowymiarowej regresji nieliniowej uzyskano nieco lepsze wyniki, ale tylko dla niektórych zanieczyszczeń.

Zaskakująco dobre rezultaty określania stężeń zanieczyszczeń zdominowanych osiągnięto z zastosowaniem modelu nieliniowego opartego na pierwszej składowej głównej wyłonionej w toku analizy PCA (rys. 11.13b). W sześciu na dziesięć rozważanych przypadków uzyskano średni błąd względny mniejszy niż 10%. Wyników takich nie udało się osiągnąć za pomocą liniowej regresji wielokrotnej składowych głównych nawet dla siedmiu składowych (rys. 11.13a). W porównaniu z regresją wielokrotną składowych głównych lepsze wyniki uzyskiwano za pomocą regresji metodą PLS dla tej samej liczby zmiennych objaśniających (por. rys. 11.13a i 11.15). Metoda ta była jednak mniej korzystna od regresji nieliniowej pierwszej składowej głównej. Otrzymane rezultaty wskazywały na istotną rolę metod nieliniowych w oznaczaniu zanieczyszczeń zdominowanych w mieszaninach.

Niezależnie od zastosowanej metody obliczeń błędy wyznaczania stężeń substancji zdominowanych było bardzo zróżnicowane. Zależały one nie tylko od rodzaju substancji zdominowanej, lecz również od rodzaju substancji dominującej.

11.2. Stężenie atomów węgla pochodzących od LZO

Miary ilościowe proponowane jako alternatywne do stężenia powinny mieć tę zaletę, że można je stosować nawet wtedy, gdy tożsamość chemiczna substancji zanieczyszczających powietrze nie jest znana. Posługiwanie się stężeniem substancji zanieczyszczającej wymaga znajomości jej tożsamości chemicznej, a koszt pozyskania tej informacji może być niewspółmierny z efektem finalnym oceny stanu zanieczyszczenia powietrza.

Jako alternatywne do stężenia substancji analizowano sumę stężeń składników mieszaniny [ppm] i stężenie atomów węgla [mmol C/m^3]. Możliwości ich wyznaczania oceniano dla:

- mieszanin substancji zanieczyszczających o znanym składzie jakościowym,
- kategorii substancji zanieczyszczających,
- kategorii mieszanin substancji zanieczyszczających.

Miary te na ogół są różne. Jednak dla rozważanych zanieczyszczeń wyniki, które przedstawiono w pracy na przykładzie stężenia atomów węgla, były pod każdym względem niemal identyczne.

11.2.1. Mieszanina substancji zanieczyszczających o znanym składzie jakościowym

Określanie stężenia atomów węgla jako alternatywę do wyznaczania stężeń składników mieszaniny o znanym składzie przeanalizowano dla dziesięciu dwuskładnikowych mieszanin substancji zanieczyszczających powietrze, zawierających substancję dominującą i zdominowaną. Były to: heksan z heptanem, oktanem, cykloheksanem, benzenem bądź toluenem, toluen z heksanem, heptanem, benzenem, ksylenem oraz z etylobenzenem.

Regresja liniowa i nieliniowa oraz wektor cech wyselekcjonowanych

Zastosowano reprezentacje rozważanych mieszanin w postaci wektorów cech wyselekcjonowanych. Rozważono wektory składające się z 1, 2, 3, 5 lub 7 cech. W celu obliczania stężenia atomów węgla posłużono się metodą regresji liniowej oraz nieliniowej (tylko dla jednoelementowych wektorów cech) odpornej. Metody te zastosowano do oceny wektorów cech w ramach selekcji w systemie opakowanym. Na potrzeby selekcji wielowymiarowej przestrzeń cech przeszukiwano metodą symulowanego wyżarzania.

Dla porównania przeprowadzono filtrację cech metodą analizy korelacji w układzie zupełnego przeglądu przestrzeni cech.

Tabela 11.8. Liczba pojedynczych cech umożliwiających określenie stężeń atomów węgla w mieszaninie substancji zanieczyszczających o znanym składzie jakościowym[a] oraz liczba cech istotnych wskazana przez kryterium filtracji[b]

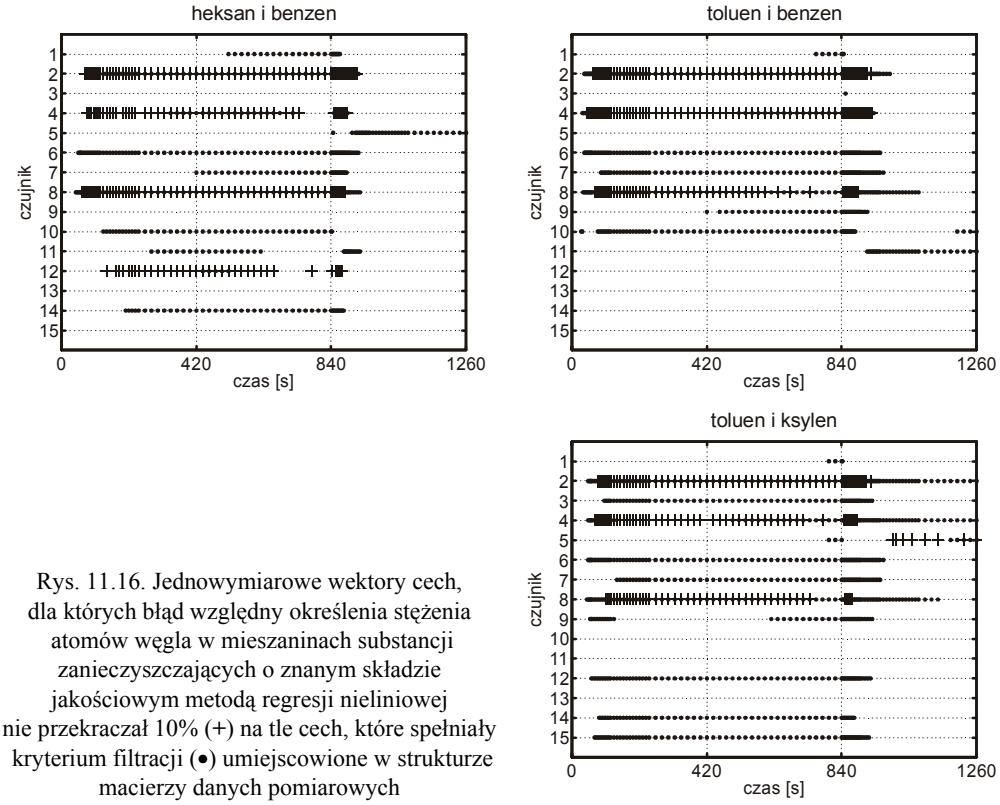
Skład jakościowy mieszaniny substancji zanieczyszczających	Selekcja cech		Filtracja cech
	Regresja liniowa	Regresja nieliniowa	
Heksan i heptan	47	436	742
Heksan i oktan	0	3	42
Heksan i cykloheksan	66	392	626
Heksan i benzen	0	235	494
Heksan i toluen	0	131	99
Toluen i heksan	10	185	448
Toluen i heptan	6	243	422
Toluen i benzen	93	239	631
Toluen i ksylen	53	197	848
Toluen i etylobenzen	39	194	775

[a]Ze średnim błędem względnym MRE ≤ 10%.
[b]Współczynnik korelacji $R \geq 0{,}95$.

W tabeli 11.8 podano liczbę jednoelementowych wektorów cech umożliwiających obliczenie stężeń atomów węgla pochodzących od mieszaniny substancji zanieczyszczających o znanym składzie jakościowym z błędem mniejszym niż 10% oraz liczbę pojedynczych cech uznanych za istotne według kryterium filtracji (współczynnik korelacji $R \geq 0{,}95$). Na rysunku 11.16 pokazano położenie cech według tabeli 11.8 w strukturze macierzy danych pomiarowych dla przykładowych mieszanin m4, m8 i m9.

Rezultaty określania stężenia atomów węgla pochodzących od mieszaniny substancji zanieczyszczających o znanym składzie jakościowym na podstawie wektorów cech o 1, 2, 3, 5 i 7 elementach wyłonionych w wyniku selekcji cech zestawiono w tabeli 11.9. Na rysunku 11.17 pokazano przynależność elementów trzydziestu wyselekcjonowanych trójelementowych wektorów cech do sygnałów czujników. Wektory te były najbardziej przydatne do obliczania stężenia atomów węgla w mieszaninach LZO o znanym składzie jakościowym. Błędy oznaczeń podano w tabeli 11.9.

W tabeli 11.10 przedstawiono listę czujników, takich że algorytm selekcji najczęściej wybierał cechy pochodzące z ich sygnałów jako elementy najlepszych wektorów cech. Wyniki selekcji we współpracy z regresją liniową zestawiono w odniesieniu do jednoelementowych i trójelementowych wektorów cech.



Rys. 11.16. Jednowymiarowe wektory cech, dla których błąd względny określenia stężenia atomów węgla w mieszaninach substancji zanieczyszczających o znanym składzie jakościowym metodą regresji nieliniowej nie przekraczał 10% (+) na tle cech, które spełniały kryterium filtracji (•) umiejscowione w strukturze macierzy danych pomiarowych

Tabela 11.9. Średni błąd względny (MRE) określania stężenia atomów węgla pochodzących od mieszaniny substancji zanieczyszczających o znanym składzie jakościowym[a]

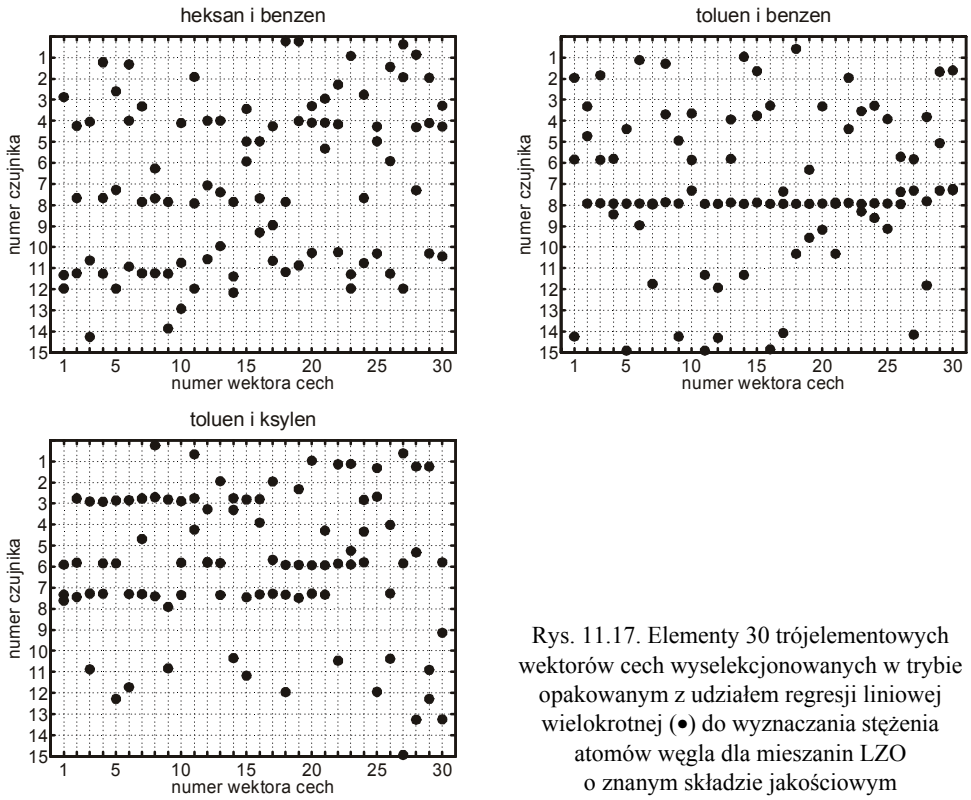
Skład jakościowy mieszaniny	Regresja liniowa					Regresja nieliniowa
	Liczba elementów wektora cech					
	1	2	3	5	7	1
Heksan i heptan	5–9	5–9	4–5	4–5	3–4	3–4
Heksan i oktan	10–12	6–13	4–7	3–6	3–5	10–12
Heksan i cykloheksan	4–5	3–5	3–4	2–3	2–3	3–4
Heksan i benzen	11–12	7–12	7–9	6–7	4–6	6–7
Heksan i toluen	12–16	7–16	6–8	4–6	4–6	6–8
Toluen i heksan	9–10	7–11	6–8	5–7	5–6	5–7
Toluen i heptan	9–11	7–10	6–8	5–7	4–6	4–5
Toluen i benzen	7–9	7–9	6–7	5–7	4–6	4–5
Toluen i ksylen	8–9	8–10	6–7	5–7	5–6	5–6
Toluen i etylobenzen	7–10	7–10	5–8	5–6	4–6	5–7

[a]Podano wartości min(MRE)–max(MRE) [%] dla 30 najlepszych wektorów cech o liczbie elementów: 1, 2, 3, 5 i 7 uzyskanych w wyniku selekcji.

Tabela 11.10. Czujniki będące najczęściej źródłem elementów wektorów cech wyselekcjonowanych do obliczania stężenia atomów węgla dla mieszanin LZO o znanym składzie jakościowym[a]

Dwa lotne związki organiczne	Liczba elementów wektora cech	
	1	3
Heksan i heptan	2, 3, 4, 8, 12	3
Heksan i oktan	–	9,2
Heksan i cykloheksan	2, 4, 8, 14, 15	3
Heksan i benzen	2, 4, 8, 12	–
Heksan i toluen	4, 2	9
Toluen i heksan	2, 4	9, 8
Toluen i heptan	2, 4, 8	8, 9
Toluen i benzen	2, 4, 8	8
Toluen i ksylen	2, 4, 8	6, 3, 8
Toluen i etylobenzen	2, 4, 8	8

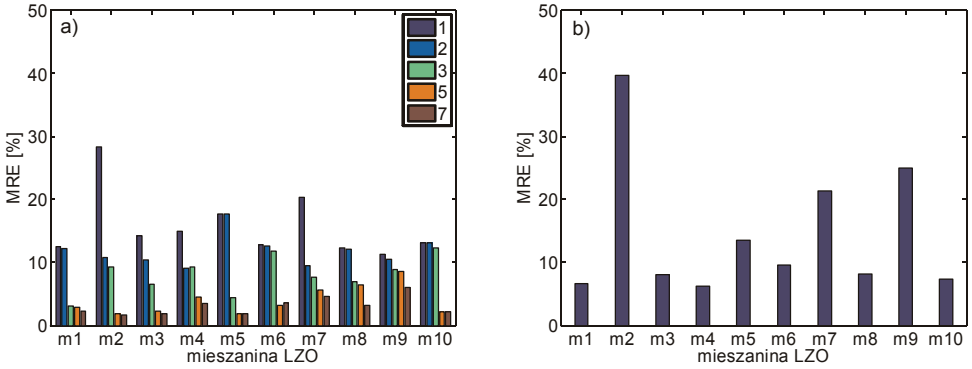
[a]Algorytm selekcji opakowanej zastosowano do modelu regresji liniowej. Numery czujników zgodne z tabelą 8.3.



