

**PRACE NAUKOWE**

Uniwersytetu Ekonomicznego we Wrocławiu

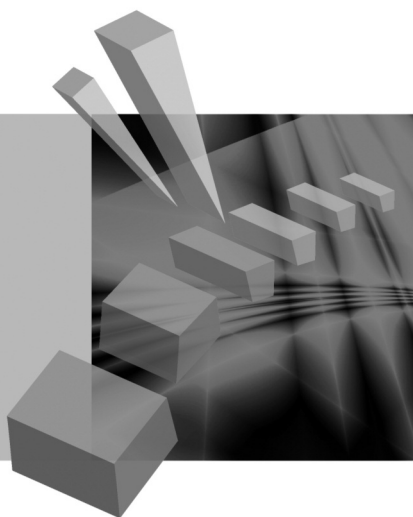
**RESEARCH PAPERS**

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](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>Sylvia 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 Dębowska</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

**Andrzej Sokółowski**

Uniwersytet Ekonomiczny w Krakowie

---

## Q UNIwersALNA MIARA ODLEGŁOŚCI

---

**Streszczenie:** W zadaniach taksonomicznych coraz częściej mamy do czynienia z danymi zawierającymi zmienne mierzone w różnych skalach. Jedną ze strategii jest wykonanie analizy osobno dla jednorodnej grupy cech, a potem scalenie wyników. Druga strategia to zastosowanie miary odległości, która ze swej natury umożliwia jednocześnie wykorzystanie cech mierzonych w różnych skalach. Najbardziej dojrzałą propozycją jest tu miara Walesiaka GDM. W proponowanym artykule przede wszystkim wykraczamy poza tradycyjny podział cech ze względu na skalę pomiaru – na nominalne, porządkowe, przedziałowe i ilorazowe. Wyróżniono 15 rodzajów cech statystycznych. Następnie zaproponowano takie przekształcenia tych cech (lub odległości), że składowe sumy w odległości typu Manhattan przyjmują wartości z przedziału  $[0,1]$  i są niemianowane. To umożliwia policzenie odległości ogólnej.

**Słowa kluczowe:** miary odległości, skale pomiaru.

W zadaniach taksonomicznych coraz częściej mamy do czynienia z danymi zawierającymi zmienne mierzone w różnych skalach. Jedną z możliwych strategii jest wykonanie analizy osobno dla jednorodnej grupy cech, a potem scalenie wyników. Druga strategia to zastosowanie miary odległości, która ze swej natury umożliwia jednocześnie wykorzystanie cech mierzonych w różnych skalach. Najbardziej dojrzałą propozycją jest miara Marka Walesiaka GDM.

W pracy wykraczamy poza tradycyjny podział cech ze względu na skalę pomiaru – na nominalne, porządkowe, przedziałowe i ilorazowe. Zaproponowano takie przekształcenia tych cech (lub odległości), że składowe sumy w odległości typu Manhattan przyjmują wartości z przedziału  $[0,1]$  i są niemianowane. To umożliwia policzenie odległości ogólnej.

Zdarza się, że niektóre cechy mają ograniczenia naturalne. Na przykład wiele cech może przyjmować tylko wartości dodatnie, więc zero jest ograniczeniem z lewej strony. Udziały procentowe przyjmują z kolei wartości z przedziału  $[0;100]$ . Te ograniczenia mogą być wykorzystane w procesie normalizacji. Przy braku naturalnych ograniczeń zmienności można wykorzystać ograniczenia empiryczne, czyli zaobserwowane wartości skrajne, ewentualnie przesunięte jeszcze o jakąś umowną stałą.

Wybieramy odległość Manhattan dlatego, że ona daje się dekomponować na poszczególne cechy. W naszych rozważaniach pomijamy wagi, choć wprowadzenie ich do procedury liczenia odległości jest możliwe i proste.

