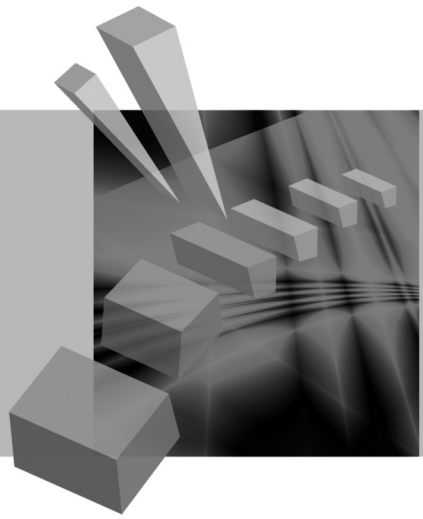
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

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 Pociecha,
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

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

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

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

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

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

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].

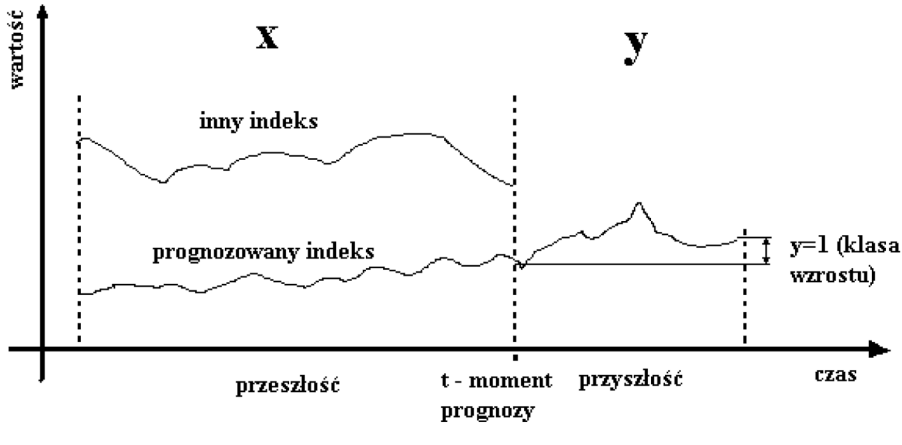
Klasyfikator Zero-R jest prostym klasyfikatorem i został użyty w artykule do celów porównawczych (jako tzw. benchmark). Nauka tego klasyfikatora polega na znalezieniu w zbiorze uczącym klasy najbardziej licznej i następnie klasyfikowaniu wszystkich nowych obiektów do tej właśnie klasy, niezależnie od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

$$0 = w_0 + w_1 x_1 + ... + w_n x_n .$$

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 $x[n] = [x_1, ..., x_n]^T$. 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 nieznana. 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 Williamsa – 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:

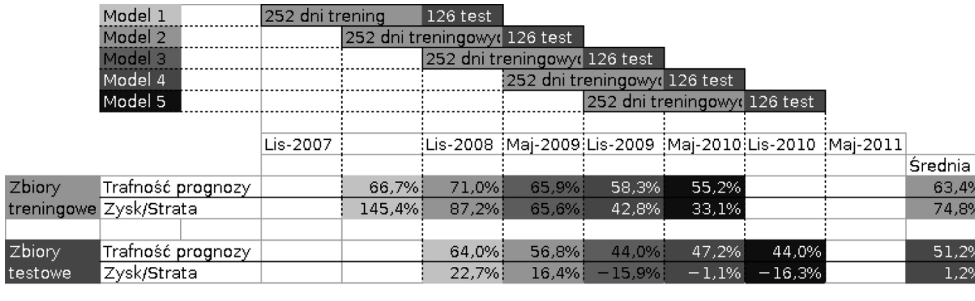
– indeksy giełd innych 14 krajów (Kanada, Meksyk, Brazylia, Australia, Japonia, Malezja, Tajwan, Korea Płd., Hongkong, Chiny, Rosja, Wielka Brytania, Szwajcaria, Niemcy),

– 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 VIX[1], 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.



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.
Łukaszuk 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.