PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1076 2005
TAKSONOMIA 12
Klasyfikacja i analiza danych – teoria i zastosowania

Marek Walesiak
Akademia Ekonomiczna we Wrocławiu

PROBLEMY SELEKCJI I WAŻENIA ZMIENNYCH W ZAGADNIENIU KLASYFIKACJI

1. Wstęp

Głównym celem klasyfikacji jest badanie podobieństwa lub odrębności obiektów i ich zbiorów. Celem tym jest więc podział zbioru obiektów na klasy zawierające obiekty podobne ze względu na obserwacje na zmiennych (tzw. klasy względnie jednorodne). Obiekty znajdujące się w różnych klasach powinny być jak najmniej podobne. Postuluje się, aby wyodrębnione klasy spełniały kryteria wewnętrznej spójności i zewnętrznej izolacji (por. [Gordon 1999, s. 3]). Wybór zmiennych jest jednym z najważniejszych, a zarazem najtrudniejszych zagadnień. Od jakości zestawu zmiennych zależy bowiem wiarygodność ostatecznych wyników klasyfikacji i trafność podejmowanych na ich podstawie decyzji. W procedurze klasyfikacji należy uwzględnić tylko te zmienne, które mają zdolność dyskryminacji zbioru obiektów. Podejście polegające na uwzględnieniu jak największej liczby zmiennych jest nieuzasadnione. Dodanie do zbioru jednej lub kilku nieistotnych zmiennych nie pozwala na odkrycie w zbiorze obiektów właściwej struktury klas (zob. [Milligan 1994; 1996, s. 348]).

W zagadnieniu wyboru zmiennych na potrzeby klasyfikacji zbioru obiektów na względnie jednorodne klasy wyróżnia się trzy podejścia [Grabiński 1992, s. 42; Gnanadesikan, Kettenring, Tsao 1995]), są nimi:

1. Wprowadzenie zróżnicowanych wag dla poszczególnych zmiennych wyrażających ich relatywną ważność.
2. Selekcja zmiennych – dobór mniejszej liczby zmiennych przez eliminację tych, które nie mają zdolności dyskryminacji zbioru obiektów. Zagadnienie selekcji zmiennych jest szczególnym przypadkiem ważenia zmiennych, ponieważ zmienne usunięte otrzymują wagę 0, a zmienne wybrane wagę 1.

3. Zastąpienie oryginalnych zmiennych nowymi „sztucznymi" zmiennymi o pożądanych właściwościach (wykorzystuje się tutaj analizę czynnikową i analizę głównych składowych).

W artykule, głównie na przykładzie wygenerowanych danych w dwuwymiarowej przestrzeni zmiennych, wskazano ograniczenia, które należy wziąć pod uwagę przy selekcji zmiennych w zagadnieniu klasyfikacji. W niektórych sytuacjach jest możliwe uogólnienie na większą liczbę wymiarów. W przeprowadzonych eksperymentach wykorzystano procedurę `NtRandMultiNorm` z programu NtRand 2.01, generującą liczby losowe odpowiednie do zadanych wektorów średnich i macierzy kowariancji. W artykule zakładać będziemy, że zmienne opisujące obiekty badania są mierzone na skali przedziałowej lub ilorazowej.

W identyfikacji problemów selekcji zmiennych wykorzystano własne spostrzeżenia oraz m.in. następujące opracowania autorów [Gnanadesikan, Kettenring,Tsao 1995; Guyon, Elisseeff 2003].

2. Podstawowe problemy selekcji zmiennych w zagadnieniu klasyfikacji

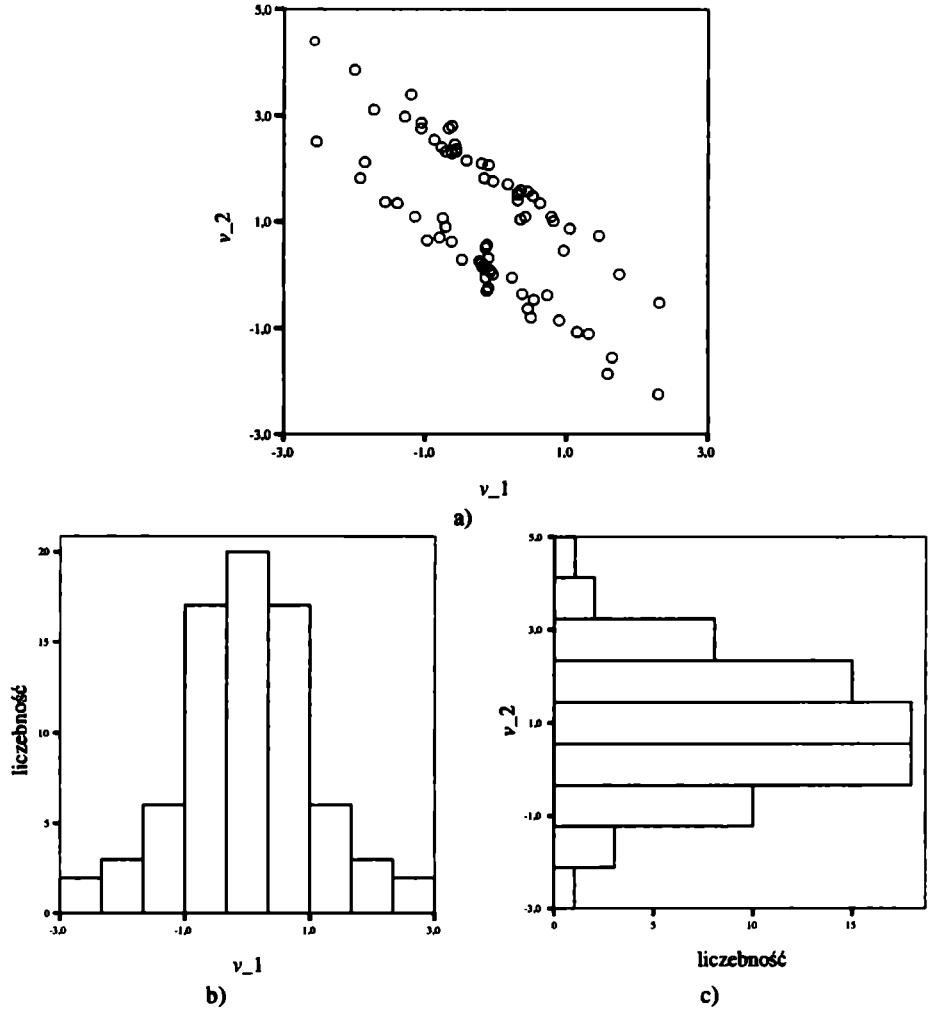
Zagadnienie selekcji zmiennych na potrzeby klasyfikacji zbioru obiektów, przeprowadzane w sensie **indywidualnego rozpatrywania przydatności poszczególnych zmiennych**, niesie ze sobą pewne ograniczenia:

1. Zmienne indywidualnie nie wykazujące zdolności do dyskryminacji zbioru obiektów (zob. rys. 1b i 1c) rozpatrywane łącznie mogą mieć taką zdolność (zob. rys. 1a).

W eksperymencie do otrzymania 76 dwuwymiarowych obserwacji zgodnych z rozkładem normalnym i reprezentujących dwa skupienia separowalne przyjęto w procedurze `NtRandMultiNorm` następujące wektory średnich i macierze kowariancji dla skupień: $\mu_1 = [0 \quad 0]^T$, $\mu_2 = [0 \quad 1{,}8]^T$, $\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & -0{,}98 \\ -0{,}98 & 1 \end{bmatrix}$. Zmienne *v_1* i *v_2* indywidualnie nie wykazują zdolności do dyskryminacji zbioru obiektów, rozpatrywane łącznie zaś pozwalają wyodrębnić dwa skupienia separowalne.

2. Zmienne rozpatrywane indywidualnie zwykle wykrywają inną strukturę klas niż zmienne rozpatrywane łącznie (zob. rys. 2).

W eksperymencie do otrzymania 60 dwuwymiarowych obserwacji, zgodnych z rozkładem normalnym, reprezentujących cztery skupienia separowalne przyjęto w procedurze `NtRandMultiNorm` następujące wektory średnich i macierze kowariancji dla skupień: $\mu_1 = [0 \quad 0]^T$, $\mu_2 = [0 \quad 10]^T$, $\mu_3 = [10 \quad 0]^T$, $\mu_4 = [10 \quad 10]^T$, $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Zmienne *v_1* i *v_2* indywidualnie wykrywają dwa skupienia (rys. 2b i 2c), rozpatrywane łącznie zaś pozwalają wyodrębnić cztery skupienia separowalne (rys. 2a).



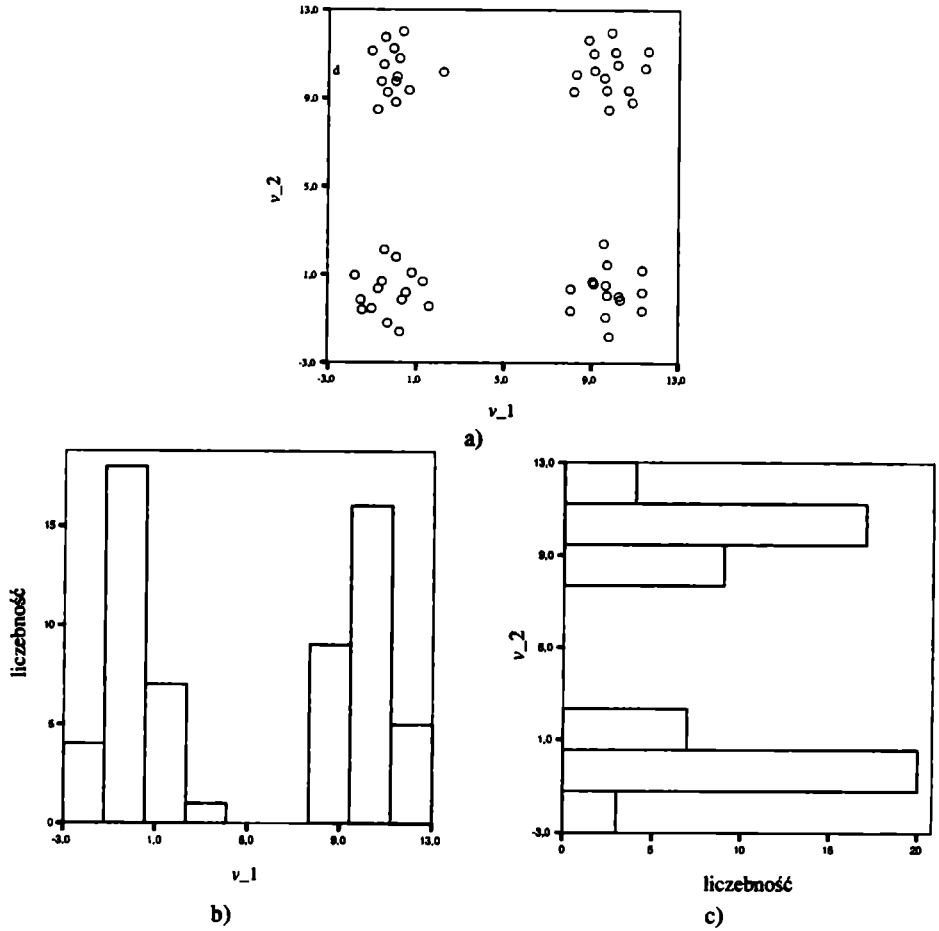
Rys. 1. a) 76 dwuwymiarowych obserwacji zgodnych z rozkładem normalnym reprezentujących dwa skupienia separowalne, b) i c) rozkład liczebności (histogram) dla zmiennych v_1 i v_2
Źródło: opracowanie własne.