Odległości indywidualne mogą być względne lub bezwzględne. Odległość bezwzględna zależy tylko od współrzędnych tych dwóch punktów, a nie od współrzędnych pozostałych punktów w analizowanym zbiorze. Odległości względne liczone są np. na danych standaryzowanych lub na normalizowanych wykorzystujących empiryczne granice zmienności.

Odległość  $Q$  jest sumą odległości indywidualnych. W poniższym wzorze  $i$  oraz  $j$  to numery obiektów,  $k$  – numer cechy,  $m$  zaś to rozmiar przestrzeni klasyfikacji. Podany wzór można oczywiście różnie przekształcić, choćby zamieniając go na średnią odległość lub wprowadzając wagi.

$$q_{ij} = \sum_{k=1}^m d_{ij}^{(k)},$$

$$0 \leq d_{ij}^{(k)} \leq 1.$$

W dalszej części pracy omówiono różne rodzaje cech statystycznych i sposoby sprowadzania ich do przedziału  $[0;1]$  (w niektórych przypadkach nie jest to przedział obustronnie domknięty) lub też sposób sprowadzania odległości do przedziału jednostkowego. Opisy są z reguły lakoniczne.

## CECHY ILOŚCIOWE CIĄGŁE

To chyba najpopularniejszy rodzaj cech statystycznych. Rozważamy tutaj tylko te cechy, które mogą być zakwalifikowane jako stymulanty lub destymulanty. Nominanty będą przekształcane w inny sposób. We wzorach przyjmujemy konwencję, wedle której  $x_i$  to oryginalne wartości cech, a  $x_i^*$  to wartości znormalizowane (przekształcone).

Stymulanty

$$x_i^* = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}.$$

Destymulanty

$$x_i^* = \frac{\max_i\{x_i\} - x_i}{\max_i\{x_i\} - \min_i\{x_i\}}.$$

Odległość Manhattan dla stymulant

$$d_{ij} = \left| \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}} - \frac{x_j - \min_j\{x_j\}}{\max_j\{x_j\} - \min_j\{x_j\}} \right|.$$

Minimum i maksimum są liczone po wszystkich obiektach, więc są identyczne. Dla uproszczenia (i dla podkreślenia, że nie muszą to być wartości zaobserwowane) będziemy używali tylko symboli  $\min$  i  $\max$ . Mamy więc

$$d_{ij} = \left| \frac{x_i - \min}{\max - \min} - \frac{x_j - \min}{\max - \min} \right| = \left| \frac{x_i - x_j}{\max - \min} \right|.$$

Dla destymulant mamy

$$d_{ij} = \left| \frac{x_j - x_i}{\max - \min} \right|.$$

Obydwa wzory dają więc tę samą odległość.

### CECHY ILOŚCIOWE SKOKOWE NIEUJEMNE (*COUNT DATA*)

Przykładem takiej cechy jest liczba dzieci w rodzinie. Ogólnie rzecz biorąc, jeżeli cecha ta przyjmuje dużo wartości, to może być przekształcana tak jak zmienna ciągła. Przy relatywnie małych wartościach najpierw szacujemy wartość przeciętną  $\lambda$ . Założymy, że dobrą aproksymacją rozkładu cechy jest rozkład Poissona z dystrybuantą  $F(x) = P(X < x)$ . Przy takiej wersji definicji dystrybuanty ważna jest jej wartość w punkcie o jeden większym niż  $x_i$ , który chcemy przekształcić. Wyjściowa miara odległości ma postać

$$d_{ij} = |F(x_i + 1) - F(x_j + 1)|.$$

Niedogodnością może być zakres zmienności wartości przekształconych punktów empirycznych. Mamy

$$x_i^* \in [F(0); 1).$$

Można te przedziały rozciągnąć, przesuwając początek zakresu zmienności do zera. Przekształcenie ma postać

$$x_i^* = \frac{F(x_i+1) - F(0)}{1 - F(0)}.$$

Wtedy odległość indywidualna to

$$d_{ij} = \frac{|F(x_i+1) - F(x_j+1)|}{1 - F(0)}.$$

Ta odległość ma zakres zmienności  $[0; 1)$ . Jest odległością względną, ponieważ zależy od  $\lambda$ , szacowanej ze wszystkich obserwacji.

