

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

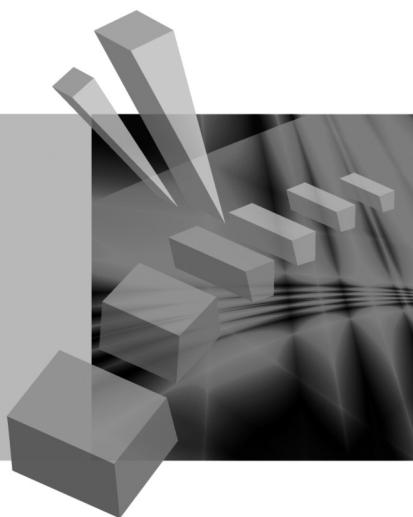
RESEARCH PAPERS

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
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 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 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

Dorota Rozmus

Uniwersytet Ekonomiczny w Katowicach

PORÓWNANIE DOKŁADNOŚCI TAKSONOMII SPEKTRALNEJ ORAZ ZAGREGOWANYCH ALGORYTMÓW TAKSONOMICZNYCH OPARTYCH NA IDEI METODY *BAGGING*

Streszczenie: Kiedy stosuje się metody taksonomiczne w jakimkolwiek zagadnieniu klasyfikacji, ważną kwestią jest zapewnienie wysokiej poprawności wyników grupowania. Stąd też w literaturze wciąż proponowane są nowe rozwiązania, które mają przynieść poprawę dokładności grupowania w stosunku do tradycyjnych metod. Przykładem mogą tu być metody polegające na zastosowaniu podejścia zagregowanego oraz algorytmy spektralne. Głównym celem tego artykułu jest porównanie dokładności zagregowanych algorytmów taksonomicznych opartych na idei metody *bagging* [Dudoit, Fridlyand 2003; Hornik 2005; Leisch 1999] oraz spektralnego algorytmu taksonomicznego zaproponowanego przez Ng i in. [2001].

Słowa kluczowe: taksonomia, zagregowane algorytmy taksonomiczne, taksonomia spektralna, dokładność.

1. Wstęp

Kiedy stosuje się metody taksonomiczne w jakimkolwiek zagadnieniu klasyfikacji, ważną kwestią jest zapewnienie wysokiej dokładności wyników grupowania. Od niej bowiem zależeć będzie skuteczność wszelkich decyzji podjętych na ich podstawie. Przez pojęcie *dokładność grupowania* należy rozumieć zdolność metody do rozpoznawania rzeczywistej struktury klas. Stąd też w literaturze wciąż proponowane są nowe rozwiązania, które mają przynieść poprawę dokładności grupowania w stosunku do tradycyjnych metod (np. *k*-średnich, metod hierarchicznych). Przykładem mogą tu być metody polegające na zastosowaniu podejścia zagregowanego oraz algorytmy spektralne. Taksonomia spektralna polega na zastosowaniu wartości własnych pochodzących ze spektralnej dekompozycji macierzy podobieństwa opisującej badane obiekty. Podejście zagregowane w taksonomii można natomiast sformułować następująco: mając wyniki wielokrotnie przeprowadzonego grupowania, znajdź zagregowany podział ostateczny.

Głównym celem tego artykułu jest porównanie dokładności zagregowanych i spektralnych algorytmów taksonomicznych. W badaniu pod uwagę wzięta zostanie tylko specyficzna klasa metod agregacji, która oparta jest na idei metody *bagging* [Dudoit, Fridlyand 2003; Hornik 2005; Leisch 1999]. Natomiast jako algorytm spektralny zastosowana będzie metoda zaproponowana przez Ng i in. [2001].

2. Metoda *bagging* w taksonomii

Metoda *bagging* w taksonomii jest pewną ogólną ideą, w ramach której narodziło się kilka szczegółowych rozwiązań. Generalnie polega ona na losowaniu B prób bootstrapowych i dokonywaniu ich grupowania w celu uzyskania podziałów składowych, które będą agregowane. Różnice w poszczególnych rozwiązaniach polegają na zastosowaniu różnych operatorów agregacji.

