

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

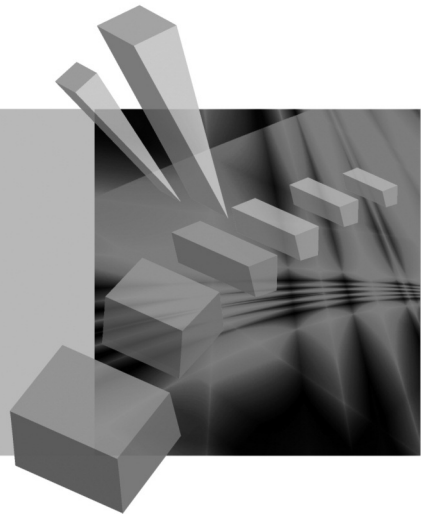
RESEARCH PAPERS

of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pocięcha,
Miroslaw Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities <http://cejsh.icm.edu.pl>
oraz w The Central and Eastern European Online Library 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

Kopowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2012

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Wstęp	13
Stanisława Bartosiewicz , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej	17
Andrzej Sokolowski , Q uniwersalna miara odległości	22
Eugeniusz Gatnar , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP)	31
Marek Walesiak , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
Krzysztof Jajuga, Marek Walesiak , XXV lat konferencji taksonomicznych – fakty i refleksje	47
Józef Pocięcha, Barbara Pawelek , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne	50
Paweł Lula , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych	58
Ewa Roszkowska , Zastosowanie metody TOPSIS do wspomaganie procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka, Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
Ewa Chodakowska, Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
Bartosz Soliński, Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
Krzysztof Szwarz, Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
Elżbieta Gołata, Grażyna Dehnel, Rejestry administracyjne w analizie przedsiębiorczości.....	202
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień, Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
Katarzyna Dębowska, Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
Alina Bojan, Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
Justyna Brzezińska, Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka, Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
Bartłomiej Jefmański, Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
Julita Stańczuk, Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
Jerzy Krawczuk, Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
Anna Czapkiewicz, Beata Basiura, Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
Radosław Pietrzyk, Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
Aleksandra Witkowska, Marek Witkowski, Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
Marcin Pelka, Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
Justyna Wilk, Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

Tomasz Bartłomowicz, Justyna Wilk , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
Kamila Migdał-Najman , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących	342
Dorota Rozmus , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	352
Krzysztof Najman , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG	361
Małgorzata Misztal , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna	370
Mariusz Kubus , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
Barbara Batóg, Jacek Batóg , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym	387
Katarzyna Wójcik, Janusz Tuchowski , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej	396
Iwona Staniec , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach	406
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami	416
Iwona Foryś , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
Ewa Genge , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
Jerzy Korzeniewski , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień	444
Andrzej Dudek , SMS – propozycja nowego algorytmu analizy skupień	451
Artur Mikulec , Metody oceny wyniku grupowania w analizie skupień.....	460
Małgorzata Machowska-Szewczyk , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych	469
Artur Zaborski , Analiza PROFIT i jej wykorzystanie w badaniu preferencji	479
Karolina Bartos , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena	488

Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych .	496
Izabela Kurzawa , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś	505
Aleksandra Łuczak, Feliks Wysocki , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych	513
Agnieszka Sompolska-Rzechuła , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim	523
Joanna Banaś, Małgorzata Machowska-Szewczyk , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego	532
Iwona Bąk , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę	541
Aneta Becker , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
Katarzyna Dębowska , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej	562
Anna Domagała , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej	580
Hanna Gruchociak , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
Tomasz Klimanek, Marcin Szymkowiak , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy	601
Jarosław Lira , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce	610
Christian Lis , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku	619
Beata Bieszk-Stolorz, Iwona Markowicz , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
Lucyna Przezbórska-Skobiej, Jarosław Lira , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
Paweł Ulman , Model rozkładu wydatków a funkcje popytu.....	646
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Zastosowanie metod analizy statystycznej w badaniach mięczaków	655