### CECHY RANGOWE

Tu wystarczy jeden wzór dla stymulant, bo dla stymulant odwrócenie należy przeprowadzić już na etapie rangowania. Przy stymulantach największa wartość otrzymuje rangę 1, a przy destymulantach tę rangę otrzymuje wartość najmniejsza. Przy takim rangowaniu obiekt  $i$  jest lepszy od  $(n - r_i)$  obiektów. Rangę obiektu o numerze  $i$  oznaczamy przez  $r_i$ . Jeżeli obiekt jest pierwszy, to jest lepszy od  $(n - 1)$  obiektów. Sprowadzenie rang do przedziału  $[0, 1]$  dokonuje się według wzoru

$$x_i^* = \frac{n - r_i}{n - 1}.$$

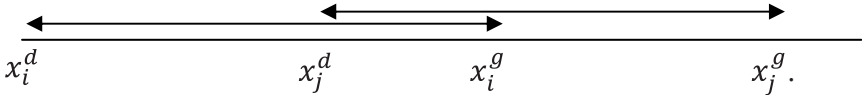
Odległość

$$d_{ij} = \left| \frac{n - r_i}{n - 1} - \frac{n - r_j}{n - 1} \right| = \left| \frac{r_j - r_i}{n - 1} \right|.$$

## CECHY PRZEDZIAŁOWE

Nie chodzi tu o cechy mierzone w skali przedziałowej, lecz o takie, dla których zamiast pojedynczej wartości podajemy przedział liczbowy. Przykładami może tu być zakres zmian temperatury czy wydatki na grupę dóbr.

Przedział definiowany jest jako  $(x_i^d, x_i^g)$



Miara odległości

$$d_{ij} = 1 - \frac{x_i^g - x_i^d + x_j^g - x_j^d}{2(\max\{x_i^g, x_j^g\} - \min\{x_i^d, x_j^d\})}$$

Miara jest równa zero, jeżeli przedziały są identyczne, zaś jeden – jeżeli są zdegenerowane do punktów.

## CECHY NOMINALNE

Taka zmienna ma kilka wariantów, które są nieuporządkowane i wzajemnie wykluczające się. Gdy są tylko dwie kategorie, to zmienna staje się cechą binarną. Przykłady to: status na rynku pracy (pełnozatrudniony, część etatu, bezrobotny, nie należy do siły roboczej), kolor oczu, płeć. Przy cechach binarnych są tylko dwie możliwe relacje: albo dwa obiekty mają ten sam wariant cechy, albo różny. W pliku danych każdy wariant ma osobną kolumnę. Jego występowanie oznaczone jest przez 1, a niewystępowanie przez zero.

$$d_{ij} = \frac{\sum_{l=1}^w |x_{il} - x_{jl}|}{2},$$

$$x_{il}, x_{jl} \in \{0, 1\},$$

$w$  – liczba wariantów cechy,  $l$  – numer wariantu.

## CECHY NOMINALNE Z WIELOKROTNYM POJEDYNCZYM WYBOREM

Przykłady: lokalizacja przerzutów nowotworowych, objawy choroby, wyposażenie gospodarstwa domowego, wskazanie ważnych wartości z listy

$$d_{ij} = \frac{\sum_{l=1}^w |x_{il} - x_{jl}|}{w}$$

Ta odległość znana jest pod nazwą *niezgodności procentowej*. Dobrze byłoby, gdyby warianty cechy jakościowej były mniej więcej takiej samej ważności (wyjściowo). Problem może pojawić przy wyposażeniu gospodarstwa domowego. Takie

elementy, jak samochód, odtwarzacz DVD, są w odległości traktowane tak samo. Można wprowadzić zmienną pomocniczą: *wartość przedmiotu*, i zamienić cechę na ilościową. Częstym rozwiązaniem jest też zamiana takiej cechy na zestaw cech zero-jedynkowych – dla każdego wariantu tworzy się osobną cechę binarną.

