

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

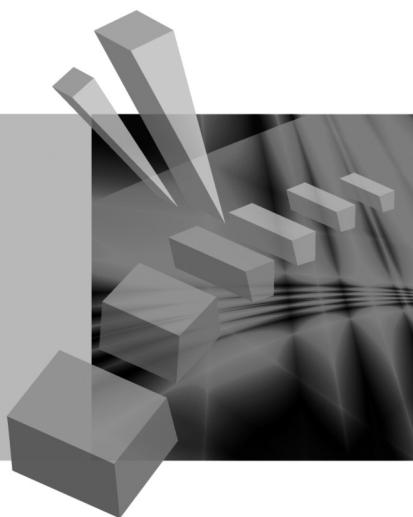
RESEARCH PAPERS

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,
Mirosław 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 wspomagania 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
Sylvia 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 Pawelczyk, 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 Szwarz , 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 Pawełczyk, 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

Andrzej Dudek

Uniwersytet Ekonomiczny we Wrocławiu

SMS – PROPOZYCJA NOWEGO ALGORYTMU ANALIZY SKUPIEŃ

Streszczenie: Klasyfikacja spektralna (*spectral clustering* [Ng, Jordan, Weiss 2002; von Luxburg 2006]) i klasyfikacja za pomocą średniej przesunięcia okna w kierunku wektora średniej (*mean shift clustering* [Wang, Xiu, Zamar 2007]) to dwa stosunkowo nowe podejścia w analizie skupień, dające, zwłaszcza dla skupień o nietypowych kształtach, lepsze rezultaty niż klasyczne metody *k*-średnich, *k*-medoidów czy hierarchiczne metody aglomeracyjne. Artykuł zawiera propozycję algorytmu o roboczej nazwie SMS (*Spectral-Mean Shift*) łączącego cechy obu podejść i wyróżniającego się wśród innych algorytmów analizy skupień m.in.:

- możliwością analizy skupień o nietypowych kształtach;
- możliwością automatycznego rozpoznawania liczby skupień;
- lepszą odpornością na zmienne zakłócające.

Słowa kluczowe: analiza skupień, klasyfikacja spektralna, SMS.

1. Wstęp

Klasyfikacja spektralna (*spectral clustering* [Ng, Jordan, Weiss 2002; von Luxburg 2006]) i klasyfikacja za pomocą przesunięcia okna w kierunku wektora średniej (*mean shift clustering* [Wang, Xiu, Zamar 2007]) to dwa stosunkowo nowe podejścia w analizie skupień, dające, zwłaszcza dla skupień o nietypowych kształtach, lepsze rezultaty niż klasyczne metody *k*-średnich, *k*-medoidów czy hierarchiczne metody aglomeracyjne. W artykule zostanie zaproponowane połączenie obu podejść w algorytm o roboczej nazwie SMS (*Spectral – Mean Shift*).

2. Podejście spektralne

Klasyfikacja spektralna ([Ng, Jordan, Weiss 2002; Karatzoglou 2006; von Luxburg 2006]) to stosunkowo nowe i szybko rozwijające się podejście w analizie skupień. Nie jest ona ściśle nowym algorytmem, ale raczej sposobem przygotowania danych do znanych metod analizy skupień (np. *k*-średnich). Podejście to daje bardzo obiecujące rezultaty zwłaszcza dla nietypowych (tj. nie pochodzących z rozkładu normalne-

go) kształtów. W większości badań technika ta jest stosowana w połączeniu z klasyczną metodą *k-średnich*, ale założenie to nie jest oparte na żadnej podstawie teoretycznej. Procedura klasyfikacji spektralnej ma wiele wariantów, dla wszystkich z nich można wyróżnić sześć najważniejszych etapów (według [Ng, Jordan, Weiss 2002, za Walesiak, Dudek 2009]):

1. Konstrukcja macierzy danych $\mathbf{X} = [x_{ij}]$ o wymiarach $n \times m$ ($i = 1, \dots, n$ – numer obiektu, $j = 1, \dots, m$ – numer zmiennej). Dla danych metrycznych należy przeprowadzić normalizację wartości zmiennych.

