METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XII/2, 2011, str. 232–240

KLASYFIKATOR LINIOWY TYPU CPL UWZGLĘDNIAJĄCY KOSZTY BŁĘDÓW KLASYFIKACJI JAKO NARZĘDZIE PROGNOZOWANIA GIEŁDY

Jerzy Krawczuk
Wydział Informatyki, Politechnika Białostocka
e-mail: j.krawczuk@pb.edu.pl

Streszczenie: Jeden z rodzajów eksploracji danych – klasyfikacja – może zostać użyty do prognozowania zmian cen na giełdzie. W najprostszym scenariuszu możemy klasyfikować dane giełdowe do jednej z dwóch klas: wzrostów bądź spadków. W standardowym podejściu przy budowie klasyfikatora maksymalizowana jest ilość prawidłowo sklasyfikowanych obiektów, jednak dla danych giełdowych lepszym wyznacznikiem jakości modelu może być osiągnięty zysk. W artykule tym opisano klasyfikator liniowy oparty o wypukłe i odcinkowo-liniowe funkcje kary (CPL) maksymalizujący wartość zysku.

Słowa kluczowe: klasyfikator liniowy, prognozowanie giełdy, funkcje typu CPL

WSTĘP

Do prognozowania zachowań cen na rynkach finansowych używanych jest wiele technik. Najstarszą z nich jest analiza techniczna [Edwards i in. 1997], jej początki datowane są na rok 1688 i pojawienie się książki Joshepa de la Vegi „Confusion of Confusions" opisującej działanie giełdy w Amsterdamie. Z kolei w Azji w XVIII wieku pojawiła się analiza za pomocą tzw. świec japońskich [Nison 2001] zapoczątkowana przez Homma Munehisa. Analiza techniczna skupia się na analizie wykresów i poszukiwaniu powtarzających się formacji. Jest to technika bardzo popularna, w codziennych komentarzach giełdowych odnajdziemy informację o pojawiających się formacjach. Również aplikacje dostarczane przez biura maklerskie zawierają wbudowane moduły analizy technicznej.

Drugim standardowym podejściem do prognozy giełdy są techniki oparte o ekonometryczne szeregi czasowe. Najczęściej są to modele autoregresyjne [Hamilton 1994], które modelują zachowanie cen jako kombinację liniową cen historycznych i czynnika losowego. Popularne modele to ARIMA, metodologie budowy takich modeli odnajdziemy w pracy [Box Jenkins 1983]. W przypadku szeregów finansowych często mamy do czynienia z tzw. heteroskedastycznością, czyli wariancją zmienną w czasie. Dla takich szeregów dodatkowo do modelowania wariancji używa się modeli z rodziny ARCH [Engle 1982] i GARCH [Bollerslev 1986].

Stosunkowo młodym podejściem w prognozowaniu giełdy jest użycie technik eksploracji danych. Możemy wyróżnić dwa podejścia. Pierwszym z nich to podejście regresyjne, które stara się przewidzieć dokładną wartość prognozowanego instrumentu. Wykorzystywane są do tego celu głównie sieci neuronowe [Zhang 1998]. Drugim jest podejście klasyfikacyjne, które stara się przewidzieć jedynie kierunek zmiany (wzrost bądź spadek) [Kim 2003].

Znaczącą grupę klasyfikatorów stanowią klasyfikatory liniowe [Duda i in. 2001]. Jednym z nich jest maszyna wektorów wspierających (SVM) [Vapnik 1995][Cortes i in. 1995] intensywnie rozwijana w ostatnich latach. Jej zastosowanie dla danych giełdowych odnajdziemy w pracy [Kim 2003] oraz [Huang i in. 2005].

