

# PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

# RESEARCH PAPERS

of Wrocław University of Economics

Nr 328

**Taksonomia 23**

**Klasyfikacja i analiza danych –  
teoria i zastosowania**

Redaktorzy naukowci

Krzysztof Jajuga, Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu  
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

[www.ibuk.pl](http://www.ibuk.pl), [www.ebscohost.com](http://www.ebscohost.com),

w Dolnośląskiej Bibliotece Cyfrowej [www.dbc.wroc.pl](http://www.dbc.wroc.pl),

The Central and Eastern European Online Library [www.ceeol.com](http://www.ceeol.com),

a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon

[http://kangur.uek.krakow.pl/bazy\\_ae/bazekon/nowy/index.php](http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php)

Informacje o naborze artykułów i zasadach recenzowania znajdują się  
na stronie internetowej Wydawnictwa

[www.wydawnictwo.ue.wroc.pl](http://www.wydawnictwo.ue.wroc.pl)

Tytuł dofinansowany ze środków Narodowego Banku Polskiego  
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Kopiowanie i powielanie w jakiegokolwiek formie  
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu  
Wrocław 2014

**ISSN 1899-3192** (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)

**ISSN 1505-9332** (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

## Spis treści

<b>Wstęp</b> .....	11
<b>Małgorzata Rószkiewicz</b> , Wykorzystanie metaanalizy w budowaniu modelu pomiarowego w przypadku braku niezmienniczości zasad pomiaru na przykładzie pomiaru zadowolenia z życia.....	13
<b>Elżbieta Sobczak</b> , Harmonijność inteligentnego rozwoju regionów Unii Europejskiej .....	21
<b>Ewa Roszkowska, Renata Karwowska</b> , Analiza porównawcza województw Polski ze względu na poziom zrównoważonego rozwoju w roku 2010.....	30
<b>Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel</b> , Analiza porównawcza wybranych filtrów w analizie synchronizacji cyklu koniunkturalnego.....	41
<b>Marcin Salamaga</b> , Próba konstrukcji tablic „wymierania scenicznego” spektakli operowych na przykładzie Metropolitan Opera.....	51
<b>Iwona Foryś</b> , Wykorzystanie analizy dyskryminacyjnej do typowania rynków podobnych w procesie wyceny nieruchomości niemieszkalnych .....	59
<b>Jerzy Korzeniewski</b> , Selekcja zmiennych w klasyfikacji – propozycja algorytmu .....	69
<b>Sabina Denkowska</b> , Testowanie wielokrotne przy weryfikacji wieloczynnikowych modeli proporcjonalnego hazardu Coxa.....	76
<b>Ewa Chodakowska</b> , Teoria równań strukturalnych w klasyfikacji zmiennych jawnych i ukrytych według charakteru ich wzajemnych oddziaływań .....	85
<b>Iwona Konarzewska</b> , Model PCA dla rynku akcji – studium przypadku .....	94
<b>Katarzyna Wójcik, Janusz Tuchowski</b> , Dobór optymalnego zestawu słów istotnych w opiniach konsumentów na potrzeby ich automatycznej analizy	106
<b>Aleksandra Łuczak</b> , Zastosowanie metody AHP-LP do oceny ważności determinant rozwoju społeczno-gospodarczego w jednostkach administracyjnych .....	116
<b>Aleksandra Witkowska, Marek Witkowski</b> , Klasyfikacja pozycyjna banków spółdzielczych według stanu ich kondycji finansowej w ujęciu dynamicznym .....	126
<b>Adam Depta</b> , Zastosowanie analizy korespondencji do oceny jakości życia ludności na podstawie kwestionariusza SF-36v2 .....	135
<b>Marek Lubicz, Maciej Zięba, Konrad Pawelczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej</b> , Indukcja reguł dla danych niekompletnych i niezbalansowanych: modele klasyfikatorów i próba ich zastosowania do predykcji ryzyka operacyjnego w torakochirurgii .....	146