Propozycja Leischa

Leisch [1999] zaproponował, by w pierwszym kroku na podstawie każdej podpróby bootstrapowej określane były rezultaty grupowania przy zastosowaniu tzw. bazowej metody taksonomicznej, którą jest jedna z metod iteracyjno-optymalizacyjnych, np. k -średnich. W kolejnym etapie ostateczne centra skupień przekształcane są w nowy zbiór danych obejmujący $B \times K$ obserwacji (K to liczba skupień w metodzie bazowej), który poddawany jest podziałowi za pomocą metod hierarchicznych. Uzyskany dendrogram jest podstawą ostatecznego podziału – obserwacje z pierwotnego zbioru przydzielane są do tej grupy, której środek ciężkości znajduje się w minimalnej odległości euklidesowej.

Szczegółowo algorytm zaproponowany przez Leischa przebiega w następujących krokach:

1. Z pierwotnego n -elementowego zbioru $G = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ należy wylosować B prób bootstrapowych $G_n^1, G_n^2, \dots, G_n^B$, losując n obserwacji przy wykorzystaniu schematu losowania ze zwracaniem.

2. Na podstawie każdego podzbioru za pomocą metod iteracyjno-optymalizacyjnych (np. k -średnich) dokonuje się podziału na grupy obserwacji podobnych do siebie, uzyskując w ten sposób $B \times K$ załączków skupień $c_{11}, c_{12}, \dots, c_{1K}, c_{21}, \dots, c_{BK}$ gdzie K oznacza liczbę skupień w metodzie bazowej, a c_{bk} jest k -tym załączkiem znalezionym na podstawie podpróby G_n^b .

3. Niech załączki skupień uzyskane na podstawie kolejnych prób bootstrapowych utworzą nowy zbiór danych $C^B = \{c_{11}, \dots, c_{BK}\}$.

4. Do tak skonstruowanego zbioru należy zastosować hierarchiczną metodę taksonomiczną, uzyskując w ten sposób dendrogram.

5. Podział na grupy pierwotnego zbioru danych określany jest w ten sposób, że dendrogram uzyskany na podstawie zbioru C^B jest cięty na określonym przez

badacza poziomie, co prowadzi do uzyskania grup obiektów podobnych C_1^B, \dots, C_m^B , gdzie $1 \leq m \leq BK$. Każda obserwacja x_i z pierwotnego zbioru danych G jest przydzielana do tej grupy, w której znajduje się najbliższy załączek $c(x_i)$.

Propozycja Dudoit i Fridlyand

Metoda *bagging* w wersji zaproponowanej przez Dudoit i Fridlyand [2003] stosuje algorytm iteracyjno- optymalizacyjny do oryginalnego zbioru danych i poszczególnych prób bootstrapowych. Następnie, po dokonaniu permutacji etykiet klas w poszczególnych podpróbach tak, by zachodziła jak największa zbieżność z podziałem obiektów z oryginalnego zbioru danych, stosuje głosowanie majoryzacyjne w celu określenia ostatecznego grupowania zagregowanego.

Kroki zaproponowanego przez nich algorytmu można ująć według poniższego schematu.

Dla założonej liczby klas K :

1. Zastosuj iteracyjno- optymalizacyjny algorytm taksonomiczny T do pierwotnego zbioru danych $G = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, uzyskując w ten sposób etykiety klas

$$T(x_i, G) = \hat{y}_i \text{ dla każdej obserwacji } x_i, i = 1, \dots, n.$$

2. Skonstruuj b -tą próbę bootstrapową $G_n^b = \{\mathbf{x}_1^b, \dots, \mathbf{x}_n^b\}$.

3. Zastosuj metodę taksonomiczną T do skonstruowanej próby bootstrapowej G_n^b , uzyskując podział na klasy: $T(x_i^b, G_n^b)$ dla każdej obserwacji w zbiorze G_n^b .

4. Dokonaj permutacji etykiet klas przyznanych obserwacjom w próbie bootstrapowej G_n^b tak, by zachodziła jak największa zbieżność z klasyfikacją obiektów z oryginalnego zbioru danych G . Niech PR_K oznacza zbiór wszystkich permutacji zbioru liczb całkowitych $1, \dots, K$. Znajdź permutację $\tau^b \in PR_K$ maksymalizującą:

$$\sum_{i=1}^n I(\tau(T(x_i^b, G_n^b)) = T(x_i^b, G)), \quad (1)$$

gdzie $I(\cdot)$ to funkcja wskaźnikowa równa 1, gdy zachodzi prawda, 0 w przypadku przeciwnym.