Proponowany w tej pracy klasyfikator jest również klasyfikatorem liniowym, opartym o wypukłą i odcinkow-liniową funkcję kryterialną (ang. convex and piece-linear CPL) [Bobrowski 2005][Krawczuk i in. 2010]

KLASYFIKATOR LINIOWY

Klasyfikator liniowy $LC(\mathbf{w}[n_k],\theta)$ może zostać zdefiniowany jako następująca reguła decyzyjna [Bobrowski 2005]:

$$\begin{aligned} &\textbf{\textit{jeżeli}}\ \ \mathbf{w}[n]^T\mathbf{x}[n] \geq \theta,\ \ \textbf{\textit{to}}\ \mathbf{x}[n] \text{ przypisujemy do klasy } \omega^+ \\ &\textbf{\textit{jeżeli}}\ \ \mathbf{w}[n]^T\mathbf{x}[n] < \theta,\ \ \textbf{\textit{to}}\ \mathbf{x}[n] \text{ przypisujemy do klasy } \omega^- \end{aligned} \tag{1}$$

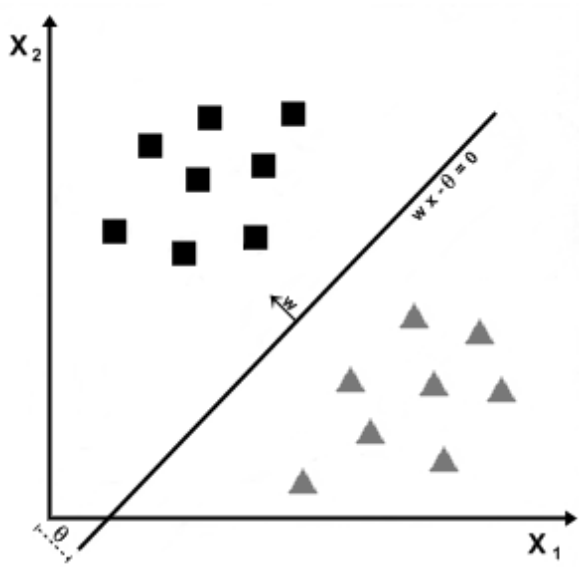
gdzie $\mathbf{w}[n] = [w_1,...,w_n]^T$ jest wektorem wag ($w_i \in R^1$) a θ jest progiem ($\theta \in R^1$). Graficzną interpretację tej reguły w przestrzeni 2-wymiarowej przedstawia rysunek 1.

Utworzenie reguły decyzyjnej (prognostycznej) (1) wymaga wyznaczenia wartości parametrów $\mathbf{w}[n]$ oraz θ. Parametry te mogą zostać obliczone na podstawie dwóch zbiorów uczących. Zbioru G^+ zawierającego przykładowe obiekty $\mathbf{x}_j[n]$ z klasy ω^+ ($j \in J^+$) oraz zbioru G^- zawierającego obiekty z klasy ω^+ ($j \in J^+$).

$$G^+ = \{\mathbf{x}_j[n]: j \in J^+\} \textit{ oraz } G^- = \{\mathbf{x}_j[n]: j \in J^-\} \tag{2}$$

W przypadku klasyfikacji danych giełdowych, zbiór G^+ zawiera takie stany giełdy $\mathbf{x}_j[n]$, po których nastąpił wzrost prognozowanego instrumentu, natomiast zbiór G^- stany po których nastąpił jego spadek. Wybór sposobu opisu stanu giełdy, czyli wybór wektora cech $\mathbf{x}_j[n]$, może być dokonany na wiele sposobów. W tej pracy użyto 53 cech, z czego 11 z nich opartych jest o historyczne notowania prognozowanego indeksu S&P500, a pozostałe 42 to notowania innych instrumentów. Szczegółowy opis wybranych cech znajduje się w kolejnych rozdziałach.

Rysunek 1. Hiperpłaszczyzna separująca dwa zbiory w przestrzeni 2-wymiarowej



Źródło: opracowanie własne

FUNKCJA KRYTERIALNA UWZGLĘDNIAJĄCA KOSZTY BŁĘDNYCH KLASYFIKACJI

Klasyfikator liniowy typu CPL [Bobrowski 2005] wyznacza optymalne wartości parametrów hiperpłaszczyzny $\mathbf{w}[n]$, θ minimalizując następującą funkcję kryterialną (funkcję kary):

