

**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](http://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](http://www.ceeol.com),  
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon [http://kangur.uek.krakow.pl/  
bazy\\_ae/bazekon/nowy/index.php](http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php)

Informacje o naborze artykułów i zasadach recenzowania znajdują się  
na stronie internetowej Wydawnictwa  
[www.wydawnictwo.ue.wroc.pl](http://www.wydawnictwo.ue.wroc.pl)

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

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

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

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

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

## Summaries

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

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



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

---

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

**Marcin Pelka**

Uniwersytet Ekonomiczny we Wrocławiu

---

## PODEJŚCIE WIELOMODELOWE Z WYKORZYSTANIEM METODY *BOOSTING* W ANALIZIE DANYCH SYMBOLICZNYCH

---

**Streszczenie:** Celem artykułu jest zaprezentowanie możliwości wykorzystania metody *boosting* w agregacji modeli dla danych symbolicznych z zastosowaniem metody  $k$ -najbliższych sąsiadów jako klasyfikatora bazowego. W artykule przedstawiono podstawowe pojęcia z zakresu analizy danych symbolicznych, metody  $k$ -najbliższych sąsiadów. W części empirycznej przedstawiono zastosowanie podejścia wielomodelowego dla danych symbolicznych dla kredytów konsumpcyjnych.

**Słowa kluczowe:** analiza danych symbolicznych, podejście wielomodelowe, *boosting*.

### 1. Wstęp

Ideą podejścia wielomodelowego jest łączenie – nazywane także agregacją – wyników  $M$  modeli bazowych  $(D_1, \dots, D_M)$  w jeden model zagregowany  $D^*$ , por. [Kuncheva 2004, s. 6-7; Walesiak, Gatnar 2009, s. 261; Gatnar 2008, s. 62]. Celem zastosowania podejścia wielomodelowego, zamiast wykorzystania pojedynczego modelu, jest zmniejszenie błędu predykcji. Oznacza to, że model połączony jest modelem bardziej dokładnym niż jakkolwiek z pojedynczych modeli, które wchodzi w jego skład, zob. [Gatnar 2008, s. 62]. Metoda *boosting* jest drugą obok metody *bagging* z bardziej znanych metod stosowanych w podejściu wielomodelowym. Metoda ta została zaproponowana pod nazwą *AdaBoost* przez Freund'a i Schapire w 1995 r., zob. [Gatnar 2008, s. 148; Freund, Schapire 1997, s. 119].

Celem artykułu jest zaprezentowanie możliwości zastosowania i modyfikacji metody *boosting* w agregacji modeli dla danych symbolicznych z wykorzystaniem metody  $k$ -najbliższych sąsiadów jako klasyfikatora bazowego. W części empirycznej przedstawiono wyniki badań z wykorzystaniem zbioru kredytów konsumpcyjnych.

### 2. Dane symboliczne

Obiekty symboliczne mogą być opisywane przez następujące rodzaje zmiennych symbolicznych [Bock, Diday 2000, s. 2-3]:

1) zmienne w ujęciu klasycznym, tj. ilorazowe, przedziałowe, porządkowe, nominalne;

2) zmienne symboliczne, tj. zmienne:

- interwałowe, których realizacją są przedziały liczbowe rozłączne lub nierozłączne;
- wielowariantowe, gdzie realizacją zmiennej jest więcej niż jeden wariant (liczba lub kategoria);
- wielowariantowe z wagami, gdzie realizacją zmiennej oprócz wielu wariantów są dodatkowo wagi (lub prawdopodobieństwa) dla każdego z wariantów zmiennej dla danego obiektu.

Niezależnie od typu zmiennej w analizie danych symbolicznych możemy mieć do czynienia ze zmiennymi strukturalnymi [Bock, Diday 2000, s. 2-3; 33-37]. Do tego typu zmiennych zalicza się **zmienne hierarchiczne** – w których *a priori* ustalone są reguły decydujące o tym, czy dana zmienna opisuje dany obiekt czy nie; **zmienne taksonomiczne** – w których ustalone są *a priori* realizacje danej zmiennej; **zmienne logiczne** – tj. takie, dla których ustalono *a priori* reguły logiczne lub funkcyjne, które decydują o wartościach zmiennej.