3. Zmienne indywidualnie mające rozkład równomierny, traktuje się jako nieprzydatne z punktu widzenia zagadnienia klasyfikacji ze względu na to, że nie wykazują zdolności do dyskryminacji zbioru obiektów. Zdolność grupowania j-tej zmiennej określa wzór (zob. [Sokołowski 1992, s. 12-13]):

$$G_j = 1 - \frac{1}{r_j} \sum_{p=1}^{n-1} \min\left\{(x_{j(p+1)} - x_{j(p)}); r_j/(n-1)\right\} \text{ dla } r_j > 0, \qquad (1)$$

gdzie: r_j – rozstęp wyznaczony z wartości j-tej zmiennej, $x_{j(p)}$ – uporządkowane niemalejąco wartości j-tej zmiennej, n – liczba badanych obiektów.

Miara G_j przyjmuje wartości z przedziału $[0; 1-1/(n-1)]$. Wyższa wartość G_j oznacza lepszą zdolność grupowania dla j-tej zmiennej.

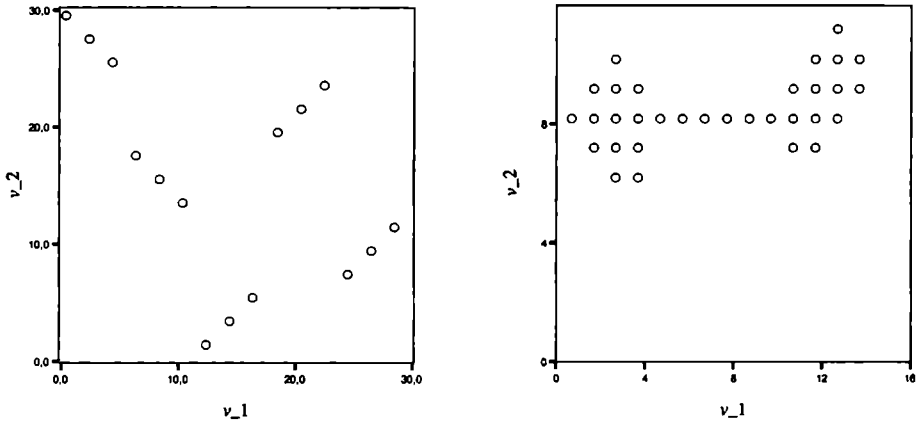


Rys. 2. a) 60 dwuwymiarowych obserwacji, zgodnych z rozkładem normalnym i reprezentujących cztery skupienia separowalne, b) i c) rozkład liczebności (histogram) dla zmiennych v_1 i v_2
Źródło: opracowanie własne.

Na rysunku 3a zaprezentowano przykład dwóch zmiennych indywidualnych o rozkładzie równomiernym:

v_1	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29
v_2	29	27	25	17	15	13	1	3	5	19	21	23	7	9	11

Zdolność grupowania dla zmiennych v_1 i v_2 według miary G_j wynosi 0. Mimo to dwuwymiarowa konfiguracja zmiennych v_1 i v_2 daje pięć skupień separowalnych (rys. 3a).



Rys. 3. Dwuwymiarowa konfiguracja zmiennych v_1 i v_2:
a) zmienne rozpatrywane indywidualnie mają rozkład równomierny, rozpatrywane łącznie reprezentują pięć skupień separowalnych;
b) między najbliższymi sąsiadami odległości (miejskie) są identyczne (na rysunku można zidentyfikować dwa skupienia)
Źródło: opracowanie własne.