<b>Małgorzata Misztal</b> , Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań.....	156
<b>Anna M. Olszewska</b> , Wykorzystanie wybranych metod taksonomicznych do oceny potencjału innowacyjnego województw .....	167
<b>Iwona Bąk</b> , Porównanie jakości grupowań powiatów województwa zachodniopomorskiego pod względem atrakcyjności turystycznej.....	177
<b>Agnieszka Kozera, Joanna Stanisławska, Romana Głowicka-Wołoszyn</b> , Segmentacja gospodarstw domowych według wydatków na turystykę zorganizowaną.....	186
<b>Agnieszka Wałęga</b> , Podejście syntetyczne w analizie spójności ekonomicznej gospodarstw domowych.....	196
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk, Bożena Mroczek</b> , Zastosowanie analizy korespondencji do badania wpływu elektrowni wiatrowych na jakość życia ludności .....	205
<b>Joanna Banaś, Krzysztof Małecki</b> , Klasyfikacja punktów pomiarów ankietowych kierowców na granicy Szczecina z wykorzystaniem zmiennych symbolicznych.....	214
<b>Aneta Becker</b> , Wykorzystanie informacji granularnej w analizie wymagań rynku pracy.....	222
<b>Katarzyna Cheba, Joanna Holub-Iwan</b> , Wykorzystanie analizy korespondencji w segmentacji rynku usług medycznych.....	230
<b>Adam Depta, Iwona Staniec</b> , Identyfikacja czynników decydujących o jakości życia studentów łódzkich uczelni.....	238
<b>Katarzyna Dębowska, Jarosław Kilon</b> , Reguły asocjacyjne w analizie wyników badań metodą Delphi.....	247
<b>Anna Domagała</b> , O wykorzystaniu analizy głównych składowych w metodzie <i>Data Envelopment Analysis</i> .....	254
<b>Alicja Grześkowiak</b> , Analiza wykluczenia cyfrowego w Polsce w ujęciu indywidualnym i regionalnym.....	264
<b>Anna M. Olszewska, Anna Gryko-Nikitin</b> , Pomiar postrzegania jakości kształcenia uczelni wyższej na danych porządkowych z wykorzystaniem środowiska R.....	273
<b>Karolina Paradysz</b> , Hierarchiczna metoda grupowania powiatów jako podejście benchmarkowe w ocenie bezrobocia według BAEL-u w wybranych typach małych obszarów .....	282
<b>Radosław Pietrzyk</b> , Porównanie metod pomiaru efektywności zarządzania portfelami funduszy inwestycyjnych.....	290
<b>Agnieszka Przedborska, Małgorzata Misztal</b> , Wybrane metody statystyki wielowymiarowej w ocenie skuteczności terapeutycznej głębokiej stymulacji elektromagnetycznej u pacjentów z chorobą zwyrodnieniową stawów.....	299

<b>Wojciech Roszka, Marcin Szymkowiak</b> , Podejście kalibracyjne w statystycznej integracji danych .....	308
<b>Iwona Skrodzka</b> , Zastosowanie wybranych metod klasyfikacji do analizy kapitału ludzkiego krajów Unii Europejskiej .....	316
<b>Agnieszka Stanimir</b> , Wielowymiarowa analiza czynników sprzyjających włączeniu społecznemu .....	326
<b>Dorota Strózik, Tomasz Strózik</b> , Przestrzenne zróżnicowanie poziomu życia w województwie wielkopolskim.....	334
<b>Izabela Szamrej-Baran</b> , Identyfikacja przyczyn ubóstwa energetycznego w Polsce przy wykorzystaniu modelowania miękkiego.....	343
<b>Janusz Tuchowski, Katarzyna Wójcik</b> , Klasyfikacja obiektów w systemie Krajowych Ram Kwalifikacji opisanych za pomocą ontologii .....	353
<b>Aleksandra Matuszewska-Janica</b> , Grupowanie krajów Unii Europejskiej ze względu na poziom feminizacji sektorów gospodarczych .....	361
<b>Monika Rozkrut, Dominik Rozkrut</b> , Identyfikacja strategii innowacyjnych przedsiębiorstw usługowych w Polsce .....	369

