

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

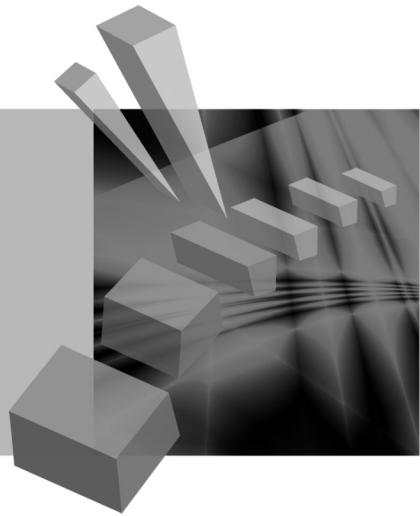
RESEARCH PAPERS

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 wspomagania procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

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

Tomasz Bartłomowicz, Justyna Wilk , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
Kamila Migdał-Najman , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących	342
Dorota Rozmus , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	352
Krzysztof Najman , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG	361
Małgorzata Misztal , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna	370
Mariusz Kubus , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
Barbara Batóg, Jacek Batóg , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym	387
Katarzyna Wójcik, Janusz Tuchowski , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej	396
Iwona Staniec , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach	406
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad 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

Tomasz Bartłomowicz, Justyna Wilk

Uniwersytet Ekonomiczny we Wrocławiu

ZASTOSOWANIE METOD ANALIZY DANYCH SYMBOLICZNYCH W PRZESZUKIWANIU DZIEDZINOWYCH BAZ DANYCH

Streszczenie: Treścią artykułu jest propozycja wykorzystania metodologii analizy danych symbolicznych w filtrowaniu dziedzinowych baz danych. Proponowane rozwiązanie, obok zmiennych klasycznych, uwzględnia zmienne symboliczne, które opisują obiekty bez utraty informacji. Ponadto w rozwiązaniu wykorzystano znormalizowaną miarę Ichino-Yaguchiego dla danych symbolicznych. W opinii autorów połączenie to umożliwia przeszukiwanie baz danych na podstawie wszystkich możliwych kryteriów bez względu na rodzaj występujących zmiennych. W praktyce oznacza to rozszerzenie możliwości występujących w dostępnych powszechnie mechanizmach filtrowania dziedzinowych baz danych, co zostało zaprezentowane na przykładzie przeszukiwania ofert nieruchomości.

Słowa kluczowe: filtrowanie baz danych, analiza danych symbolicznych, pomiar odległości, oferty nieruchomości.

1. Wstęp

Przeszukiwanie rozumiane głównie jako filtrowanie¹ dziedzinowych baz danych to jedno z podstawowych zagadnień informatyki polegające na selekcji zbioru danych względem pewnych zmiennych (cech, atrybutów) charakterystycznych dla każdego z elementów tego zbioru². Czynność ta realizowana niejednokrotnie intuicyjnie stanowi podstawową praktykę wszędzie tam, gdzie ilość prezentowanego materiału przekracza możliwości jego pełnego poznania. Przykładem mogą być tutaj katalogi biblioteczne, serwisy aukcyjne, portale społecznościowe.

Podstawą przeszukiwania jest najczęściej zestaw kryteriów, które powinny spełniać wyszukane obiekty³. Ze względu na brak sprecyzowanych preferencji, chęć

¹ Oprócz filtrowania równie popularną czynnością jest sortowanie rekordów bazy danych.

² Opis na podstawie internetowego słownika języka polskiego, <http://sjp.pwn.pl>.