Rys. 11.17. Elementy 30 trójelementowych wektorów cech wyselekcjonowanych w trybie opakowanym z udziałem regresji liniowej wielokrotnej (•) do wyznaczania stężenia atomów węgla dla mieszanin LZO o znanym składzie jakościowym

‚Regresja liniowa i nieliniowa oraz wektor cech wyekstrahowanych

Określano stężenie atomów węgla pochodzących od mieszanin substancji zanieczyszczających o znanym składzie jakościowym na podstawie wektora cech wyekstrahowanych. Jego elementy stanowiły składowe główne utworzone w wyniku przekształcenia wektora cech będącego rozwinięciem całej macierzy cech. Rozważono wektory cech zbudowane z 1, 2, 3, 5 lub 7 pierwszych składowych głównych. Podobnie jak dla wektorów cech wyselekcjonowanych zastosowano metodę regresji liniowej oraz nieliniowej jednowymiarowej. Łączny udział składowych głównych w wyjaśnianiu zmienności danych wielowymiarowych dotyczących rozważanych par substancji organicznych przedstawiono na rys. 11.9.

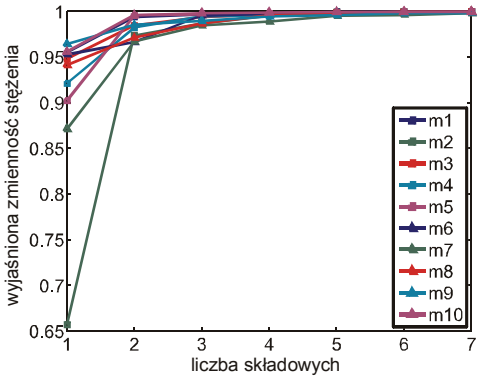


Rys. 11.18. Średni błąd względny (MRE) określenia stężenia atomów węgla w mieszaninach lotnych związków organicznych o znanym składzie jakościowym z zastosowaniem: a) regresji liniowej 1, 2, 3, 5 i 7 składowych głównych, b) regresji nieliniowej z pierwszą składową główną jako zmienną objaśniającą

Na rysunku 11.18 pokazano średni błąd względny określenia stężenia atomów węgla za pomocą regresji liniowej składowych głównych z wykorzystaniem 1, 2, 3, 5 i 7 składowych oraz regresji nieliniowej z pierwszą składową główną jako zmienną objaśniającą.

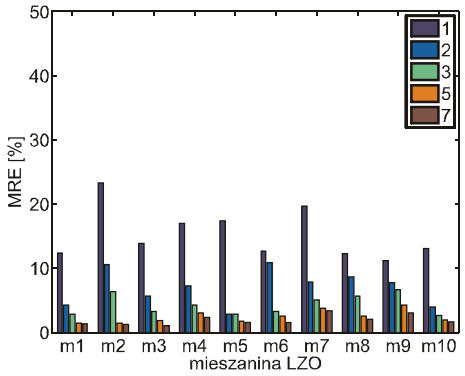
Regresja metodą cząstkowych najmniejszych kwadratów

Do określania stężeń atomów węgla w mieszaninach substancji zanieczyszczających o znanym składzie jakościowym zastosowano regresję metodą cząstkowych najmniejszych kwadratów.



Rys. 11.19. Łączny udział składowych w wyjaśnianiu zmienności stężenia atomów węgla dla mieszanin zanieczyszczeń o znanym składzie jakościowym

Na rysunku 11.19 przedstawiono łączny udział różnej liczby składowych w opisie stężenia atomów węgla dla poszczególnych mieszanin, oszacowane błędy oznaczeń natomiast pokazano na rysunku 11.20.



Rys. 11.20. Średni błąd względny (MRE) określenia stężenia atomów węgla dla mieszanin zanieczyszczeń o znanym składzie jakościowym za pomocą regresji cząstkowych najmniejszych kwadratów, na podstawie następującej liczby składowych: 1, 2, 3, 5 i 7

Dyskusja

Analiza danych czujnikowych pokazała, że dla mieszanin o znanym składzie jakościowym liczba cech typu *P* istotnie skorelowanych ze stężeniem atomów węgla była na ogół większa niż liczba cech skorelowanych ze stężeniem substancji dominujących w poszczególnych mieszaninach (por. tabela 11.8 i tabela 11.4). Taką samą prawidłowość stwierdzono dla selekcji jednowymiarowej.

Na podstawie porównania tabeli 11.5 i tabeli 11.9 łatwo zauważyć, że dla mieszanin dwóch substancji zanieczyszczających wyniki określania stężenia atomów węgla były lepsze niż wyniki określania stężeń poszczególnych składników tych mieszanin. Natomiast cechy najbardziej przydatne do obliczenia stężenia atomów węgla pochodziły w dużej mierze z sygnałów tych samych czujników co cechy najbardziej przy-

datne do określenia stężenia substancji dominującej (por rys. 11.7 i 11.16 oraz rys. 11.8 i 11.17).

Z jednym wyjątkiem (mieszanina heksan i oktan) uzyskano bardzo dobre rezultaty obliczeń stężenia atomów węgla (MRE ≤ 8%) z zastosowaniem regresji jednowymiarowej nieliniowej (tabela 11.9). Były one porównywalne z wynikami dla wielokrotnej regresji liniowej z trójelementowymi wektorami cech wyselekcjonowanych. Błędy oznaczeń były o 1–2% mniejsze niż w wypadku określania substancji dominującej (tabela 11.5). Wiele czujników tych samych co dla substancji dominującej było najczęściej selekcjonowanych jako źródło informacji (por. tabelea 11.10 i 11.6).

Niskowymiarowa reprezentacja badanych gazów w postaci wektora cech wyekstrahowanych sprawdzała się gorzej. Dotyczyło to zarówno regresji liniowej 1–3 składowych głównych (rys. 11.18a), jak i regresji nieliniowej pierwszej składowej głównej (rys. 11.18b). Posługiwanie się bardziej złożoną reprezentacją (5–7-elementowy wektor cech) pozwalało natomiast obliczać stężenia atomów węgla z błędem ok. 5%.

Z porównania modeli różnego rodzaju, zbudowanych dla ustalonej liczby cech (większej niż jeden), wynika, że najmniejsze błędy określania stężenia atomów węgla uzyskano metodą cząstkowych najmniejszych kwadratów (rys. 11.20). Optymalne dla tego rodzaju modelu było zastosowanie trzech składowych.

11.2.2. Kategoria substancji zanieczyszczających

Analizowano możliwości określania stężenia atomów węgla pochodzących od pojedynczej substancji zanieczyszczającej, o której wiadomo tylko, że należy ona do określonej grupy zanieczyszczeń. Odniesiono się do sytuacji, gdy kategorię stanowiły: i) wszystkie badane LZO, tj. węglowodory alifatyczne (heksan, heptan i oktan), związek cykliczny (cykloheksan) i związki aromatyczne (benzen, toluen, ksylen i etylobenzen), ii) węglowodory alifatyczne (heksan, heptan i oktan), iii) związki aromatyczne (benzen, toluen, ksylen i etylobenzen).

Regresja liniowa i nieliniowa oraz wektor cech wyselekcjonowanych

W wyniku przeprowadzonych analiz stwierdzono brak jednoelementowych wektorów cech umożliwiających obliczenie stężenia atomów węgla dla grupy pojedynczych substancji zanieczyszczających z błędem mniejszym niż 10%. W rezultacie filtracji wykazano brak cech wyraźnie skorelowanych ($R \geq 0{,}95$) ze stężeniem atomów węgla, gdy grupa obejmowała wszystkie rozważane substancje zanieczyszczające. Jeżeli grupę ograniczono do związków alifatycznych, to stwierdzono istnienie 165 takich cech, a dla grupy związków aromatycznych było ich 13.

Tabela 11.11. Średni błąd względny (MRE) określania stężenia atomów węgla w grupie pojedynczych substancji zanieczyszczających[a]

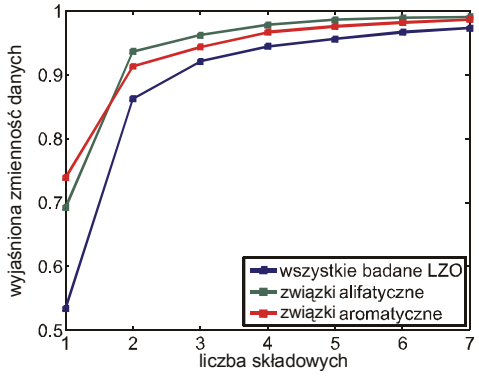
Grupa pojedynczych substancji zanieczyszczających	Regresja liniowa					Regresja nieliniowa
	Liczba elementów wektora cech					
	1	2	3	5	7	1
Wszystkie badane LZO	34–38	24–28	21–25	20–23	19–22	31–32
Węglowodory alifatyczne	20–31	13–16	11–15	8–11	6–9	14–19
Związki aromatyczne	19–24	18–20	17–19	16–19	15–18	26–29

[a]Podano wartości min(MRE)–max(MRE) [%] dla 30 najlepszych wektorów cech o następującej liczbie elementów: 1, 2, 3, 5 i 7 uzyskanych w wyniku selekcji.

W tabeli 11.11 przedstawiono rezultaty obliczeń stężenia atomów węgla dla grupy pojedynczych substancji zanieczyszczających z zastosowaniem regresji liniowej wielokrotnej odpornej na podstawie wektorów cech o następującej liczbie elementów: 1, 2, 3, 5 i 7. Wyniki dotyczą trzydziestu wyselekcjonowanych wektorów cech.

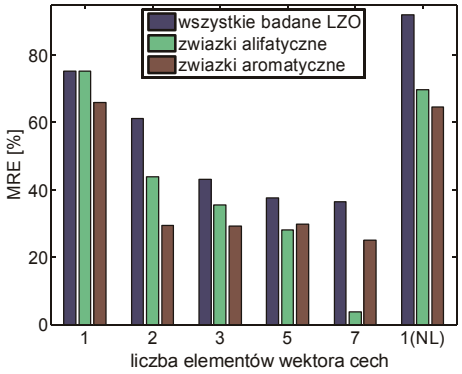
Regresja liniowa i nieliniowa oraz wektor cech wyekstrahowanych

Łączny udział składowych głównych w wyjaśnianiu zmienności danych wielowymiarowych dotyczących rozważanych grup substancji organicznych przedstawiono na rysunku 11.21.



Rys. 11.21. Łączny udział składowych głównych w wyjaśnianiu zmienności danych wielowymiarowych dotyczących różnych grup substancji zanieczyszczających

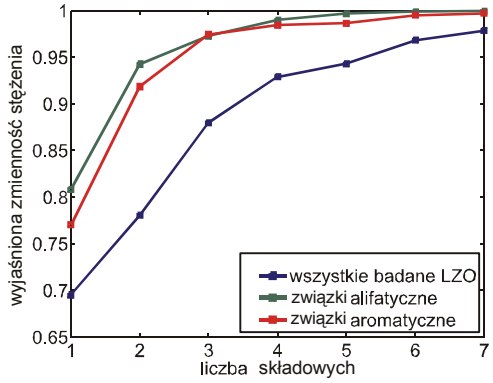
Na rysunku 11.22 pokazano średni błąd względny określenia stężenia atomów węgla dla grup pojedynczych substancji zanieczyszczających z zastosowaniem regresji liniowej 1, 2, 3, 5 i 7 składowych głównych oraz regresji nieliniowej z pierwszą składową główną jako zmienną objaśniającą.



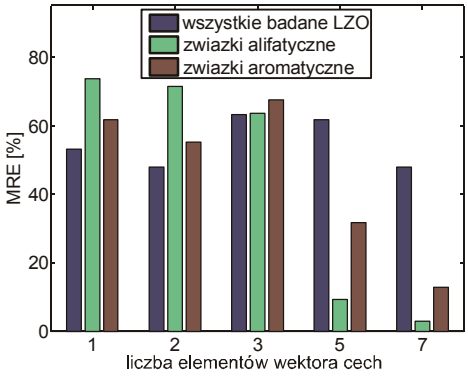
Rys. 11.22. Średni błąd względny (MRE) określenia stężenia atomów węgla dla różnych grup substancji zanieczyszczających z zastosowaniem regresji liniowej 1, 2, 3, 5 i 7 składowych głównych oraz regresji nieliniowej z pierwszą składową główną jako zmienną objaśniającą

Regresja metodą cząstkowych najmniejszych kwadratów

Zastosowano regresję metodą cząstkowych najmniejszych kwadratów do wyznaczania stężenia atomów węgla dla różnych kategorii substancji zanieczyszczających. Udział określonej liczby składowych w opisie stężenia atomów węgla dla poszczególnych grup LZO pokazano na rys. 11.23. Rezultaty oznaczeń z wykorzystaniem modeli opartych na różnej liczbie składowych przedstawiono na rys. 11.24.