Summaries

Stanisława Bartosiewicz , The effects of subjectivism in multivariate analysis revisited.....	21
Andrzej Sokółowski , Q universal distance measure	30
Eugeniusz Gatnar , Data quality in central banks' statistical systems (NBP example)	38
Marek Walesiak , Distance measures for ordinal data – strategies of proceedings.....	46
Krzysztof Jajuga, Marek Walesiak , XXV years of taxonomic conferences – some facts and remarks.....	49
Józef Pocięcha, Barbara Pawelek , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
Paweł Lula , Learning-based systems of information extraction from textual resources	67
Ewa Roszkowska , The application of the TOPSIS method to support the negotiation process	75
Andrzej Młodak , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
Andrzej Bąk , Models for unordered categories in preference analysis.....	95
Kowalewski Jacek , An integrated model of optimizing statistical surveys	105
Jan Paradysz, Karolina Paradysz , Areas of unemployment in Poland – benchmark problem	115
Tomasz Szubert , How to play to lose the least? Classification of systems in sports bets	125
Izabela Szamrej-Baran , Classification of EU member states in view of fuel poverty	134
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , An attempt to use the gravity model in the analysis of commuters.....	143
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households	152
Hanna Dudek , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study	172
Ewa Chodakowska , Selected methods of classification in schools' rating.....	181
Bartosz Soliński , Renewable energy sector in the European Union – classification in the light of change management strategy	191
Krzysztof Szwarc , Classification of Wielkopolska voivodeship due to the demographic situation	201

Elżbieta Gołata, Grażyna Dehnel , Administrative registers in business analysis.....	211
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
Katarzyna Dębowska , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
Alina Bojan , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
Justyna Brzezińska , Log-linear analysis in the study of mortality in EU.....	246
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Latent class analysis in student satisfaction surveys.....	254
Bartłomiej Jefmański , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
Julita Stańczuk , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
Jerzy Krawczuk , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
Anna Czapkiewicz, Beata Basiura , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
Radosław Pietrzyk , Timing and selectivity in mutual funds performance measurement.....	305
Aleksandra Witkowska, Marek Witkowski , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
Marcin Pelka , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
Justyna Wilk , Comparative study of symbolic data classification software.....	332
Tomasz Bartłomowicz, Justyna Wilk , Application of symbolic data analysis methods for domain database searching.....	341
Kamila Migdał-Najman , A proposal of hybrid clustering method based on self-learning networks.....	351
Dorota Rozmus , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
Krzysztof Najman , A dynamic grouping based on self-learning GNG networks.....	369
Małgorzata Misztal , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
Mariusz Kubus , The application of pre-conditioning of explanatory variable for feature selection.....	386
Barbara Batóg, Jacek Batóg , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395

Katarzyna Wójcik, Janusz Tuchowski , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
Iwona Staniec , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawelczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
Iwona Foryś , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
Ewa Genge , Trimming approach to the mixtures of normal distributions.....	443
Jerzy Korzeniewski , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
Andrzej Dudek , SMS – proposal of new clustering algorithm.....	459
Artur Mikulec , Evaluation methods for the grouping result in cluster analysis.....	468
Małgorzata Machowska-Szewczyk , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
Artur Zaborski , PROFIT analysis and its using in the research of preferences.....	487
Karolina Bartos , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
Izabela Kurzawa , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
Aleksandra Luczak, Feliks Wysocki , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
Agnieszka Sompolska-Rzechuła , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
Joanna Banaś, Małgorzata Machowska-Szewczyk , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
Iwona Bąk , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
Aneta Becker , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
Katarzyna Dębowska , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

Anna Domagała , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Statistical analysis in demand research of ICT services in mobile networks.....	589
Hanna Gruchociak , Construction of regression estimator for two-level data	600
Tomasz Klimanek, Marcin Szymkowiak , Application of spatial models in indirect estimation of some labor market characteristics	609
Jarosław Lira , Forecasting of hog livestock production profitability in Poland	618
Christian Lis , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports	627
Beata Bieszk-Stolorz, Iwona Markowicz , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers	636
Lucyna Przezbórska-Skobiej, Jarosław Lira , Agritourism space of Poland and its valuation.....	645
Paweł Ulman , Model of expenses distribution and demand functions.....	654
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Methods of statistical analysis in research of molluscs	663

Ewa Genge

Uniwersytet Ekonomiczny w Katowicach

ANALIZA SKUPIEŃ OPARTA NA MIESZANKACH UCIĘTYCH ROZKŁADÓW NORMALNYCH

Streszczenie: W artykule przedstawione zostało najnowsze podejście modelowe, polegające na usuwaniu obserwacji oddalonych w trakcie estymacji parametrów mieszanki. W podejściu tym na macierze kowariancji nakładane są pewne ograniczenia, wyznaczane są funkcje wiarygodności uciętych zmiennych losowych, a parametry takiej mieszanki szacowane są za pomocą zmodyfikowanej wersji algorytmu EM, tj. algorytmu TCLUS_T [García-Escudero i in. 2010, s. 89-109]. Przeprowadzone zostały również badania porównawcze z bardziej popularnym, odpornym modelem mieszanek rozkładów normalnych.