2. Zastosowanie estymatora jądrowego do obliczenia macierzy podobieństw $\mathbf{A} = [A_{ik}]$ (*affinity matrix*) między obiektami.

3. Konstrukcja znormalizowanej macierzy Laplace’a $\mathbf{L} = \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$ (\mathbf{D} – diagonalna macierz wag, w której na głównej przekątnej znajdują się sumy każdego wiersza z macierzy $\mathbf{A} = [A_{ik}]$, a poza główną przekątną są zera). W rzeczywistości znormalizowana macierz Laplace’a przyjmuje postać: $\mathbf{I} - \mathbf{L}$.

4. Obliczenie wartości własnych i odpowiadających im wektorów własnych (o długości równej jeden) dla macierzy \mathbf{L} . Uporządkowanie wektorów własnych według malejących wartości własnych. Pierwsze u wektorów własnych (u – liczba klas) tworzy macierz $\mathbf{E} = [e_{ij}]$ o wymiarach $n \times u$.

5. Przeprowadzenie normalizacji tej macierzy zgodnie ze wzorem $y_{ij} = e_{ij} / \sqrt{\sum_{j=1}^u e_{ij}^2}$ ($i = 1, \dots, n$ – numer obiektu, $j = 1, \dots, u$ – numer zmiennej, u – liczba klas). Dzięki tej normalizacji długość każdego wektora wierszowego macierzy $\mathbf{Y} = [y_{ij}]$ jest równa jeden.

6. Macierz \mathbf{Y} stanowi punkt wyjścia zastosowania klasycznych metod analizy skupień.

Poszczególne warianty algorytmów w podejściu spektralnym mogą się różnić (por. [Walesiak, Dudek 2009]):

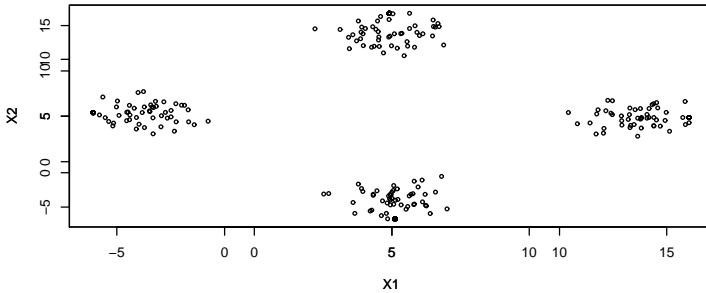
- typem estymatora jądrowego w kroku 2;
- formułą konstrukcji macierzy Laplace’a w kroku 4 (zob. np. [von Luxburg 2006]);
- określaniem wartości parametru σ – szerokości pasma (*kernel width*) liczby skupień.

W oryginalnej propozycji Ng, Jordan i Weiss [2002] postulują użycie algorytmu *k-średnich* jako ostatecznego algorytmu analizy skupień w kroku szóstym.

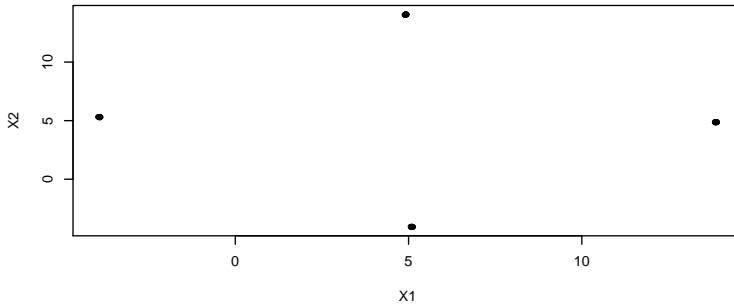
3. Klasyfikacja za pomocą przesunięcia okna w kierunku wektora średniej w kierunku wektora średniej

Klasyfikacja za pomocą przesunięcia okna w kierunku wektora średniej to nieparametryczna technika analizy skupień iteracyjnie przesuwająca punkty m -wymiarowej przestrzeni euklidesowej w kierunku środków ciężkości skupień.