Rys. 11.23. Łączny udział składowych w wyjaśnianiu zmienności stężenia atomów węgla dla różnych grup substancji zanieczyszczających



Rys. 11.24. Średni błąd względny (MRE) określenia stężenia atomów węgla dla różnych grup substancji zanieczyszczających z zastosowaniem regresji metodą cząstkowych najmniejszych kwadratów z 1, 2, 3, 5 i 7 składowymi

Dyskusja

Jak wynika z przedstawionych analiz, jeżeli przyjmie się stężenie atomów węgla za miarę ilościową zanieczyszczenia w warunkach zupełnej nieznajomości rodzaju substancji zanieczyszczających występujących w powietrzu, to należy się liczyć z błędem oceny ok. 25%. Pokazano, że ograniczenie grupy zanieczyszczeń do jednego rodzaju związków, np. węglowodorów alifatycznych, umożliwiało zmniejszenie błędu poniżej 10%.

Najkorzystniejszym z rozważanych sposobów określania stężenia atomów węgla, gdy w powietrzu występują nieznane pojedyncze substancje zanieczyszczające, była regresja wielokrotna na wieloelementowych wektorach cech wyselekcjonowanych.

Rezultaty uzyskane z zastosowaniem regresji składowych głównych i regresji cząstkowych najmniejszych kwadratów były zaskakująco niekorzystne (rys. 11.23 i 11.24). Jednak również w przypadku tych metod stwierdzono poprawę wyników wynikającą z ograniczenia grupy rozważanych zanieczyszczeń do związków tego samego typu.

Na podstawie wykonanej analizy danych stwierdzono, że wektor cech wyekstrahowanych był nieodpowiednią reprezentacją badanych gazów, gdy poszukiwano ich opisu w kategoriach stężenia atomów węgla.

11.2.3. Kategoria mieszanin substancji zanieczyszczających

Analizowano możliwości określania stężenia atomów węgla pochodzących od mieszanin substancji zanieczyszczających dla mieszanin o nieznanym składzie jakościowym. Przyjęto, że znana jest tylko przynależność mieszaniny do określonej grupy. Wydzielono następujące grupy (patrz tabela 8.2): i) wszystkie badane mieszaniny LZO, ii) mieszaniny LZO o składzie zdominowanym przez heksan, iii) mieszaniny LZO o składzie zdominowanym przez toluen.

Regresja liniowa i nieliniowa oraz wektor cech wyselekcjonowanych

Stwierdzono brak jednoelementowych wektorów cech umożliwiających obliczenie stężenia atomów węgla dla tak zdefiniowanych grup mieszanin substancji zanieczyszczających z błędem mniejszym niż 10%. Podobnie liczba cech uznanych za istotne według kryterium filtracji (współczynnik korelacji $R \geq 0{,}95$) wynosiła zero.

Rezultaty obliczeń stężenia atomów węgla dla grup mieszanin substancji zanieczyszczających z zastosowaniem regresji liniowej odpornej dla najlepszych trzydzie-

stu wektorów cech liczących 1, 2, 3, 5 lub 7 elementów oraz jednowymiarowej regresji nieliniowej przedstawiono w tabeli 11.12.

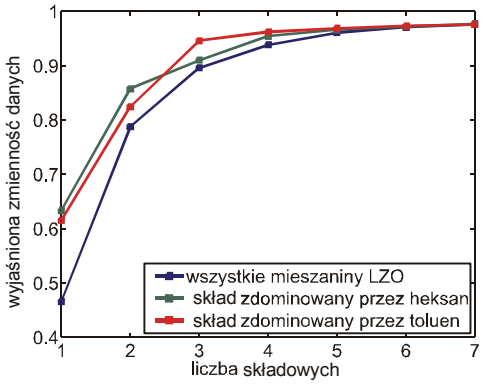
Tabela 11.12. Średni błąd względny (MRE) określania stężenia atomów węgla[a]

Grupa pojedynczych substancji zanieczyszczających	Regresja liniowa					Regresja nieliniowa
	Liczba elementów wektora cech					
	1	2	3	5	7	1
Wszystkie badane mieszaniny LZO	68–68	20–23	17–20	15–20	14–15	24–27
Mieszaniny o składzie zdominowanym przez heksan	19–21	18–20	15–17	12–15	12–14	12–15
Mieszaniny o składzie zdominowanym przez toluen	29–22	15–20	14–16	12–14	12–13	16–20

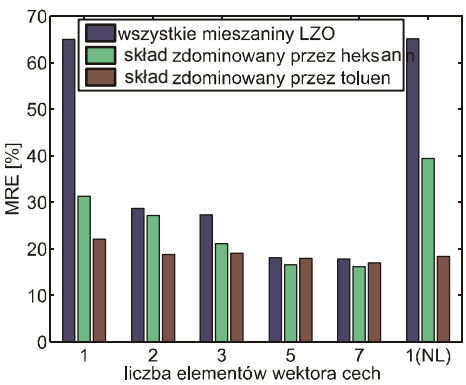
[a]Podano wartości min(MRE)–max(MRE) [%] dla grupy mieszanin substancji zanieczyszczających dla 30 najlepszych wektorów cech o liczbie elementów 1, 2, 3, 5 i 7 uzyskanych w wyniku selekcji.

Regresja liniowa i nieliniowa oraz wektor cech wyekstrahowanych

Łączny udział składowych głównych w wyjaśnianiu zmienności danych wielowymiarowych dotyczących rozważanych grup mieszanin lotnych związków organicznych pokazano na rys. 11.25.



Rys. 11.25. Łączny udział składowych głównych w wyjaśnianiu zmienności danych wielowymiarowych dotyczących poszczególnych grup mieszanin substancji zanieczyszczających

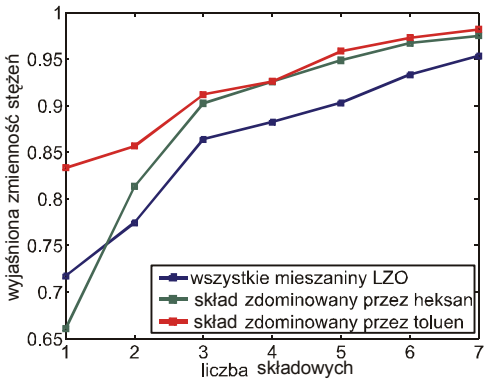


Rys. 11.26. Średni błąd względny (MRE) określenia stężenia atomów węgla w poszczególnych grupach mieszanin substancji zanieczyszczających z zastosowaniem regresji liniowej 1, 2, 3, 5 i 7 składowych głównych oraz regresji nieliniowej z pierwszą składową główną jako zmienną objaśniającą

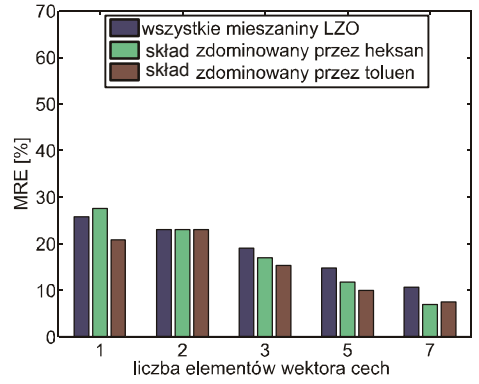
Na rysunku 11.26 przedstawiono wyniki określenia stężenia atomów węgla dla grup mieszanin substancji zanieczyszczających, z zastosowaniem regresji liniowej 1, 2, 3, 5 i 7 składowych głównych oraz regresji nieliniowej z pierwszą składową główną jako zmienną objaśniającą.

Regresja metodą cząstkowych najmniejszych kwadratów

Zastosowano regresję metodą cząstkowych najmniejszych kwadratów. Udział określonej liczby składowych w opisie stężenia atomów węgla dla poszczególnych grup mieszanin LZO pokazano na rys. 11.27. Rezultaty oznaczeń z wykorzystaniem modeli, w których zastosowano różną liczbę składowych przedstawiono natomiast na rys. 11.28.



Rys. 11.27. Łączny udział składowych głównych w wyjaśnianiu zmienności stężenia atomów węgla dla wyróżnionych grup mieszanin substancji zanieczyszczających



Rys. 11.28. Średni błąd względny (MRE) określenia stężenia atomów węgla dla wyróżnionych grup mieszanin substancji zanieczyszczających z zastosowaniem regresji metodą cząstkowych najmniejszych kwadratów z 1, 2, 3, 5 i 7 składowymi

Dyskusja

Na podstawie rezultatów analizy danych stwierdzono, że posługiwanie się miarą ilościową w postaci stężenia atomów węgla w odniesieniu do kategorii mieszanin

substancji zanieczyszczających i kategorii pojedynczych zanieczyszczeń dawało porównywalne rezultaty (por. tabele 11.12 i 11.11).

Dla mieszanin miarę tę można było wyznaczyć z błędem mniejszym niż 20% dowolną z rozważanych metod obliczeniowych na podstawie trzech zmiennych objaśniających. Podejście jednowymiarowe należało uznać za nieskuteczne.

Najlepsze rezultaty otrzymano, stosując regresję metodą cząstkowych najmniejszych kwadratów (rys. 11.28) i regresję wielokrotną na wektorze cech wyselekcjonowanych (tabela 11.12). Dla metody PLS wraz ze wzrostem liczby składowych błędy systematycznie się zmniejszały, a posługując się siedmioma składowymi można było uzyskać błąd mniejszy niż 10%. Dla regresji składowych głównych błąd określania stężenia wynosił ok. 20% i był praktycznie niewrażliwy na liczbę składowych, jeśli przekraczała trzy (rys. 11.28).

Podobnie jak dla grup pojedynczych zanieczyszczeń kilkuprocentową poprawę dokładności oznaczenia można było uzyskać przez ograniczenie zróżnicowania mieszanin opisywanych tym samym modelem (tabela 11.12, 11.26, rys. 11.28). W rozważnym przypadku efekt ten zaobserwowano dla mieszanin o składzie zdominowanym przez heksan i toluen razem i osobno. Biorąc pod uwagę różnorodność składów w obrębie tych grup, można uznać, że określanie ilościowe mieszanin gazów z zastosowaniem stężenia atomów węgla jako miary może być przeprowadzane w dość dużym zakresie składów jakościowych.

11.3. Wnioski z analizy danych czujnikowych ze względu na informację ilościową o zanieczyszczeniach

Dane z pomiarów czujnikowych zawierają informację dotyczącą właściwości ilościowych zanieczyszczeń powietrza. Podobnie jak dla informacji jakościowej efektywność jej pozyskiwania zależy od doboru elementów systemu rozpoznawania wzorców odpowiednio do rodzaju poszukiwanej informacji.

W pracy zajmowano się informacją ilościową w postaci stężeń zanieczyszczeń oraz miar ilościowych innych niż stężenie. Możliwość pozyskania informacji o stężeniach analizowano w odniesieniu do zanieczyszczeń występujących w powietrzu pojedynczo, zanieczyszczeń dominujących ilościowo oraz zanieczyszczeń ilościowo zdominowanych. Za miarę inną niż stężenie substancji zanieczyszczającej przyjęto stężenie atomów węgla pochodzących od LZO. Miarę tę zastosowano do ilościowego opisu mieszanin substancji zanieczyszczających o znanym składzie jakościowym, kategorii substancji zanieczyszczających oraz kategorii mieszanin substancji zanieczyszczających.

W tabelach 11.13–11.18 przedstawiono podsumowanie rezultatów pozyskiwania poszczególnych rodzajów informacji ilościowej o zanieczyszczeniach wszystkimi rozważanymi metodami analizy danych. Podano zakres średniego błędu względnego (MRE) określenia miary ilościowej za pomocą najlepszych rozwiązań w danej grupie metod.

Tabela 11.13. Błąd uzyskiwania informacji o stężeniu substancji zanieczyszczającej występującej w powietrzu pojedynczo metodą regresji liniowej (RL), nieliniowej (RNL) oraz cząstkowych najmniejszych kwadratów (RPLS); zakres MRE dla najlepszych rozwiązań [%]

Liczba cech	Selekcja cech		Ekstrakcja cech		RPLS
	RL	RNL	RL	RNL	
1	7–19	3–16	20–45	10–30	10–45
2	3–8		4–25		4–18
3	3–7		1–10		3–5
5	3–7		2–13		2–4
7	3–6		2–7		2–4

Tabela 11.14. Błąd uzyskiwania informacji o stężeniu dominującej substancji zanieczyszczającej metodą regresji liniowej (RL), nieliniowej (RNL) oraz metodą cząstkowych najmniejszych kwadratów (RPLS); zakres MRE dla najlepszych rozwiązań [%]

Liczba cech	Selekcja cech		Ekstrakcja cech		RPLS
	RL	RNL	RL	RNL	
1	8–17	3–15	14–32	5–40	14–25
2	6–17		10–20		8–15
3	5–11		6–15		4–10
5	5–10		5–10		2–6
7	5–9		3–8		2–4

Tabela 11.15. Błąd uzyskiwania informacji o stężeniu zdominowanej substancji zanieczyszczającej metodą regresji liniowej (RL), nieliniowej (RNL) oraz metodą cząstkowych najmniejszych kwadratów (RPLS); zakres MRE dla najlepszych rozwiązań [%]

Liczba cech	Selekcja cech		Ekstrakcja cech		RPLS
	RL	RNL	RL	RNL	
1	28–70	13–66	75–85	6–40	50–75
2	26–58		41–80		42–126
3	21–45		35–55		22–48
5	18–41		15–42		10–35
7	18–40		10–42		7–20

Tabela 11.16. Błąd uzyskiwania informacji o stężeniu atomów węgla dla mieszaniny o znanym składzie jakościowym metodą regresji liniowej (RL), nieliniowej (RNL) oraz metodą cząstkowych najmniejszych kwadratów (RPLS); zakres MRE dla najlepszych rozwiązań [%]

Liczba cech	Selekcja cech		Ekstrakcja cech		RPLS
	RL	RNL	RL	RNL	
1	7–16	3–12	12–28	6–40	12–24
2	3–16		10–17		4–12
3	3–9		3–12		3–7
5	2–7		2–9		2–5
7	2–6		2–6		2–4

Tabela 11.17. Błąd uzyskiwania informacji o stężeniu atomów węgla dla kategorii substancji zanieczyszczających metodą regresji liniowej (RL), nieliniowej (RNL) oraz metodą cząstkowych najmniejszych kwadratów (RPLS); zakres MRE dla najlepszych rozwiązań [%]

Liczba cech	Selekcja cech		Ekstrakcja cech		RPLS
	RL	RNL	RL	RNL	
1	19–38	14–32	62–75	62–90	52–73
2	13–28		29–60		49 –70
3	11–25		29–42		60–65
5	8–23		29–39		9–60
7	6–22		2–39		3–50

Tabela 11.18. Błąd uzyskiwania informacji o stężeniu atomów węgla dla kategorii mieszanin substancji zanieczyszczających metodą regresji liniowej (RL), nieliniowej (RNL) oraz metodą cząstkowych najmniejszych kwadratów (RPLS). Podano zakres MRE dla najlepszych rozwiązań [%]

Liczba cech	Selekcja cech		Ekstrakcja cech		RPLS
	RL	RNL	RL	RNL	
1	19–68	12–27	20–65	19–65	20–28
2	18–23		19–29		23–23
3	14–20		20–27		17–20
5	12–20		19–20		10–15
7	12–15		19–20		8–11

W wyniku przeprowadzonej analizy danych stwierdzono, że na podstawie pomiarów czujnikowych można pozyskiwać różne informacje ilościowe o zanieczyszczeniach, stosując odpowiednią reprezentację badanego gazu oraz metodę obliczeniową.

