

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

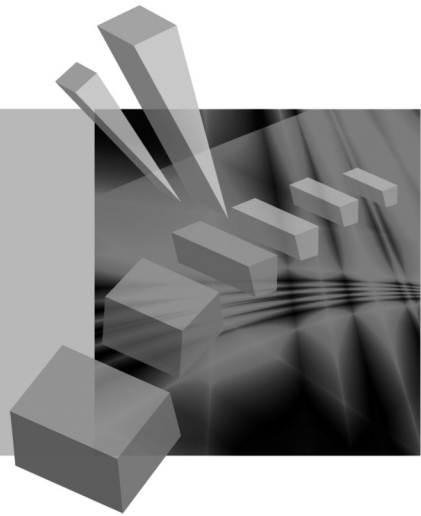
RESEARCH PAPERS

of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pocięcha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities <http://cejsh.icm.edu.pl>
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon [http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php](http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php)

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Kopowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2012

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Wstęp	13
Stanisława Bartosiewicz , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej	17
Andrzej Sokolowski , Q uniwersalna miara odległości	22
Eugeniusz Gatnar , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP)	31
Marek Walesiak , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
Krzysztof Jajuga, Marek Walesiak , XXV lat konferencji taksonomicznych – fakty i refleksje	47
Józef Pocięcha, Barbara Pawelek , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne	50
Paweł Lula , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych	58
Ewa Roszkowska , Zastosowanie metody TOPSIS do wspomaganie procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
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 Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczak , 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

Jerzy Krawczuk

Politechnika Białostocka

SKUTECZNOŚĆ METOD KLASYFIKACJI W PROGNOZOWANIU KIERUNKU ZMIAN INDEKSU GIEŁDOWEGO S&P500

Streszczenie: Kluczowe w prognozowaniu zachowania rynków finansowych jest określenie kierunku zmiany, czyli określenie, czy nastąpi wzrost czy spadek. Narzędziem, które może zostać wykorzystane do tego celu, są używane w eksploracji danych klasyfikatory. Klasyfikator na podstawie zbioru uczącego (danych historycznych) może przydzielić klasę wzrostu bądź spadku. W artykule zbadano skuteczność takiej klasyfikacji (prognozy) dla indeksu giełdy amerykańskiej S&P500. Użyto siedmiu popularnych klasyfikatorów, w tym drzew decyzyjnych i klasyfikatora SVM.

Słowa kluczowe: prognozowanie giełdy, finansowe szeregi czasowe, uczenie maszynowe, klasyfikacja.

1. Wstęp

Prognozowanie indeksów czy też innych instrumentów notowanych na giełdach, takich jak ceny akcji, kursy walut, jest zadaniem trudnym, dla którego zastosowanie znalazło wiele technik. Standardowym podejściem jest analiza techniczna [Edwards, Magee 1997] oparta na analizie wykresów. Wiele aplikacji dostarczanych przez biura maklerskie ma rozbudowane moduły analizy technicznej. Również w komentarzach prasowych odnajdziemy informacje o pojawiających się formacjach cenowych. Jedną z pierwszych technik analizy wykresów były tzw. świece japońskie, których powstanie datuje się na wiek XVIII [Nison 2001].

Podejściem bardziej współczesnym jest przewidywanie zmian cen na podstawie ekonometrycznych szeregów czasowych [Hamilton 1994]. Popularne są modele ARIMA [Box, Jenkins 1983] modelujące wartość oczekiwaną szeregu, jak również modele ARCH [Engle 1982], GARCH [Bollerslev 1986], w których wariancja jest zmienna w czasie. Modelowanie wariancji zmiennej w czasie jest istotne z punktu widzenia ryzyka finansowego. Rozwój tej dziedziny spotkał się w roku 2003 z uznaniem szwedzkiej Królewskiej Akademii Nauk, która przyznała Nagrodę Nobla Robertowi Engle'owi za rozwój tych właśnie technik.