## CECHY NOMINALNE Z WIELOKROTNYM WYBOREM CZĘSTOŚCI

Tu dobrym przykładem jest wyposażenie gospodarstwa domowego. Można mieć trzy telewizory, dwa samochody itd. Przez wartości trzeba przejść na sumaryczną zmienną ilościową lub stworzyć osobną cechę dla każdego wariantu. Wtedy będą to cechy *ilościowe punktowe*.

## KATEGORIE UPORZĄDKOWANE LINIOWO

Wykluczające się, wyczerpujące kategorie uporządkowane. Nie są zdefiniowane odległości pomiędzy kategoriami. Przykłady to: zadowolenie z pracy, rating agencji, odpowiedzi w skali Likerta, NYHA, indeks oftalmopatii.

Wydaje się, że są trzy możliwości:

1. Porangowanie kategorii i postępowanie takie jak w przypadku zmiennych rangowych – jest to raczej „wyjście rozpaczy”.

2. Wykorzystanie zmiennej pomocniczej – np. przy wykształceniu jest nią średnia płaca osób z danym wykształceniem. Wtedy klasy: co najwyżej podstawowe, zasadnicze zawodowe, średnie zawodowe, średnie ogólne, wyższe nie są kodowane liczbami naturalnymi od 1 do 5.

3. Opinie ekspertów, które pozwolą ustalić odległości między wariantami, biorąc pod uwagę ogólne kryterium grupowania.

## KATEGORIE UPORZĄDKOWANE NIELINIOWO

Przykład: klasyfikacja TNM – stopnia zaawansowania nowotworu. Klasa zależy od rozmiaru guza (*Tumor*), zajęcia węzłów chłonnych (*Nodes*), przerzutów odległych (*Metastasis*). Przykład ocen w tej skali to IIA, IIB, IIIA. Tutaj odpada pierwsza – wspomniana powyżej – możliwość prostego numerowania, natomiast praktycznie łączy się wykorzystanie zmiennej pomocniczej z opiniami ekspertów. Praktycznie dla wszystkich nowotworów znane są średnie przeżycia (dla ustalonego okresu  $t^*$ ) i to może być wykorzystane do kodowania wariantów cechy. Przy oznaczeniu funkcji przeżycia przez  $S(t)$  mamy następujący wzór na odległość

$$d_{ij} = |S_i(t^*) - S_j(t^*)|.$$

## UDZIAŁY i PROPORCJE

Wartości tego typu cech mieszczą się z natury w przedziale  $[0,1]$ . Wobec tego odległość to po prostu

$$d_{ij} = |x_i - x_j|.$$

## CECHY NIEUJEMNE Z DUŻĄ LICZBĄ ZER

Klasyczny przykład rozważany przez Tobina [1958]: wydatki na pewien rodzaj dóbr w gospodarstwie domowym, w określonym czasie. Cecha nieujemna z dużą liczbą zer jest moim zdaniem cechą jakościowo-ilościową. Tu dobitnym przykładem jest liczba zajętych węzłów chłonnych (inny przykład: tzw. indeks wiązania w cytologii, w którym inną wartością jakościową jest 1, a inną liczby większe od 1). Inny dobry przykład to liczba miesięcy odsiedzianych w więzieniu. Wzory podajemy dla zmiennej z dużą liczbą zer. Najpierw należy przekształcić zmienną oryginalną według wzoru

$$x_i^* = \begin{cases} 0 & \text{dla } x_i = 0 \\ \frac{x_i + \max\{x_i\}}{2\max\{x_i\}} & \text{dla } x_i > 0 \end{cases}$$

Dalej odległość indywidualna to po prostu

$$d_{ij} = |x_i^* - x_j^*|.$$

## PRZEŻYCIE (CZAS TRWANIA ZJAWISKA)

Zakładamy możliwość wykorzystania danych cenzurowanych. Zazwyczaj są one jeszcze prawostronnie ucinane. Na przykład jeżeli interesują nas przeżycia 5-letnie, to wszystkie dłuższe przeżycia skracamy do 60 miesięcy. Oznaczmy tę wartość graniczną jako  $x_g$ . Przeżycie to specyficzna zmienna, która „składa się” z dwóch kolumn. Jedna to czas przeżycia, a druga to zmienna wskaźnikowa (dwustanowa) informująca, czy na końcu okresu zapisanego w pierwszej kolumnie zaszło badane zdarzenie (zgon, upadek przedsiębiorstwa, awaria) czy też nie. Są tu możliwe dwa podejścia.