³ Istnieje możliwość przeszukiwania baz danych na zasadzie negacji, tj. wyszukiwania rekordów, które nie spełniają określonych kryteriów.

porównania kilku obiektów, tudzież ich specyfikę informacje zamieszczone w bazach danych lub dostępne kryteria wyszukiwania są niejednokrotnie mało precyzyjne lub złożone. Mogą mieć formę nie tylko pojedynczej kategorii lub wartości liczbowej, ale także zbioru kategorii, przedziału wartości lub struktury udziałowej. Tego rodzaju dane w literaturze przedmiotu określane są jako dane symboliczne, a metodami umożliwiającymi ich analizę są metody analizy danych symbolicznych (por. [Bock, Diday 2000; Diday, Noirhomme-Fraiture 2008; Billard, Diday 2006; Gatnar, Walesiak 2011]).

Treścią artykułu jest propozycja wykorzystania metodologii analizy danych symbolicznych do filtrowania zawartości dziedzinowych baz danych. W szczegółowym zakresie w artykule zostały zaprezentowane format i źródła danych symbolicznych w zastosowaniach bazodanowych oraz proponowany mechanizm przeszukiwaniu baz danych z wykorzystaniem metodologii analizy danych symbolicznych. Całość proponowanego rozwiązania ilustruje przykład selekcji ofert rynku nieruchomości.

2. Definicja problemu – dane symboliczne

Filtrowanie baz danych wiąże się z określeniem kryteriów wyszukiwania i znalezieniem obiektów spełniających te kryteria. Kryteria wyszukiwania oraz cechy obiektów bazodanowych mogą mieć formę danych klasycznych bądź symbolicznych. W ujęciu klasycznym obiekty opisywane są za pomocą zmiennych, których realizacjami są pojedyncze wartości liczbowe, np. rok produkcji samochodu lub kategorie, np. płeć.

W ujęciu symbolicznym obiekty mogą być również charakteryzowane zmiennymi o realizacjach w postaci:

- przedziałów liczbowych (skokowych lub ciągłych, rozłącznych lub nierozłącznych) ze zbioru liczb rzeczywistych, np. zakres cenowy produktu, zakres wiekowy osoby;
- zbiorów kategorii (nominalnych lub porządkowych) bądź zbiorów wartości (o charakterze metrycznym bądź przedziałów liczbowych), np. słowa kluczowe w publikacji, zastosowanie produktu;
- struktur procentowych, tj. zbiorów kategorii (nominalnych lub porządkowych) lub wartości (metrycznych lub przedziałów liczbowych) z przypisanymi współczynnikami wagowymi, np. skład produktu.

Uwzględnia się także występowanie logicznych (hierarchicznych lub taksonomicznych) relacji między zmiennymi, np. marka i model samochodu.

W artykule przewodnie jest spostrzeżenie, iż filtrowanie informacji zawartych w dziedzinowych bazach danych odbywa się głównie z wykorzystaniem kryteriów w postaci zmiennych klasycznych. Dla obiektów w postaci nieruchomości jest to zazwyczaj cena nieruchomości, tudzież inne metryczne cechy, jak powierzchnia nieruchomości, liczba pokoi, piętrowa lokalizacja lokalu itp. Rzadziej na potrzeby filtrowania wykorzystywane są zmienne klasyczne mierzone na słabych skalach pomiaru. Przykładem może być tutaj lokalizacja nieruchomości występująca jako zmienna mierzona na skali nominalnej.

W przypadku zmiennych symbolicznych złożona struktura danych w ujęciu symbolicznym sprawia, iż nie można w sposób bezpośredni stosować metod statystycznych opracowanych dla danych w ujęciu klasycznym (por. [Everitt, Dunn 2001; Hair i in. 2006]). Oznacza to brak mechanizmów filtrowania baz danych na podstawie kryteriów wyszukiwania o charakterze zmiennych symbolicznych, a w przypadku wyszukiwania nieruchomości brak możliwości uwzględnienia m.in. takiej zmiennej występującej w postaci listy kategorii, jaką jest uzbrojenie terenu. Realizacjami tej zmiennej nie są pojedyncze wartości, lecz zbiór kategorii. Oznacza to możliwość występowania dowolnej kombinacji możliwych kategorii dla tej zmiennej, np.: „prąd, woda, gaz”, w sytuacji gdy dla innej nieruchomości dowolne z wymienionych czynników nie występują lub są zastępowane innymi.