Słowa kluczowe: model mieszanek, algorytm EM, algorytm TCLUS_T.

1. Wstęp

Analiza skupień oparta na mieszankach rozkładów normalnych nie pozwala na osiągnięcie zadowalających wyników podziału w przypadku, gdy w zbiorze danych występują obserwacje nietypowe. Do tej pory problem ten rozwiązywano dzięki modyfikacji modelu mieszanek, polegającej na dodaniu dodatkowej klasy (rozkładu składowego mieszanki) będącej reprezentantem obserwacji wyraźnie różniących się od obserwacji typowych [Dasgupta, Raftery 1998, s. 294-302].

W artykule przedstawione zostanie najnowsze podejście modelowe, polegające na usuwaniu obserwacji oddalonych w trakcie estymacji parametrów mieszanki. W podejściu tym na macierze kowariancji nakładane są pewne ograniczenia, wyznaczane są funkcje wiarygodności uciętych zmiennych losowych, a parametry takiej mieszanki szacowane są za pomocą zmodyfikowanej wersji algorytmu EM, tj. algorytmu TCLUS_T [García-Escudero i in. 2008, s. 1324-1345; 2010, s. 89-109]. W artykule dokonane zostanie również porównanie wyników klasyfikacji dla prezentowanego podejścia oraz wspomnianej już modyfikacji modelu mieszanek rozkładów normalnych.

Obliczenia zostaną przeprowadzone m.in. za pomocą pakietów: `tclust`, `clusterSim`, `mclust` oraz `mlbench` programu **R**.

2. Ucinanie obserwacji w podejściu modelowym

Gallegos [2002] oraz Gallegos i Ritter [2005] zaproponowali probabilistyczny model dla obserwacji oddalonych (*spurious outliers model*). Funkcję wiarygodności takiego modelu dla obserwacji \mathbf{x}_i można zapisać jako:

$$L = \left[\prod_{s=1}^u \prod_{i \in P_s} f_s(\mathbf{x}_i | \Theta_s) \right] \left[\prod_{i \in P_0} g_i(\mathbf{x}_i) \right], \quad (1)$$

\mathbf{x}_i – wektor obserwacji, $\mathbf{x}_i = [x_{i1}, \dots, x_{im}]$,

$f(\mathbf{x}_i | \Theta_s)$ – funkcja gęstości rozkładu składowego P_s , $P_s = P_1, \dots, P_u$,

$g(\mathbf{x}_i)$ – funkcja gęstości obserwacji nietypowych,

P_0 – rozkład składowy obserwacji nietypowych, $\#P_0 = [n\alpha]$,

α – parametr przycięcia (*trimming parameter*), tj. procent obserwacji nietypowych.

Liczebność zbioru obserwacji właściwych wyznaczana jest ze wzoru:

$$\#P_s = [n(1 - \alpha)]. \quad (2)$$

Zakłada się, że obserwacje nietypowe pochodzą z innego rozkładu prawdopodobieństwa o funkcji gęstości oznaczonej symbolem $g(\mathbf{x}_i)$. Obserwacje te nie biorą udziału w maksymalizacji funkcji wiarygodności danej wzorem (1) wtedy, gdy spełniony jest poniższy warunek:

$$\arg \max_{P_\alpha} \max_{\mu_s, \Sigma} \prod_{s=1}^u \prod_{i \in P_s} f_s(\mathbf{x}_i | \Theta_s) \subseteq \arg \max_{P_\alpha} \prod_{i \in \bigcup_{s=1}^u P_s} g_i(\mathbf{x}_i), \quad (3)$$

gdzie: $P_\alpha = \{ \{P_0, P_1, \dots, P_u\} : \bigcup_{s=0}^u P_s = \{1, \dots, n\}, P_w \cap P_r = \emptyset \text{ dla } w \neq r \text{ i } \#P_0 = [n\alpha] \}$ (por. [García-Escudero i in. 2011]).