Czynione są również próby stosowania innych narzędzi, w tym narzędzi eksploracji danych, takich jak sieci neuronowe [Egeli i in. 2003] i algorytmy genetyczne [Kim, Han 2000]. Niniejszy artykuł opisuje użycie jednej z metod eksploracji danych, jaką jest klasyfikacja do prognozy kierunku zmiany indeksu S&P500. W eksperymencie porównano wyniki osiągnięte za pomocą siedmiu różnych klasyfikatorów.

2. Klasyfikacja

Jedną z metod eksploracji danych jest klasyfikacja. Jest to metoda uczenia z nadzorem (z nauczycielem). Jej zadaniem jest określenie przynależności obiektu na podstawie wartości jego atrybutów do jednej z istniejących klas. Odpowiedni model (klasyfikator) budowany jest na podstawie zbioru danych uczących, gdzie przynależności do klas są znane.

Klasyfikator:

$$h: X \rightarrow Y$$

budowany jest na podstawie zbioru uczącego

$$\{(x[n]_1, y_1), \dots, (x[n]_N, y_N)\},$$

gdzie: N – liczba obserwacji,

n – wymiar wektora cech (liczba zmiennych objaśniających).

Gdy pojawia się nowa obserwacja x_{n+1} , klasyfikator h może zostać użyty do nadania jej etykiety klasy $y \in Y$. Dla danych giełdowych obiektem może być liczbowy opis stanu giełdy danego dnia, a klasą wzrost lub spadek indeksu w kolejnym dniu. Podobne podejście do prognozowania danych giełdowych odnajdziemy w pracy [Kim 2003]. Autor prognozuje kierunek zmiany indeksu giełdy koreańskiej, wykorzystując klasyfikator SVM (*Support Vector Machine*), sieć neuronową i metodę najbliższych sąsiadów (K-NN z *K Nearest Neighbours*). W innej pracy [Huang i in. 2005] autorzy za pomocą klasyfikatora SVM prognozują kierunek zmiany indeksu giełdy japońskiej NIKKEI.

W niniejszej pracy do prognozy indeksu giełdy amerykańskiej S&P500 użyto siedmiu klasyfikatorów. Implementacja sześciu standardowych algorytmów pochodzi z pakietu do analizy danych WEKA [Hall i in. 2009]:

- metoda najbliższych sąsiadów (K-NN) [Cover, Hart 1967] (Weka – IB1),
- drzewa decyzyjne – algorytm C 4.5 [Quinlan 1993] (Weka – J48),
- regresja logistyczna [Hosmer, Lemeshow 2000] (Weka – Logistic),
- naiwny klasyfikator bayesowski [Duda i in. 2001] (Weka – NaiveBayes),
- maszyna wektorów wspierających (SVM) [Cortes, Vapnik 1995] (Weka – SMO),
- zero-R (Weka – ZeroR),

- klasyfikator liniowy oparty na wypukłej i odcinkowo-liniowej funkcji kryterialnej CPL [Bobrowski 2005], implementacja [Łukaszuk 2010].

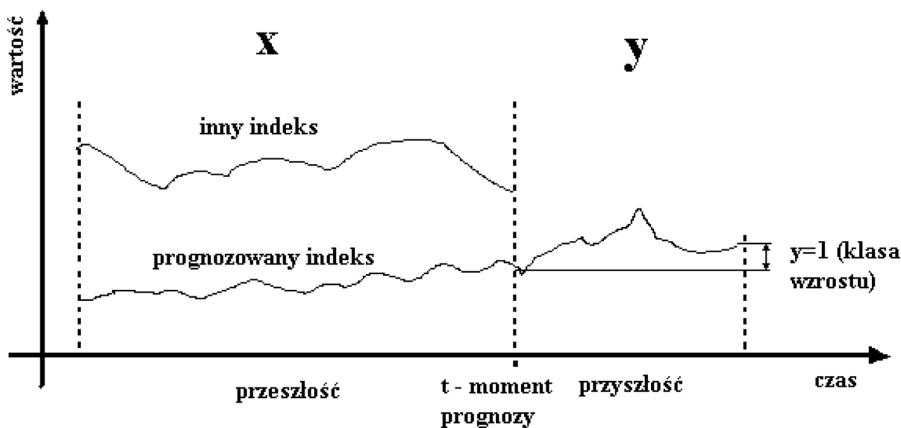
Klasyfikator Zero-R jest prostym klasyfikatorem i został użyty w artykule do celów porównawczych (jako tzw. benchmark). Nauka tego klasyfikatora polega na znalezieniu w zbiorze uczącym klasy najbardziej licznej i następnie klasyfikowaniu wszystkich nowych obiektów do tej właśnie klasy, niezależnie od wartości ich atrybutów. Klasyfikatory SVM i CPL są klasyfikatorami liniowymi, które poszukują hiperpłaszczyzny separującej zbiory o postaci:

$$0 = w_0 + w_1x_1 + \dots + w_nx_n.$$

Klasyfikator SVM poszukuje hiperpłaszczyzny z maksymalnym marginesem separującym, natomiast zasada działania klasyfikatora CPL oparta jest na wypukłej i odcinkowo-liniowej (*convex and piecewise linear*) funkcji kary, która podobna jest do perceptronowej funkcji kryterialnej [Rosenblatt 1958].

3. Reprezentacja danych giełdowych

Stan giełdy możemy opisać za pomocą n -wymiarowego wektora cech $x[n] = [x_1, \dots, x_n]^T$. Cechy to liczby rzeczywiste ($x_i \in R^1$), którymi mogą być ceny historyczne, średnie cen czy też wybrane wskaźniki analizy technicznej. Konstruując wektor cech x i etykietując obserwacje klasą wzrostu bądź spadku y , należy pamiętać o zależności czasowej (rys. 1).



Rys. 1. Czasowa zależność pomiędzy zmienną objaśnianą y a zmienną objaśniającą x

Źródło: opracowanie własne.

W chwili t wszystkie cechy wektora x powinny być wyliczone z bieżącej i przeszłych (historycznych) cen indeksu prognozowanego, jak również innych indeksów.

Natomiast klasa y powinna zostać określona na podstawie przyszłej zmiany ceny od chwili t . W przeprowadzonym eksperymencie wektor x został określony dla każdego dnia notowań giełdy amerykańskiej z okresu od listopada 2007 do maja 2011 r. Prognoza stawiana jest na otwarciu giełdy, zatem wartością przewidywaną jest zmiana wartości indeksu do kolejnego otwarcia:

$$y = 0, \text{ jeżeli otwarcie } (t + 1) < \text{otwarcie } (t),$$

$$y = 1, \text{ jeżeli otwarcie } (t + 1) > \text{otwarcie } (t).$$

Wartości atrybutów wektora x (zmiennych objaśniających) to np. otwarcie (t), zamknięcie ($t - 1$), otwarcie ($t - 1$). Niedozwolone jest np. użycie ceny zamknięcia z dnia t , gdyż w momencie stawiania prognozy jest ona jeszcze nieznaną. Wektor x został skonstruowany z dwóch grup cech. Pierwsza grupa to historyczne ceny i wskaźniki analizy technicznej dla prognozowanego instrumentu, indeksu S&P500. Skonstruowano w ten sposób 18 cech (tab. 1).

Tabela 1. Cechy wektora x opisującego stan giełdy, uzyskane z historycznych notowań indeksu S&P500