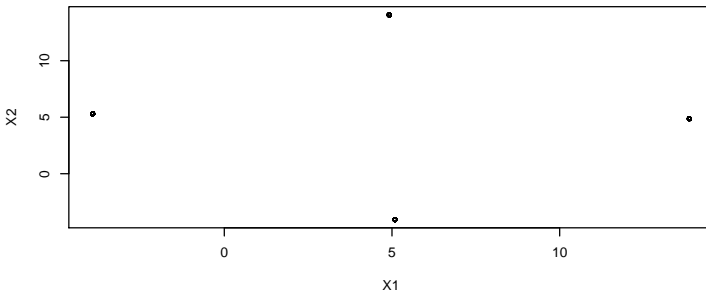
Dane pierwotne



Dane przesunięte w pierwszym kroku algorytmu



Dane przesunięte w drugim kroku algorytmu



Rys. 1. Dane w kolejnych rundach klasyfikacji za pomocą przesunięcia okna w kierunku wektora średniej

Źródło: opracowanie własne.

Niech $\{x_i\} i=1, 2, \dots, n$ będzie zbiorem danych w m -wymiarowej przestrzeni euklidesowej. Wektor przesunięcia średniej jest zdefiniowany (w ogólnej postaci) jako (1):

$$ms_h(x) = \frac{\sum_{i=1}^n -x_i k'_x \left(\left(\frac{x-x_i}{h} \right)^2 \right)}{\sum_{i=1}^n -k'_x \left(\left(\frac{x-x_i}{h} \right)^2 \right)} - x, \quad (1)$$

gdzie k' jest pochodną funkcji jądrowej (gaussowskiej, Bessela, Epanechnikova itd.). W szczególnym przypadku dla jądra Epanechnikova (por. [Comaniciu, Meer 1999; Korzeniewski 2005]) przyjmuje ona formę (2):

$$ms_h(x) = \sum_{k_x(x_i) > 0} x_i - x, \quad (2)$$

gdzie:

$$k_x(x_i) = \begin{cases} \frac{3}{4} \left(1 - \sum_{j=j}^m \left(\frac{x_j - x_{ij}}{h} \right)^2 \right) & d(x, x_i) < h \\ 0 & d(x, x_i) \geq h \end{cases}$$

Wektor przesunięcia średniej zawsze skierowany jest w kierunku maksymalnego zwiększenia gęstości. W procedurze klasyfikacji z przesunięciem okna w kierunku wektora średniej w kolejnych krokach algorytmu najpierw obliczany jest wektor przesunięcia, a następnie współrzędne punktów są przesuwane w kierunku wyznaczonym przez ten wektor, aż do osiągnięcia konwergencji (punktów stacjonarnych). Schematyczny obraz procedury klasyfikacji z przesunięciem okna w kierunku wektora średniej przedstawia rys. 1.

4. Algorytm SMS

Algorytm *Spectral – Mean Shift clustering* łączy cechy obu podejść. Jego idea polega na tym, aby punkty wyznaczone przez znormalizowaną macierz wartości własnych macierzy Laplace'a iteracyjnie przesuwać w kierunku wyznaczonym przez wektor przesunięcia średniej. W postaci ogólnej algorytm ten można więc zapisać w siedmiu podstawowych krokach:

1. Konstrukcja macierzy danych.
2. Obliczenia macierzy podobieństw **A**.
3. Konstrukcja znormalizowanej macierzy Laplace'a.
4. Obliczenie wartości własnych i odpowiadających im wektorów własnych macierzy **L**. Utworzenie macierzy **E** $\mathbf{E} = [e_{ij}]$.

5. Obliczenie znormalizowanej macierzy Y .

6. Iteracyjne przesuwanie punktów macierzy Y w kierunku wyznaczonym przez wektor przesunięcia średniej aż do osiągnięcia konwergencji (punktów stacjonarnych).