Stężenia substancji zanieczyszczających występujących w powietrzu pojedynczo z powodzeniem określano za pomocą regresji nieliniowej jednej, wyselekcjonowanej cechy typu *P* lub regresji liniowej wielokrotnej trzech wyselekcjonowanych cech tego typu. Należało się wówczas liczyć z błędem wynoszącym kilka procent.

Określenie stężeń substancji zanieczyszczających występujących w powietrzu w mieszaninach zależało od ich proporcji ilościowych. Stężenia zanieczyszczeń dominujących (udział 60–95%) można było wyznaczać z małym błędem, podobnym jak dla substancji występujących w powietrzu pojedynczo. W takich zadaniach wystarczającą okazała się regresja nieliniowa jednej, wyselekcjonowanej cechy typu *P* lub regresja liniowa wielokrotna trzech wyselekcjonowanych cech. Stosując takie rozwiązanie, należało akceptować błąd ok. 10%. Zwiększenie wymiaru przestrzeni cech do pięciu pozwało zmniejszyć go do wartości mniejszej niż 5% pod warunkiem zastosowania regresji metodą cząstkowych najmniejszych kwadratów.

Większe trudności nastręczało obliczenie stężenia substancji zdominowanych (udział na poziomie 5–40%). Błędy ok. 10% uzyskano wyłącznie po zastosowaniu regresji wykładniczej pierwszej składowej głównej (PCA) i to nie dla wszystkich LZO, które rozważano w pracy. Rozbudowywanie wektora cech w ramach regresji liniowej wielokrotnej nie przynosiło poprawy, nawet dla regresji metodą cząstkowych najmniejszych kwadratów. Można sądzić, że wynik ten wskazywał na potrzebę zastosowania podejścia nieliniowego do określania stężenia substancji zdominowanej, jednak z wykorzystaniem większej reprezentacji informacji niż jedna cecha typu *P*.

Przeprowadzona analiza danych pokazała, że dobrą alternatywą dla poszukiwania informacji o poszczególnych składnikach mieszaniny jest określenie jej zbiorczą miarą, np. stężeniem atomów węgla pochodzących od składników mieszaniny, gdy znany jest skład jakościowy mieszaniny. Błędy określania stężenia atomów węgla były w tej sytuacji nieznacznie mniejsze od błędu określania stężenia składnika dominującego z zastosowaniem tych samych metod obliczeniowych i tak samo złożonej reprezentacji informacji. Wynika stąd, że taka miara zbiorcza dobrze oddaje fakt, że odpowiedź czujnika łączy informację o wielu substancjach zanieczyszczających, na które jest on eksponowany.

Z oczywistych względów nie można stosować miary zbiorczej, jaką jest stężenie zanieczyszczenia, do określania zanieczyszczeń nieznanego rodzaju, np. grupy pojedynczych substancji zanieczyszczających lub grupy mieszanin substancji zanieczyszczających. Pokazano, że dla efektywnej ilościowej oceny zanieczyszczenia za pomocą stężenia atomów węgla duże znaczenie ma podobieństwo chemiczne substancji w obrębie grupy. W najlepszym z uzyskanych rozwiązań (regresja PCA lub PLS na siedmioelementowym wektorze cech) związki alifatyczne określano z błędem mniejszym niż 5%. Wzrost zróżnicowania zanieczyszczeń w grupie powodował również pogorszenie dokładności. Można jednak przyjąć, że błąd tego typu oznaczeń w umiarkowanie wielowymiarowej przestrzeni cech (3−5) będzie wynosił kilkanaście procent.

Uprzywilejowaną metodą okazała się regresja liniowa wielokrotna odporna na wektorze cech wyselekcjonowanych.

Rezultaty określania stężenia atomów węgla dla grup mieszanin zanieczyszczeń organicznych były podobne jak w grupach pojedynczych zanieczyszczeń. Wydaje się, że stosując stężenie atomów węgla jako miarę ilościową, gdy w powietrzu występuje wiele substancji zanieczyszczających, należy się liczyć z błędem na poziomie kilkunastu procent. Jest wówczas konieczne zastosowanie trój- lub wyżej wymiarowych przestrzeni cech. Regresja cząstkowych najmniejszych kwadratów w siedmiowymiarowej przestrzeni pozwoliła uzyskać błąd kilku procent.

Z metodologicznego punktu widzenia warto podkreślić oczywistą przewagę podejścia opakowanego nad filtracją jako metodą wyboru najlepszej reprezentacji informacji o zanieczyszczeniu, również jednowymiarowej. Przyjęcie dużego stopnia skorelowania cech jako formy preselekcji z miarą ilościową określającą zanieczyszczenie, doprowadziłoby do wyeliminowania z rozważań szeregu cech dobrze współpracujących z modelami nieliniowym.

Bardzo dobrymi właściwościami jako reprezentacja informacji ilościowej o zanieczyszczeniach wykazały się wieloelementowe wektory cech wyselekcjonowanych. Zastosowanie cech wyselekcjonowanych jako zmiennych objaśniających w modelach regresji liniowej wielokrotnej opakowanej pozwalało uzyskać małe błędy predykcji różnych miar ilościowych zanieczyszczeń. Jedyny wyjątek dotyczył określania stężenia substancji zanieczyszczającej zdominowanej. Wniosek ten jest bardzo korzystny, gdyż taka reprezentacja informacji jest wygodna w praktyce. Po wykonaniu pomiaru dostęp do wektorów danych odpowiadających wektorom cech typu P jest natychmiastowy, gdyż nie jest wymagana żadna wstępna obróbka sygnału czujnikowego. Modele regresji liniowej można natomiast stosunkowo łatwo i szybko parametryzować. Ponadto wykazano, że każdy rozważany problem ilościowy można było rozwiązać ze zbliżoną efektywnością na podstawie co najmniej kilkudziesięciu różnych wektorów cech. Równoczesne zastosowanie wielu modeli umożliwia poprawę właściwości statystycznych oszacowania stężenia zanieczyszczenia na postawie pomiaru czujnikowego.

Zauważono, że elementy wektorów cech wyselekcjonowanych jako najlepsze do pozyskania różnych informacji, pochodziły z sygnałów różnych czujników, a ponadto z różnych fragmentów tych sygnałów. Potwierdziło to zalety posługiwania się złożonymi danymi pomiarowymi dostarczanymi przez matrycę czujników pracującą w trybie dynamicznym do pozyskania zróżnicowanej informacji o zanieczyszczeniach gazowych.

12. Podsumowanie

Dane pomiarowe są źródłem najbardziej wiarygodnych informacji o środowisku. W związku z rozwojem technologicznym zakres oraz stopień różnorodności dostępnej informacji są coraz większe. W dużej mierze efekt ten jest związany ze złożonością danych dostarczanych w wyniku stosowania współczesnych metod pomiarowych. Miejsce danych jednowymiarowych coraz częściej zajmują dane wielowymiarowe, a nawet wielokierunkowe.

Przekształcenie tego rodzaju danych w użyteczną informację wymaga przeprowadzenia analizy z zastosowaniem odpowiednich metod obliczeniowych. Wskutek tego niezbywalnym elementem nowoczesnych systemów pozyskiwania informacji o środowisku staje się analiza danych. Jej rola jest szczególnie ważna w systemach czujnikowych przeznaczonych do takich celów. Wynika to z właściwości pomiarowych czujników.

Przedmiotem tej monografii była analiza danych czujnikowych umożliwiająca jakościowe i ilościowe określenie zanieczyszczenia powietrza. Celem analizy było efektywne uzyskanie poprawnych wartości różnych miar charakteryzujących zanieczyszczenie w sposób jakościowy oraz ilościowy.

Z dotychczasowych badań oraz praktyki wynika, że dobrą strategią analizy danych w czujnikowych pomiarach gazów jest rozpoznawanie wzorców. Tezę tę potwierdzono w pracy w odniesieniu do czujnikowych pomiarów zanieczyszczeń powietrza. Zasadniczymi elementami tego podejścia są: uzyskanie skompresowanej reprezentacji badanego gazu na podstawie danych pomiarowych, a następnie przypisanie tej reprezentacji etykiety jakościowej lub wartości zmiennej ilościowej, określającej właściwości jakościowe lub ilościowe zanieczyszczenia. Zachodzi to przez rozwiązanie problemu klasyfikacji lub regresji w przestrzeni cech, której poszczególne wymiary stanowią zarazem składowe reprezentacji badanego gazu.

W wyniku przeprowadzonej analizy danych wykazano, że dla efektywnego pozyskania informacji o zanieczyszczeniu zasadnicze znaczenie ma dobór konfiguracji – reprezentacja zanieczyszczenia i metoda odczytu informacji, odpowiednia do rodzaju poszukiwanej informacji. Dotyczy to nie tylko zasadniczej różnicy między informacją jakościową i ilościową, lecz również zróżnicowanej skali trudności poszukiwania określonych informacji jednego czy drugiego typu.

Metodologia rozpoznawania wzorców odnosi się przede wszystkim do problemów wielowymiarowych i posługuje się głównie metodami wielowymiarowej analizy danych. W pracy badano również rozwiązania jednowymiarowe. W dobrze zdefiniowanych problemach pomiarowych są one dopuszczalne mimo przesłanek teoretycznych, dyskredytujących ich zastosowanie w pomiarach gazów czujnikami częściowo selektywnymi. Z uzyskanych rezultatów wynika jednak jednoznacznie konieczność posługiwania się reprezentacją wielowymiarową oraz metodami analizy wielowymiarowej w celu zapewnienia odpowiedniej dokładności pozyskania informacji o zanieczyszczeniach gazowych na podstawie danych czujnikowych. Skok jakościowy między tymi dwoma rodzajami podejść jest oczywisty, dużo bardziej widoczny w dziedzinie problemów jakościowych.

Mimo bezsprzecznej przewagi podejścia wielowymiarowego wykazano, że zadowalające rozwiązania w dziedzinie rozpoznawania zanieczyszczeń, jak również ich ilościowego określania można uzyskać, konstruując relatywnie niskowymiarowe reprezentacje badanych gazów. Znaleziono zadowalające, tzn. bezbłędne lub prawie bezbłędne rozwiązania wszystkich rozważanych w pracy problemów jakościowych w dwu- lub co najwyżej trójwymiarowych przestrzeniach cech. Podobnie w zagadnieniach o charakterze ilościowym, dwu- lub trójwymiarowe wektory cech na ogół były wystarczające do określania ilościowego zanieczyszczeń z błędem mniejszym niż kilka procent. Stwierdzono, że uzyskanie zadowalających rozwiązań bardziej złożonych problemów wymagało zastosowania przestrzeni cech o większej liczbie wymiarów. Było to np. obliczanie stężeń substancji zdominowanych w mieszaninach gazów, czy wyznaczanie miar ilościowych dla kategorii gazów o składzie jakościowym z bardzo szerokiego zakresu.

Podstawą do opracowania zwartych i efektywnych reprezentacji gazów była koncepcja cechy, rozumianej jako pojedyncza wartość sygnału czujnika zarejestrowana w określonej chwili podczas ekspozycji na badany gaz. Wykazano, że koncepcja ta bardzo dobrze sprawdza się w zastosowaniu do kompresji złożonych danych pomiarowych, pochodzących z pomiarów wykonanych matrycą czujników w określonym trybie pracy.

Stwierdzono, że pozyskanie konkretnej informacji o zanieczyszczeniach zależy od sposobu konstrukcji przestrzeni cech. Dotyczy to przede wszystkim wyboru między selekcją cech i mapowaniem. Z uzyskanych rezultatów wynika, że wektory cech wyselekcjonowanych lepiej współpracowały z metodami liniowymi, zarówno w wypadku klasyfikacji, jak i regresji. Bardzo dobre efekty przyniosła nadzorowana ekstrakcja cech na potrzeby regresji metodą cząstkowych najmniejszych kwadratów oraz nienadzorowana ekstrakcja dla klasyfikacji metodą *k*-najbliższych sąsiadów.

Zestawienie składu wektorów cech wyselekcjonowanych, najlepszych do scharakteryzowania wybranego zanieczyszczenia pod względem jakościowym i ilościowym wskazuje na zasadniczo różne nośniki tych dwóch rodzajów informacji. Wynika stąd, że jakościowe oraz ilościowe określenie zanieczyszczenia powinno korzystać z róż-

nych reprezentacjach badanego gazu, choć zbudowanych na podstawie tych samych danych pomiarowych. Podobnie konkretne rodzaje informacji jakościowej czy ilościowej uprzywilejowują różne reprezentacje badanego gazu.

Zasadnicze znaczenie dla efektywności pozyskania informacji o zanieczyszczeniach miał wybór metody odczytu informacji w danej przestrzeni cech. Wykonane analizy wykazały przewagę metody nieliniowej w rozwiązywaniu problemów jakościowego określania zanieczyszczeń. Dla zagadnień ilościowych przydatność porównywanych metod zależała od rodzaju poszukiwanej informacji.