Tę niedogodność częściowo eliminuje wielowymiarowe uogólnienie zaproponowane przez Sokołowskiego [1992, s. 50-51]. Dotyczy to jednak tylko metody najbliższego sąsiada. Wprowadzono tam statystykę pozwalającą porównać rzeczywistą strukturę dendrytu znormalizowanego z dendrytem teoretycznym zawierającym równe wiązadła:

$$S = 1 - \frac{1}{\sum_{i=1}^{n-1} a_i} \sum_{i=1}^{n-1} \min\left\{a_i; \sum_{i=1}^{n-1} a_i / n-1\right\} \quad \text{dla } \sum_{i=1}^{n-1} a_i > 0, \qquad (2)$$

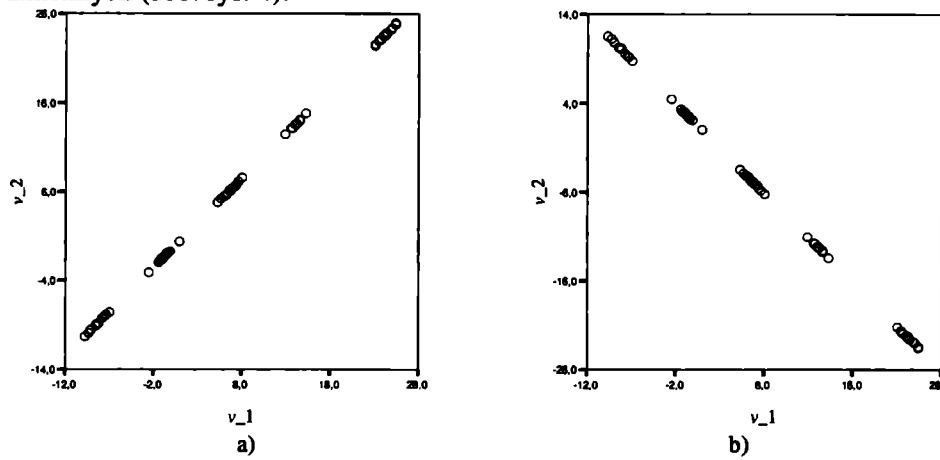
gdzie: a_i – i-te wiązadło w dendrycie, $i = 1,\ldots,n-1$ – numer wiązadła, n – liczba wierzchołków w dendrycie.

Statystyka S przyjmuje wartości z przedziału $[0; 1-1/(n-1)]$ i mierzy „odległość" pomiędzy strukturą dendrytu empirycznego i teoretycznego. Im bardziej wartości tej miary oddalają się od zera, tym większe są możliwości wykrycia struktury klas badanej zbiorowości obiektów za pomocą metody najbliższego sąsiada.

Na rysunku 3b zaprezentowano 32 obiekty badania, dla których odległości miejskie między najbliższymi sąsiadami są identyczne. Zastosowanie do klasyfikacji tego zbioru obiektów metody najbliższego sąsiada nie pozwoli na wykrycie istniejącej struktury klas (statystyka $S = 0$). Zastosowanie do tych samych danych, np. metody najdalszego sąsiada, pozwala zidentyfikować dwa skupienia.

Do zagadnienia selekcji zmiennych na potrzeby klasyfikacji zbioru obiektów w sensie **rozpatrywania parami przydatności poszczególnych zmiennych ze względu na ich skorelowanie** należy przystępować bardzo ostrożnie:

1. Zmienne skorelowane ściśle ($r_{12} = \pm 1$), z punktu widzenia wyników klasyfikacji, nic nowego nie wnoszą do struktury klas właściwej tylko dla jednej ze zmiennych (zob. rys. 4).