7. Właściwa klasyfikacja według reguły sekwencyjnej: *jeśli odległość obiektu od prototypu klasy jest mniejsza niż zadana wartość, obiekt zostaje dołączony do danego skupienia, jeżeli nie, tworzy prototyp nowej klasy.*

5. Rezultaty klasyfikacji za pomocą nowego algorytmu dla nietypowych kształtów skupień

W procedurze symulacyjnej analizy porównawczej porównano rezultaty klasyfikacji za pomocą nowego algorytmu z innymi algorytmami klasyfikacji dla podejścia spektralnego¹. W eksperymencie wykorzystano sześć modeli zbiorów danych:

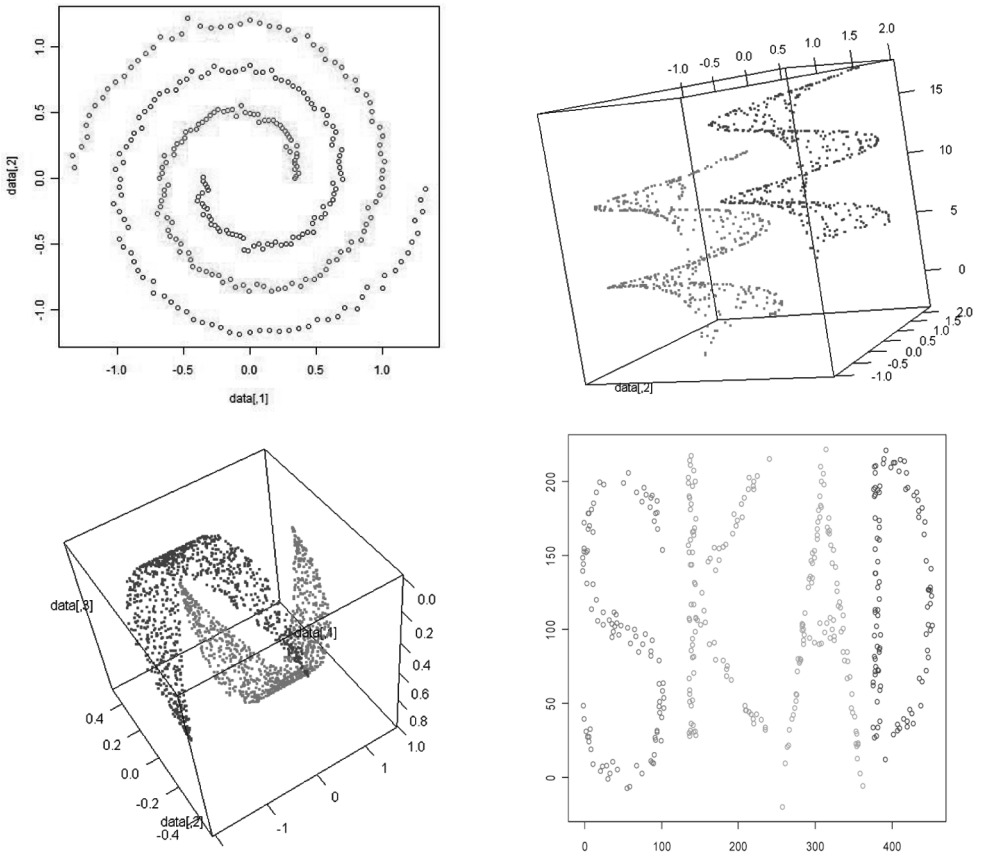
- zbiory *spirale* i *dini* pochodzą z pakietu `mlbench` środowiska **R**,
- zbiory *clusterSim1* i *clusterSim2* pochodzą z pakietu `clusterSim` środowiska **R**,
- zbiory *banany* i *SKAD* tworzone są przez autorskie funkcje programu **R**, niedołączone jeszcze do żadnej biblioteki tego środowiska.

Dla poszczególnych modeli w każdym eksperymencie wygenerowano 50 zbiorów danych, przeprowadzono procedurę klasyfikacyjną i porównano otrzymane rezultaty klasyfikacji ze znaną strukturą klas za pomocą skorygowanego indeksu Randa [Hubert, Arabie 1985]).