Lp.	Nazwa atrybutu	Opis
1	Cena otwarcia	Cena z otwarcia notowań 9:30 (Eastern Time)
2	Cena zamknięcia	Cena z zamknięcia notowań 16:00 (Eastern Time)
3	Luka otwarcia	Procentowa różnica pomiędzy ceną otwarcia a poprzednią ceną zamknięcia
4, 5, 6	1-, 2-, 5-dniowa zmiana ceny	Procentowa różnica ceny bieżącego otwarcia i otwarcia sprzed 1 dnia, 2 i 5 dni
7, 8	Jednodniowa zmiana z wczoraj i z przedwczoraj	Jednodniowe historyczne procentowe zmiany cen
9, 10, 11	Średnia ruchoma 9-, 12-, 26-dniowa	Średnia notowań z cen otwarcia
12	Luka otwarcia do potęgi drugiej	Element nieliniowy
13, 14	Disparity z 5 i 10 dni	Odległość pomiędzy ceną bieżącą a 5- i 10-dniową średnią
15	MACD(12, 26)	Wskaźnik zbieżności średnich ruchomych
16	PercentK(14)	Wskaźnik oddaje zależność pomiędzy ceną bieżącą i minimalną i maksymalną z ostatnich 14 dni
17	PercentR(10)	%R Williama – oscylator pokazujący zależność pomiędzy bieżącą ceną a zakresem cen z ostatnich 10 dni
18	RSI(14)	Wskaźnik względnej siły, określający siłę trendu

Źródło: opracowanie własne.

Druga grupa to notowania innych instrumentów finansowych, takich jak:

- indeksy giełd innych 14 krajów (Kanada, Meksyk, Brazylia, Australia, Japonia, Malezja, Tajwan, Korea Płd., Hongkong, Chiny, Rosja, Wielka Brytania, Szwajcaria, Niemcy),

- surowce i metale szlachetne (ropa naftowa, gaz ziemny, srebro i złoto),
- kursy walut (5 walut do dolara USA: dolar australijski, dolar kanadyjski, funt brytyjski, euro oraz jen).

Razem są to 23 instrumenty finansowe. Każdy z nich został reprezentowany przez jeden atrybut, tzw. lukę otwarcia, czyli różnicę procentową pomiędzy ceną otwarcia a poprzednią ceną zamknięcia. Dodatkowym 24. atrybutem w tej grupie jest wartość otwarcia indeksu VIX¹, który wyraża oczekiwaną w najbliższych 30 dniach zmienność indeksu S&P500.

4. Eksperyment

W pracy tej cały zbiór danych został podzielony na okresy uczące o długości jednego roku (252 obiekty) i okresy testowe o trzech długościach 6, 3 miesięcy i 1 miesiąca. Przykład pojedynczego eksperymentu dla 6-miesięcznego okresu testowego i klasyfikatora CPL pokazuje rys. 2.

		Model 1		Model 2		Model 3		Model 4		Model 5								
		252 dni trening		126 test		252 dni treningowy		126 test		252 dni treningowy		126 test		252 dni treningowy		126 test		
		Lis-2007		Lis-2008		Maj-2009		Lis-2009		Maj-2010		Lis-2010		Maj-2011		Średnia		
Zbiory treningowe	Trafność prognozy		66,7%	71,0%	65,9%	58,3%	55,2%										63,4%	
	Zysk/Strata		145,4%	87,2%	65,6%	42,8%	33,1%											74,8%
Zbiory testowe	Trafność prognozy			64,0%	56,8%	44,0%	47,2%	44,0%										51,2%
	Zysk/Strata			22,7%	16,4%	-15,9%	-1,1%	-16,3%										1,2%

Rys. 2. Podział danych na zbiory treningowe i testowe oraz wyniki uzyskane dla klasyfikatora CPL

Źródło: opracowanie własne.

Na przykład klasyfikator zbudowany na zbiorze danych uczących z okresu listopad 2007-listopad 2008 (model 1) prawidłowo klasyfikuje 66,7% tych danych.

Dodatkową miarą oprócz liczby poprawnie sklasyfikowanych wzrostów/spadków jest miara potencjalnego zysku/straty modelu. Wiąże się ona z tym, iż każdy wzrost i spadek to konkretna wartość procentowa, np. wzrost o 1% bądź spadek o 1%. Zmiana ta może być większa, np. 3%, jak również mniejsza, np. 0,2%. Dla inwestora ważny jest nie tylko sam fakt trafności prognozy kierunku zmiany, ale również jej wielkość. Z większą zmianą wiąże się większy zysk bądź strata. Miara zysku/straty to właśnie uwzględnia i została obliczona jako suma procentowych zmian trafnie przewidzianych (zysków) minus suma zmian błędnie przewidzianych

¹ <http://www.cboe.com/micro/VIX/vixintro.aspx>.