W analizie danych symbolicznych wyróżnia się dwa typy obiektów symbolicznych:

- **obiekty symboliczne pierwszego rzędu** – obiekty rozumiane w sensie „klasycznym” (obiekty elementarne), np. konsument, przedsiębiorstwo, produkt, pacjent czy gospodarstwo domowe,
- **obiekty symboliczne drugiego rzędu** – obiekty utworzone w wyniku agregacji zbioru obiektów symbolicznych pierwszego rzędu, np. grupa konsumentów preferująca określony produkt, region geograficzny (jako wynik agregacji podregionów).

### 3. Idea metody *boosting*

Drugą, obok metody *bagging*, popularną metodą łączenia modeli bazowych jest metoda *boosting*, zob. [Gatnar 2008, s. 145-154; Polikar 2006, s. 28-32; Kuncheva 2004, s. 212-222]. Metoda ta realizuje w swej konstrukcji architekturę szeregową modeli zagregowanych (zob. rys. 1). Oznacza to, że wyniki kolejnych modeli zależą od wyników modeli poprzednich.



Rys. 1. Architektura szeregową

Źródło: [Gatnar 2008, s. 69].

Metoda *boosting* polega na poprawianiu (inaczej wzmacnianiu) dokładności predykcji modelu zagregowanego  $D^*$  w rezultacie modyfikacji kolejnych modeli ba-

zowych  $D_1, \dots, D_M$ , por. [Gatnar 2008, s. 145]. Poprawę dokładności predykcji uzyskuje się poprzez zastosowanie podwójnego systemu wag. Pierwszy dotyczy obserwacji i polega na tym, że obserwacje, które błędnie sklasyfikował  $i$ -ty model  $D_i$ , otrzymują wyższe wagi. Drugi system wag polega na przydzieleniu każdemu z modeli wag proporcjonalnych do błędu jego predykcji. Obserwacje są losowane do każdego z  $M$  modeli bazowych zgodnie z przypisanymi im wagami, których suma dla obiektu musi wynosić jeden, zob. [Gatnar 2008, s. 145; Kuncheva 2004, s. 216; Polikar 2006, s. 29].

Algorytm metody *boosting* zostanie omówiony na przykładzie algorytmu *Ada-Boost* (nazwa pochodzi od *Adaptive Boosting*) [por. Gatnar 2008, s. 146]. Algorytm ten składa się z czterech kroków [Gatnar 2008, s. 146; Polikar 2006, s. 29-30; Kuncheva 2004, s. 216]:

1. Ustalenie liczby modeli bazowych  $M$ .

2. Ustalenie początkowych wag obserwacji ze zbioru uczącego  $U$  :

a) wagi mogą być odwrotnie proporcjonalne do liczby obiektów w zbiorze uczącym:

$$\forall_{i=1, \dots, N} w_i^{(1)} = \frac{1}{N}, \quad (1)$$

b) wagi mogą być zależne od potencjału opisowego obiektów symbolicznych – propozycję takiego rozwiązania zawarł w swojej pracy A. Dudek, zob. [2009, s. 33-40]:

$$\forall_{i=1, \dots, N} w_i^{(1)} = \frac{\pi(O_i)}{\pi(O_E)}, \quad (2)$$

gdzie:  $O_i$  –  $i$ -ty obiekt symboliczny ze zbioru uczącego,

$O_E$  – syntetyczny obiekt symboliczny opisujący wszystkie obiekty ze zbioru uczącego,

$\pi$  – potencjał obiektu symbolicznego liczony zgodnie ze wzorem [Bock, Diday 2000, s. 176]:

$$\pi(O_i) = \prod_{j=1}^P \mu(v_{ij}), \quad (3)$$

gdzie:  $j = 1, \dots, P$  – numer zmiennej symbolicznej,

$\mu(v_{ij})$  – długość przedziału dla zmiennych interwałowych, dla zmiennych wielowariantowych – liczba elementów (wariantów)  $j$ -tej zmiennej dla  $i$ -tego obiektu symbolicznego.