## Summaries

<b>Małgorzata Rószkiewicz</b> , The use of meta-analysis in building the measurement model in case of the absence of measurement invariance on the example of measuring of life satisfaction.....	20
<b>Elżbieta Sobczak</b> , Harmonious smart growth of European Union regions.....	29
<b>Ewa Roszkowska, Renata Karwowska</b> , The comparative analysis of Polish voivodeships with respect to sustainable development in 2010 .....	40
<b>Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel</b> , Comparative analysis of chosen filters in business cycles analysis .....	50
<b>Marcin Salamaga</b> , The attempt of construction of the life tables for opera works on the example of the Metropolitan Opera .....	58
<b>Iwona Foryś</b> , Using discriminant analysis to select similar markets in non-residential property valuation process.....	68
<b>Jerzy Korzeniewski</b> , Variable selection in classification – algorithm proposal .....	75
<b>Sabina Denkowska</b> , Multiple testing in the verification process of multifactorial Cox proportional hazards models .....	84
<b>Ewa Chodakowska</b> , The theory of structural equations modelling in the classification of observed variables and latent constructs according to the character of their relationship.....	93
<b>Iwona Konarzewska</b> , Modelling stock market by PCA factor model – case study .....	105

<b>Katarzyna Wójcik, Janusz Tuchowski</b> , Selection of the optimal set of relevant words in consumers opinions in the context of the opinion mining ..	115
<b>Aleksandra Łuczak</b> , Application of AHP-LP to the evaluation of importance of determinants of socio-economic development in the administrative units .....	125
<b>Aleksandra Witkowska, Marek Witkowski</b> , A dynamic approach to the ranking of cooperative banks by their financial condition .....	134
<b>Adam Depta</b> , Application of correspondence analysis for the measurement of quality of life – questionnaire SF-36v2 based research .....	145
<b>Marek Lubicz, Maciej Zięba, Konrad Pawelczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej</b> , Classification rules extraction for missing and imbalance data: models of classifiers and initial results in the rules-based thoracic surgery risk prediction.....	155
<b>Małgorzata Misztal</b> , Selected methods for assessing the performance of classifiers – an overview and examples of applications.....	166
<b>Anna M. Olszewska</b> , The application of selected quantitative methods to the evaluation of voivodeship innovation level potential.....	176
<b>Iwona Bąk</b> , The comparison of the quality of groupings of poviats of West Pomeranian Voivodeship in terms of tourism attractiveness .....	185
<b>Agnieszka Kozera, Joanna Stanisławska, Romana Głowicka-Wołoszyn</b> , Household segmentation with respect to the expenditure on organized tourism.....	195
<b>Agnieszka Wałęga</b> , Synthetic approach in the analysis of economic coherence of households .....	204
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk, Bożena Mroczek</b> , Using the correspondence analysis to examine the impact of wind turbines on the quality of life.....	213
<b>Joanna Banaś, Krzysztof Małecki</b> , Classification of measurement survey points of drivers on the boundary of Szczecin using symbolic variables...	221
<b>Aneta Becker</b> , The use granular information in the analysis of the requirements of the labor market.....	229
<b>Katarzyna Cheba, Joanna Hołub-Iwan</b> , The application of the correspondence analysis of patients segmentation on the medical service market .....	237
<b>Adam Depta, Iwona Staniec</b> , Identification of the factors that determine the quality of students life at universities in Lodz.....	246
<b>Katarzyna Dębkowska, Jarosław Kilon</b> , Association rules in the analysis of research results the Delphi method .....	253
<b>Anna Domagała</b> , About using Principal Component Analysis in Data Envelopment Analysis .....	263
<b>Alicja Grześkowiak</b> , Analysis of the digital divide in Poland at the individual and regional level .....	272