(strat). Na przykład, gdy przewidziano prawidłowo zmianę o 3% i błędnie zmianę o 1%, wówczas wartość tej miary wyniesie 2%.

Miara zysku/straty dla pierwszego okresu testowego wynosi 145,4%. Model 1 został następnie użyty na danych testowych z kolejnych 6 miesięcy, uzyskując jakość klasyfikacji 64% i zysk 22,7%. Kolejne modele na danych testowych wypadły już jednak gorzej, np. model 3 przewidział poprawnie jedynie 44% kierunków zmian, co odpowiada stracie w wysokości 15,9%. Wartości obu miar zostały uśrednione dla wszystkich 5 modeli, dając średnią jakość klasyfikacji 51,2% i średni zysk 1,2%.

Analogiczne obliczenia wykonano dla pozostałych sześciu klasyfikatorów (tab. 2). Dla okresu 6-miesięcznego najlepszy wynik uzyskał klasyfikator oparty na regresji logistycznej 54,50%, jednak był to wynik jedynie nieznacznie lepszy od 54,24% uzyskanych przez klasyfikator benchmarkowy (Zero-R). Podobnie wyglądają wyniki dla 3-miesięcznego okresu testowego, ponownie najlepszy wynik 56,77% uzyskał klasyfikator oparty na regresji logistycznej, drugi wynik 54,68% osiągnął klasyfikator benchmarkowy. Najwyższą trafność prognozy dla okresu miesięcznego uzyskał klasyfikator Zero-R (56,57%).

Miara zysku/straty w tab. 2 została przeskalowana do okresu rocznego, tzn. np. 1,2% uzyskane na 6 miesiącach odpowiada 2,4% w skali roku. Wartości zysku/straty dla 3-miesięcznych okresów testowych zostały pomnożone przez 4, a miesięcznych przez 12.

Tabela 2. Jakość klasyfikacji oraz zysku/straty na zbiorach testowych, uzyskane dla różnych klasyfikatorów i różnych okresów testowych (w %)

Klasyfikator	Jakość klasyfikacji			Roczny zysk/strata		
	1 miesiąc	3 miesiące	6 miesięcy	1 miesiąc	3 miesiące	6 miesięcy
K-NN	53,06	50,97	48,00	53,73	1,68	-1,85
Regresja log.	53,26	56,77	54,50	9,90	37,29	18,39
Drzewa C 4.5	54,67	54,19	53,60	-1,95	12,55	1,85
Naiwny Bayes	55,73	50,16	47,52	7,02	-13,63	-8,65
SVM	51,85	53,87	49,60	-13,40	6,40	-5,70
CPL	52,53	53,02	52,22	21,25	20,20	7,73
Zero-R	56,57	54,68	54,24	4,86	6,69	6,69

Źródło: opracowanie własne.

Generalnie możemy zaobserwować poprawę wyników wraz ze skróceniem okresu testowego. Dla okresu 3-miesięcznego każdy z 7 klasyfikatorów uzyskał wynik lepszy niż dla okresu 6-miesięcznego. Kolejne skrócenie okresu do miesiąca poprawiło wynik 4 klasyfikatorów.

5. Podsumowanie

Celem pracy było sprawdzenie skuteczności metod klasyfikacji w prognozowaniu danych giełdowych. Wykonane obliczenia pokazały, iż wyniki uzyskiwane różnymi metodami są do siebie zbliżone i nieznacznie przekraczają 50%. Jest to wynik niewiele lepszy od losowego. Jednak w przypadku danych giełdowych raczej nie należy oczekiwać wyników na poziomie 90%. Jak pokazują wyniki na zbiorach treningowych, już uzyskanie poziomu 67% przekłada się na roczny zysk w wysokości aż 145%.