W części empirycznej artykułu wykorzystane zostaną zarówno wagi zależne od potencjału opisowego, jak i wagi odwrotnie proporcjonalne do liczby obiektów.

3. Wykonanie dla każdego  $m = 1, \dots, M$  następujących czynności:

a) wylosowanie ze zbioru uczącego  $U$  do próby uczącej  $U_m$  obiektów zgodnie z rozkładem ich wag,

b) zbudowanie modelu bazowego  $D_m$  na podstawie próby uczącej  $U_m$  i obliczenie błędu predykcji (jako błędu resubstytucji):

$$e(D_m) = \sum_{i=1}^N w_i^{(m)} I(\hat{D}_m(O_i) \neq y_i), \quad (4)$$

c) jeżeli  $e(D_m) = 0$  lub  $e(D_m) \geq 0,5$ , należy przerwać działanie algorytmu,

d) w przeciwnym razie obliczana jest waga dla modelu bazowego  $D_m$  :

$$\beta_m = \frac{e(D_m)}{1 - e(D_m)}, \quad (5)$$

e) zmodyfikowanie wag obserwacji zgodnie ze wzorem:

$$w_i^{(m+1)} = \frac{w_i^{(m)} \beta_m^{I(\hat{D}_m(O_i)=y_i)}}{\sum_{k=1}^N w_k^{(m)} \beta_m^{I(\hat{D}_m(O_k)=y_k)}}, \quad (6)$$

f) powrót do podpunktu 3a.

4. Dokonanie predykcji modelu zagregowanego dla obserwacji  $O_i$  za pomocą modeli bazowych z wykorzystaniem ważonego głosowania:

$$\hat{D}^*(O_i) = \arg \max_j \left\{ \sum_{m=1}^M \ln \left( \frac{1}{\beta_m} \right) I(\hat{D}_m(O_i) = C_j) \right\}. \quad (7)$$

Jak wspomniano we wprowadzeniu, klasyfikatorem bazowym w przykładzie empirycznym jest metoda  $k$ -najbliższych sąsiadów dla danych symbolicznych. Algorytm tej metody można wyrazić za pomocą następujących kroków (zob. [Malerba, D'Amato, Esposito, Monopoli 2003; Malerba, Esposito, D'Amato, Appice 2004; Malerba, Esposito, D'Amato, Appice 2006; Pełka 2010]):

1. Wybór liczby sąsiadów branych pod uwagę w dalszej części algorytmu ( $k$ ).

2. Obliczenie odległości między obiektami symbolicznymi (ze zbioru uczącego i testowego).

3. Wybór  $k$  obiektów ze zbioru uczącego najbliższych  $i$ -temu obiektowi ze zbioru testowego.

4. Obliczenie prawdopodobieństw *a posteriori* przydzielenia obiektu ze zbioru testowego do każdej z klas zbioru uczącego. Prawdopodobieństwo to obliczane jest zgodnie ze wzorem:

$$P(O_i | C_j) = \frac{\frac{K_j}{K} \cdot \Omega_j}{\sum_{j=1}^J \frac{K_j}{K} \cdot \Omega_j}, \quad \forall j = 1, \dots, J, \quad (8)$$

gdzie:  $\Omega_j = \sum_{i=1}^J w_i \cdot \delta(C_j, C_k)$ ,

$w_i = \frac{1}{d(O_i, O_k)}$  – wagi, które są odwrotnością odległości między  $i$ -tym obiektem ze zbioru testowego a  $k$ -tym sąsiadem ze zbioru uczącego,

$\delta(C_j, C_k) = 1$  – jeżeli klasa, do której należy  $k$ -ty sąsiad, jest taka sama jak klasa, do której przyporządkowujemy  $i$ -ty obiekt,

$\delta(C_j, C_k) = 0$  – jeżeli klasa, do której należy  $k$ -ty sąsiad, jest inna niż klasa, do której przyporządkowujemy  $i$ -ty obiekt,

$K_j$  – liczba sąsiadów należących do  $j$ -tej klasy,

$j = 1, \dots, J$  – numer klasy.