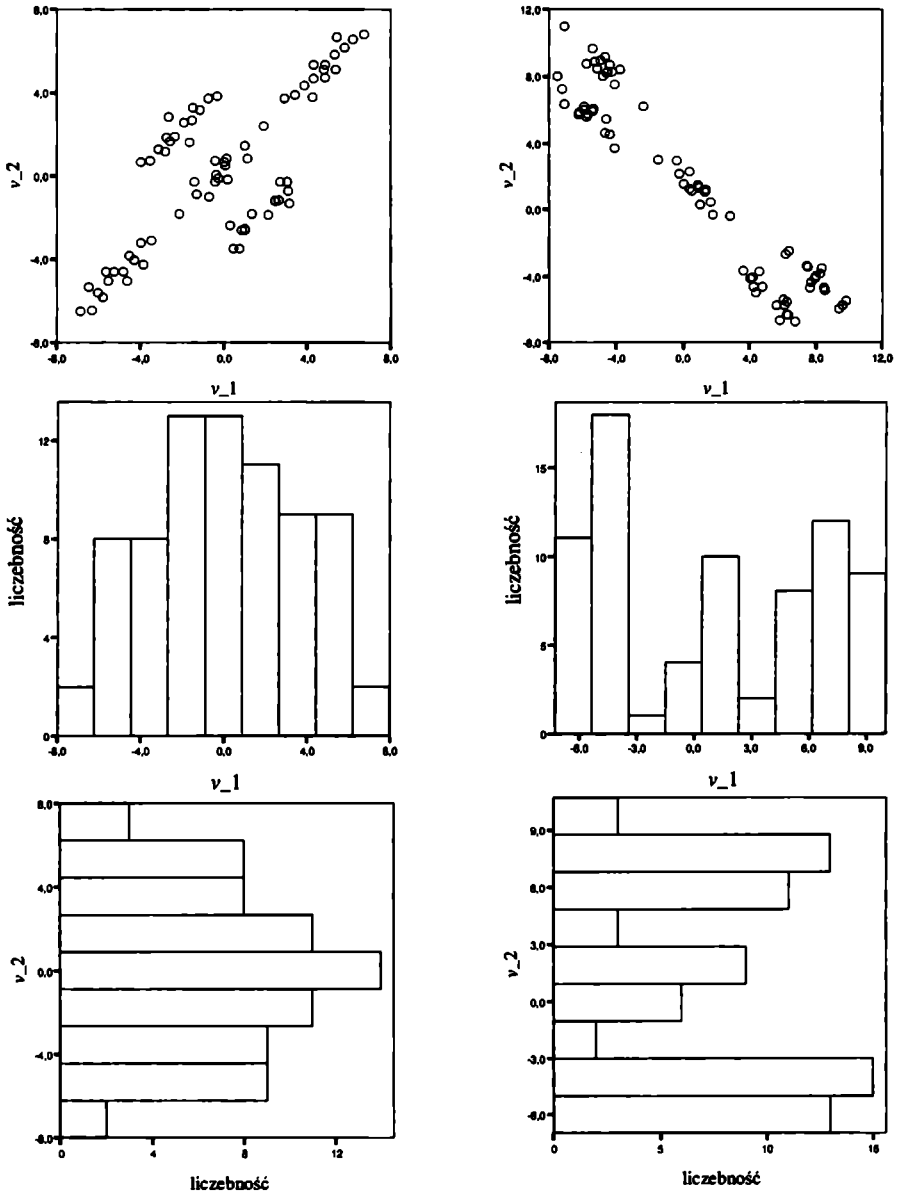


Rys. 4. 75 dwuwymiarowych obserwacji zgodnych z rozkładem normalnym i reprezentujących pięć skupień separowalnych: a) dla $r_{12} = 1,0$; b) dla $r_{12} = -1,0$
Źródło: opracowanie własne.

W eksperymencie do otrzymania 75 dwuwymiarowych obserwacji zgodnych z rozkładem normalnym i reprezentujących pięć skupień separowalnych przyjęto w procedurze `NtRandMultiNorm` następujące wektory średnich i macierze kowariancji dla skupień:

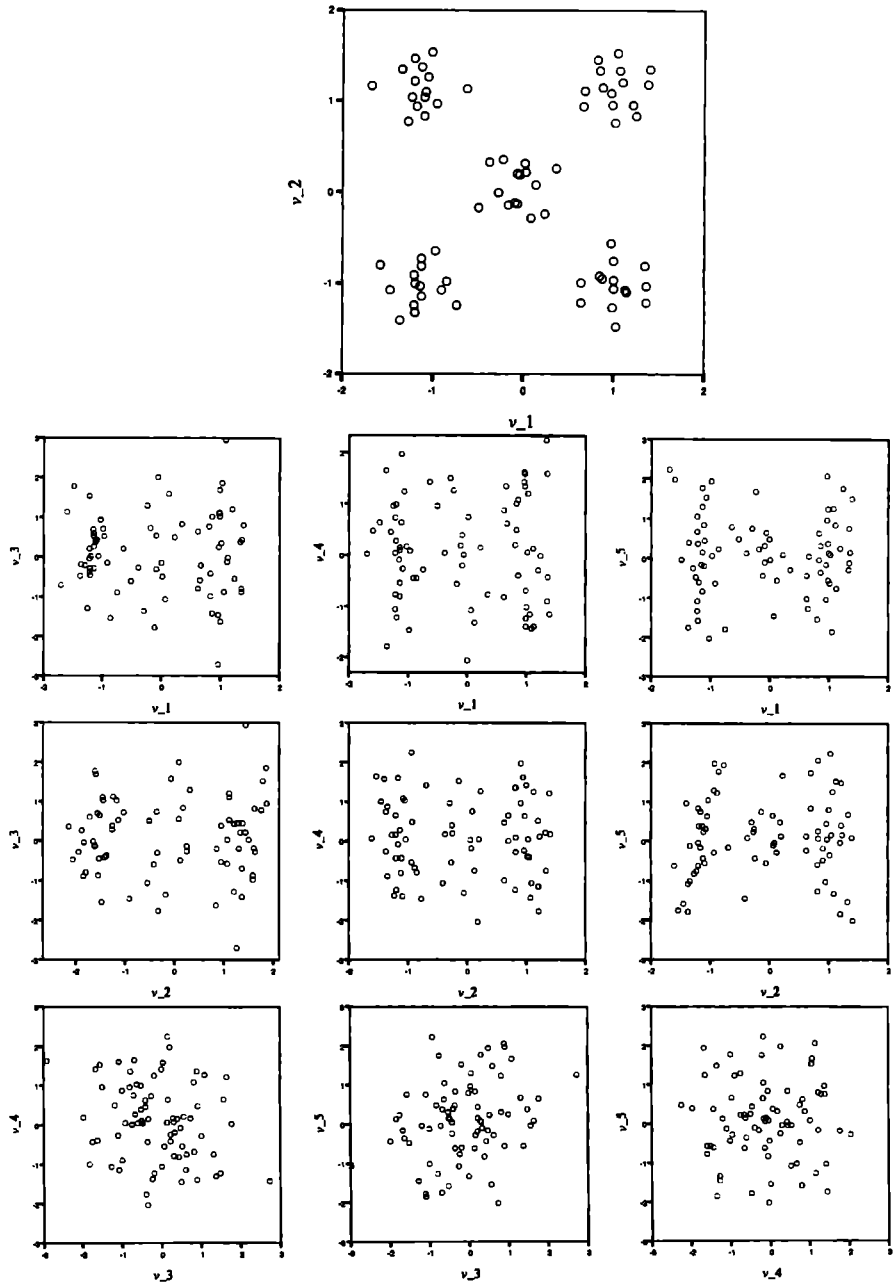
- dla $r_{12} = 1,0$: $\mu_1 = [-7,5 \quad -9,5]^T$, $\mu_2 = [0 \quad -2]^T$, $\mu_3 = [7,5 \quad 5,5]^T$, $\mu_4 = [15 \quad 13]^T$, $\mu_5 = [25 \quad 23]^T$, $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = \Sigma_5 = \begin{bmatrix} 0,5 & 0,5 \\ 0,5 & 0,5 \end{bmatrix}$;
- dla $r_{12} = -1,0$: $\mu_1 = [-7,5 \quad 9,5]^T$, $\mu_2 = [0 \quad 2]^T$, $\mu_3 = [7,5 \quad -5,5]^T$, $\mu_4 = [15 \quad -13]^T$, $\mu_5 = [25 \quad -23]^T$, $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = \Sigma_5 = \begin{bmatrix} 0,5 & -0,5 \\ -0,5 & 0,5 \end{bmatrix}$.