<b>Anna M. Olszewska, Anna Gryko-Nikitin</b> , Assessment of perception of quality of teaching at an institution of higher learning based on the ordinal data with the utilization of R environment.....	281
<b>Karolina Paradysz</b> , The hierarchical method of grouping poviats as a benchmark approach in the assessment of unemployment by BAEL in selected types of small areas .....	289
<b>Radosław Pietrzyk</b> , Comparison of methods of measuring the performance of investment funds portfolios.....	298
<b>Agnieszka Przedborska, Małgorzata Misztal</b> , Selected multivariate statistical analysis methods in the evaluation of efficacy of deep electromagnetic stimulation in patients with degenerative joint disease .....	307
<b>Wojciech Roszka, Marcin Szymkowiak</b> , A calibration approach in statistical data integration .....	315
<b>Iwona Skrodzka</b> , Application of some methods of classification to the analysis of human capital in the European Union.....	325
<b>Agnieszka Stanimir</b> , Multivariate analysis of social inclusion factors.....	333
<b>Dorota Strózik, Tomasz Strózik</b> , Spatial differentiation of the standard of living in Great Poland Voivodeship .....	342
<b>Izabela Szamrej-Baran</b> , Identification of fuel poverty causes in Poland using soft modelling .....	352
<b>Janusz Tuchowski, Katarzyna Wójcik</b> , Classification of objects in the National Classification Framework described by the ontology.....	360
<b>Aleksandra Matuszewska-Janica</b> , Clustering of European Union states taking into consideration the levels of feminization of economic sectors..	368
<b>Monika Rozkrut, Dominik Rozkrut</b> , Identification of service sector innovation strategies in Poland.....	379

**Jerzy Korzeniewski**

Uniwersytet Łódzki

---

## SELEKCJA ZMIENNYCH W KLASYFIKACJI – PROPOZYCJA ALGORYTMU

---

**Streszczenie:** Selekcja zmiennych w klasyfikacji obiektów ze zbiorem uczącym jest ważna zarówno w przypadku metod pojedynczych, jak i zagregowanych. Najprostszym sposobem selekcji jest sprawdzenie korelacji każdej zmiennej z prawidłową klasyfikacją obiektów na zbiorze uczącym. Ten naturalny sposób ma jednak poważne ograniczenia wynikające z tego, że im słabsza skala pomiaru wartości zmiennej, tym trudniej mierzyć siłę korelacji. W artykule zaproponowana jest metoda pomiaru siły korelacji za pomocą współczynnika korelacji liniowej pomiędzy odległościami pomiędzy parami obiektów na badanej zmiennej i na zmiennej reprezentującej etykiety klas. Zmienne, które mają siłę korelacji poniżej ustalonego progu, są eliminowane. Efektywność takiej metody selekcji jest zbadana na zbiorach danych empirycznych z repozytorium UCI Uniwersytetu Kalifornijskiego (*UCI Machine Learning Repository*). Wyniki są porównane z wynikami procedur *stepclass* oraz *Boruta* dostępnymi w języku *R*.

**Słowa kluczowe:** klasyfikacja, zbiór uczący, korelacja zmiennych.

### 1. Wstęp

W klasyfikacji ze zbiorem uczącym [por. Gatnar, Walesiak 2004], podobnie jak w wielu innych dziedzinach statystyki, bardzo istotny jest początkowy etap wyboru zmiennych, które zostaną użyte w dalszych etapach procedury. Takich metod w klasyfikacji opracowano bardzo wiele [por. Dash, Liu b.d.w.; Mahdi, Fazekas 2011]. Metody te można podzielić na dwie grupy: metody filtrujące zbiór wszystkich zmiennych (*filter methods*) oraz metody zależne od sposobu klasyfikowania obiektów (*wrapper methods*). Cechą metod pierwszej grupy jest ocenianie wybranych podzbiorów zmiennych bez klasyfikowania obiektów – nie ma bezpośredniego odniesienia wyników badania podzbiorów zmiennych do efektów klasyfikowania obiektów. Cechą metod drugiej grupy jest to, że ocenianie wybranych podzbiorów zmiennych odbywa się na podstawie otrzymanej klasyfikacji obiektów. Rezultatem takiego podejścia jest wysoka jakość selekcji, ale również brak ogólności.



Celem niniejszego artykułu jest zaproponowanie metody filtrującej zbiór zmiennych, która bez przyjmowania żadnych założeń pozwoli selekcjonować zmienne, które (o ile takie się znajdują) uczestniczą w tworzeniu struktury klas obiektów.