Podsumowując, należy zauważyć, iż dostępne, głównie na stronach internetowych, mechanizmy filtrowania baz danych są co najmniej niewystarczające, a nawet – co stanowi tezę niniejszego artykułu – nieadekwatne do zawartości dziedzinowych baz danych. W przekonaniu autorów, z racji występowania wspomnianych zmiennych symbolicznych, dedykowane rozwiązania w postaci „klasycznych” mechanizmów filtrowania uszczuplają możliwości filtrowania baz danych. Tym samym celem artykułu jest prezentacja rozwiązania pozwalającego wymienione wady wyeliminować, a przynajmniej w wysokim stopniu je ograniczyć.

3. Proponowane rozwiązanie

Rozwiązanie postawionego w artykule problemu polega na zaproponowaniu sposobu filtrowania dziedzinowych baz danych z wykorzystaniem porządkowania obiektów na podstawie odległości od obiektu wzorca z uwzględnieniem odpowiedniej miary podobieństwa. Warto podkreślić, iż ze względu na obiekty opisane m.in. zmiennymi symbolicznymi w rozwiązaniu zdecydowano się wykorzystać znormalizowaną miarę odległości Ichino-Yaguchiego U_3 . Konstrukcję miar odległości danych symbolicznych zaprezentowano m.in. w pracach: [Bock, Diday 2000, s. 165-185; Diday, Norihomme-Fraiture 2008, s. 126-129; Wilk 2005; 2006; Malerba i in. 2001; Malerba, Esposito, Monopoli 2002].

Miara Ichino-Yaguchiego U_3 w swojej konstrukcji wykorzystuje operatory połączenia „ \oplus ” oraz przekroju „ \otimes ”. Operator połączenia jest definiowany jako (por. [Bock, Diday 2000, s. 170-171]):

$$v_{ik} \oplus v_{jk} = \begin{cases} \left[\min\{\underline{v}_{ik}, \underline{v}_{jk}\}, \max\{\overline{v}_{ik}, \overline{v}_{jk}\} \right] & \text{zmienna w postaci przedziału liczbowego,} \\ v_{ik} \cup v_{jk} & \text{zmienna w postaci zbioru kategorii,} \end{cases}$$

gdzie: i, j – i -ty, j -ty obiekt,
 v_{ik}, v_{jk} – realizacja k -tej zmiennej dla odpowiednio i -tego i j -tego obiektu,
 $\underline{v}_{ik}, \underline{v}_{jk}$ ($\overline{v}_{ik}, \overline{v}_{jk}$) – dolne (górne) końce przedziału k -tej zmiennej, odpowiednio dla obiektów i oraz j .

Operator przekroju dla wszystkich typów zmiennych definiowany jest jako:

$$v_{ik} \otimes v_{jk} = v_{ik} \cap v_{jk} \cdot +.$$

W strukturze miary U_3 wyróżnia się odległości składowe oraz odległość agregatową. Odległości obiektów względem k -tej zmiennej wyznacza się w następujący sposób (por. [Bock i Diday 2000]):

$$d_{ijk} = \frac{\mu(v_{ik} \oplus v_{jk}) - \mu(v_{ik} \otimes v_{jk}) + \gamma \mathcal{V}(v_{ik}, v_{jk})}{\mu(V_k)}, \quad (1)$$

gdzie: $\mathcal{V}(v_{ik}, v_{jk}) = 2\mu(v_{ik} \otimes v_{jk}) - \mu(v_{ik}) - \mu(v_{jk})$,

γ – parametr, $\gamma \in [0; 0,5]$,

$\mu(v_{ik})$ – rozpiętość przedziału k -tej zmiennej dla i -tego obiektu lub liczba kategorii k -tej zmiennej dla i -tego obiektu,

$\mu(v_{ik} \oplus v_{jk})$ – wartość bezwzględna z różnicy wartości minimalnej i maksymalnej wyrażenia $v_{ik} \oplus v_{jk}$ lub liczba kategorii dla v_{ik} oraz v_{jk} ,