5. Powtórz kroki 2-4 B razy. Ostatecznie zaklasyfikuj i -tą obserwację, stosując głosowanie majoryzacyjne, zatem przydzielając ją do tej klasy, dla której zachodzi:

$$\arg \max_{1 \leq k \leq K} \sum_{b: x_i \in G_n^b} I(\tau^b(T(x_i, G_n^b)) = k). \quad (2)$$

Propozycja Hornika

W metodzie tej po skonstruowaniu B prób bootstrapowych i zastosowaniu do nich algorytmu taksonomicznego uzyskuje się podziały składowe. Grupowanie za-

gregowane natomiast jest uzyskiwane za pomocą tzw. podejścia optymalizacyjnego, które ma za zadanie zminimalizować funkcję o postaci:

$$\sum_{b=1}^B \text{dist}(c, c_b)^2 \Rightarrow \min_{c \in C}, \quad (3)$$

gdzie: C – zbiór wszystkich możliwych podziałów zagregowanych,
 dist – odległość euklidesowa,
 (c_1, \dots, c_B) – grupowania wchodzące w skład podziału zagregowanego.

3. Taksonomia spektralna

Taksonomia spektralna polega na zastosowaniu wartości własnych pochodzących ze spektralnej dekompozycji macierzy podobieństwa opisującej badane obiekty. Następnie największe wartości własne oraz odpowiadające im wektory własne są wykorzystywane do ostatecznego podziału obserwacji. W literaturze zaproponowano kilka metod spektralnych, a każda z nich w nieco inny sposób stosuje wektory własne [Kannan i in. 2004; Ng i in. 2001; Shi, Malik 2000]. W niniejszym badaniu zastosowana zostanie metoda zaproponowana przez Ng i in. [2001].

Dany jest zbiór obserwacji $G = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ w przestrzeni R^l , który należy podzielić na k grup.

1. Skonstruuj macierz podobieństwa (*affinity matrix*) $A \in R^{n \times n}$, której elementy są zdefiniowane jako:

$$A_{ij} = \exp\left(-\|x_i - x_j\|^2 / 2\sigma^2\right), \quad (4)$$

gdzie $i \neq j$ oraz $A_{ii} = 0$. σ to parametr skalujący dobierany przez badacza.

2. Zdefiniuj D jako macierz diagonalną, której element (i, i) jest sumą i -tego wiersza macierzy A i na jej podstawie skonstruuj macierz:

$$L = D^{-1/2} A D^{-1/2}. \quad (5)$$

3. Znajdź k pierwszych wektorów własnych $(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k)$ macierzy L i zestawiając je w kolumny, skonstruuj macierz:

$$Z = [\mathbf{z}_1, \dots, \mathbf{z}_k] \in R^{n \times k}. \quad (6)$$

4. Skonstruuj macierz Y przez normalizację każdego wiersza macierzy Z tak, by miały jednakową długość, tj.:

$$y_{ij} = z_{ij} / \left(\sum_j z_{ij}^2\right)^{1/2}. \quad (7)$$

5. Traktując każdy wiersz macierzy Y jako punkt w przestrzeni R^k , podziel je na k grup z zastosowaniem metody k -średnich (lub innej).

6. Ostatecznie przydziel każdą pierwotną obserwację x_i do j -tej grupy wtedy i tylko wtedy, gdy i -ty wiersz macierzy Y został przydzielony do j -tej grupy.

4. Badania empiryczne

W celu porównania dokładności grupowania badanych metod zastosowano miarę opartą na indeksie Randa:

$$Acc = \frac{1}{Z} \sum_{z=1}^Z R(P_z, P^T), \quad (8)$$

gdzie: Z – liczba badanych podziałów,

R – indeks Randa,

P_z – grupowanie na podstawie z -tego podziału,

P^T – rzeczywiste etykiety klas.

Miara ta jest uśrednioną po wszystkich badanych podziałach miarą dokładności i ocenia podobieństwo między ostatecznym grupowaniem zagregowanym a prawdziwymi etykietami klas.