García-Escudero i in. [2008] dokonali pewnej modyfikacji modelu określonego za pomocą wzoru (1), zakładając różne macierze kowariancji (Σ_s) oraz różne wagi dla poszczególnych klas ($\pi_s \geq 0, \sum_{s=1}^u \pi_s = 1$). Funkcję wiarygodności takiego modelu można zapisać jako:

$$L = \left[\prod_{s=1}^u \prod_{i \in P_s} \pi_s f_s(\mathbf{x}_i | \Theta_s) \right] \left[\prod_{i \in P_0} g_i(\mathbf{x}_i) \right]. \quad (4)$$

W modelu tym zakłada się również, że iloraz największej i najmniejszej wartości własnej jest mniejszy od pewnej stałej c bądź jej równy:

$$M_n / m_n \leq c, \quad (5)$$

$$M_n = \max_{s=1, \dots, u} \max_{j=1, \dots, m} \lambda_{j,s}, \quad (6)$$

$$m_n = \min_{s=1, \dots, u} \min_{j=1, \dots, m} \lambda_{j,s}, \quad (7)$$

gdzie $\lambda_{j,s}$ to wartości własne macierzy Σ_s . Podobne ograniczenia na macierz kowariancji nałożone zostały przez Hathaway [1985] w przypadku mieszanek o jednowymiarowej zmiennej losowej. Zbiór wszystkich parametrów Θ , które spełniają warunek (5), oznaczany jest jako Θ_s . W praktyce maksymalizuje się najczęściej logarytm funkcji wiarygodności określonej jako:

$$L_c(\alpha, s) = \sum_{s=1}^u \sum_{i \in P_s} \log(\pi_s f(\mathbf{x}_i; \boldsymbol{\mu}_s, \Sigma_s)). \quad (8)$$

Stała c pozwala w pewien sposób kontrolować ograniczenia nakładane na macierz kowariancji. W przypadku gdy $c=1$, macierz kowariancji jest najbardziej ograniczona, co odpowiada ważonej metodzie uciętych k -średnich (*weighted version of trimmed k-means*) [Cuesta-Albertos i in. 1997]. García-Escudero i in. [2010] pokazali, że dla $c=50$ możliwe jest znalezienie klas o największym rozproszeniu (obserwacji właściwych).

Oceny parametrów funkcji wiarygodności danej wzorem (8) wyznaczone są za pomocą algorytmu TCLUST [García-Escudero i in. 2008]. Algorytm ten jest pewną modyfikacją algorytmu EM [Dempster i in. 1977] składającego się z kroku estymacji parametrów oraz kroku maksymalizacji funkcji największej wiarygodności, rozszerzoną o tzw. krok koncentracji (*concentration step*).

Algorytm TCLUST

1. Wyznacz wartości początkowe dla wartości przeciętnych $\hat{\mu}_1^0, \dots, \hat{\mu}_s^0$, macierzy kowariancji $\hat{\Sigma}_1^0, \dots, \hat{\Sigma}_s^0$ oraz początkowe wagi $\hat{\pi}_1^0, \dots, \hat{\pi}_s^0$ dla każdego z rozkładów składowych.

2. W l -tej iteracji:

- Dla każdej obserwacji wyznacz prawdopodobieństwa *a posteriori*:

$$d_i = \pi_s^l f(\mathbf{x}_i; \boldsymbol{\mu}_s^l, \Sigma_s^l). \quad (9)$$

- Usuń $[n\alpha]$ obserwacji o najniższych wartościach d_i (przypisz je do klasy P_0). Pozostaw („skoncentruj”) pozostałe obserwacje i przydziel je do jednej z klas $P_s = P_1, \dots, P_u$, aby spełniony był warunek (10):

$$d_i^* = \max_{s=1, \dots, u} \pi_s^l f(\mathbf{x}_i; \boldsymbol{\mu}_s^l, \boldsymbol{\Sigma}_s^l). \quad (10)$$

- Dla każdej klasy P_s wyznacz liczebność n_s , wartość przeciętną $\boldsymbol{\mu}_s$ oraz macierz kowariancji $\boldsymbol{\Sigma}_s$. Jeżeli wartości własne wyznaczonej w ten sposób macierzy kowariancji nie spełniają warunków określonych równaniem (5), problem musi zostać rozwiązany ponownie.
 - Wyznacz nowe wartości parametrów $\Theta^{l+1} = (\boldsymbol{\pi}^{l+1}, \boldsymbol{\mu}^{l+1}, \boldsymbol{\Sigma}^{l+1})$, tak by funkcja wiarygodności dana wzorem (8) osiągnęła maksimum.
3. Procedurę iteracyjną powtarzaj kilkakrotnie.