#### 4. Przykład empiryczny

Bank BGŻ SA osiągnął w 2004 r. dziesiąte miejsce pod względem sumy udzielonych kredytów, a jedenaste pod względem funduszy własnych, por. [Adomski 2005, s. 6]. Jednocześnie bank ten w rankingu 50 największych banków w Polsce został uznany w 2004 r. za drugi bank, po BPH SA, w kategorii banków uniwersalnych, por. [Adomski 2005, s. 22].

Zadłużenie z tytułu kredytów konsumpcyjnych według danych na koniec 2004 r. stanowiło 69% całego bankowego zadłużenia gospodarstw domowych, zob. [Penczar i in. 2005, s. 19].

W przykładzie empirycznym wykorzystano dwa zbiory danych dotyczący kredytów konsumpcyjnych udzielonych w 2004 r. przez Bank Gospodarki Żywnościowej SA w Kłodzku. Jako metodę doboru próby do badania wybrano dobór nielosowy (wybór kwotowy), zob. [Szreder 2004, s. 53-60]. Kwotami w tym przypadku były poszczególne rodzaje kredytów konsumpcyjnych udzielone przez BGŻ SA. Oddział w Kłodzku. Zbiór danych zawiera 100 obiektów podzielonych na dwie klasy (72 decyzje o udzieleniu kredytu – klasa 1, 28 decyzji o odrzuceniu wniosku kredytowego – klasa 2). Obydwie klasy opisuje trzynaście zmiennych:

1. Średnie wpływy na rachunek bieżący – zmienna interwałowa.
2. Staż pracy kredytobiorcy – zmienna interwałowa.
3. Czas trwania kredytu w miesiącach – zmienna interwałowa.
4. Dochody kredytobiorcy – zmienna interwałowa.
5. Wnioskowana kwota kredytu – zmienna interwałowa.
6. Historia kredytowa – zmienna wielowariantowa.
7. Staż klienta w banku BGŻ SA – zmienna interwałowa.

8. Wskazanie poręczyciela – zmienna wielowariantowa.
9. Ocena poręczyciela – zmienna wielowariantowa.
10. Inne proponowane zabezpieczenia – zmienna wielowariantowa.
11. Ocena klienta w BGŻ SA – zmienna wielowariantowa.
12. Lojalność klienta wobec BGŻ SA – zmienna wielowariantowa.
13. Udzielona informacja o sytuacji kredytowej – zmienna wielowariantowa.

Zbiór danych podzielono na dwa podzbiory – zbiór uczący stanowiło 75 obiektów, a zbiór testowy 25 obiektów.

Wyniki otrzymane przy zastosowaniu podejścia wielomodelowego *boosting* z zastosowaniem wag odwrotnie proporcjonalnych do liczby obiektów zestawiono w tab. 1. Wyniki otrzymane przy zastosowaniu wag zależnych od potencjału opisowego obiektów symbolicznych zestawiono w tab. 2.

**Tabela 1.** Wyniki obliczeń dla wag odwrotnie proporcjonalnych do liczby obiektów.

		Klasyfikacja rzeczywista	
		klasa 1	klasa 2
Decyzja KNN	klasa 1	12	2
	klasa 2	1	10

Źródło: obliczenia własne w programie R.

Najmniejszy błąd klasyfikacji (12%) otrzymano dla 38 modeli bazowych. W tym przypadku błędnie sklasyfikowano 3 spośród 25 obiektów w zbiorze testowym.

**Tabela 2.** Wyniki obliczeń dla wag zależnych od potencjału opisowego obiektów

		Klasyfikacja rzeczywista	
		klasa 1	klasa 2
Decyzja KNN	klasa 1	12	0
	klasa 2	1	12

Źródło: obliczenia własne w programie R.

Najmniejszy błąd klasyfikacji (4%) uzyskano dla 26 modeli bazowych. W tym przypadku jedynie jeden obiekt (o pozytywnej decyzji kredytowej) został sklasyfikowany jako obiekt, który nie powinien otrzymać kredytu.