$$\Phi(\mathbf{w}[n], \theta) = \sum_{j \in J+} \alpha_j \varphi_j^+(\mathbf{w}[n], \theta) + \sum_{j \in J-} \alpha_j \varphi_j^-(\mathbf{w}[n], \theta) \quad (3)$$

gdzie:
$(\forall \mathbf{x}_j[n] \in G^+)$:

$$\varphi_j^+(\mathbf{w}[n],\theta) = \begin{cases} \theta + 1 - \mathbf{w}[n]^T \mathbf{x}_j[n] & \textit{if} \quad \mathbf{w}[n]^T \mathbf{x}_j[n] < \theta + 1 \\ 0 & \textit{if} \quad \mathbf{w}[n]^T \mathbf{x}_j[n] \geq \theta + 1 \end{cases} \tag{4}$$

oraz

$(\forall \mathbf{x}_j[n] \in G^-)$

$$\varphi_j^-(\mathbf{w}[n], \theta) = \begin{cases} \theta - 1 + \mathbf{w}[n]^T \mathbf{x}_j[n] & \textit{if} \quad \mathbf{w}[n]^T \mathbf{x}_j[n] > \theta - 1 \\ 0 & \textit{if} \quad \mathbf{w}[n]^T \mathbf{x}_j[n] \leq \theta - 1 \end{cases} \tag{5}$$

Nieujemne parametry α_j reprezentują koszty związane z każdym z obiektów. Często przyjmuje się wartość parametrów α_j równą 1 dla wszystkich obiektów. Pomija się w ten sposób indywidualne koszty błędnej klasyfikacji każdego obiektu, przyjmując ich równowartość. Gdy α_j=1 klasyfikator maksymalizuje ilość poprawnie klasyfikowanych obiektów. Gdy koszty α_j są różne dla różnych obiektów, wówczas ilość poprawnie sklasyfikowanych obiektów może zmaleć. Jednak poprawnie sklasyfikowane obiekty będą miały większe koszty.

W przypadku danych giełdowych w sposób naturalny możemy przypisać koszt błędnej klasyfikacji jako wartość zmiany prognozowanego instrumentu. Jako przykład rozważmy wzrost o 0,1% i wzrost o 2,5%. Koszt błędnej klasyfikacji obiektu reprezentującego 0,1% wzrost powinien być znacząco mniejszy od kosztu błędnej klasyfikacji obiektu reprezentującego wzrost o 2,5%. W pracy tej przyjęto dokładnie wartości 0,1 i 2,5 jako koszty α_j. Również w przypadku spadków kosztem jest wartość spadku (liczba dodatnia). Zatem ze wzrostem o np. 2,0% i spadkiem o 2,0% związany jest ten sam koszt α_j=2,0.

EKSPERYMENT

Dane

W przeprowadzonym eksperymencie obliczeniowym prognozowana była zmiana głównego indeksu giełdy amerykańskiej S&P500, reprezentowany przez instrument (ETF) o symbol SPY. Starano się przewidzieć wzrost bądź spadek tego indeksu na jeden dzień do przodu bazując na opisie stanu giełdy za pomocą następujących 53 atrybutów:

11 atrybutów opisujących indeks S&P500

1. Cena otwarcia
2. Wczorajsza cena zamknięcia
3. Luka, zmiana zamknięcie - otwarcie
4. Zmiana jednodniowa

5. Zmiana dwudniowa
6. Zmiana pięciodniowa
7. Jednodniowa zmiana z wczoraj
8. Jednodniowa zmiana z przedwczoraj
9. 9-dniowa średnia
10. 12-dniowa średnia
11. 26-dniowa średnia

Atrybut 12 to wartość indeksu zmienności ^VIX. Pozostałe 41 atrybutów to procentowe zmiany wartości innych 41 instrumentów finansowych notowanych na giełdzie amerykańskiej, opisujących sytuację na giełdach w innych krajach, na giełdach towarowych, metali szlachetnych jak również kursów walut. Zestaw wszystkich użytych instrumentów przedstawiono w tabeli 1.