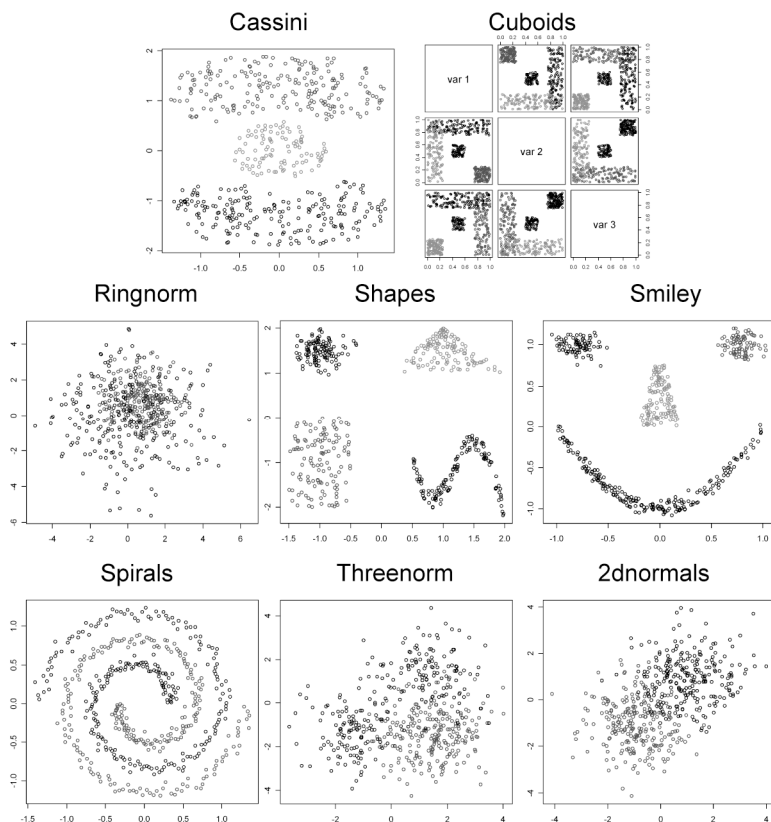
W badaniach zastosowano sztucznie generowane zbiory danych, które standardowo wykorzystywane są w badaniach porównawczych w taksonomii¹. Są to takie zbiory, w których przynależność obiektów do klas jest znana. Ich krótka charakterystyka znajduje się w tab. 1, natomiast struktura przedstawiona jest na rys. 1.

Tabela 1. Charakterystyka zastosowanych zbiorów danych

Zbiór danych	Liczba obiektów	Liczba cech	Liczba klas
<i>Cassini</i>	500	2	3
<i>Cuboids</i>	500	3	4
<i>Ringnorm</i>	500	2	2
<i>Shapes</i>	500	2	4
<i>Smiley</i>	500	2	4
<i>Spirals</i>	500	2	2
<i>Threenorm</i>	500	2	2
<i>2dnormals</i>	500	2	2

Źródło: opracowanie własne.

¹ Zbiory zaczerpnięte zostały z pakietu `mlbench` z programu **R**.



Rys. 1. Struktura zastosowanych zbiorów danych

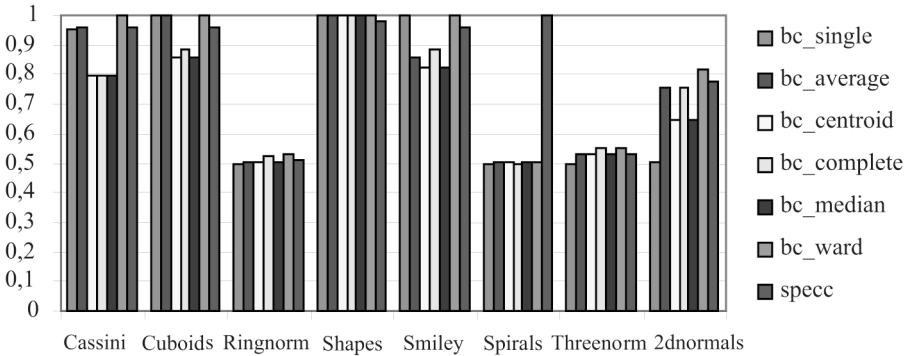
Źródło: opracowanie własne na podstawie programu R.