Uwzględniono następujące metody klasyfikacji: 1. `spec-kmeans` – klasyfikacja spektralna z metodą *k-średnich* użytą jako algorytm analizy skupień w kroku 6.; 2. `spec-ward` – klasyfikacja spektralna z hierarchiczną aglomeracyjną metodą powiększonej sumy kwadratów odległości użytą jako algorytm analizy skupień w kroku 6.; 3. `spec-mcquitty` – klasyfikacja spektralna z hierarchiczną aglomeracyjną metodą ważonej średniej klasowej użytą jako algorytm analizy skupień w kroku 6.; 4. `spec-pam` – klasyfikacja spektralna z metodą *k-medoidów* użytą jako algorytm analizy skupień w kroku 6.; 5. `spec-kmeans` – klasyfikacja spektralna z metodą *k-średnich* użytą jako algorytm analizy skupień w kroku 6.; 6. SMS – nowy algorytm.

Tabela 1 prezentuje uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa policzonego z 50 symulacji. W pięciu na sześć przypadkach nowy algorytm dał najlepsze wyniki. W szóstym przypadku wszystkie algorytmy podejścia spektralnego dały ten sam rezultat.

¹ Porównanie rezultatów klasyfikacji dla nietypowych skupień pomiędzy algorytmami podejścia spektralnego a „klasycznymi” algorytmami analizy skupień można znaleźć np. w pracach [Dudek 2009; Waleśiak, Dudek 2010].



Rys. 2. Przykładowe zbiory danych utworzone z wykorzystaniem funkcji pakietu mlbench (*spirale*, *dini*) oraz zbiorów własnych (*banany*, *SKAD*)

Źródło: opracowanie własne z wykorzystaniem programu R.

Tabela 1. Uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa

Metoda	Średnia wartość indeksu Randa/kolejność	Zbiory danych												
		cluster-Sim1		cluster-Sim2		spirale	banany	dini	SKAD					
spec-kmeans	0,878	4	1	–	1	–	0,926	–	0,839	2-4	0,568	4	0,936	2-4
spec-ward	0,867	5	1	–	1	–	0,926	–	0,838	5	0,591	5	0,847	6
spec-mcquitty	0,883	2	1	–	1	–	0,926	–	0,838	6	0,6	6	0,936	2-4
spec-pam	0,877	3	1	–	1	–	0,926	–	0,839	2-4	0,564	2	0,936	2-4
spec-diana	0,866	6	1	–	1	–	0,926	–	0,839	2-4	0,567	3	0,868	5
SMS	0,893	1	1	–	1	–	0,926	–	0,842	1	0,632	1	0,945	1

Źródło: obliczenia własne z wykorzystaniem programu R.

Należy również nadmienić, iż nowy algorytm jako jedna z metod podejścia spektralnego daje lepsze rezultaty niż tradycyjne metody klasyfikacji w przypadku zbiorów zawierających zmienne zakłócające (por. np. [Dudek 2009]).

6. Wykrywanie liczby skupień

Algorytm SMS można zaproponować również w wersji znajdującej optymalną liczbę skupień dla danego zbioru danych. W tym celu należy nieznacznie zmodyfikować punkty 4-7 algorytmu i zastąpić je punktami 4'-8':

Obliczenie wartości własnych i odpowiadających im wektorów własnych dla macierzy L . Utworzenie macierzy E z pierwszych u wektorów własnych. Powtarzanie kroków 5-7 dla $u = 2, 3, \dots, n/2$.