Wybór pomiarów czujnikowych jako źródła informacji o środowisku umożliwia posługiwanie się różnymi sposobami jakościowego i ilościowego opisu zanieczyszczeń. Możliwe jest znaczne poszerzenie zakresu poszukiwanej informacji w stosunku do tradycyjnego zainteresowania tożsamością chemiczną i stężeniem substancji zanieczyszczających.

Stwierdzono, że na podstawie danych z pomiarów czujnikowych istnieje możliwość bezbłędnego określenia zanieczyszczeń pod względem jakościowym w zakresie następujących rodzajów informacji: tożsamość chemiczna substancji zanieczyszczających, skład jakościowy mieszanin zanieczyszczeń, przynależność do określonej kategorii substancji zanieczyszczających oraz przynależność do klasy mieszanin substancji zanieczyszczających charakteryzujących się wspólnym składnikiem dominującym.

Poszukując w tego rodzaju danych klasycznej informacji analitycznej o stężeniu zanieczyszczeń, należy liczyć się z błędem od kilku do kilkudziesięciu procent w zależności od udziału oznaczanej substancji w mieszaninie zanieczyszczającej. Im większy udział zanieczyszczenia, tym mniejszy błąd oznaczenia. Zastosowanie alternatywnej miary, np. stężenia atomów węgla pochodzących od lotnych związków organicznych, wymaga zaakceptowania błędu w zbliżonym zakresie, którego wartość zależy od zróżnicowania zanieczyszczeń objętych wspólnym modelem ilościowym.

W pomiarach czujnikowych dokładność określenia zanieczyszczeń pod względem ilościowym w dużym stopniu zależy od zakresu dostępnej informacji o ich właściwościach. Nakazuje to przywiązywać dużą wagę do analizy danych umożliwiającej bezbłędne rozpoznanie zanieczyszczenia.

W monografii pokazano, że właściwie przeprowadzona analiza danych pozwala pozyskać z zadowalającą dokładnością różnorodne informacje o zanieczyszczeniu powietrza na podstawie pomiarów czujnikowych. Z przeprowadzonych badań wynikają przesłanki do projektowania systemów analizy danych w systemach czujnikowych do pomiarów zanieczyszczeń. Ze względu na szereg zalet takich rozwiązań należy się spodziewać, że w niedalekiej przyszłości staną się one podstawą powszechnie stosowanych systemów monitorujących środowisko.

Literatura

[1] WEBSTER F., *Theories of the Information Society*, Routledge, 2002.

[2] MAS S., DE JUAN A., TAULER R., OLIVIERI A.C., ESCANDAR G.M., *Application of chemometric methods to environmental analysis of organic pollutants – A review*, Talanta, 80 (3) (2010), 1052–1067.

[3] HIERLEMANN A., GUTIERREZ-OSUNA R., *Higher-order chemical sensing*, Chem. Rev., 108 (2008), 563–613.

[4] YAMAZOE N., SHIMANOE K., *New perspectives of gas sensor technology*, Sens. Actuators B, 138 (2009), 100–107.

[5] THEORORIDIS S., KOUTROUMBAS K., *Pattern recognition*, Academic Press, USA, 1999.

[6] SNOPOK B.A., KRUGLENKO I.V., *Multisensor system for chemical analysis – state-of-the-art in Electronic Nose technology and new trends in machine olfaction*, Thin Solid Films, 418 (2002), 21–41.

[7] CHARRABARTI S., NADEAU T.P., COX E., NEAPOLITAN R.E., FRANK E., PYLE D., GÜTING R.H., REFAAT M., HAN J., SCHNEIDER M., JIANG X., TEOREY T.J., KAMBER M., WITTEN I.H., LIGHTSTONE S.S., *Data mining*, Elsevier, 2009.

[8] WEBB A., *Statistical pattern recognition*, Arnold, UK, 1999.

[9] SZCZUREK A., **MACIEJEWSKA M.,** *Recognition of benzene, toluene and xylene using TGS array integrated with linear and non-linear classifier*, Talanta, 64 (2004), 609–617.

[10] NEGRI R.M., REICH S., *Identification of pollutant gases and its concentrations with a multisensor array*, Sens. Actuators B, 75 (2001), 172–178.

[11] SZCZUREK A.**, MACIEJEWSKA M.,** R*elationship between odour intensity assessed by human assessor and TGS sensor array response*, Sens. Actuators B,106 (2005), 13–19.

[12] MAEKAWA T., SUZUKI K., TAKADA T., KOBAYASHI T., EGASHIRA M., *Odor identification using sno$_2$-based sensor array*, Sens. Actuators B, 80 (2001), 51–58.

[13] **MACIEJEWSKA M.,** KOLODZIEJCZAK K., SZCZUREK A., *Discrimination of coatings on wooden materials using the gas sensor system*, Talanta, 68 (2005), 138–145.

[14] SZCZUREK A., **MACIEJEWSKA M.,** FLISOWSKA-WIERCIK B., BODZOJ Ł., *Application of a sensor system for determining the kind and quantity of two component VOC mixtures in air after the use of solvents*, J. Environ. Monit., 11 (2009), 1942–1951.

[15] MENZEL R., GOSCHNICK J., *Gradient gas sensor microarray for on-line process control – A new dynamic classification model for fast and reliable air quality assessment*, Sens. Actuators B, 68 (2000), 115–122.

[16] WESCHLER C.J., C*hemistry in indoor environments – 20 years of research*, Indoor Air, 21 (2011), 205–218.

[17] **MACIEJEWSKA M.,** SZCZUREK A., BODZOJ Ł., FLISOWSKA-WIERCIK B., *Sensor array and stop-flow mode applied to discrimination and quantification of gas mixtures*, Sens. Actuators B, 150 (2010), 93–98.

[18] GAO D., LIU F., WANG J., *Quantitative analysis of multiple kinds of volatile organic compounds using hierarchical models with an electronic nose*, Sens. Actuators B, 161 (2012), 578–586.

[19] MØLHAVE L., NIELSEN G. D., *Interpretation and limitations of the concept Total Volatile Organic Compounds (TVOC), as an indicator of human responses to exposures of Volatile Organic Compounds (VOC), in indoor air*, Indoor Air, 2 (2) (1992), 65–77.
[20] BITTER F., MÜLLER B., MÜLLER D., *Estimation of odour intensity of indoor air pollutants from building materials with a multi-gas sensor system*, Building and Environment, 45 (2010), 197–204.
[21] MÜLLER R., *Chemical sensing and pattern recognition*, Proceedings of CompEuro '89, VLSI and Computer Peripherals, VLSI and Microelectronic Applications in Intelligent Peripherals and their Interconnection Networks, Hamburg, Germany, 3/57–3/61.
[22] GUTIERREZ-OSUNA R., *Pattern analysis for machine olfaction – A review*, IEEE Sensors Journal, 2 (3) (2002), 189–202.
[23] BERMAK A., BELHOUARI S.B., SHI M., MARTINEZ D., *Pattern recognition techniques for odor discrimination in gas sensor array*, Encyclopedia of Sensors, 10 (2006), 1–17.
[24] JAIN A.K., DUNIN R.P.W., MAO J., *Statistical pattern recognition – A review*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2 (1) (2000), 4–37.
[25] PERERA A., SUNDIC T., PARDO A., GUTIERREZ-OSUNA R., MARCO S., *A Portable Electronic Nose Based on Embedded PC Technology and GNU/Linux – Hardware, Software and Applications*, IEEE Sensors Journal, 2 (3) (2002), 235–246.
[26] OLIVAS E.S., GUERRERO J. D.M., MARTINEZ-SOBER M., RAFAEL J., MAGDALENA-BENEDITO J.R., SERRANO LOPEZ A.J., *Handbook of Research on Machine Learning Applications and Trends – Algorithms, Methods and Techniques*, Information Science Reference, 2009.
[27] VAPNIK V., *Statistical learning theory*, Wiley, 1998.
[28] PARDO M., SBERVEGLIERI G., Learning from data – *A tutorial with emphasis on modern pattern recognition methods*, IEEE Sensors Journal, 2 (3) (2002), 203–217.
[29] SCOTT S.M., JAMES D., ALI Z., *Data analysis for electronic nose systems*, Microchim. Acta, 156 (2007), 183–207.
[30] BRERETON R., *Chemometrics for pattern recognition*, Wiley, 2009.
[31] OSTROUF O., STROUF O., *Chemical pattern recognition*, Wiley, 1986.
[32] HULANICKI A., GŁĄB S., INGMAN F., *Chemical sensors definitions and classificatio*n, Pure App. Chem., 63 (9) (1991), 1247–1250.
[33] SZCZUREK A., *Pomiary lotnych związków organicznych rezystancyjnymi czujnikami gazów*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2006.
[34] MÜLLER R., *High electronic selectivity obtainable with nonselective chemosensors*, Sens. Actuators B, 4 (1991), 35–39.
[35] ZAROMB S., STETTER J.R., *Theoretical basis for identification and measurement of air contaminants using an array of sensors having partly overlapping selectivity*, Sensors and Actuators, 6 (4) (1984), 225–243.
[36] GÖPEL W., *Chemical imaging I – Concepts and visions for electronic and bioelectronic noses*, Sens. Actuators B, 52 (1998), 125–142.
[37] PEARCE T.C., SCHIFFMAN S.S., NAGLE H.T., GARDNER J.W., *Handbook of Machine Olfaction – Electronic nose technology*, Wiley, 2003.
[38] JURS P.C., BAKKEN G.A., MCCLELLAND H.E., *Computational methods for the analysis of chemical sensor array data from volatile analytes*, Chem. Rev., 100 (2000), 2649–2678.
[39] SZCZUREK A., **MACIEJEWSKA M.,** *Gas sensor array with broad applicability. Sensor array*, W. Yang (Ed.), Rijeka, InTech, 2012, 81–108.
[40] ALBERT K.J., LEWIS N.S., SCHAUER C.L., SOTZING G.A., STITZEL S.E., VAID T. P., WALT D.R., *Cross-reactive chemical sensor arrays*, Chem. Rev., 100 (2000), 2595–2626.
[41] ALTHAINZ P., GOSCHNICK J., EHRMANN S., ACHE H.J., *Multisensor microsystem for contaminants in air*, Sens. Actuators B, 33 (1996), 72–76.

[42] ZAROMB S., STETTER J.R., *Theoretical basis for identification and measurement of air contaminants using an array of sensors having partly overlapping sensitivities*, Sens. Actuators B, 6 (1984), 225– 243.

[43] DE VITO S., MASSERA E., PIAGA M., MARTINOTTO L., DI FRANCIA G., *On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario*, Sens. Actuators B, 129 (2008), 750–757.

[44] DE VITO S., PIGA M., MARTINOTTO L., DI FRANCIA G., *CO, NO_2 and NO_x urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization*, Sens. Actuators B, 143 (2009), 182–191.

[45] SZCZUREK A., **MACIEJEWSKA M.,** *Assessment of VOCs in air using sensors array under various exposure conditions*, 2012 IEEE Sensors applications symposium, Brescia, Italy, February 7–9, 2012, Proceedings (2012), 1–5.

[46] CAROTTA M.C., MARTINELLI G., CREMA L., MALAGU C., MERLI M., GHIOTTI G., TRAVERSA E., *Nanostructured thick-film gas sensors for atmospheric pollutant monitoring – quantitative analysis on field tests*, Sens. Actuators B, 76 (2001), 336–342.

[47] BERMAK A., BELHOUARI S.B., SHI M., MARTINEZ D., *Pattern recognition techniques for odor discrimination in gas sensor array*, [in:] *Encyclopedia of Sensors*, C.A. Grimes, E.C. Dickey, M.V. Pishko (Eds.), Vol. 10, American Scientific Publishers, USA, 2006, 1–17.

[48] GARDNER J.W., BARTLETT P.N., *Performance definition and standardization of electronic noses*, Sens. Actuators B, 33 (1996), 60–67.

[49] DAQI G., WEI CH., *Simultaneous estimation of odor classes and concentrations using an electronic nose with function approximation model ensembles*, Sens. Actuators B, 120 (2007), 584–594.

[50] KARLIK B., YÜKSEK K., *Fuzzy clustering neural networks for real-time odor recognition systems*, Journal of Automated Methods and Management in Chemistry (2007), 1–6.

[51] BURLB M.C., DOLEMANA B.J., SCHAFFER A., LEWIS N.S., *Assessing the ability to predict human percepts of odor quality from the detector responses of a conducting polymer composite-based electronic nose*, Sens. Actuators B, 72 (2001), 149–159.

[52] ZABIEGAŁA B., *Jakość powietrza wewnętrznego – lotne związki organiczne jako wskaźnik jakości powietrza*, *Polska Inżynieria Środowiska pięć lat po wstąpieniu do Unii Europejskiej*, 2 (2009), 303–315.

[53] MØLHAVE L., CLAUSEN G., BERGLUND B., DE CEAURRIZ J., KETTRUP A., LINDVALL T., MARONI M., PICKERING A.C., RISSE U., ROTHWEILER H., SEIFERT B., YOUNES M., *Total volatile organic compounds (TVOC), in indoor air quality investigations*, Indoor Air, 7 (4) (1997), 225–240.

[54] ARNOLD CH., HARMS M., GOSCHNICK J., A*ir quality monitoring and fire detection with the Karlsruhe electronic micronose KAMINA*, IEEE Sensors Journal, 2 (3) (2002), 179–188.

[55] ZAMPOLLI S., ELMI I., AHMED F., PASSINI M., CARDINALI G.C., NICOLETTI S., DORI L., *An electronic nose based on solid state sensor arrays for low-cost indoor air quality monitoring applications*, Sens. Actuators B, 101 (2004), 39–46.

[56] MENG F.L., ZHANG L., JIA Y., LIU J.Y., SUNG Y.F., LUO T., LI M.Q., LIU J.H., HUANG X.J., *Electronic chip based on self-oriented carbon nanotube microelectrode array to enhance the sensitivity of indoor air pollutants capacitive detection*, Sens. Actuators B, 153 (2011), 103–109.