1. Na podstawie danych szacujemy funkcję przeżycia  $S(t)$ . Jeżeli przez  $x_i$  oznaczymy przeżycie  $i$ -tego obiektu, to przekształcenie ma postać:

$$x_i^* = 1 - S(x_i).$$

Ta metoda nie bierze pod uwagę faktu, czy przeżycie jest kompletne czy nie. Naprawdę wszystkie przeżycia traktowane są jako cenzurowane. Jest to dość karkołomna interpretacja, że sekundy przed zgonem ktoś jednak żył i wtedy nastąpiło cenzurowanie.

2. W tym podejściu zakłada się, że istnieje jednak fundamentalna różnica pomiędzy tymi, co zmarli, a tymi, co jeszcze żyją. Odcinek  $[0,1]$  zostaje podzielony na trzy równe części. Pierwsza 1/3 to ci, którzy żyją, druga 1/3 to strefa buforowa, która oddziela żyjących od zmarłych, i wreszcie ostatnia 1/3, czyli zmarli.

Przekształcenie odbywa się według wzoru

$$x_i^* = \begin{cases} \frac{x_i}{3x_g} & \text{dla przeżyć kompletnych} \\ \frac{2x_i}{3x_g} & \text{dla przeżyć cenzurowanych} \end{cases}$$



## WYBÓR Z PORZĄDKOWANIEM

Na przykład z listy 15 wartości trzeba wybrać pięć i uporządkować je od najważniejszej. Załóżmy, że to porządkowanie odbywa się przez przyznawanie punktów. Widać dwa sposoby:

1. Jeżeli należy wybrać  $s$  wartości, to ta uznana za najważniejszą otrzymuje  $s$  punktów, następna ( $s - 1$ ) itd., a wartości niewybrane „otrzymują” po 0 punktów. Oznaczmy przez  $x_l$  liczbę punktów przyznanych wariantowi o numerze  $l$ . Odległość jest wtedy dana wzorem

$$d_{ij} = 1 - \frac{2 \sum_{l=1}^s \min \{x_{il}, x_{jl}\}}{s(s+1)}.$$

2. Zliczamy punkty z wszystkich i tworzymy hierarchię wartości. Punkty można przeliczyć na wartości sumujące się do jedności. Każdej wartości wybranej przyporządkowana jest wartość według tych punktów. Maksimum punktów to wybranie  $s$  pierwszych w klasyfikacji, a minimum to  $s$  ostatnich.

$$\begin{aligned} \max &= \sum_{(l)=1}^s x_{(l)} \\ \min &= \sum_{(l)=s-1}^s x_{(l)}. \end{aligned}$$

## WYNIKI SPOTKAŃ SPORTOWYCH

Jest to typowa zmienna ilościowo-jakościowa, gdyż w przykładowym wyniku meczu piłkarskiego 3:2 mamy zawartą informację o tym, że mecz wygrała drużyna gospodarzy, a także informację o liczbie bramek strzelonych przez poszczególne drużyny. Propozycję odległości sformułował Sokołowski [2007]. Odległość liczona tu jest odległością względną. Najpierw wyniki są przekształcane według następujących wzorów

$$\text{jeżeli } x_H > x_A \rightarrow \begin{cases} x_H^* = x_H + 2 \cdot \max \\ x_A^* = x_A \end{cases}$$

$$\text{jeżeli } x_H = x_A \rightarrow \begin{cases} x_H^* = x_H + \max \\ x_A^* = x_A + \max \end{cases}$$