$\mu(v_{ik} \otimes v_{jk})$ – wartość bezwzględna części wspólnej przedziałów v_{ik} i v_{jk} lub liczba wspólnych kategorii dla v_{ik} oraz v_{jk} ,

$\mu(V_k)$ – rozpiętość zbioru realizacji k -tej zmiennej $\mu(V_k) = |\max\{\overline{v_k}\} - \min\{\underline{v_k}\}|$ lub liczba wszystkich kategorii k -tej zmiennej.

Odległość agregatową obiektu i -tego oraz j -tego wyznacza się w następujący sposób:

$$d_{ij} = \left[\sum_{k=1}^p (\omega_k d_{ijk})^\lambda \right]^{1/\lambda}, \quad (2)$$

gdzie: $d_{ij} \in [0,1]$,

λ – parametr, $\lambda \geq 1$,

ω_k – waga k -tej zmiennej ($k=1, \dots, p$), $\omega_k > 0$, $\sum_{k=1}^p \omega_k = 1$.

Sposób zastosowania znormalizowanej miary Ichino-Yaguchiego U_3 do rozwiązania postawionego w artykule problemu filtrowania dziedzinowych baz danych prezentuje poniższy przykład.

4. Filtrowanie dziedzinowych baz danych na przykładzie przeszukiwania ofert niezabudowanych nieruchomości gruntowych

W proponowanym przykładzie zawartość dziedzinowej bazy danych stanowi 38 wariantów (profilów) niezabudowanych nieruchomości gruntowych (działek budowlanych) opisanych zestawem ośmiu cech (atrybutów) w postaci następujących zmiennych klasycznych: cena nieruchomości, powierzchnia działki, lokalizacja oraz nasłonecznienie nieruchomości. W przypadku zmiennych symbolicznych są to: funkcja w miejscowym planie zagospodarowania przestrzennego, forma władania nieruchomością, uzbrojenia terenu oraz rodzaj dojazdu do nieruchomości. Należy zauważyć, iż każdą ze zmiennych charakteryzują określone realizacje (por. tab. 1). W przypadku ceny oraz powierzchni działki są to pojedyncze wartości liczbowe; w przypadku lokalizacji oraz nasłonecznienia nieruchomości są to wybrane, pojedyncze kategorie. W odniesieniu do lokalizacji są to następujące strefy: „centralna”, „śródmiejska” lub „peryferyjna”, w przypadku nasłonecznienia nieruchomości – wybrane, pojedyncze kategorie: „bardzo słoneczna” lub „słoneczna”.

Tabela 1. Cechy niezabudowanych nieruchomości gruntowych na jeleniogórskim rynku nieruchomości z oferty wybranego biura pośrednictwa w obrocie nieruchomościami

Nazwa cechy	Rodzaj cechy	Możliwe realizacje
Cena nieruchomości	ilorazowa	42.000-1.457.000 zł
Powierzchnia działki	ilorazowa	437-49.970 m ²
Lokalizacja nieruchomości	nominalna	centralna
		śródmiejska
		peryferyjna
Funkcja w miejscowym planie zagospodarowania przestrzennego	symboliczna (lista kategorii)	mieszkaniowa
		usługowa
		rolna
Forma władania	symboliczna (lista kategorii)	własność
		użytkowanie wieczyste
		hipoteka
Uzbrojenie terenu	symboliczna (lista kategorii)	kanalizacja
		prąd
		siła
		woda
		gaz
Dojazd do nieruchomości	symboliczna (lista kategorii)	droga asfaltowa
		droga utwardzona
		droga nieutwardzona
Nasłonecznienie działki	nominalna	bardzo słoneczna
		słoneczna