## 2. Selekcja zmiennych oparta na korelacji odległościowej

Jeśli jakieś dwa rozłączne zbiory zmiennych przyczyniają się do tworzenia struktury klas, to włączenie pewnych podzbiorów obiektów do obszaru o większej gęstości, czyli do klasy, powinno odbywać się „na obu” zbiorach zmiennych. Innymi słowy, jeśli w zbiorze danych istnieje wyraźna struktura klas, to odległości pomiędzy parami obserwacji obliczane w oparciu o te dwa zbiory zmiennych powinny być ze sobą skorelowane dodatnio.

Definicja: Współczynnik korelacji odległościowej pomiędzy zbiorami  $A$ ,  $B$  zmiennych dany jest wzorem:

$$WKO(A, B, l) = \frac{\frac{1}{l} \sum_{i=1}^l d_i^A d_i^B - \bar{d}^A \bar{d}^B}{s^A s^B}, \quad (1)$$

gdzie:  $1 \leq l \leq n$  oznacza liczbę par obserwacji wybranych w sposób zależny spośród wszystkich par obiektów;  $d_i^A$ ,  $d_i^B$  oznaczają odległości dla  $i$ -tej pary obliczone w oparciu o zmienne ze zbioru, odpowiednio,  $A$ ,  $B$ ;  $\bar{d}^A$ ,  $\bar{d}^B$ ,  $s^A$ ,  $s^B$  są to, odpowiednio, średnie arytmetyczne i odchylenia standardowe obliczone dla wszystkich  $l$  odległości dla zbiorów, odpowiednio,  $A$ ,  $B$ .

Podejście oparte na korelacji odległościowej dało dobre rezultaty w kontekście analizy skupień [por. Korzeniewski 2012]. Standardowo, jeżeli ustalimy liczbę  $l$  par obiektów dla wszystkich zbiorów  $A$  i  $B$ , to parametr  $l$  we wzorze (1) można pominąć. W najprostszy sposób, mając do dyspozycji zbiór uczący, można wykorzystać współczynnik korelacji odległościowej, znajdując współczynniki  $WKO$  pomiędzy pojedynczymi zmiennymi a zmienną reprezentującą numery klas. Zmienną numerów klas oznaczmy przez  $nry\_klas$ . Zaproponujemy następujący algorytm.

Algorytm selekcji zmiennych:

1. Dla każdej zmiennej  $u$  opisującej obiekty znajdujemy  $WKO(u, nry\_klas)$ .
2. Odrzucamy wszystkie zmienne, dla których wartość współczynnika nie przekracza 0,1.
3. Do zbioru wyselekcjonowanych zmiennych dołączamy iteracyjnie każdą zmienną, która ma wartość współczynnika powyżej 0,1 z jakąkolwiek zmienną będącą już w zbiorze zmiennych wyselekcjonowanych.

Tak sformułowany algorytm jest najprostszym algorytmem z możliwych. Odległości dla zmiennej reprezentującej numery klas mogą być obliczane zgodnie z

formułą Sokala-Michenera [por. Gatnar, Walesiak 2004]. Można go modyfikować na wiele sposobów. Na przykład, do wyselekcjonowanego zbioru zmiennych dołączać zmienne, dla których współczynnik korelacji liniowej (gdy taki istnieje) ma wartość wyższą od ustalonego progu. Można też badać wartość współczynnika niekoniecznie na całym zbiorze danych, lecz na przykład na połowie klas, na innych wybranych podzbiorach klas. Można również badać wartość współczynnika dla podzbiorów kilku zmiennych.

### 3. Badanie porównawcze efektywności nowego algorytmu

Przeprowadzone zostało badanie na kilkunastu zbiorach danych z repozytorium UCI. Charakterystyki badanych zbiorów zawiera tabela 1. W badaniu tym obliczono wartość współczynnika *WKO* dla  $l = 30$  wylosowanych zależnie par obiektów. Losowanie to powtarzane było 200 razy, z powtórzeń tych ostateczną wartością współczynnika była średnia arytmetyczna uzyskanych 200 wartości. Najwyższe wartości współczynnika, które uwzględniono przy selekcji zmiennych, przedstawione są w tabelach 2-11.