Tabela 1. Użyte w eksperymencie instrumenty finansowe i ich symbole na giełdzie amerykańskiej

Grupa	Sym.	Opis	Grupa	Sym.	Opis
EUROPA	EWU	Wielka Brytania	AMERYKA	SPY	USA
	EWL	Szwajcaria		EWC	Kanada
	EWI	Włochy		^VIX	Indeks zmienności
	EWD	Szwecja		EWW	Meksyk
	EWG	Niemcy		EWZ	Brazylia
	EWP	Hiszpania		GML	Kraje wschodzące
	EWQ	Francja	AFRYKA	EZA	RPA
	EWO	Austria		GAF	Bliski wschód
	EWK	Belgia	TOWARY	USO	Ropa naftowa
	EWN	Holandia		UNG	Gaz ziemny
	RSX	Rosja		DBA	Rolnictwo
	GUR	Wschodząca Europa		GLD	Złoto
AZJA	FEU	stoxx50		SLV	Srebro
	EWJ	Japonia	WALUTY	FXA	Dolar Australijski
	EWM	Malezja		FXB	Funt Brytyjski
	EWT	Tajwan		FXC	Dolar Kanadyjski
	EWY	Korea Płd.		FXE	Euro
	EWH	Hong Kong		FXY	Jen
	FXI	Chiny		UDN	Spadek $ USA
	EWA	Australia		UUP	Wzrost $ USA
	EWS	Singapur		DBV	Koszyk walut G10

Źródło: opracowanie własne

Zakres danych użytych w obliczeniach zawiera się w przedziale od listopada 2007 do maja 2011.

Wyniki

Do zbudowania klasyfikatora (modelu prognostycznego) użyto danych z okresu 1 roku (252 dni roboczych). Następnie klasyfikatora tego użyto do prognozowania indeksu S&P500 w okresie kolejnych 6 miesięcy (126 dni roboczych). Po upływie pół roku budowany był nowy model na podstawie nowych danych. Dla posiadanego okresu danych, zostało zbudowanych 5 takich modeli. Szczegóły przedstawia rysunek 2.

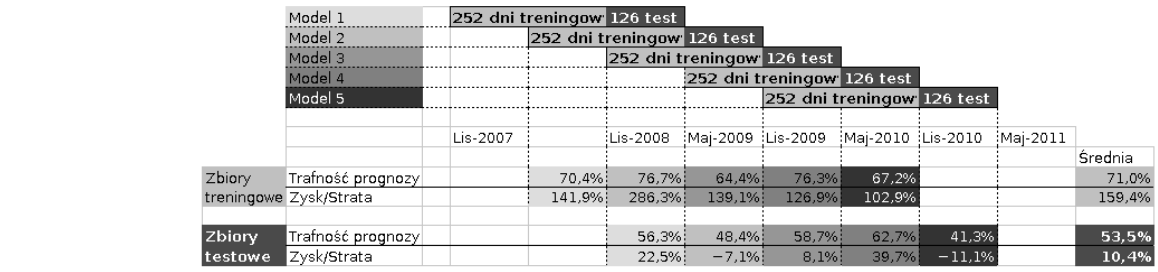
Przykładowo model zbudowany z danych z okresu listopad 2007 – listopad 2008, poprawnie klasyfikuje 70,4% tych danych, jednak użyty w przyszłości na nowych danych od listopada 2008 do maja 2009 sklasyfikował poprawnie jedynie 56,3% obiektów.

Dodatkową miarą oprócz ilości poprawnie sklasyfikowanych wzrostów/spadków jest miara potencjalnego zysku/straty modelu. Zysk/strata zostały obliczone jako suma prawidłowo przewidzianych procentowych zmian, minus suma błędnie przewidzianych zmian.

Przykładowo w pierwszym okresie testowym przy 56,3% prawidłowo przewidzianych wzrostów/spadków potencjalny zysk wyniósłby 22,5%. W kolejnym okresie testowym byłaby to już strata w wysokości 7,1%. Wartość tej miary jest bardzo wysoka w okresie treningowym i przykładowo w pierwszym okresie wynosi aż 141,9%.

Dla tak zbudowanych 5 modeli wyliczono wartości średnie obu miar. Dla zbioru testowego prawidłowo przewidziano kierunek zmiany w 53,5% przypadków, co średnio odpowiada zyskowi w wysokości 10,4% Wartości średnie dotyczą okresu półrocznego.