Szczegółowe informacje na temat algorytmu TCLUST można znaleźć w pracy [García-Escudero i in. 2008; 2010; 2011].

3. Analiza empiryczna

Celem analizy empirycznej jest zbadanie jakości podziału uzyskanego za pomocą prezentowanego w artykule podejścia modelowego, polegającego na usuwaniu obserwacji oddalonych. Podejście to zostało dodatkowo porównane z bardziej popularną, odporną analizą skupień opartą na mieszankach rozkładów normalnych [Dasgupta, Raftery 1998; Fraley, Raftery 2002; Witek 2008]¹. W badaniu wykorzystano 4 zbiory generowane w przestrzeni dwuwymiarowej oraz 4 zbiory sztuczne, wykorzystywane w badaniach porównawczych metod taksonomicznych, dostępne w bibliotekach programu **R**. Zbiory te cechują się różnym stopniem separowalności (wśród zbiorów można znaleźć zbiory trudno i łatwo separowalne) o różnej liczbie klas. Poniżej przedstawiono krótką charakterystykę analizowanych zbiorów:

1. `6dnormals` – 6 niełatwo separowalnych klas, zawierających łącznie 500 obserwacji, wygenerowanych za pomocą funkcji `mlbench.2dnormals` z pakietu `mlbench`.

2. `Zb_gen_1` – 5 łatwo separowalnych klas o kształcie sferycznym, zawierającym łącznie 2500 obserwacji, generowanych za pomocą modelu 8 funkcji `cluster.Gen` z pakietu `clusterSim`.

3. `Zb_gen_2` – 5 niełatwo separowalnych klas, zawierających łącznie 2500 obserwacji, generowanych za pomocą modelu 6 funkcji `cluster.Gen` z pakietu `clusterSim`.

¹ W podejściu tym dokonywana jest modyfikacja modelu mieszank, polegająca na dodaniu dodatkowej klasy (rozkładu składowego mieszanki) będącej reprezentantem obserwacji wyraźnie różniących się od obserwacji typowych. Obserwacje dodatkowej klasy przyjmują zwykle rozkład jednostajny lub rozkład Poissona.

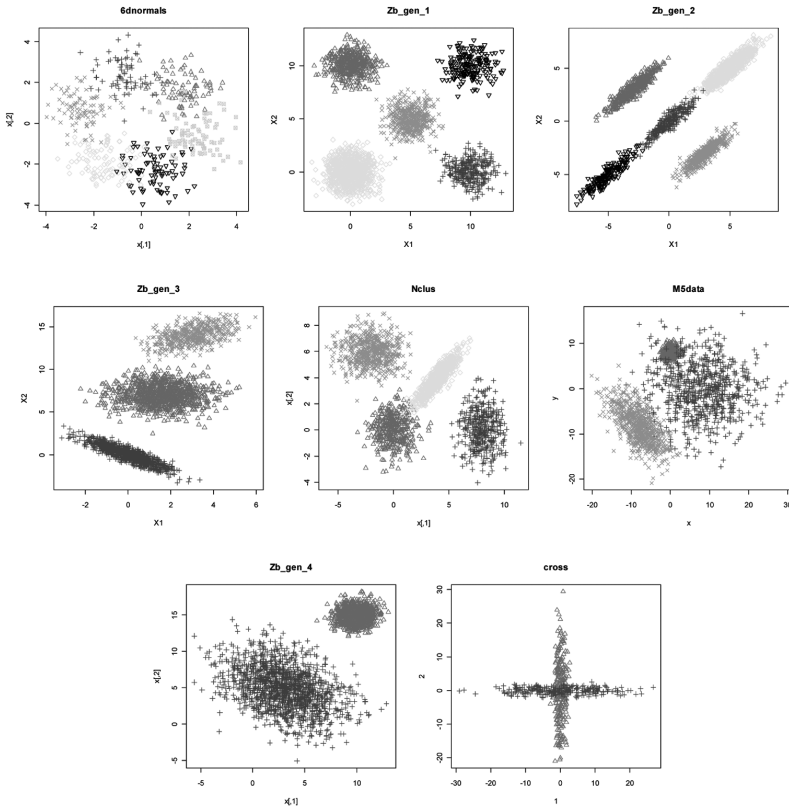
4. `Nclus` – 4 niełatwo separowalne klasy, zawierające łącznie 2200 obserwacji, zbiór dostępny w pakiecie `flexmix`².