W przypadku występujących w przykładzie zmiennych symbolicznych realizacjami jest zbiór kategorii, tj. dowolna kombinacja pojedynczych kategorii dostępnych dla zmiennej. W przypadku funkcji w miejscowym planie zagospodarowania przestrzennego dostępne kategorie to: „mieszkaniowa”, „usługowa” oraz „rolna”; dla formy władania nieruchomością – „własność”, „użytkowanie wieczyste” i „hipoteka”; w przypadku uzbrojenia terenu – „kanalizacja”, „prąd”, „siła”, „woda” i „gaz”; odnośnie do dojazdu do nieruchomości – „droga asfaltowa”, „droga utwardzona” i „droga nieutwardzona” (por. tab. 1). Źródło danych stanowią strony internetowe jeleniogórskich biur pośrednictwa w obrocie nieruchomościami⁴.

Przedstawione w artykule rozwiązanie w miejsce filtrowania obiektów (ofert nieruchomości) z wykorzystaniem wyłącznie zmiennych klasycznych, co oznacza automatyczną eliminację kryteriów „symbolicznych”, proponuje mechanizm polegający każdorazowo na wykorzystaniu obu rodzajów zmiennych – klasycznych oraz symbolicznych. Połączenie takie oznacza automatycznie brak możliwości zastosowania „klasycznych” metod wielowymiarowej analizy statystycznej na rzecz metodologii analizy danych symbolicznych. Z uwzględnieniem odpowiedniej miary odległości Ichino-Yaguchiego dla danych symbolicznych proponuje się obliczenie odległości od obiektu wzorca do każdego z obiektów bazy danych. W proponowanym rozwiązaniu umożliwia to wyróżnienie obiektów najbardziej do wzorca podobnych, których odpowiednią liczbę uznać należy za wynik przeszukiwania bazy danych.

Na potrzeby prezentacji proponowanego rozwiązania przyjmuje się założenie, iż użytkownika bazy danych (strony internetowej z ofertami nieruchomości) szczególnie interesuje na własność wariant nieruchomości w postaci bardzo słonecznej działki mieszkaniowej w cenie do 300 000 zł o powierzchni między 700 a 3000 m², charakteryzujący się „śródmiejską” lub „peryferyjną” lokalizacją z dostępem do „kanalizacji”, „prądu”, „wody” oraz „gazu”, posiadający dojazd w postaci „drogi asfaltowej” i/lub „drogi utwardzonej”. Jednocześnie zakłada się, iż użytkownika interesuje pewien zbiór obiektów – działek budowlanych – w przykładzie są to 3 działki budowlane, które spełniają wszystkie kryteria lub jak najwięcej kryteriów wyszukiwania.

Aby możliwe było wyszukanie obiektów spełniających jak najwięcej zadanych kryteriów, w proponowanym rozwiązaniu przeszukiwanie bazy danych polega na wyznaczeniu macierzy odległości między obiektami z wykorzystaniem wzorów (1) oraz (2), a następnie odpowiednim ich uszeregowaniu. Obliczenia, wykorzystując w tym celu polecenie `dist.SDA` z modułu `symbolicDA`, zrealizowano w środowisku R, otrzymując następujący zestaw wyników dla miary Ichino-Yaguchiego:

```
> library(symbolicDA)
> x<-parse.SO("nieruchomosci")
> d<-dist.SDA(x, type="U_3", gamma=0.2, power=2)
```