**Tabela 1.** Charakterystyka zbiorów danych poddanych badaniu

Zbiór danych	Liczba obiektów	Liczba klas	Charakterystyka zbioru zmiennych
<i>Adult</i>	1000	2	6 zmiennych ciągłych, 8 zmiennych nominalnych
<i>Australiancredit</i>	690	2	6 zmiennych ciągłych, 8 nominalnych
<i>Balance</i>	625	3	4 zmienne porządkowe
<i>Blood</i>	748	2	4 zmienne ciągłe
<i>Concrete</i>	1030	5	8 zmiennych ciągłych
<i>Glass</i>	214	6	9 zmiennych ciągłych
<i>Hayes</i>	132	3	2 zmienne porządkowe, jedna nominalna
<i>Housing</i>	506	5	12 zmiennych ciągłych, jedna binarna
<i>Ionosphere</i>	351	2	33 zmienne ciągłe
<i>Iris</i>	150	3	4 zmienne ciągłe
<i>Votes</i>	435	2	16 zmiennych binarnych
<i>Wines</i>	178	3	13 zmiennych ciągłych

Źródło: obliczenia własne.

W celu oceny efektywności nowego algorytmu zaproponowano porównanie go z dwiema metodami selekcji zmiennych dostępnymi w programie *R*: pakietem *Boruta* oraz funkcją *stepclass*. Selekcja zmiennych w pakiecie *Boruta* przebiega w następujący sposób. Powiększamy zbiór zmiennych, dodając taką samą liczbę zmiennych i permutując wartości dodanych zmiennych. Klasyfikujemy obiekty za pomocą metody *random forest* i oceniamy jakość tej klasyfikacji. Znajdujemy maksymalną wartość oceny dla zmiennych dodanych MZSA i zapamiętujemy wszystkie zmienne, które uzyskały lepszą ocenę od MZSA. Porównujemy ocenę

uzyskaną dla każdej zmiennej oryginalnej z MZSA za pomocą testu dla dwóch średnich. Zmienne, które uzyskały ocenę istotnie wyższą, uznajemy za wybrane, zaś zmienne, które uzyskały ocenę istotnie niższą, uznajemy za odrzucone (por. instrukcja Package *Boruta* z programu R).

Selekcja zmiennych w pakiecie *Stepclass* przebiega w następujący sposób. Ta metoda jest uzależniona od ustalonej metody klasyfikacyjnej, np. *lda* (*linear discriminant analysis* – ta metoda była stosowana w badaniu). Polega ona na krokowym dołączaniu lub odrzucaniu pojedynczych zmiennych z aktualnego zbioru zmiennych wybranych. Dla aktualnego zbioru zmiennych budujemy model klasyfikacyjny i oceniamy go metodą walidacji krzyżowej. Jeśli przyjęte kryterium oceny jest lepsze od dotychczasowej wartości kryterium plus *improvement* (w badaniu przyjęto standardową wartość *improvement* = 0,05), to zachowujemy aktualny zbiór zmiennych. Jeśli nie, to wyrzucamy zmienną ostatnio dołączoną (lub dołączamy ostatnio wyrzuconą) i próbujemy dołączać lub wyrzucać inną zmienną (por. instrukcja Package *klaR* z programu R).

Należy zaznaczyć, że obie metody, tj. *Boruta* i *stepclass*, są typowymi metodami klasy *wrapper*. Wobec tego przyjęto następującą zasadę oceniania efektywności. Jeżeli jakaś z dwóch porównywanych metod wskazuje na bardzo wysoką (powyżej 90%) zgodność klasyfikacji obiektów dla wyselekcjonowanego zbioru *S* zmiennych, to zbiór zmiennych wybranych przez nowy algorytm porównujemy ze zbiorem *S*.

**Tabela 2.** Wartości *WKO* wyższe od 0,1 dla zbioru *Adult*

Para zmiennych	{11,nry}	{12,nry}
<i>WKO</i>	0,18	0,11

Źródło: obliczenia własne.

**Tabela 3.** Wartości *WKO* wyższe od 0,1 dla zbioru *Australiancredit*

Para zmiennych	{8,nry}	{9,nry}	{10,nry}	{7,10}	{2,7}	{2,3}
<i>WKO</i>	0,47	0,21	0,15	0,26	0,25	0,20

Źródło: obliczenia własne.