5. `Zb_gen_3` – 3 łatwo separowalne klasy o kształcie eliptycznym, zawierające łącznie 2500 obserwacji, generowanych za pomocą modelu 13 funkcji `cluster.Gen` z pakietu `clusterSim`.

6. `M5data` – 3 trudno separowalne klasy, zawierające łącznie 1800 obserwacji, zbiór dostępny w pakiecie `tclust`.

7. `Zb_gen_4` – 2 niełatwo separowalne klasy, zawierające łącznie 2500 obserwacji, generowanych za pomocą funkcji `rmvnorm` z pakietu `mvtnorm`.

8. `cross` – 2 nakładające się na siebie klasy (nierozłączne), zawierające łącznie 500 obserwacji, zbiór dostępny w pakiecie `mclust`.



Rys. 1. Zbiory danych wykorzystane w analizie empirycznej

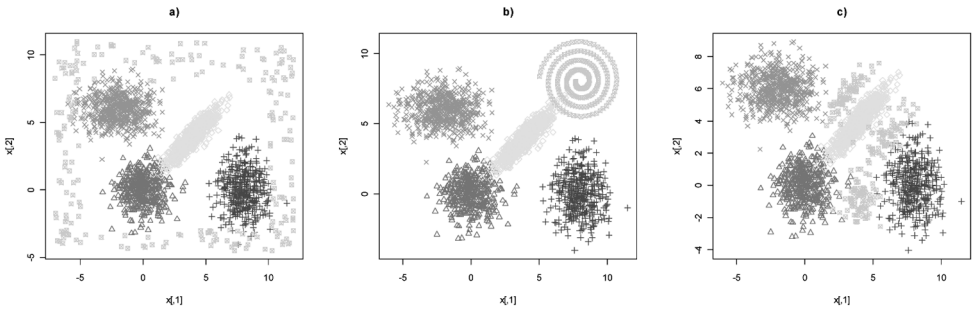
Źródło: obliczenia własne.

² W badaniu wygenerowano zbiór za pomocą funkcji `ExNclus()` o tej samej strukturze, lecz czterokrotnie większej liczbie obserwacji niż zbiór `Nclus`.

Wykorzystane w analizie zbiory zostały przedstawione na rys. 1. Do każdego z tych zbiorów dodano 10% obserwacji nietypowych, które znajdowały się:

- a) poza obszarem wszystkich klas,
- b) tworzyły dodatkową klasę o kształcie nieregularnym (spirali),
- c) pomiędzy klasami (na granicy ich seperowalności).

Rozważane warianty dodawania obserwacji oddalonych dla jednego z analizowanych zbiorów zilustrowano graficznie na rys. 2.



Rys. 2. Trzy sposoby dodawania (generowania) obserwacji oddalonych dla zbioru Nclus

Źródło: obliczenia własne.

Tabela 1. Miary Randa (R_{HA}) dla zbiorów danych

Zbiór	Obserwacje nietypowe	mclust	tclust
6dnormals	a)	0,74	0,71
	b)	0,59	0,70
	c)	0,66	0,66
Zb_gen_1	a)	0,98	0,98
	b)	0,96	0,98
	c)	0,90	0,96
Zb_gen_2	a)	0,98	0,98
	b)	0,94	0,98
	c)	0,92	0,95
Nclus	a)	0,95	0,95
	b)	0,89	0,95
	c)	0,87	0,92
Zb_gen_3	a)	0,98	0,98
	b)	0,88	0,98
	c)	0,88	0,97
M5data	a)	0,83	0,89
	b)	0,68	0,94
	c)	0,78	0,82
Zb_gen_4	a)	0,96	0,96
	b)	0,80	0,94
	c)	0,81	0,85
cross	a)	0,76	0,73
	b)	0,76	0,74
	c)	0,71	0,65

Źródło: obliczenia własne.

W analizowanych zbiorach znana jest przynależność obiektów do poszczególnych klas. Informacja ta jest jednak traktowana jako informacja *a priori*, która wykorzystana została do badania zgodności podziałów uzyskanych za pomocą obu odpornych podejść w modelowej analizie skupień.