⁴ Główne źródło danych o nieruchomościach stanowiła strona WWW biura pośrednictwa w obrocie nieruchomościami „Nieruchomości Kopczyński”, <http://www.nkop.pl>.


```

      1      2      3      4      5      6      7      8      9     10     11     ...     36     37     38
2  0.41
3  1.01 0.92
4  0.73 0.80 0.80
5  0.86 0.76 0.53 0.80
6  1.01 1.00 0.38 0.80 0.84
7  1.01 1.00 1.13 0.80 1.36 0.92
8  0.82 0.60 0.60 0.84 0.80 0.60 0.80
9  0.82 0.97 1.10 0.38 1.10 1.10 0.97 1.07
10 1.08 0.86 1.26 1.10 1.00 1.36 1.47 1.16 1.03
11 0.67 0.93 0.54 0.80 0.76 0.85 0.85 0.96 0.80 1.25
12 0.98 0.96 0.80 0.39 0.80 0.60 0.80 0.93 0.76 1.29 1.11
13 1.04 0.97 0.26 0.83 0.60 0.45 1.17 0.65 1.12 1.27 0.59 ...
14 1.05 1.23 0.80 0.76 1.10 0.96 0.61 1.13 0.84 1.53 0.60 ...
...
37 0.82 0.96 1.22 0.66 1.10 1.22 1.10 1.19 0.76 0.90 0.97 ... 0.01
38 0.82 0.96 1.22 0.66 1.10 1.22 1.10 1.19 0.76 0.90 0.97 ... 0.01 0.00
39 0.90 1.10 0.80 0.65 0.96 1.03 0.71 1.13 0.65 1.34 0.60 ... 0.85 0.85 0.85.

```

Ze względu na potrzebę uszeregowania profili nieruchomości od najbardziej do najmniej podobnego względem profilu wzorca w obliczeniach stosuje się dodatkowo polecenie `sort`, co pozwala uzyskać następującą, ostateczną postać wyników:

```

> library(symbolicDA)
> x<-parse.SO("nieruchomosci")
> d<-dist.SDA(x, type="U_3", gamma=0.2, power=2)
> d<-as.matrix(d)
> round(sort(d[,]), 4)

```

```

      1      29      2      16      11      4      17      8      9      30      31
0.0000 0.4137 0.4142 0.4922 0.6750 0.7257 0.7257 0.8179 0.8179 0.8179 0.8179
32      33      34      35      36      37      38      5      20      21      22
0.8179 0.8179 0.8179 0.8179 0.8179 0.8179 0.8179 0.8602 0.8602 0.8602 0.8964
25      39      12      3      6      7      13      27      14      28      10
0.9006 0.9006 0.9764 1.0122 1.0122 1.0123 1.0428 1.0467 1.0470 1.0572 1.0802
23      15      24      18      19      26
1.1043 1.1441 1.2717 1.2902 1.4325 1.4987.

```

Rezultaty przeszukiwania jednocześnie wskazują, iż wariantami nieruchomości najbardziej podobnymi do profilu wzorcowego (w liczbie 3 sztuk, nie licząc wariantu wzorcowego opatrzonego numerem 1) są profile nr 29, 2 oraz 16. Analiza wariantów nieruchomości wskazuje, iż w przypadku wariantu nr 29 wszystkie cechy mieszczą się w granicach kryteriów wariantu wzorca, w przypadku dwóch pozostałych wariantów różnice dotyczą pojedynczych cech. I tak w przypadku wariantu nr 2 zauważyć można niezgodność w odniesieniu do powierzchni działki, gdzie powierzchnia wynosi „437 m²” w miejsce wymaganych co najmniej „700 m²”; w przypadku wariantu nr 16 różnica dotyczy uzbrojenia terenu i polega na występowaniu dodatkowej kategorii w postaci „siły”, tj. *de facto* kompletnego uzbrojenia na tle wzorca w postaci kombinacji „kanalizacji, prądu, wody, gazu”.

Warto w tym miejscu zauważyć, iż dla przykładu wariant najbardziej odległy (nr 26) od wzorcowego profilu nieruchomości różni się względem wzorca w zakresie 4 cech. Różnice te obejmują następujące realizacje (w odniesieniu do obiektu wzorca): cena – „1 457 000 zł” (wzorzec – maksymalnie „300 000 zł”), powierzchnia – „26 489 m²” (wzorzec – „3300 m²”), funkcja w miejscowym planie zagospodarowania przestrzennego – „rolna” (wzorzec – „mieszkaniowa”), uzbrojenie – „prąd” (wzorzec – „kanalizacja, prąd, woda, gaz”).