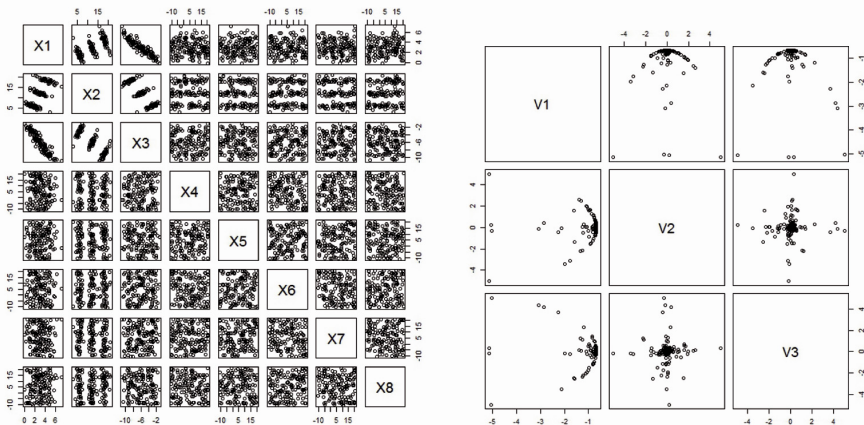
Obliczenie znormalizowanej macierzy Y .

Iteracyjne przesuwanie współrzędnych punktów z macierzy Y w kierunku wyznaczonym przez wektor przesunięcia średniej aż do osiągnięcia konwergencji (punktów stacjonarnych).

Właściwa klasyfikacja według reguły sekwencyjnej: *jeśli odległość obiektu od prototypu klasy jest mniejsza niż zadana wartość, obiekt zostaje dołączony do danej skupienia, jeżeli nie, tworzy prototyp nowej klasy.*

Wybór u , dla którego zadana liczba skupień (p. 4) jest równa rzeczywistej liczbie skupień (p. 7).

W przypadku gdy w kroku ósmym algorytmu nie można określić takiego u , że zadana liczba skupień jest równa rzeczywistej liczbie skupień, lub można znaleźć więcej niż jedną taką wartość u , do wyznaczenia optymalnej liczby klas należy wykorzystać jeden z indeksów jakości klasyfikacji i wskazać tę liczbę klas, dla której wartość indeksu jest optymalna (por. np. [Walesiak, Dudek 2007; Walesiak 2009, s. 418]).



Rys. 3. Przykładowy zbiór danych ze zmiennymi zakłócającymi przed transformacją spektralną i po niej

Źródło: opracowanie własne z wykorzystaniem programu R.

Należy jednak zaznaczyć, iż sytuacja taka występuje najczęściej, gdy dane pierwotne nie mają czytelnej struktury klas lub gdy występują w nich zmienne zakłócające. Przykładową taką sytuację przedstawia rys. 3.

Tabela 2. Wartości indeksu wskaźnikowego (S) dla liczby skupień $u = 2, \dots, 10$ dla danych z rys. 3

u	2	3	4	5	6	7	8	9	10
S	0,209	0,223	-0,068	-0,084	-0,101	-0,122	-0,122	-0,107	-0,107

Źródło: obliczenia własne z wykorzystaniem programu R.

Do wyznaczenia liczby klas dla danych z rys. 3. wykorzystany został indeks sylwetkowy. Tabela 3 przedstawia wartości tego indeksu dla liczby skupień $u = 2, 3, \dots, 10$. Wartość optymalna (maksymalna) indeksu została osiągnięta dla 3 skupień, jednak jego niska wartość świadczy o słabej stabilności otrzymanych skupień.

7. Złożoność czasowa algorytmu

Nowy algorytm, jak wszystkie algorytmy podejścia spektralnego, wymaga obliczenia wartości własnych macierzy \mathbf{L} . Jest to najbardziej złożona obliczeniowo część algorytmu. Metody obliczenia wartości własnych często w praktyce wykorzystują dekompozycję *svd* [Anderson i in. 1999, s. 573], czas tej czynności jest wprost proporcjonalny do rozmiaru macierzy \mathbf{L} ($n \times n$). Tabela 3 przedstawia przybliżone średnie czasy tej czynności dla $n = 10, 100, \dots, 1\,000\,000$ w środowisku R.

Tabela 3. Średni przybliżony czas znalezienia wartości własnych macierzy \mathbf{L}

n	10	100	1000	10 000	100 000	1 000 000
Średni czas (w min.)*	0,028	2,83	283	28 333	283 333 333	28 333 333 333

* Dla $n > 1000$ estymowane wartości przybliżone.