## 5. Podsumowanie

Metoda *boosting* może znaleźć zastosowanie w klasyfikacji różnych zbiorów danych symbolicznych. Podejście wielomodelowe analizy danych symbolicznych, podobnie jak podejście wielomodelowe dla danych klasycznych, pozwala osiągnąć mniejszy błąd klasyfikacji niż zastosowanie pojedynczego modelu.



Na potrzeby badań empirycznych opracowano w programie R skrypt realizujący algorytm metody *boosting* z zastosowaniem metody  $k$ -najbliższych sąsiadów jako klasyfikatora bazowego.

W metodzie *boosting* można zastosować dwa sposoby ważenia obiektów symbolicznych – zależny od liczby obiektów oraz zależny od potencjału opisowego. W obydwu przypadkach nieco lepsze wyniki uzyskano, stosując ważenie obiektów symbolicznych zależne od ich potencjału opisowego.

Etapem dalszych prac będzie porównanie wyników otrzymywanych przy różnych sposobach ważenia obiektów symbolicznych oraz porównanie metody *boosting* z innymi metodami podejścia wielomodelowego (np. *bagging*).

## Literatura

- Adomski G., *Każdemu według potrzeb*, „Bank” 2005, nr 4(150).
- Bock H.-H., Diday E (red.), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin 2000.
- Dudek A., *Tworzenie zagregowanych modeli dyskryminacyjnych dla obiektów symbolicznych – wybrane problemy*, [w:] J. Pocięcha, *Współczesne problemy statystyki, ekonometrii i matematyki stosowanej*, Studia i Prace Uniwersytetu Ekonomicznego w Krakowie, Kraków 2009.
- Freund Y., Schapire R.E., *A decision-theoretic generalization of on-line learning and an application to boosting*, „Journal of Computer and System Sciences” 1997, vol. 55, no 1.
- Gatnar E., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa 2008.
- Kuncheva L.I., *Combining Pattern Classifiers. Methods and Algorithms*, Wiley, New Jersey 2004.
- Malerba D., D’Amato C., Esposito F., Monopoli M., *Extending the K-Nearest Neighbour Classification Algorithm to Symbolic Objects*, Atti del Convegno Intermedio della Società Italiana di Statistica „Analisi Statistica Multivariata per le scienze economico-sociali, le scienze naturali e la tecnologia”, Napoli 2003.
- Malerba D., Esposito F., D’Amato C., Appice A., *K-nearest Neighbor Classification for Symbolic Objects*, [w:] *Symbolic and Spatial Data Analysis: Mining Complex Data Structures*, P. Brito, M. Noirhomme-Fraiture (red.), University of Pisa, Pisa 2004.
- Malerba D., Esposito F., D’Amato C., Appice A., *Classification of symbolic objects: A lazy learning approach*, „Intelligent Data Analysis” 2006, vol. 10, no 4.
- Polikar R., *Ensemble based systems in decision making*, „IEEE Circuits and Systems Magazine” 2006, vol. 6, no 3.
- Pełka M., *K-nearest neighbour classification for symbolic data*, „Acta Universitatis Lodzianensis. Folia Oeconomica” 2010, nr 235.
- Penczar M., Lepczyński B., Gostomski E. (red.), *Zadłużenie konsumentów w bankach i instytucjach finansowych*, Instytut Badań nad Gospodarką Rynkową, Gdańsk.
- Szreder M., *Metody i techniki sondażowych badań opinii*, PWE, Warszawa 2004.
- Walesiak M., Gatnar E. (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa 2009.

## ENSEMBLE LEARNING WITH THE APPLICATION OF *BOOSTING* IN SYMBOLIC DATA ANALYSIS

**Summary:** The aim of this paper is to present the application of boosting method in ensemble learning for symbolic data with the application of  $k$ -nearest neighbour method as the base classifier. The article presents basic terms of symbolic data,  $k$ -nearest neighbour classification rule for symbolic data. In the empirical part the results of application of ensemble learning for symbolic data applied for credit data set are presented.

**Keywords:** symbolic data analysis, ensemble learning, *boosting*.