[57] HERBERGER S., HEROLD M., ULMER H., BURDACK-FREITAG A., MAYER F., *Detection Of Human Effluents By A MOS Gas Sensor In Correlation To VOC Quantification By GC/MS*, Building and Environment, 45 (2010), 2430–2439.

[58] SASAHARA T., KATO H., SAITO A., NISHIMURA M., EGASHIRA M., *Development of a ppb-level sensor based on catalytic combustion for total volatile organic compounds in indoor air*, Sens. Actuators B, 126 (2007), 536–543.

[59] **MACIEJEWSKA M.,** *Rozpoznawanie zanieczyszczeń powietrza wewnętrznego metodami analizy wielowymiarowej*, *Nowoczesne rozwiązania w inżynierii i ochronie środowiska*. T. 2,

S. Anisimov (red.), Instytut Klimatyzacji i Ogrzewnictwa. Wydział Inżynierii Środowiska, Politechnika Wrocławska, Wrocław 2011, 49–54.
[60] WEIMAR U., GÖPEL W., *Chemical imaging II – Trends in practical multiparameter sensor systems*, Sens. Actuators B, 52 (1998), 143–161.
[61] OLIVIERI A.C., FABER N.M., FERRÉ J., BOQUÉ R., KALIVAS J.H., MARK H., *Uncertainty estimation and figures of merit for multivariate calibration*, Pure Appl. Chem., 78 (3) (2006), 633–661.
[62] WOYCZYŃSKI W.A., *A first course in statistics for signal analysis*, Birkhäuser, USA, 2006.
[63] SRIVASTAVA A.K., DRAVID V.P., *On the performance evaluation of hybrid and mono-class sensor arrays in selective detection of VOCs – A comparative study*, Sens. Actuators B, 117 (2006), 244–252.
[64] DUDA R.O., HART P.E., STORK D.G., *Pattern classification*, Wiley, New York 2001.
[65] BREZMES J., LLOBET E., AL-KHALIFA S., MALDONADO S., GARDNER J.W., *Gas sensing using support vector machines*, Studies in Fuzziness and Soft Computing, 177 (2005), 365–386.
[66] ROUSSEL S., FORSBERG G., GRENIER P., BELLON-MAUREL V., *Optimisation of electronic nose measurements.* Part II – *Influence of experimental parameters*, J. Food Eng., 39 (1999), 9–15.
[67] DISTANTE C., LEO M., SICILIANO P., PERSAUD K.C., *On the study of feature extraction methods for an electronic nose*, Sens. Actuators B, 87 (2002), 274– 288.
[68] DABLE B.K., BOOKSH K.S., CAVICCHI R., SEMANCIK S., *Calibration of microhotplate conductometric gas sesnors by non-linear multivariate regression methods*, Sens. Actuators B, 101 (2004), 284–294.
[69] **MACIEJEWSKA M.,** SZCZUREK A., OCHROMOWICZ Ł., *The charact.eristics of a "stop–flow" mode of sensor array operation using data with the best classification performance*, Sens. Actuators B, 141 (2009), 417–423.
[70] GARDNER J.W., *Detection of vapours and odours from a multisensor array using pattern recognition*. Part. 1 – *Principal component and cluster analysis*, Sens. Actuators B, 4 (1991), 109–115.
[71] BARBRI N.E., DURAN C., BREZMES J., CANELLAS N., RAMIREZ J.L., BOUCHIKHI B., LLOBET E., *Selectivity enhancement in multisensor systems using flow modulation technique*, Sensors, 8 (2008), 7369–7379.
[72] GUTIERREZ-OSUNA R., NAGLE H.T., *A method for evaluating data-preprocessing techniques for odor classification with an array of gas sensors*, IEEE Transactions on Systems, Man, and Cybernetics – Part B – Cybernetics, 29 (5) (1999), 626–632.
[73] HOSSEIN-BABAEI F., GHAFARINIA V., *Compensation for the drift-like terms caused by environmental fluctuations in the response of chemoresistive gas sensors*, Sens. Actuators B, 143 (2010), 641–648.
[74] ARTUSSON T., EKLÖV T., LUNDSTRÖM I., MÅRTENSSON P., SJÖSTRÖM M., HOLMBERG M., *Drift correction for gas sensors using multivariate methods*, Journal of Chemometrics, 14 (5–6) (2000), 711–723.
[75] ZIYATDINOV A., MARCO S., CHAUDRY A., PERSAUD K., CAMINAL P., PERERA A., *Drift compensation of gas sensor array data by common principal component analysis*, Sensors and Actuators B, Chemical, 146 (2) (2010), 460–465.
[76] VERGARA A., VEMBU S., AYHAN T., RYAN M.A., HOMER M.L., HUERTA R., *Chemical gas sensor drift compensation using classifier ensembles*, Sens. Actuators B, 166–167 (2012), 320–329.
[77] WOLFRUM E.J., MEGLEN R.M., PETERSON D., SLUITER J., *Metal oxide sensor arrays for the detection, differentiation and quantification of volatile organic compounds at sub-parts-per-million concentration levels*, Sens. Actuators B, 115 (2006), 322–329.
[78] CLIFFORD P.K., TUMA D.T., *Characteristics of semiconductor gas sensors I.steady state gas response*, Sens. Actuators B, 3 (1982/83), 233–254.
[79] LLOBET E., VILANOVA X., BREZMES J., SUEIRAS J.E., CORREIG X., *Transient response of thick-film tin oxide gas-sensors to multicomponent gas mixtures*, Sens. Actuators B, 47 (1998), 104–112.

[80] SZCZUREK A., KRAWCZYK B., **MACIEJEWSKA M.,** *VOCs classification based on committee of classifiers coupled with single sensor signals*, submitted to Journal of Chemometrics and Intelligent Laboratry Systems, 2012.

[81] **MACIEJEWSKA M.,** SZCZUREK A., Discrimination abilities of transient signal originating from single gas sensor, Proceedings of 5th WSEAS International Conference on Sensors and Signals (SENSIG '12), Sliema, Malta, September 7–9, 2012.

[82] RAMAN B., GUTIERREZ-OSUNA R., *Chemosensory processing in a spiking model of the olfactory bulb – Chemotopic convergence and center surround inhibition*, Neural Information Processing Systems, Vancouver, BC, Dec. 13–16, 2004.

[83] ZHANG W.M., HU J.S., SONG W.G., WAN L.J., D*etection of VOCs and their concentrations by a single SnO_2 sensor using kinetic information*, Sens. Actuators B, 123 (2007), 454–460.

[84] IONESCU R., LLOBET E., AL-KHALIFA S., GARDNER J.W., VILANOVA X., BREZMES J., CORREIG X., *Response model for thermally modulated tin oxide-based microhotplate gas sensor*, Sens. Actuators B, 95 (2003), 203–211.

[85] VERGARA A., MUEZZINOGLU M.K., RULKOV N., HUERTA R., *Information-theoretic optimization of chemical sensors*, Sens. Actuators B, 148 (2010), 298–306.

[86] MUEZZINOGLU M.K., VERGARA A., HUERTA R., RABINOVICH M.I., *A sensor conditioning principle for gas identification*, Sens. Actuators B, 146 (2010), 472–476.

[87] RAMAN B., HERTZ J.L., BENKSTEIN K.D., SEMANCIK S., *Bioinspired methodology for artificial olfaction*, Anal. Chem., 80 (2008), 8364–8371.

[88] GRAMM A., SCHÜTZE A., *High performance solvent vapor identification with a two sensor array using temperature cycling and pattern classification*, Sens. Actuators B, 95 (2003), 58–65.

[89] MONTOLIU I., TAULER R., PADILLA M., PARDO A., MARCO S., *Multivariate curve resolution applied to temperature-modulated metal oxide gas sensors*, Sens. Actuators B, 145 (2010), 464–473.

[90] VILANOVA X., LLOBET E., ALCUBILLA R., SUEIRAS J.E., CORREIG X., *Analysis of the conductance transient in thick-film tin oxide gas sensors*, Sens. Actuators B, 31 (1996), 175–180.

[91] LLOBET E., BREZMESA J., VILANOVA X., SUEIRASB J.E., CORREIGA X., *Qualitative and quantitative analysis of volatile organic compounds using transient and steady-state responses of a thick-film tin oxide gas sensor array*, Sens. Actuators B, 41 (1997), 13–21.

[92] MUEZZINOGLU M.A., VERGARA A.V., HUERTA R., RULKOV N., RABINOVICH M.I., SELVERSTON A., ABARBANEL H.D.I., *Acceleration of chemo-sensory information processing using transient features*, Sens. Actuators B, 137 (2009), 507–512.

[93] GARDNER J.W., *A non-linear diffusion-reaction model of electrical conduction in semiconductor gas sensor*, Sens. Actuators B, 1 (1990), 166–170.

[94] BOTRE B.A., GHARPURE D.C., SHALIGRAM A.D., *Embedded electronic nose and supporting software tool for its parameter optimization*, Sens. Actuators B,146 (2010), 453–459.

[95] HELLI O., SIADAT M., LUMBRERAS M., *Qualitative and quantitative identification of H_2S/NO_2 gaseous components in different reference atmospheres using a metal oxide sensor array*, Sens. Actuators B, 103 (2004), 403–408.

[96] ALEIXANDRE M., SAYAGO I., HORRILLO M.C., FERNÁNDEZ M.J., ARÉS L., GARCÍA M., GUTIÉRREZ J., *Analysis of neural networks and analysis of feature selection with genetic algorithm to discriminate among pollutant gas*, Sens. Actuators B, 103 (2004), 122–128.

[97] ROUSSEL S., FORSBERG G., STEINMETZ V., GRENIER P., BELLON-MAUREL V., *Optimisation of electronic nose measurements*. Part I – *Methodology of output feature selection*, J. Food Eng., 37 (1998), 207–222.

[98] GUALDRÓN O., LLOBET E., BREZMES J., VILANOVA X., CORREIG X., *Coupling fast variable selection method to neural network-based classifiers – Application to multisensor systems*, Sens. Actuators B, 114 (2006), 522–529.

[99] PARDO M., SBERVEGLIERI G., *Comparing the performance of different features in sensor arrays*, Sens. Actuators B, 123 (2007), 437–443.
[100] KLINGVALL R., LUNDSTROM I., ERIKSSON M., *Robust gas detection at sub ppm concentrations*, Sens. Actuators B, 160 (2011), 571–579.
[101] GUO D., ZHANG D., ZHANG L., *Sparse representation-based classification for breath sample identification*, Sensors and Actuators B, 158 (2011), 43–53.
[102] GUTIERREZ-OSUNA R., NAGLE H.T., SCHIFFMAN S.S., *Transient response analysis of an electronic nose using multi-exponential models*, Sens. Actuators B, 61 (1999), 170–182.
[103] CARMEL L., LEVY S., LANCET D., HAREL D., *A feature extraction model for chemical sensors in electronic noses*, Sens. Actuators B, 93 (2003), 67–76.
[104] VEMBU S., VERGARA A., MUEZZINOGLU M.K., HUERTA R., *On time series features and kernels for machine olfaction*, Sensors and Actuators B, in press.
[105] DI NATALE C., MARCO S., DAVIDE F., D'AMICO A., *Sensor array calibration time reduction by dynamic modelling*, Sens. Actuators B, 24–25 (1995), 578–583.
[106] IONESCU R., LLOBET E., VILANOVA X., BREZMES J., SUEIRAS J. E., CALDERER J., CORREIG X., *Quantitative analysis of NO_2 in the presence of CO using a single tungsten oxide semiconductor sensor and dynamic signal processing*, Analyst, 127 (2002), 1237–1246.
[107] MARTINELL E., FALCONI C., D'AMICO A., DINATALE C., *Feature extraction of chemical sensors in phase space*, Sens. Actuators B, 95 (2003), 132–139.
[108] ZHANG S., XIE CH., HU M., LI H., BAI Z., ZENG D., *An entire feature extraction method of metal oxide gas sensors*, Sens. Actuators B, 132 (2008), 81–89.
[109] ZHANG S., XIE CH., LI H., BAI Z., XIA X., ZENG D., *A reaction model of metal oxide gas sensors and a recognition method by pattern matching*, Sens. Actuators B, 135 (2009), 552–559.
[110] VERGARA A., LLOBET E., MARTINELLI E., DI NATALE C., D'AMICO A., CORREIG X., *Feature extraction of metal oxide gas sensors using dynamic moments*, Sens. Actuators B, 122 (2007), 219–226.
[111] SZCZUREK A., **MACIEJEWSKA M.,** *Single gas sensor measurement system with different signal subsampling approaches*, Proceedings of 5th WSEAS International Conference on Sensors and Signals (SENSIG '12), Sliema, Malta, September 7–9, 2012.
[112] HINES E.L., LLOBET E., GRADNER J.W., *Electronic noses – a review of signal processing techniques*, IEE Proc. Circuits Devices Syst., 146 (6) (1999), 297–310.
[113] PHAISANGITTISAGUL E., NAGLE H. T., AREEKUL V., *Intelligent method for sensor subset selection for machine olfaction*, Sens. Actuators B, 145 (2010), 507–515.
[114] VERGARA A., LLOBET E., *Feature selection and sensor array optimization in machine olfaction*, [in:] *Intelligent systems for machine olfaction – tools and methodologies*, E.L. Hines, M.S. Leeson (Eds.), IGI Global, USA, 2011.
[115] GUYON I., ELISSEEFF A., *An introduction to variable and feature selection*, Journal of Machine Learning Research, 3 (2003), 1157–1182.
[116] SHI M., BERMAK A., BELHOUARI S.B., CHAN P.C.H., *Gas identification based on committee machine for microelectronic gas sensor*, IEEE Trans. on Instrumentation and Measurement, 55 (5) (2006), 1786–1793.
[117] GUALDRON O., BREZMES J., LLOBET E., AMARI A., VILANOVA X., BOUCHIKHI B., CORREIG X., *Variable selection for support vector machine based multisensor systems*, Sens. Actuators B, 122 (2007), 259–268.
[118] CORCORAN P., ANGLESEA J., ELSHAW M., *The application of genetic algorithms to sensor parameter selection for multisensor array configuration*, Sens. Actuators B, 76 (1999), 57–66.
[119] ELKÖV T., MÅRTENSSON P., LUNDSTRÖM I., *Selection of variables for interpreting multivariate gas sensor data*, Anal. Chim. Acta, 381 (1999), 221–232.