Źródło: obliczenia własne z wykorzystaniem programu R.

Z przedstawionej symulacji wynika, iż praktyczne stosowanie algorytmu należy ograniczyć do zbiorów rzędu nie większego niż 1000 obiektów.

8. Wnioski i problemy otwarte

W artykule przedstawiona została propozycja nowego algorytmu analizy skupień łączącego cechy podejścia spektralnego i klasyfikacji za pomocą przesunięcia okna w kierunku wektora średniej. Prezentowany algorytm cechuje się:

- Możliwością osiągania lepszych rezultatów analizy skupień niż tradycyjne algorytmy klasyfikacyjne, zwłaszcza w przypadku nietypowych kształtów skupień.
- Możliwością automatycznego określenia liczby skupień.

- Dobrą odpornością na zmienne zakłócające.
Problemem otwartym jest klasyfikacja danych z dużych (>1000 obiektów) zbiorów danych.

Literatura

- Anderson E., Bai Z., Bischof C., Blackford S., Demmel J., Dongarra J., Du Croz J., Greenbaum A., Hammarling A., McKenney A., Sorensen D., *LAPACK User's Guide*, SIAM, Philadelphia 1999.
- Comanicu D., Meer P., *Mean Shift Analysis and Applications*, IEEE Int. Conf. Computer Vision (ICCV'99), Kerkyra, Greece 1999.
- Dudek A., *Klasyfikacja spektralna a tradycyjne metody analizy skupień*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 37, Wydawnictwo UE, Wrocław 2009.
- Hubert L.J., Arabie P., *Comparing partitions*, "Journal of Classification" 1985, no 1.
- Karatzoglou A., *Kernel methods. Software, algorithms and applications*, Dissertation, Technical University, Wien 2006.
- Korzeniewski J., *Propozycja nowego algorytmu wyznaczającego liczbę skupień*, [w:] Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, Taksonomia 12, Wydawnictwo AE, Wrocław 2005.
- Ng A., Jordan M., Weiss Y., *On Spectral Clustering: Analysis and an Algorithm*, [w:] *Advances in Neural Information Processing Systems 14*, T. Dietterich, S. Becker, Z. Ghahramani (red.), MIT Press, 2002.
- von Luxburg U., *A Tutorial on Spectral Clustering*, Max Planck Institute for Biological Cybernetics, Technical Report TR-149, 2006.
- Walesiak M., *Analiza skupień*, [w:] *Statystyczna analiza danych z wykorzystaniem programu R*, M. Walesiak, E. Gatnar (red.), Wydawnictwo Naukowe PWN, Warszawa 2009.
- Walesiak M., Dudek A., *Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych – charakterystyka problemu*, Zeszyty Naukowe Uniwersytetu Szczecińskiego nr 450, Szczecin 2007.
- Walesiak M., Dudek A., *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 84, Wydawnictwo UE, Wrocław 2009.
- Wang S., Qiu W., Zamar R. H., CLUES: *A non-parametric clustering method based on local shrinking*, „Computational Statistics & Data Analysis“ 2007, vol. 52, issue 1.

SMS – PROPOSAL OF NEW CLUSTERING ALGORITHM

Summary: Spectral clustering [Ng, Jordan, Weiss 2002; von Luxburg 2006] and mean shift clustering [Wang, Xiu, Damar 2007] are two relatively new approaches in cluster analysis, giving, especially for clusters of unusual shapes, better results than classical methods such as *k-means*, *k-medoids* or hierarchical agglomerative methods. The article contains a proposal for algorithm with the working name – SMS (*Spectral-Mean Shift*) that combines features of both approaches, distinguishing among other cluster analysis algorithms with:

- the possibility of cluster analysis of unusual shapes,
- the ability to automatically identify the number of clusters,
- better resistance to interference (noisy) variables.

Keywords: cluster analysis, spectral clustering, SMS.