**Tabela 4.** Wartości *WKO* wyższe od 0,1 dla zbioru *Balance*

Para zmiennych	{1,nry}	{2,nry}	{3,nry}	{4,nry}
<i>WKO</i>	0,13	0,15	0,14	0,12

Źródło: obliczenia własne.

**Tabela 5.** Wartości *WKO* wyższe od 0,1 dla zbioru *Blood*

Para zmiennych	{1,nry}	{2,nry}	{3,2}	{4,3}
<i>WKO</i>	0,14	0,104	1,00	0,43

Źródło: obliczenia własne.

**Tabela 6.** Wartości *WKO* wyższe od 0,1 dla zbioru *Concrete*

Para zmiennych	{1,nry}	{4,nry}	{5,nry}	{8,nry}	{3,5}	{7,8}	{1,6}	{2,7}
<i>WKO</i>	0,17	0,14	0,14	0,13	0,18	0,17	0,13	0,11

Źródło: obliczenia własne.

**Tabela 7.** Wartości *WKO* wyższe od 0,1 dla zbioru *Glass*

Para zmiennych	{2,nry}	{3,nry}	{4,nry}	{6,nry}	{8,nry}	{1,2}	{1,5}	{1,7}
<i>WKO</i>	0,20	0,28	0,23	0,14	0,19	0,31	0,49	0,73

Źródło: obliczenia własne.

**Tabela 8.** Wartości *WKO* wyższe od 0,1 dla zbioru *Hayes*

Para zmiennych	{1,nry}	{2,nry}	{3,nry}
<i>WKO</i>	0,13	0,12	0,096

Źródło: obliczenia własne.

**Tabela 9.** Wartości *WKO* wyższe od 0,1 dla zbioru *Housing*

Para zmiennych	{2,nry}	{3,nry}	{5,nry}	{6,nry}	{7,nry}	{10,nry}	{11,nry}	{13,nry}
<i>WKO</i>	0,14	0,17	0,16	0,28	0,24	0,13	0,19	0,28
Para zmiennych	{1,10}	{8,2}	{9,1}	{12,1}				
<i>WKO</i>	0,43	0,50	0,52	0,34				

Źródło: obliczenia własne.

Jeśli w zbiorze *Housing* numer klasy potraktujemy jako zmienną porządkową (można to zrobić, bo numer klasy może być miarą atrakcyjności nieruchomości), to korelacje odległościowe są o wiele wyraźniejsze, np. 8. i 9. zmienna mają współczynnik wyraźnie wyższy od 0,1.

### Zbiór *Ionosphere*

Mniej więcej połowa zmiennych ma wartości współczynnika korelacji odległościowej powyżej 0,1 (na ogół znacznie). Pozostałe mają te wartości „na granicy” 0,1, ale wszystkie są bardzo silnie skorelowane odległościowo z większością pozostałych zmiennych.

**Tabela 10.** Wartości *WKO* wyższe od 0,1 dla zbioru *Iris*

Para zmiennych	{1,nry}	{2,nry}	{3,nry}	{4,nry}
<i>WKO</i>	0,39	0,19	0,70	0,71

Źródło: obliczenia własne.

**Tabela 11.** Wartości *WKO* wyższe od 0,1 dla zbioru *Votes*

Para zmiennych	{1,nry}	{3,nry}	{4,nry}	{5,nry}	{7,nry}	{8,nry}	{9,nry}
<i>WKO</i>	0,11	0,54	0,82	0,56	0,25	0,42	0,35

Źródło: obliczenia własne.

#### Zbiór *Wines*

Wszystkie zmienne oprócz piątej mają wysokie wartości *WKO*, około 0,4. Ale piąta zmienna jest wystarczająco silnie skorelowana odległościowo z trzecią zmienną, gdyż  $WKO\{3,5\} = 0,124$ .