2. Zmienne silnie skorelowane ze sobą indywidualnie nie pozwalają na wykrycie struktury klas, rozpatrywane zaś łącznie pozwalają wykryć strukturę klas (rys. 5). Usunięcie jednej ze zmiennych silnie skorelowanych spowoduje utratę struktury klas z pierwotnej przestrzeni dwuwymiarowej.



a) $r_{12} = 0{,}738$ b) $r_{12} = -0{,}949$

Rys. 5. 75 dwuwymiarowych obserwacji zgodnych z rozkładem normalnym i reprezentujących pięć skupień separowalnych: a) dla $r_{12} = 0{,}738$; b) dla $r_{12} = -0{,}949$
Źródło: opracowanie własne.



Rys. 6. 75 obserwacji pięciu zmiennych w układach dwuwymiarowych
Źródło: opracowanie własne.

W eksperymencie do otrzymania 75 dwuwymiarowych obserwacji zgodnych z rozkładem normalnym i reprezentujących pięć skupień separowalnych przyjęto w

procedurze NtRandMultiNorm następujące wektory średnich i macierze kowariancji dla skupień:

- dla $r_{12}=0{,}738$: $\mu_1=[5\ \ 5]^T$, $\mu_2=[-2\ \ 2]^T$, $\mu_3=[2\ \ -2]^T$, $\mu_4=[0\ \ 0]^T$, $\mu_5=[-5\ \ -5]^T$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_4=\Sigma_5=\begin{bmatrix}1 & 0{,}9\\ 0{,}9 & 1\end{bmatrix}$;
- dla $r_{12}=-0{,}949$: $\mu_1=[-5{,}5\ \ 5{,}5]^T$, $\mu_2=[-4{,}5\ \ 8{,}3]^T$, $\mu_3=[1\ \ 1]^T$, $\mu_4=[8{,}3\ \ -4{,}5]^T$, $\mu_5=[5{,}5\ \ -5{,}5]^T$. $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_4=\Sigma_5=\begin{bmatrix}1 & -0{,}9\\ -0{,}9 & 1\end{bmatrix}$.

3. Na rys. 6 zaprezentowano 75 dwuwymiarowych obserwacji zgodnych z rozkładem normalnym i reprezentujących pięć skupień separowalnych w układzie zmiennych *v*_1 i *v*_2 ($r_{12}=0{,}0$). W procedurze NtRandMultiNorm przyjęto następujące wektory średnich i macierze kowariancji dla skupień: $\mu_1=[0\ \ 0]^T$, $\mu_2=[0\ \ 10]^T$, $\mu_3=[5\ \ 5]^T$, $\mu_4=[10\ \ 0]^T$, $\mu_5=[10\ \ 10]^T$, $\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_4=\Sigma_5=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$.

Do analizy wprowadzono dodatkowe trzy zmienne *v*_3, *v*_4 i *v*_5, zakłócające istniejącą w układzie dwuwymiarowym strukturę klas (tzw. *noisy variables*). 75 trójwymiarowych obserwacji zgodnych z rozkładem normalnym wygenerowano, przyjmując w eksperymencie następujący wektor średnich i macierz kowariancji: $\mu=[5\ \ 5\ \ 7{,}5]^T$, $\Sigma=\begin{bmatrix}5 & 2 & 6\\ 2 & 1 & -5\\ 6 & -5 & 2\end{bmatrix}$. W ostatnim etapie obserwacje na wszystkich zmiennych podano klasycznej standaryzacji.

Następnie porównano za pomocą miary Randa [Rand 1971] i skorygowanej miary Randa [Hubert i Arabie 1985] wyniki podziału zbioru obiektów na 5 klas w przestrzeni 5-wymiarowej i 2-wymiarowej dla wybranych metod klasyfikacji (zob. tab. 1).

Tabela 1. Porównanie wyników podziału zbioru obiektów na 5 klas w przestrzeni 5-wymiarowej i 2-wymiarowej dla wybranych metod klasyfikacji

Wartość miary/ metoda	1	2	3	4	5	6	7	8	9
Randa	0,74234	0,25766	0,72505	0,25405	0,32324	0,74486	0,61477	0,76865	0,39820
Skorygowana Randa*	0,17521	0,00470	0,13717	-0,00013	0,01449	0,28258	0,16217	0,31402	0,02506

* W obliczeniach tego indeksu wykorzystano wzór podany przez Kriegera i Greena [1999, s. 67]
1 – *k*-średnich; 2 – najbliższego sąsiada; 3 – najdalszego sąsiada; 4 – centroidalna; 5 – mediany; 6 – Warda; 7 – średniej odległości pomiędzy skupieniami; 8 – średniej odległości wewnątrz skupień; 9 – dwustopniowe grupowanie.