Warto zwrócić uwagę, iż większa jakość prognozy (np. miesiąc Zero-R 56,57%, a K-NN 53,06%) nie musi koniecznie przekładać się na większy zysk (Zero-R 4,86% a K-NN 53,73%). Jest to zrozumiałe, gdyż wielkość zysku zależy od tego, jak dużą zmianę procentową trafnie przewidzimy. W przeprowadzonym eksperymencie wzrosty indeksu o 0,2% i 2,0% trafiały do tej samej klasy wzrostów. Na przykład gdyby prawidłowo przewidziano tylko jedną z tych dwóch wartości, wówczas jakość klasyfikacji wyniosłaby 50%, natomiast zysk/strata mógłby wynieść +1,8% lub -1,8%. Preferowany powinien być klasyfikator z wynikiem +1,8%. Uwzględnienie zysku/straty w procesie budowy klasyfikatora np. za pomocą algorytmu MetaCost [Domingos 1999] to kolejny możliwy etap badań nad zastosowaniem klasyfikatorów do prognozowania danych giełdowych.

Literatura

- Bobrowski L., *Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych*, Wydawnictwa Politechniki Białostockiej, 2005.
- Bollerslev T., *Generalized autoregressive conditional heteroskedasticity*, „Journal of Econometrics” 1986, no 31.
- Box G.E.P., Jenkins G.M., *Analiza szeregów czasowych*, Państwowe Wydawnictwo Naukowe, 1983.
- Cortes C., Vapnik V., *Support-vector networks*, “Machine Learning” 1995, no 20.
- Cover T.M., Hart P.E., *Nearest neighbor pattern classification*, „IEEE Transactions on Information Theory” 1967, no IT-13.
- Domingos P., *MetaCost: A General Method for Making Classifiers Cost-Sensitive*, Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.
- Duda O.R., Hart P.E., Stork D.G., *Pattern Classification*, Wiley, New York 2001.
- Edwards R.D., Magee J., *Technical Analysis of Stock Trends*, AMACOM 7th edition, 1997.
- Egeli B., Ozturan M., Badur B., *Stock Market Prediction Using Artificial Neural Networks*, „Proceedings of the 3rd International Conference on Business”, 2003.
- Engle R.F., *Autoregressive conditional heteroskedasticity with the estimates of the variance of U.K. inflation*, „Econometrica” 1982, no 4.
- Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H., *The WEKA data mining software: an update*, „SIGKDD Explorations” 2009, vol. 11, Issue 1.
- Hamilton J.D., *Time Series Analysis*, Princeton University Press, 1994.
- Hosmer D.W., Lemeshow S., *Applied Logistic Regression*, Wiley, 2000.
- Huang W., Nakamori Y., Wang S.Y., *Forecasting stock market movement direction with support vector machine*, „Computers & Operations Research” 2005, vol. 32, Issue 10.

- Kim K.J., *Financial time series forecasting using support vector machines*, „Neurocomputing” 2003, vol. 55, Issues 1-2.
- Kim K.J., Han I., *Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index*, „Expert Systems with Applications”, 2000, vol. 19.
- Lukaszuk T., *Techniki eksploracji danych oparte na funkcjach kryterialnych typu CPL w informatycznym systemie pracy zdalnej*, rozprawa doktorska, Politechnika Białostocka, 2010.
- Nison S., *Japanese Candlestick Charting Techniques*, Prentice Hall Press, 2nd edition, 2001.
- Quinlan R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo 1993.
- Rosenblatt F., *The perceptron: a probabilistic model for information storage and organization in the brain*, „Cornell Aeronautical Laboratory, Psychological Review” 1958, vol. 65, no 6.

EFFECTIVENESS OF CLASSIFICATION METHODS IN S&P500 STOCK INDEX DIRECTION CHANGES FORECASTING

Summary: The most important factor in the forecasting of financial market is to determine the direction of the market – will the market go up or will there be the descent. A tool which can be used for this purpose is a classifier – a frequently used data mining technique. Based on a learning dataset (historical data), the classifier will determine if a new observation falls into the class of increase or the class of decline. In this article the accuracy of such a classification (or forecast) has been analyzed for the American stock exchange index – S&P500. Seven popular classifiers have been used, including the decision trees and the SVM classifier.

Keywords: stock market prediction, financial time series, machine learning, classification.