Wyniki selekcji zmiennych dla dwóch spośród trzech metod zostały zebrane w tabeli 12. W tabeli tej nie ma wyników selekcji uzyskanych przez procedurę *stepclass*, gdyż spisała się ona bardzo słabo. W przypadku większości zbiorów nie dało się jej zastosować (jako funkcji dyskryminującej *lda*), natomiast w przypadku tych, dla których możliwe było użycie procedury, wyniki były bardzo złe. Na przykład, ze zbioru *Iris*, który jak wiadomo ma bardzo wyraźną strukturę klas, procedura *stepclass* wybrała tylko czwartą zmienną, a ze zbioru *Wines* – tylko dwie spośród (jak wskazują dwie inne metody) uczestniczących w tworzeniu struktury klas trzynastu zmiennych.

**Tabela 12.** Wyniki selekcji zmiennych dla porównywanych metod

Zbiór danych	Pakiet <i>Boruta</i>	Nowy algorytm
<i>Adult</i>	wszystkie oprócz 2, 3 i 9	11, 12
<i>Auscredit</i>	wszystkie oprócz 2, 12, 13	2, 3, 7, 8, 9, 10
<i>Balance</i>	wszystkie 4	wszystkie 4
<i>Blood</i>	wszystkie 4	wszystkie 4
<i>Concrete</i>	wszystkie 8	wszystkie 8
<i>Glass</i>	wszystkie 9	wszystkie oprócz 9
<i>Hayes</i>	wszystkie 3	1 i 2
<i>Housing</i>	wszystkie 13	wszystkie oprócz 4
<i>Ionosphere</i>	wszystkie 33	wszystkie 33
<i>Iris</i>	wszystkie 4	wszystkie 4
<i>Votes</i>	wszystkie oprócz 2, 7, 11	wszystkie oprócz 2, 11, 20
<i>Wines</i>	wszystkie 13	wszystkie 13

Źródło: obliczenia własne.

## 4. Wnioski

W przeprowadzonym badaniu zaproponowany algorytm okazał się szczególnie pożyteczny, mimo że jest on tylko typu filtrującego. Na ogół bardzo dobrze selekcjonuje zmienne, które przyczyniają się do tworzenia struktury klas w danym zbiorze. Wyniki selekcji są w dużym stopniu podobne do wyselekcjonowanych zbiorów zmiennych otrzymanych przy użyciu pakietu *Boruta*, które cechują się bardzo

wysoką (powyżej 90%) zgodnością klasyfikacji. Korelacja odległościowa jest wyjątkowo elastyczna – można ją stosować do wszystkich skal pomiarowych, zarówno słabych, jak i silnych. Zastosowany algorytm jest bardzo prostą wersją – może być modyfikowany, co powinno pozwolić na uzyskanie lepszych wyników. Prosta modyfikacją może być, na przykład, zbadanie skorelowania odległościowego zmiennych ze zmienną etykiet niekoniecznie na całym zbiorze uczącym, a tylko na wybranych podzbiorach niektórych klas. Ponadto uzyskane rezultaty selekcji zmiennych, które uczestniczą w tworzeniu struktury klas, można potraktować jako zbiór startowy do jakiejś metody, za pomocą której można próbować optymalizować ten zbiór.

## Literatura

- Dash M., Liu H. (b.d.w.), *Feature Selection for Classification*, unpublished manuscript.
- Gatnar E., Walesiak M. (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo AE we Wrocławiu, Wrocław.
- Korzeniewski J. (2012), *Metody selekcji zmiennych w analizie skupień. Nowe procedury*. Wydawnictwo Uniwersytetu Łódzkiego.
- Mahdi E., Fazekas G. (2011), *Feature Selection as an Improving Step for Decision Tree Construction*, 2009 International Conference on Machine Learning and Computing, IPCSIT, Singapore.

### VARIABLE SELECTION IN CLASSIFICATION – ALGORITHM PROPOSAL

**Summary:** Selection of variables in classification is important both in the case of single and aggregated methods. The simplest way of selecting variables is to check their correlation with the proper classification of objects on the training set. This natural way, however, has serious limitations stemming from the fact that for weak measurement scales finding correlation is troublesome. The paper proposes a method of measuring the strength of correlation by means of the linear correlation coefficient based on the distances between pairs of observations for arbitrary single attribute and the class labels attribute. The attributes with correlation below a certain threshold are rejected. The efficiency of the method is investigated on UCI data sets. The results are compared with *stepclass* and *Boruta* procedures available in R language.

**Keywords:** classification, training set, variable correlation.