[120] ALIZADEH T., *Chemiresistor sensors array optimization by using the method of coupled statistical techniques and its application as an electronic nose for some organic vapors recognition*, Sens. Actuators B, 143 (2010), 740–749.
[121] ZHANG S., XIE CH., ZENG D., LI H., LIU Y., CAI S., *A sensor array optimization method for electronic noses with sub-arrays*, Sens. Actuators B, 142 (2009), 243–252.
[122] CHAUDRY A.N., HAWKINS T.M., TRAVERS P.J., *A method for selecting an optimum sensor array*, Sens. Actuators B, 69 (2000), 236–242.
[123] XU Z., SHI X., LU S., *Integrated sensor array optimization with statistical evaluation*, Sens. Actuators B, 149 (2010), 239–244.
[124] GENG Z., YANG F., WU N., *Optimum design of sensor arrays via simulation-based multivariate calibration*, Sens. Actuators B, 156 (2011), 854–862.
[125] SZCZUREK A., **MACIEJEWSKA M.,** FLISOWSKA-WIERCIK B., *Method of gas mixtures discrimination based on sensor array, temporal response and data driven approach*, Talanta, 83 (2011), 916–923.
[126] GARDNER J.W., BOILOT P., HINES E.L., E*nhancing electronic nose performance by sensor selection using a new integer-based genetic algorithm approac*h, Sens. Actuators B, 106 (2005), 114–121.
[127] SZCZUREK A., **MACIEJEWSKA M.,** FLISOWSKA-WIERCIK B., BODZOJ Ł., *The stop-flow mode of operation applied to a single chemiresistor*, Sens. Actuators B, 148 (2010), 522–530.
[128] XU Z., LU S., *Multi-objective optimization of sensor array using genetic algorithm*, Sensors and Actuators B, 160 (2011), 278–286.
[129] DASH M., LIU H., *Feature selection for classification, Intelligent data analysis*, 1 (1997), 131–156.
[130] SINKOV N.A., HARYNUK J.J., *Cluster resolution – A metric for automated, objective and optimized feature selection in chemometric modeling*, Talanta, 83 (2011), 1079–1087.
[131] CHO J.H., KURUP P.U., *Decision tree approach for classification and dimensionality reduction of electronic nose data*, Sens. Actuators B, 6 (2011), 542–548.
[132] PARDO M., SBERVEGLIERI G., *Random forests and nearest shrunken centroids for the classification of sensor array data*, Sens. Actuators B, 131 (2008), 93–99.
[133] BENEDETTI S., BURATTI S., SPINARDI A., MANNINO S., MIGNANI I., *Electronic nose as a non-destructive tool to characterize peach cultivars and to monitor their ripening stage during shelf-life*, Postharvest Biology and Technology, 47 (2008), 181–188.
[134] CAMPAGNOLI A., CHELI F., POLIDORI C., ZANINELLI M., ZECCA O., SAVOINI G., PINOTTI L., DELL'ORTO V., *Use of the electronic nose as a screening tool for the recognition of durum wheat naturally contaminated by deoxynivalenol – A preliminary approach*, Sensors, 11 (2011), 4899–4916.
[135] SZCZUREK A., **MACIEJEWSKA M.,** OCHROMOWICZ Ł., *Sensor array data profiling for gas identification*, Talanta, 78 (2009), 840–845.
[136] SIVAKUMAR A., KANNAN K., *A novel feature selection technique for number classification problem using PNN-A plausible scheme for boiler flue gas analysis*, Sens. Actuators B, 139 (2009), 280–286.
[137] LLOBET E., BREZMES J., GUALDRÓN O., VILANOVA X., CORREIG X., *Building parsimonious fuzzy ARTMAP models by variable selection with a cascaded genetic algorithm – application to multisensor systems for gas analysis*, Sens. Actuators B, 99 (2004), 267–272.
[138] NISHIKAWA T., HAYASHI T., NAMBO H., KIMURA H., OYABU T., *Feature extraction of multi-gas sensor responses using genetic algorithm*, Sens. Actuators B, 64 (2000), 2–7.
[139] YOUNG R.C., BUTTNER W.J., LINNELL B.R., RAMESHAM R., *Electronic nose for space program applications*, Sens. Actuators B, 93 (2003), 7–16.
[140] NÆS T., ISAKSSON T., FEARN T., DAVIES T., *A user-friendly guide to multivariate calibration and classification*, Chichester, NIR Publications, 2002.

[141] GOODNER K.L., DREHER J.G., ROUSEFF R.L., *The danger of creating false classifications due to noise in electronic nose and similar multivariate analyses*, Sens. Actuators B, 80 (2001), 261–266.
[142] GARDNER J.W., BARTLETT P.N., *A brief history of electronic noses*, Sens. Actuators B, 18–19 (1994), 211–220.
[143] YING Z., JIANG Y., DU X., XIE G., YU J., TAI H., *Polymer coated sensor array based on quartz crystal microbalance for chemical agent analysis*, European Polymer Journal, 44 (2008), 1157–1164.
[144] PARDO M., SISK B.C., SBERVEGLIERI G., LEWIS N.S., *Comparison of Fisher's linear discriminant to multilayer perceptron networks in the classification of vapors using sensors array data*, Sens. Actuators B, 115 (2006), 647–655.
[145] TAURINO A.M., DISTANTE C., SICILIANO P., VASANELLI L., *Qantitative and qualitative analysis of VOCs mixtures by means of a microsensors array and different evaluation methods*, Sens. Actuators B, 93 (2003), 117–125.
[146] SZCZUREK A., **MACIEJEWSKA M.,** *Humidity as a discriminative factor in alcohols recognition*, Environment Protection Engineering, 29 (2) (2003), 125–139.
[147] HAMMOND M.H., JOHNSON K.J., ROSE-PEHRSSON S.L., ZIEGLER J., WALKER H., CAUDY K., GARY D., TILLETT D., A *novel chemical detector using cermet sensors and pattern recognition methods for toxic industrial chemicals*, Sens. Actuators B, 116 (2006), 135–144.
[148] **MACIEJEWSKA M.,** SZCZUREK A., KERENYI Z., *Utilisation of first principal component extracted from gas sensor measurements as a process control variable in wine fermentation*, Sens. Actuators B, 115 (2006), 170–177.
[149] WEI G., TANG Z., CHAN P.C.H., YU J., *A blind source separation based micro gas sensor array modeling method*, Lecture Notes in Computer Science, 3173 (2004), 696–701.
[150] DI NATALE C., MARTINELLI E., D'AMICO A., *Counteraction of environmental disturbances of electronic nose data by independent component analysis*, Sens. Actuators B, 82 (2002), 158–165.
[151] KERMIT M., TOMIC O., *Independent component analysis applied on gas sensor array measurement data*, IEEE Sensors Journal, 2 (3) (2003), 218–228.
[152] MENG X., I*CA algorithm based on intelligent electronic nose in the mixed gas of feature extraction*, 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), 23–25 September 2010, 1–4
[153] SHAO X., WANG W., HOU Z., CAI W., *A new regression method based on independent component analysis*, Talanta, 69 (2006), 676–680.
[154] FALASCONI M., PARDO M., VEZZOLI M., SBERVEGLIERI G., *Cluster validation for electronic nose data*, Sens. Actuators B, 125 (2007), 596–606.
[155] DUTTA R., DASA A., STOCKS N.G., MORGAN D., *Stochastic resonance-based electronic nose – A novel way to classify bacteria*, Sens. Actuators B, 115 (2006), 17–27.
[156] DUTTA R., HINES E.L., GARDNER J.W., BOILOT P., *Bacteria classification using Cyranose 320 electronic nose*, BioMedical Engineering OnLine, 2002, 1–4.
[157] BAUERSFELD N., KRAMER K.D., PATZWAHL S., *Methods of computational intelligence to give qualitative and quantitative statements of gas concentrations at a high temperature sensor*, HIS 2005, Proceedings of the Fifth International Conference on Hybrid Intelligent Systems, Rio de Janeiro, Brazil, 6–9 November 2005, 1–6.
[158] BARKÓ G., ABONYI J., HLAVAY J., *Application of fuzzy clustering and piezoelectric chemical sensor array for investigation on organic compounds*, Analytica Chimica Acta, 398 (1999), 219–226.
[159] LI CH., SCHMIDT N.E., GITAITIS R., *Detection of onion postharvest diseases by analyses of headspace volatiles using a gas sensor array and GC-MS, LWT*, Food Science and Technology, 44 (2011), 1019–1025.
[160] LI CH., KREWERB G.W., JI P., SCHERMD H., KAYSE S.J., *Gas sensor array for blueberry fruit disease detection and classification*, Postharvest Biology and Technology, 55 (2010), 144–149.

[161] FEND R., BESSANT C., WILLIAMS A.J., WOODMANA A.C., *Monitoring haemodialysis using electronic nose and chemometrics*, Biosensors and Bioelectronics, 19 (2004), 1581–1590.
[162] SAHGAL N., MAGAN N., *Fungal volatile fingerprints – Discrimination between dermatophyte species and strains by means of an electronic nose*, Sens. Actuators B, 131 (2008), 117–120.
[163] HOLMBERG M., WINQUIST F., LUNSTRÖM I., GARDNER J.W., HINES E.L., *Identification of paper quality using a hybrid electronic nose*, Sens. Actuators B, 26–27 (1995), 246–249.
[164] PARDO M., SBERVEGLIERI G., GARDINI S., DALCANALE E., *A hierarchical classification scheme for an electronic nose*, Sens. Actuators B, 69 (2000), 359–365.
[165] SZCZUREK A., **MACIEJEWSKA M.,** BODZOJ Ł., FLISOWSKA-WIERCIK B., *A Concept of a sensor system for determining composition of organic solvents*, IEEE Sensors Journal, 10 (5) (2010), 924–933.
[166] DUMITRESCU D., LAZZERINI B., MARCELLONI F., *A fuzzy hierarchical classification system for olfactory signals*, Pattern Analysis and Applications, 3 (4) (2000), 325–334.
[167] VERGARA A., LLOBET E., BREZMES J., IVANOV P., CANE C., GRACIA I., VILANOVA X., CORREIG X., *Quantitative gas mixture analysis using temperature-modulated micro-hotplate gas sensors – Selection and validation of the optimal modulating frequencies*, Sens. Actuators B, 123 (2007), 1002–1016.
[168] GUO D., ZHANG D., ZHANG L., *An LDA based sensor selection approach used in breath analysis system*, Sensors and Actuators B, 157 (2011), 265–274.
[169] POLIKAR R., SHINAR R., UDPA L., PORTER M.D., *Artificial Intelligence methods for selection of an optimized sensor array for identification of volatile organic compounds*, Sens. Actuators B, 80 (2001), 243–254.
[170] SHI M., GUO B., BERNAK A., *Redundancy analysis of tin oxide gas sensor array*, Proceedings of the 3rd IEEE International Workshop on Electronic Design, Test and Applications, Delta 2006, Kuala Lumpur, Malaysia, 17–19 January 2006, 1–4.
[171] BICEGO M., TESSARI G., TECCHIOLLI G., BETTINELLI M., *A comparative analysis of basic pattern recognition techniques for the development of small size electronic nose*, Sens. Actuators B, 85 (2002), 137–144.
[172] SHI M., BERMAK A., BELHOUARI S.B., CHAN P.C.H., G*as identification based on committee machine for microelectronic gas sensor*, IEEE Trans. on Instrumentation and Measurement, 55 (5) (2006), 1786–1793.
[173] ALIPPI C., PELOSI G., ROVERI M., *Computational intelligence techniques to detect toxic gas presence*, Proceedings of CIMSA 2006 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, La Coruna, Spain, 12–14 July 2006.
[174] RONCAGLIA A., ELMI I., DORI L., RUDAN M., *Adaptive K-NN for the Detection of Air Pollutants With a Sensor Array*, IEEE Sensors Journal, 4 (2) (2004), 248–256.
[175] BRAHIM-BELHOUARI S., BERMAK A., WEI G., CHAN P.C.H., A *comparative study of density models for gas identification using microelectronic gas sensor*, Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology, 2003, ISSPIT 2003, 131–141.
[176] BRAHIM-BELHOUARI S., BERMAK A., *Gas identification using density models*, Pattern Recognition Letters, 26 (2005), 699–706.
[177] BRAHIM-BELHOUARI S., BERMAK A., SHI M., CHAN P.C.H., *Fast and robust gas identification system using an integrated gas sensor technology and Gaussian mixture models*, IEEE Sensors Journal, 5 (6) (2005), 1433–1444.
[178] DISTANTE C., ANCONA N., SICILIANO P., *Support vector machines for olfactory signals recognition*, Sens. Actuators B, 88 (2003), 30–39.
[179] PARDO M., SBERVEGLIERI G., *Classification of electronic nose data with support vector machines*, Sens. Actuators B, 107 (2005), 730–737.

[180] BRUDZEWSKI K., OSOWSKI S., MARKIEWICZ T., ULACZYK J., *Classification of gasoline with supplement of bio-products by means of an electronic nose and SVM neural network*, Sens. Actuators B, 113 (2006), 135–141.

[181] WANG X., YE M., DUANMU C.J., *Classification of data from electronic nose using relevance vector machines*, Sens. Actuators B, 140 (2009), 143–148.

[182] TRINCAVELLI M., CORADESHI S., LOUTFI A., *Odour classification system for continuous monitoring applications*, Sens. Actuators B, 139 (2009), 265–273.

[183] KUNCHEVA L., BEZDEK J.C., DUIN R.P.W., *Decision templates for multiple classifier fusion – An experimental comparison*, Pattern Recognition, 34 (2) (2001), 299–314.