Źródło: opracowanie własne z wykorzystaniem programu SPSS for Windows oraz programu opracowanego przez A. Dudka.

Wyniki zawarte w tab. 1 jednoznacznie wskazują, że zmienne v_3, v_4 i v_5 zakłócają istniejącą w układzie dwuwymiarowym strukturę klas.

Tabela 2. Macierz korelacji między zmiennymi $v\_1 - v\_5$

Zmienne	v_1	v_2	v_3	v_4	v_5
v_1	1,000	0,000	–0,003	–0,017	–0,017
v_2	0,000	1,000	0,042	–0,055	0,076
v_3	–0,003	0,042	1,000	–0,218	0,268
v_4	–0,017	–0,055	–0,218	1,000	–0,002
v_5	–0,017	0,076	0,268	–0,002	1,000

Źródło: opracowanie własne.

Macierz korelacji zmiennych v_1, v_2, v_3, v_4 i v_5 pokazuje, że metody bazujące na niepowielaniu informacji (miarą powielania informacji jest tutaj współczynnik korelacji liniowej Pearsona) nie eliminują zmiennych v_3, v_4 i v_5 zakłócających istniejącą w układzie dwuwymiarowym strukturę klas.

3. Metody selekcji i ważenia zmiennych

Do rozwiązania zagadnienia wyboru zmiennych służą zasadniczo dwa ujęcia: wybór merytoryczny w ścisłym tego słowa znaczeniu, wybór merytoryczno-formalny. Oba ujęcia obejmują dwie fazy. Faza I jest taka sama w obu ujęciach, różnice zaś występują w fazie II. Punktem wyjścia obu ujęć (faza I) jest skonstruowanie wstępnej listy zmiennych na podstawie własnej hipotezy roboczej badacza (wynikającej z jego znajomości przedmiotu badania oraz wiedzy płynącej z szeroko rozumianej teorii ekonomii) oraz współpracy z przedstawicieli odpowiednich dyscyplin naukowych (eksperci).

Redukcja wstępnej listy zmiennych z wykorzystaniem analizy merytorycznej (faza II) jest działaniem w głównej mierze subiektywnym. Dokonuje się jej na podstawie własnej znajomości przedmiotu badania, wykorzystując współpracę przedstawicieli odpowiednich dyscyplin naukowych (ekspertów) oraz opierając się na szeroko pojętej teorii ekonomii.

Redukcja wstępnej listy zmiennych z wykorzystaniem metod doboru zmiennych (faza II) polega na tym, że najpierw usuwa się zmienne, charakteryzujące się małą zmiennością. Następnie do tak zredukowanej liczby zmiennych stosuje się formalny algorytm wyboru zmiennych, w których miernikiem niepowielania informacji jest współczynnik korelacji liniowej Pearsona lub odległości bazujące na nim. Na przykład w dualnych metodach taksonometrycznych celem zastosowania tych algorytmów jest wybór takiego zestawu zmiennych, w którym zmienne są wzajemnie niezależne oraz są zależne od zmiennych nie wchodzących do wybranego zestawu. Przegląd metod doboru zmiennych w polskiej literaturze statystycznej zawiera praca Grabińskiego [1992, s. 42-62].

W tej pracy [Grabiński 1992, s. 34-35] proponuje się wprowadzenie ważenia zmiennych proporcjonalnie do stopnia ich skorelowania z pozostałymi zmiennymi. System ten nadaje wyższe wagi zmiennym silniej skorelowanym z pozostałymi zmiennymi.

Podejście bazujące na niepowielaniu informacji niekoniecznie prowadzi do właściwych rezultatów. P.H.A. Sneath (zob. [Milligan 1996, s. 348]) pokazał, że redukcja przestrzeni klasyfikacji (w wyniku zastosowania np. analizy głównych składowych) może spowodować utratę struktury klas z pierwotnej przestrzeni. Z tego względu w literaturze są proponowane inne podejścia pozwalające określić zdolność zmiennych do dyskryminacji zbioru obiektów:

a) metody selekcji zmiennych:

- Fowlkes, Gnanadesikan i Kettenring [1987; 1988] zaproponowali procedurę doboru zmiennych, zwaną w analizie regresji procedurą selekcji „w przód", powiązaną z metodami hierarchicznymi;
- Sokołowski ([1992, s. 12-13, 50-51]) zaproponował miarę zdolności grupowania dla indywidualnych zmiennych i zestawu zmiennych (zob. formuły (1) i (2));
- Carmone, Kara i Maxwell [1999] zaproponowali heurystyczną procedurę doboru zmiennych (*HINoV*) powiązaną z metodą *k*-średnich i skorygowanym indeksem Randa;

b) metody ważenia zmiennych:

- Desarbo i in. [1984] zaproponowali ważoną metodę *k*-średnich, w której poszukuje się wag dla zmiennych optymalizujących wyniki klasyfikacji w sensie funkcji *stress* Kruskala;
- De Soete [1986; 1988], bazując na ważonej odległości euklidesowej, zaproponował metodę, w której poszukuje się wag dla zmiennych optymalizujących spełnianie w macierzy odległości własności ultrametryki ($d_{ik} \leq \max\{d_{il}; d_{kl}\}$, dla wszystkich trójek obiektów o numerach i,k,l) lub addytywności ($d_{ik} + d_{lp} \leq \max\{d_{il} + d_{kp}; d_{ip} + d_{kl}\}$, dla wszystkich czwórek obiektów o numerach i,k,l,p);
- Makarenkov i Legendre [2001], bazując na ważonej odległości euklidesowej, zaproponowali metodę, w której poszukuje się wag dla zmiennych optymalizujących spełnianie kryterium podziału w metodzie *k*-średnich na ustaloną liczbę klas;
- Friedman i Meulman [2004] opracowali algorytm ważenia zmiennych w ramach wyodrębnionych klas, a nie całej zbiorowości badanych obiektów. W algorytmie tym poszczególne klasy mogą być opisane innym zestawem zmiennych lub tymi samymi zmiennymi jednak o innej strukturze wag.