Porównania jakości klasyfikacji dokonano za pomocą skorygowanej miary Randa (R_{HA}) [Hubert, Arabie 1985, s. 198]. Obliczenia wykonano w programie **R** z zastosowaniem m.in. funkcji `mclust` (podejście modelowe z klasą dodatkową) oraz `tclust` (podejście modelowe polegające na ucinaniu obserwacji nietypowych). Uzyskane wyniki przedstawiono w tab. 1.

Jeżeli chodzi o jakość podziałów dla zbiorów, w których obserwacje nietypowe generowane były z rozkładu jednostajnego poza obszarem klas, obie metody dały porównywalne wyniki. Jakość podziału wyrażona za pomocą miary Randa dla analizy skupień opartej na mieszkankach uciętych rozkładów normalnych jest wyraźnie lepsza w przypadku, gdy obserwacje nietypowe tworzą dodatkową klasę o nieregularnym kształcie lub znajdują się pomiędzy klasami. Metoda ta jednak daje nieco gorsze wyniki w przypadku zbiorów o klasach nierozłącznych (zbiór „cross”).

4. Podsumowanie

Podejście modelowe, polegające na usuwaniu obserwacji oddalonych, pozwala na osiągnięcie zadowalających wyników podziału w przypadku, gdy w zbiorze danych występują obserwacje nietypowe. W przypadku obserwacji nietypowych tworzących dodatkową klasę o nieregularnym kształcie lub znajdujących się pomiędzy klasami metoda ta dała znacznie lepsze wyniki w przypadku zbiorów niejednorodnych (z wyjątkiem zbioru „cross”). W dalszych badaniach analizie poddane zostaną również zbiory o różnych wymiarach i różnych udziałach obserwacji nietypowych.

Literatura

- Cuesta-Albertos J.A., García-Escudero L.A., Gordaliza A., *Trimmed k-means: an attempt to robusify quantizers*, „The Annals of Statistics” 1997, no 25.
- Dasgupta A., Raftery A.E., *Detecting features in spatial point processes with clutter via model-based clustering*, „Journal of the American Statistical Association” 1998, no 93.
- Dempster A.P., Laird N.M., Rubin D.B., *Maximum likelihood for incomplete data via the EM algorithm (with discussion)*, „Journal of the Royal Statistical Society” 1977, no B (39).
- Fraley C., Raftery A.E., *Model-based clustering, discriminant analysis, and density estimation*, „Journal of the American Statistical Association” 2002, no 97.
- Gallegos M.T., *Maximum Likelihood Clustering with Outliers*, [w:] K. Jajuga, A. Sokołowski, H. Bock, *Classification, Clustering and Data Analysis: Recent Advances and Applications*, Springer-Verlag, 2002.
- Gallegos M.T., Ritter G., *A robust method for cluster analysis*, „Annals of Statistics” 2005, no 33(1).
- García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Isacar A., *A general trimming approach to robust cluster analysis*, „The Annals of Statistics” 2008, no 36(4).

- García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Iscar A., *A review of clustering methods*, „Advances of data analysis and classification”, Springer 2010, no 4.
- García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Iscar A., *Exploring the number of groups in robust model-based clustering*, “Statistics and Computing” 2011, no 21(4).
- Hathaway R.J., *A constrained formulation of maximum likelihood estimation for normal mixture distributions*, “The Annals of Statistics” 1985, no 13.
- Hubert L.J., Arabie P., *Comparing partitions*, „Journal of Classification” 1985, no 1.
- Witek E., *Obserwacje nietypowe w analizie skupień – podejście modelowe*, [w:] J. Harasim, Warsztaty Doktoranckie’07, *Zarządzanie–Finanse–Ekonomia*, Wydawnictwo AE, Katowice 2008.

TRIMMING APPROACH TO THE MIXTURES OF NORMAL DISTRIBUTIONS

Summary: The paper presents trimming approach to the mixtures of normal distributions. In this approach the proportion of the most outlying observations is trimmed (during parameter estimation), different constraints on the cluster scatter matrices are assumed and trimming likelihood is defined. To estimate the parameter of this kind of mixture mostly the modified version of EM, i.e. TCLUS algorithm is applied [García-Escudero et al. 2010, p. 89-109]. The results of clustering given by trimming approach and well known robust mixtures of normal distributions are compared.

Keywords: mixture model, EM algorithm, TCLUS algorithm.