[184] WOZNIAK M., ZMYSLONY M., *Combining classifiers using trained fuser – analytical and experimental results*, Neural Network World, 13 (7) (2010), 925–934.

[185] HIRAYAMA V., RAMIREZ-FERNANDEZ F.J., SALCEDO W.J., *Committee machine for LPG calorific power classification*, Sens. Actuators B, 116 (2006), 62–65.

[186] SHI M., BERMAK A., CHANDRASEKARAN S., AMIRA A., BRAHIM-BELHOUARI S., *A committee machine gas identification system based on dynamically reconfigurable FPGA*, IEEE Sensors Journal, 8 (4) (2008), 403–414.

[187] MUTIHAC L., MUTIHAC R., *Mining in chemometrics*, Analytica Chimica Acta, 612 (2008), 1–18

[188] CARMEL L., SEVER N., LANCET D., HAREL D., *An eNose algorithm for identifying chemicals and determining their concentration*, Sens. Actuators B, 93 (2003), 77–83.

[189] QUIN S.J., WU Z.J., *A new approach to analyzing gas mixtures*, Sens. Actuators B, 80 (2001), 85–88.

[190] HIROBAYASHI S., KADIR M.A., YOSHIZAWA T., YAMABUCHI T., *Verification of a logarithmic model for estimation of gas concentrations in a mixture for a tin oxide gas sensor response*, Sens. Actuators B, 92 (2003), 269–278.

[191] BOEKER P., HORNER G., RÖSLER S., *Monolithic sensor array based on a quartz microbalance transducer with enhanced sensitivity for monitoring agricultural emissions*, Sens. Actuators B, 70 (2000), 37–42.

[192] ZHANG H., CHANG M., WANG J., YE S., *Evaluation of peach quality indices using an electronic nose by MLR, QPST and BP network*, Sens. Actuators B, 134 (2008), 332–338.

[193] ZHANG P., LEE CH., VERWEIJ H., AKBAR S.A., HUNTER G., DUTTA P.K., *High temperature sensor array for simultaneous determination of O_2, CO, and CO_2 with kernel ridge regression data analysis*, Sens. Actuators B, 123 (2007), 950–963.

[194] FRANK M.L., FULKERSON M.D.,. PATTON B.R, DUTTA P.K., *TiO_2-based sensor arrays modeled with nonlinear regression analysis for simultaneously determining CO and O_2 concentrations at high temperatures*, Sens. Actuators B, 87 (2002), 471–479.

[195] DOUGHERTY A.W., BEACH E., MORRIS P.A., PATTON B.R., *Efficient orthogonalization in gas sensor arrays using reciprocal kernel support vector regression*, Sens. Actuators B, 149 (2010), 264–271.

[196] CRUZ ORTIZ M., SARABIA L.A., HERRERO A., *Robust regression techniques – A useful alternative for the detection of outlier in chemical analysis*, Talanta, 70 (2006), 499–512.

[197] ALSTRØMA T.S., LARSENA J., NIELSEN C.H., LARSEN N.B., *Data-driven modeling of nano-nose gas sensor arrays*, Signal Processing, Sensor Fusion, and Target Recognition, 19, 7697 (1), 76970U, The International Society for Optical Engineering, 2010.

[198] GETINO J., HORRILLO M.C, GUTIERREZ J., ARES L., ROBLA J.I., GARCIA C., SAYAGO I., *Analysis of VOCs with a tin oxide sensor array*, Sens. Actuators B, 43 (1997), 200–205.

[199] CAPONE S., SICILIANO P., BARSAN N., WEIMAR U., VASANELLI L., *Analysis of CO and CH_4 gas mixtures by using a micromachined sensor array*, Sens. Actuators B, 78 (2001), 40–48.

[200] WANG Q., ZHANG H., CHENG Y., *Design and Implementation a real-time electronic nose system*, Proceedings of I2MTC 2009, International Instrumentation and Measurement Technology Conference, Singapore, 5–7 May 2009.

[201] NIEBLING G., MÜLLER R., *Non-linear signal evaluation with linear regression techniques for redundant signals*, Sens. Actuators B, 24–25 (1995), 805–807.

[202] THEN D., VIDIC A., ZIEGLER CH., *A highly sensitive self-oscillating cantilever array for the quantitative and qualitative analysis of organic vapor mixtures*, Sens. Actuators B, 117 (2006), 1–9.

[203] LOZANO J., SANTOS J.P., ARROYO T., AZNAR M., CABELLOS J.M., GIL M., HORRILLO M.C., *Correlating e-nose responses to wine sensorial descriptors and gas chromatography–mass spectrometry profiles using partial least squares regression analysis*, Sens. Actuators B, 127 (2007), 267–276.

[204] SONG S., ZHANG X., HAYAT K., JIA CH., XIA S., ZHONG F., XIAO Z., TIAN H., NIU Y., *Correlating chemical parameters of controlled oxidation tallow to gas chromatography–mass spectrometry profiles and e-nose responses using partial least squares regression analysis*, Sens. Actuators B, 147 (2010), 660–668.

[205] AISHIMA T., *Correlating sensory attributes to gas chromatography–mass spectrometry profiles and e-nose responses using partial least squares regression analysis*, Journal of Chromatography A, 1054 (2004), 39–46.

[206] BOHOLT K., ANDREASEN K., DEN BERG F., HANSEN T., *A new method for measuring emission of odour from a rendering plant using the Danish Odour Sensor System (DOSS), artificial nose*, Sens. Actuators B, 106 (2005), 170–176.

[207] SOHN J.H., DUNLOP M., HUDSON N., KIM T.I., YOO Y.H., *Non-specific conducting polymer-based array capable of monitoring odour emissions from a biofiltration system in a piggery building*, Sens. Actuators B, 135 (2009), 455–464.

[208] RIVERA D., ALAM M.K., YELTON W.G., STATON A.W., SIMONSON R.J., *Use of classical east squares/partial east squares (CLS/PLS), hybrid algorithm for calibration and calibration maintenance of surface acoustic wave (SAW), devices*, Sens. Actuators B, 99 (2004), 480-490.

[209] NIEBLING G., *Identification of gases with classical pat tern-recognition methods and artificial neural networks*, Sens. Actuators B, 18–19 (1994), 259–263.

[210] MICONE P.G., GUY C., *Odour quantification by a sensor array – An application to landfill gas odours from two different municipal waste treatment works*, Sens. Actuators B, 120 (2007), 628–637.

[211] BRUDZEWSKI K., OSOWSKI S., *Gas analysis system composed of a solid-state sensor array and hybrid neural network structure*, Sens. Actuators B, 55 (1999), 38–46.

[212] DAQI G., SHUYAN W., YAN J., *An electronic nose and modular radial basis function network classifiers for recognizing multiple fragrant materials*, Sens. Actuators B, 97 (2004), 391–401.

[213] GAO D., MINGMING CH., JI Y., *Simultaneous estimation of classes and concentrations of odors by an electronic nose using combinative and modular multilayer perceptrons*, Sens. Actuators B, 107 (2005), 773–781.

[214] LEE D.S., JUNG J.K., LIM J.W., HUH J.S., LEE D.D., *Recognition of volatile organic compounds using SnO_2 sensor array and pattern recognition analysis*, Sens. Actuators B, 77 (2001), 228–236.

[215] HAMMOND J., MARQUIS B., MICHAELS R., OICKLE B., SEGEE B., VALETINO J., BUSHWAY A., CAMIRE M.E., DAVIS-DENTICI K., *A semiconducting metal-oxide array for monitoring fish freshness*, Sens. Actuators B, 84 (2002), 113–122.

[216] ROSE-PEHRSSON S.L., SHAFFER R.E., HART S.J., WILLIAMS F.W., GOTTUK D.T., STREHLEN B.D., HILL S.A., *Multi-criteria fire detection systems using a probabilistic neural network*, Sens. Actuators B, 69 (2000), 325–335.

[217] GULBAG A., TEMURTAS F., YUSUBOV I., *Quantitative discrimination of the binary gas mixtures using a combinational structure of the probabilistic and multilayer neural networks*, Sens. Actuators B, 131 (2008), 196–204.

[218] FERNANDEZ M.J., FONTECHA J. L., SAYAGO I., ALEIXANDRE M., LOZANO J., GUTIERREZ J., GRACIA I., CANE C., HORRILLO M.C., *Discrimination of volatile compounds through an electronic nose based on ZnO SAW sensors*, Sens. Actuators B, 127 (2007), 277–283.

[219] DAVIDE F.A.M., DI NATALE C., D'AMICO A., *Self-organising sensory maps in odor classification mimicking*, Biosensors Bioelectronics, 10 (1995), 203–218.
[220] CAPONE S., ZUPPA M., PRESICCE D.S., FRANCIOSO L., CASINO F., SICILIANO P., *Metal oxide gas sensor array for the detection of diesel fuel in engine oil*, Sens. Actuators B, 131 (2008), 125–133.
[221] ORTEGA A., MARCO S., SUNDIC T., SAMITIER J., *New pattern recognition system designed for electronic noses*, Sens. Actuators B, 69 (2000), 302–307.
[222] DI NATALE C., MACAGNANO A., D'AMICO A., DAVIDE F., *Electronic-nose modeling and data analysis using a self-organizing map*, Meas. Sci. Technol., 8 (1997), 1236–1243.
[223] SIRIPATRAWAN U., *Self-organizing algorithm for classification of packaged fresh vegetable potentially contaminated with foodborne pathogens*, Sens. Actuators B, 128 (2008), 435–441.
[224] ORTEGA A., MARCO S., PERERA A., SUNDIC T., PARDO A., SAMITIER J., *An intelligent detector based on temperature modulation of a gas sensor with a digital signal processor*, Sens. Actuators B, 78 (2001), 32–39.
[225] ZUPPA M., DISTANTE C., SICILIANO P., PERSAUD K.C., *Drift counteraction with multiple self-organising maps for an electronic nose*, Sens. Actuators B, 98 (2004), 305–317.
[226] LLOBET E., BREZMES J., IONESCU R., VILANOVA X., AL-KHALIFA S., GARDNER J.W., BÂRSAN N., CORREIG X., *Wavelet transform and fuzzy ARTMAP-based pattern recognition for fast gas identification using a micro-hotplate gas sensor*, Sens. Actuators B, 83 (1–3) (2002), 238–244.
[227] LLOBET E., HINES E.L., GARDNER J.W, BARTLETT P.N., MOTTRAM T.T, *Fuzzy ARTMAP based electronic nose data analysis*, Sens. Actuators B, 61 (1–3), 183–190.
[228] DISTANTE C., SICILIANO P., VASANELLI L., *Odor discrimination using adaptive resonance theory*, Sens. Actuators B, 69 (2000), 248–252.
[229] BRO R., PARAFAC. T*utorial and applications, chemometrics and intelligent laboratory systems*, 38 (1997), 149–171.
[230] PADILLA M., MONOLIU I., PARDO A., PERERA A., MARCO S., *Feature extraction on three way enose signals*, Sens. Actuators B, 116 (2006), 145–150.
[231] SKOV T., BRO R., *A new approach for modeling sensor based data*, Sens. Actuators B, 106 (2005), 719–729.
[232] BURIAN C., BREZMES J., VINAIXA M., CANELLASS N., LLOBET E., VILANOVA X., CORREIG X., *MS-electronic nose performance improvement using the retention time dimension and two-way and three way data processing methods*, Sens. Actuators B, 143 (2010), 759–768.
[233] KORONACKI J., ĆWIK J., *Statystyczne systemy uczące się*, WNT, Warszawa 2005.
[234] WETCHAKUN K., SAMERJAI T., TAMAEKONG N., LIEWHIRAN C., SIRIWONG C., KRUEFU V., WISITSORAAT A., TUANTRANONT A., PHANICHPHANT S., *Semiconducting metal oxides as sensors for environmentally hazardous gases*, Sens. Actuators B, 160 (2011), 580–591.
[235] FINE G.F., CAVANAGH L.M., AFONJA A., BINIONS R., *Metal oxide semi-conductor gas sensors in environmental monitoring*, Sensors, 10 (2010), 5469–5502.
[236] KAMIONKA M., BREUIL P., PIJOLAT C., *Calibration of a multivariate gas sensing device for atmospheric pollution measurement*, Sens. Actuators B, 118 (2006), 323–327.
[237] ELMI I., ZAMPOLLI S., COZZANI E., MANCARELLA F., CARDINALI G.C., *Development of ultra-low-power consumption MOX sensors with ppb-level VOC detection capabilities for emerging applications*, Sens. Actuators B, 135 (2008), 342–351.

Data analysis in sensor measurements of air pollutants

Reliable information about the environment has its value in the world today. The objective source of such information is the measurement data. More and more frequently modern measuring instruments provide complex data sets. The range of available information is increased in this way, because its content is proportional to data complexity. However, professional data analysis is required to achieve the transition between this kind of data and the information. The objective of the monograph was to demonstrate that wide spectrum of information about air pollution may be retrieved from the gas sensor measurement data with the appropriate data analysis methods. Gas sensor arrays composed of different partially selective sensors are an example of the source of complex measurement data. The analyzed data was provided by gas sensor array which consisted of fifteen commercially available semiconductor gas sensors (TGS). The pollution of interest was caused by volatile organic compounds. This choice allowed considering many kinds of environmental information. The chemical identity and the concentration of contaminants were addressed. Additionally, the alternative indicators of air pollution were investigated, e.g. the categorization of pollutants and concentration of organic carbon atoms. Proposing new indexes of this kind is crucial for the development of novel sensor systems dedicated to environmental measurements. The methodology of data analysis utilized the concept of pattern recognition. In the descriptive part of the monograph, the pattern recognition approaches were discussed applied in the domain of gas sensor data analysis. They are adaptable to air pollution measurement with sensor systems. The second part of the work focused on the data analysis. The effectiveness of environmental information retrieval was investigated in function of the design of pattern recognition system. Various strategies of feature space construction were tested in combination with a number of classification and regression methods. Simple but efficient solutions were searched for. Based on the obtained results, the gas sensor array coupled with the data analysis allows attaining wide range of information about air pollution. This combination may become the best environmental monitoring solution in the near future.