W metodzie *bagging* według Leischa jako metodę bazową zastosowano metodę *k*-średnich, natomiast ostatecznego grupowania dokonano z zastosowaniem: metody najbliższego sąsiedztwa (*single*), najdalszego sąsiedztwa (*complete*), średniej odległości między skupieniami (*average*), środka ciężkości (*centroid*), mediany (*median*), Warda (*Ward*). W metodzie Dudoit i Fridlyand oraz Hornika utworzono 50 prób bootstrapowych, a na ich podstawie określano podziały składowe z zastosowaniem metody *k*-średnich oraz *c*-średnich, która jest rozmytą wersją metody *k*-średnich opracowaną przez Bezdeka [1981]. Natomiast agregacja przebiegała z zastosowaniem równania 2 w metodzie Dudoit i Fridlyand oraz 3 w metodzie Hornika².

² Na rysunkach 3 i 4 stosowano skróty `cl_bagg_k` i `cl_consensus_k`, jeżeli grupowania składowe określone były z zastosowaniem metody *k*-średnich, oraz `cl_bagg_c` i `cl_consensus_c`, gdy wykorzystywano metodę *c*-średnich.

W taksonomii spektralnej macierz Y grupowana była z zastosowaniem metody k -średnich.

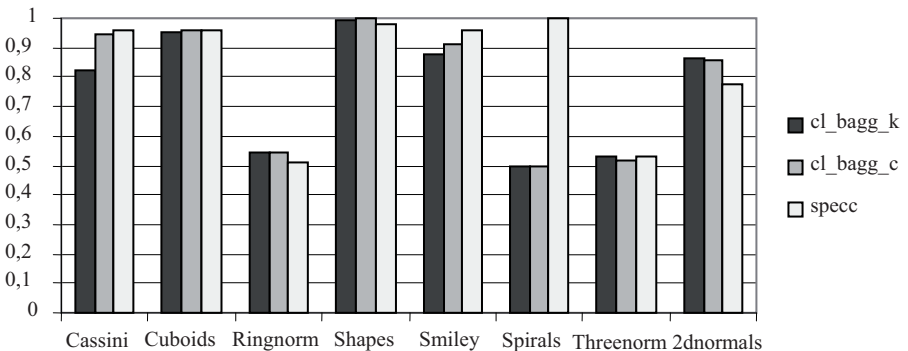
Każde podejście stosowano 50 razy i ich wyniki były potem badane pod względem dokładności.



Rys. 2. Porównanie dokładności metody *bagging* według Leischa oraz podejścia spektralnego

Źródło: opracowanie własne.

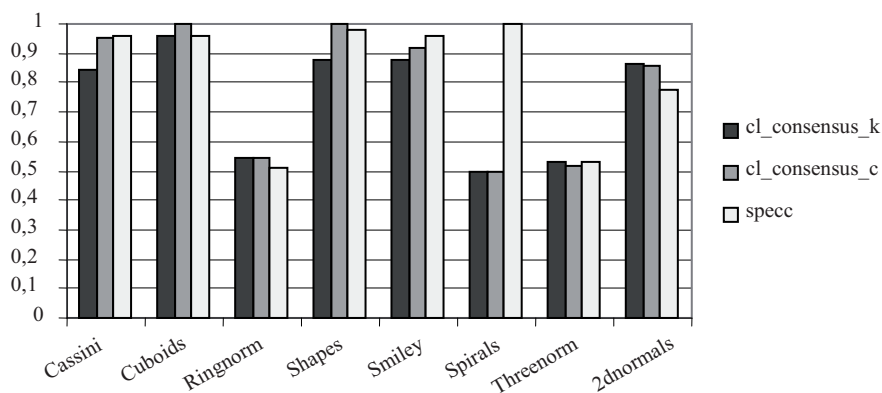
W przypadku metody *bagging* według Leischa (rys. 2) można zaobserwować, że taksonomia spektralna (*specc*) zawsze daje niższą dokładność niż metoda *bc_ward*, z wyjątkiem zbioru *Spirals*. Dla zbiorów *Cassini*, *Cuboids*, *Smiley* oraz *2dnormals* *specc* daje wyższą dokładność niż zagregowane metody taksonomiczne *bc_centroid*, *bc_complete* oraz *bc_median*. Porównywalną dokładność taksonomii spektralnej z metodami zagregowanymi można zaobserwować dla zbiorów *Ringnorm*, *Shapes* oraz *Threenorm*. Natomiast dla zbioru *Spirals* taksonomia spektralna daje znacznie lepszą dokładność w porównaniu z metodami zagregowanymi.