Analizę porównawczą wybranych metod ważenia i selekcji zmiennych zawierają prace: [Milligan 1989; Gnanadesikan, Kettenring, Tsao 1995; Makarenkov, Legendre 2001].

4. Wnioski

W artykule na przykładzie wygenerowanych danych w dwuwymiarowej przestrzeni zmiennych wskazano ograniczenia, które należy wziąć pod uwagę przy selekcji zmiennych w zagadnieniu klasyfikacji. Ponadto scharakteryzowano proponowane w literaturze podejścia pozwalające określić zdolność zmiennych do dyskryminacji zbioru obiektów, polegające na ważeniu zmiennych lub ich selekcji.

Literatura

Carmone F.J., Kara A., Maxwell S., *HINoV: a New Method to Improve Market Segment Definition by Identifying Noisy Variables*, „Journal of Marketing Research" 1999, listopad, vol. 36, s. 501-509.

Desarbo W.S., Carroll J.D., Clark L.A., Green P.E., *Synthesized Clustering: a Method for Amalgamating Clustering Bases with Differential Weighting of Variables* „Psychometrika" 1984, vol. 49, nr 1, s. 57-78.

De Soete G., *Optimal Variable Weighting for Ultrametric and Additive Tree Clustering*, „Quality & Quantity" 1986, 20, s. 169-180.

De Soete G., *OVWTRE: a Program for Optimal Variable Weighting for Ultrametric and Additive Tree Fitting*, „Journal of Classification" 1988, vol. 5, s. 101-104.

Fowlkes E.B., Gnanadesikan R., Kettenring J.R., *Variable Selection in Clustering and other Contexts*, [w:] *Design, Data, and Analysis*, red. C.L. Mallows, Wiley, New York, Toronto 1987.

Fowlkes E.B., Gnanadesikan R., Kettenring J.R., *Variable Selection in Clustering*, „Journal of Classification" 1988, vol. 5, s. 205-228.

Friedman J.H., Meulman J.J., *Clustering Objects on Subsets Attributes*, http://www-stat.stanford.edu/~jhf/ftp/cosa.pdf, 2004.

Gnanadesikan R., Kettenring J.R., Tsao S.L., *Weighting and Selection of Variables for Cluster Analysis*, „Journal of Classification" 1995, vol. 12, s. 113-136.

Gordon A.D., *Classification*, Chapman and Hall/CRC, London 1999.

Grabiński T., *Metody taksonometrii*, AE, Kraków 1992.

Guyon I., Elisseeff A., *An Introduction to Variable and Feature Selection*, „Journal of Machine Learning Research" 2003, 3, s. 1157-1182.

Hubert L.J., Arabie P., *Comparing Partitions*, „Journal of Classification" 1985, nr 1, s. 193-218.

Krieger A.M., Green P.E., *A Generalized Rand-Index Method for Consensus Clustering of Separate Partitions of the Same Data Base*, „Journal of Classification" 1999, nr 1, s. 63-89.

Makarenkov V., Legendre P., *Optimal Variable Weighting for Ultrametric and Additive Trees and K-Means Partitioning Methods and Software*, „Journal of Classification" 2001, vol. 18, s. 245-271.

Milligan G.W., *A Validation Study of a Variable Weighting Algorithm for Cluster Analysis*, „Journal of Classification" 1989, vol. 6, s. 53-71.
Milligan G.W., *Issues in Applied Classification: Selection of Variables to Cluster*, Classification Society of North America Newsletter, listopad, Issue 37, 1994.
Milligan G.W., *Clustering Validation: Results and Implications for Applied Analyses*, [w:] *Clustering and Classification*, red. P. Arabie, L.J. Hubert, G. de Soete, World Scientific, Singapore 1996, s. 341-375.
Rand W.M., *Objective Criteria for the Evaluation of Clustering Methods*, „Journal of the American Statistical Association" 1971, nr 336, s. 846-850.
Sokołowski A., *Empiryczne testy istotności w taksonomii*, Zeszyty Naukowe AE w Krakowie, seria specjalna: Monografie nr 108, AE, Kraków 1992.

VARIABLE SELECTION AND WEIGHTING PROBLEMS IN CLUSTER ANALYSIS

Summary

Choice of variables is the one of the most important steps in a cluster analysis. Variables used in applied clustering should be selected and weighted carefully. In cluster analysis we should include only those variables that are believed to help discriminate the data.

In article:

- main aspects of selection and weighting of variables to cluster were characterised,
- point at limitations of variable selection for cluster analysis based on data generated from normal distribution,
- main approaches to variable selection and weighting for cluster analysis were discussed.