Należy podkreślić, iż proponowany mechanizm filtrowania oznacza wyszukanie zadeklarowanej liczby obiektów, które ze wszystkich z dostępnych w bazie danych w największym stopniu odpowiadają ofercie porównywanej. Oznacza to, iż proponowany mechanizm przeszukiwania, co należy uznać za jego zaletę, pozwala bez względu na objętość bazy danych wyszukiwać obiekty najbardziej, choć nie do końca, spełniające kryteria obiektu wzorcowego. Filtrowanie w „klasycznym” rozumieniu tego pojęcia takiej możliwości nie daje.

5. Podsumowanie

Jak to zostało zauważone w artykule, dostępne narzędzia przeszukiwania (filtrowania) bazują głównie na zmiennych klasycznych opisujących obiekty będące przedmiotem baz danych. Przedstawione w artykule rozwiązanie obok zmiennych klasycznych uwzględnia dodatkowo zmienne symboliczne, które opisują obiekty w sposób pełny, bez utraty informacji o obiektach, których dotyczą. W opinii autorów artykułu umożliwia to przeszukiwanie baz danych na podstawie dowolnej, w tym pełnej liczby kryteriów bez względu na rodzaj występujących zmiennych. W praktyce oznacza to bardziej efektywne przeszukiwanie baz danych z uwzględnieniem możliwości do tej pory niewystępujących w dostępnych mechanizmach filtrowania dziedzinowych baz danych.

Literatura

- Billard L., Diday E., *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Wiley, Chichester 2006.
- Bock H.H., Diday E. (red.), *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin, Heidelberg 2000.
- Diday E., Noirhomme-Fraiture M. (red.), *Symbolic Data Analysis and the Sodas Software*, John Wiley & Sons, Chichester 2008.
- Everitt B.S., Dunn G., *Applied Multivariate Data Analysis*, Arnold, London 2001.
- Gatnar E., Walesiak M. (red.), *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck, Warszawa 2011.
- Hair J.F., Black W.C., Babin B.J., Anderson R.E., Tatham R.L., *Multivariate Data Analysis*, Pearson Prentice Hall, New Jersey 2006.
- Malerba D., Esposito F., Giovalle V., Tamma V., *Comparing Dissimilarity Measures for Symbolic Data Analysis*, [w:] *New Techniques and Technologies for Statistics and Exchange of Technology and Know-how*, P. Nanopoulos (red.), ETK-NTTS'01 2001.

- Malerba D., Esposito F., Monopoli M., *Comparing Dissimilarity Measures for Probabilistic Symbolic Objects*, [w:] *Data Mining III. Series Management Information Systems*, A. Zanasi, C.A. Brebbia, N.F.F. Ebecken, P. Melli (red.), vol. 6, WIT Press, Southampton 2002.
- Wilk J., *Miary odległości obiektów opisanych zmiennymi symbolicznymi z wagami*, [w:] *Taksonomia 13, Klasyfikacja i analiza danych – teoria i zastosowania*, K. Jajuga, M. Walesiak (red.), Prace Naukowe Akademii Ekonomicznej we Wrocławiu, Wrocław 2005.
- Wilk J., *Problemy klasyfikacji obiektów symbolicznych. Symboliczne miary odległości*, [w:] *Ilościowe i jakościowe metody badania rynku. Pomiar i jego skuteczność*, J. Garczarczyk (red.), ZN AE nr 71, Wydawnictwo AE, Poznań 2006.

APPLICATION OF SYMBOLIC DATA ANALYSIS METHODS FOR DOMAIN DATABASE SEARCHING

Summary: The paper presents the application of symbolic data analysis methods for domain database searching. Because it is better to use more information about objects, the proposition of authors' database searching includes classical and symbolic variables and Ichino-Yaguchi dissimilarity measure. In authors' opinion it means the most effective database filtering. The authors illustrate the presented solution using symbolic data analysis method on an empirical example.

Keywords: database searching, symbolic data analysis, distance measurement, real estate offers.