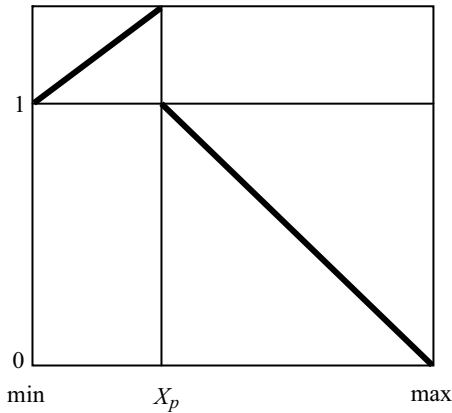
$$\text{jeżeli } x_H < x_A \rightarrow \begin{cases} x_H^* = x_H \\ x_A^* = x_A + 2 \cdot \max \end{cases}$$

Subskrypt H oznacza drużynę gospodarzy, A – drużynę gości, a  $\max$  jest największą liczbą zdobytych bramek (punktów) przez jedną drużynę w zbiorze wyników. Znormalizowana odległość między dwoma wynikami to

$$d_{ij} = \left| \frac{x_{Hi}^* - x_{Hj}^* + x_{Ai}^* - x_{Aj}^*}{6 \cdot \max} \right|.$$

## CECHA CIĄGŁA Z PUNKTEM PRZEŁAMANIA

Jest to coś w rodzaju nominanty, ale po przekroczeniu tzw. punktu przełamania wartości cechy są oceniane jako coraz gorsze, i to gorsze niż jakakolwiek wartość mniejsza od punktu przełamania. Z taką cechą miałem do czynienia przy ocenie gospodarstw rolnych, gdzie istniała optymalna dawka nawozu na hektar, a jej przekroczenie było oceniane gorzej, niż gdy stosowano mniej niż dawkę optymalną, a nawet gorzej, niż gdy nawozu w ogóle nie stosowano.



Schematycznie taką sytuację przedstawiono powyżej. Wartość  $x_i$  przyjmującą wartości z przedziału  $[min, max]$  należy przekształcić na  $0 \leq x_i^* \leq 1$ . Można to zrobić według następujących wzorów

$$x_i^* = \frac{max-min+x_i-x_p}{max-min} \quad \text{dla } x_i < x_p$$

$$x_i^* = \frac{max-x_i}{max-min} \quad \text{dla } x_i \geq x_p.$$

Podsumowując, podkreślmy, że ideą uniwersalnej miary odległości jest wykorzystanie dystansu typu Manhattan, w którym odległości „po poszczególnych wymiarach” mogą być liczone dla różnego rodzaju cech. Wartości tych cech są wstępnie sprowadzane do przedziału  $[0;1]$  bądź do tego przedziału sprowadzana jest sama odległość cząstkowa.

## Literatura

- Sokołowski A., *The Football Distance*, [w:] Taksonomia 14, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Akademii Ekonomicznej im. Oskara Langego we Wrocławiu nr 1169, Wydawnictwo AE, Wrocław 2007.
- Tobin J., *Estimation of relationship for limited dependent variables*, „Econometrica” 1958, no 26.
- Walesiak M., *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R*, Wydawnictwo UE, Wrocław 2011.

## Q UNIVERSAL DISTANCE MEASURE

**Summary:** In clustering tasks we usually have to use variables which are measured in different scales. One of possible strategies is to cluster objects separately for each scale and then try to combine them into final partition. The second strategy is to use the distance measure which allows for calculation of distances in the classification space formed by attributes measured in different scales. This second approach is possible while applying Walesiak GDM distance. In the paper we go beyond the classical four scales – nominal, order, interval and rational. Fifteen different types of variables have been defined. Then we transfer them into  $[0,1]$  interval, or such a transformation is applied to the individual (univariate) distance. The overall distance is calculated through Manhattan approach which makes possible to agglomerate distances on individual axes.

**Keywords:** distance measures, measurement scales.