Rys. 3. Porównanie dokładności metody *bagging* według Duidoit i Fridlyand oraz podejścia spektralnego

Źródło: opracowanie własne.

Rysunek 3 porównujący rezultaty dla metody *bagging* według Dudoit i Fridlyand oraz taksonomii spektralnej pozwala stwierdzić, że *specc* daje lepsze rezultaty niż obydwa rozpatrywane warianty zagregowane tylko w przypadku zbiorów *Cassini*, *Smiley* i *Spirals*. Niższą dokładność taksonomii spektralnej w porównaniu z metodami zagregowanymi można zaobserwować dla zbiorów *Ringnorm*, *Shapes* i *2dnormals*. Natomiast dla zbiorów *Cuboids* i *Threenorm* obydwa podejścia dają porównywalne rezultaty.



Rys. 4. Porównanie dokładności metody *bagging* według Hornika oraz podejścia spektralnego

Źródło: opracowanie własne.

W przypadku metody *bagging* według Hornika i taksonomii spektralnej (rys. 4) można stwierdzić, że *specc* daje zdecydowanie wyższą dokładność dla zbioru *Spirals* i nieco lepszą dla zbiorów *Smiley* i *Cassini*. Dla pozostałych zbiorów dokładność *specc* można określić jako porównywalną lub nieco gorszą niż dokładność metod zagregowanych.

5. Podsumowanie

W świetle uzyskanych wyników można wyciągnąć następujące wnioski:

1. Dla zbiorów o strukturze podobnej do zbioru *Spirals* taksonomia spektralna jest bezkonkurencyjna z punktu widzenia dokładności.
2. Warto zastąpić zagregowane metody *bc_complete*, *bc_centroid* oraz *bc_median* przez taksonomię spektralną, bo na pewno uzyskamy nie gorszą dokładność.
3. Dla zbiorów z wyraźnie separowalnymi klasami (np. *Cassini*, *Smiley*) taksonomia spektralna niejednokrotnie jest bardziej dokładna niż metody zagregowane *cl_bag_k* oraz *cl_consensus_k*.
4. Dla zbiorów z trudno separowalnymi klasami (np. *Ringnorm*, *Threenorm*, *2dnormals*) taksonomia spektralna daje porównywalne lub nieco gorsze rezultaty niż podejście zagregowane.

Literatura

- Bezdek J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York 1981.
- Dudoit S., Fridlyand J., *Bagging to improve the accuracy of a clustering procedure*, "Bioinformatics" 2003, vol. 19, no 9.
- Hornik K., *A CLUE for CLUster ensembles*, "Journal of Statistical Software" 2005, no 14.
- Kannan R., Vempala S., Vetta A., *On clustering – good, bad and spectral*, "Journal of the ACM" 2004, vol. 51, no 3.
- Leisch F., *Bagged clustering*, „Adaptive Information Systems and Modeling in Economics and Management Science", Working Papers, SFB, 1999, no 51.
- Ng A.Y., Jordan M.I., Weiss Y., *On spectral clustering: Analysis and an algorithm*, "Advances in Neural Information Processing Systems" 2001.
- Shi J., Malik J., *Normalized cuts and image segmentation*, "IEEE Transactions on Pattern Analysis and Machine Intelligence" 2000, vol. 22, no 8, <http://www-2.cs.cmu.edu/~jshi/Grouping/>.

COMPARISON OF ACCURACY OF SPECTRAL CLUSTERING AND CLUSTER ENSEMBLES STABILITY BASED ON BAGGING IDEA

Summary: High accuracy of the results is a very important task in any grouping problem (clustering). Therefore in the literature there are proposed methods and solutions that main aim is to give more accurate results than traditional clustering algorithms. The examples of such solutions can be cluster ensembles or spectral clustering algorithms. The main aim of the article is to compare the accuracy of spectral clustering and cluster ensembles. There will be considered cluster ensembles based on bagging idea [Dudoit, Fridlyand 2003; Hornik 2005; Leisch 1999] and spectral algorithm proposed by Ng et al. [2001].

Keywords: taxonomy, cluster ensemble, spectral clustering, accuracy.