Rysunek 2. Wyniki eksperymentu dla zbiorów uczących i testowych



Źródło: obliczenia własne

Przedstawione na rysunku 2 szczegółowe wyliczenia zostały wykonane dla standardowego klasyfikatora, który nie uwzględniał kosztów błędnej klasyfikacji

(wartości współczynników α_j=1). Identyczne obliczenia wykonane zostały również z uwzględnieniem kosztów błędnej klasyfikacji. Przyjęto wartości współczynników α_j równe bezwzględnej wartości prognozowanej zmiany. Średnia jakość klasyfikacji na zbiorze treningowym spadła z 70,99% do 69,41%, jednak wartość potencjalnego zysku wzrosła z 159,43% do 199,21%. Zatem klasyfikator zgodnie z oczekiwaniami nauczył się przewidywać większe ruchy, kosztem ilości poprawnych decyzji. Niestety efekt wzrostu zysku zaobserwowano jedynie na zbiorach treningowych, dla zbiorów testowych nastąpił zarówno spadek jakości klasyfikacji z 53,49% do 50,00%, jak również wartości zysku z 10,40% do 8,77%.

Analogiczne obliczenia wykonano również dla danych przekształconych. Wykonano dwa przekształcenia:

- standaryzacja : każdą cechę ustandaryzowano do średniej 0 i wariancji 1
- przeskalowanie: każdą cechę przeskalowano do przedziału <-1,1>

Wyniki wszystkich obliczeń zestawiono w tabeli 2 i 3. Wyniki dla danych standaryzowanych są podobne do wyników dla danych nie przekształconych, przy czym osiągnięta jakość klasyfikacji i zysk jest nieco niższy.

Tabela 2. Wyniki dla zbiorów treningowych

	Jakość Klasyfikacji		Zysk/Strata	
Z kosztami	NIE	TAK	NIE	TAK
Nie przekształcone	70,99%	69,41%	159,43%	199,21%
Standaryzowane	71,70%	70,99%	154,08%	214,47%
Przeskalowane	72,73%	69,49%	172,13%	202,25%

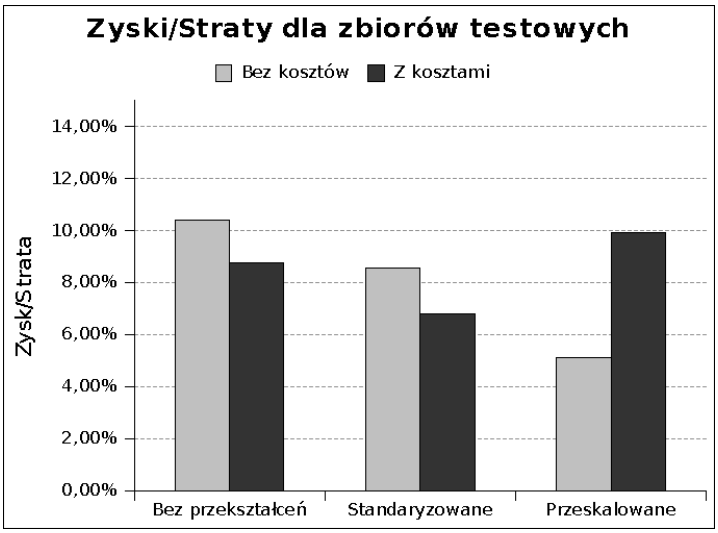
Źródło: opracowanie własne

Tabela 3. Wyniki dla zbiorów testowych

	Jakość Klasyfikacji		Zysk/Strata	
Z kosztami	NIE	TAK	NIE	TAK
Nie przekształcone	53,49%	50,00%	10,40%	8,77%
Standaryzowane	52,06%	49,84%	8,56%	6,80%
Przeskalowane	51,11%	50,32%	5,11%	9,94%

Źródło: opracowanie własne

Rysunek 3. Jakość modeli zmierzona na zbiorach testowych za pomocą miary zysków i strat



Źródło: obliczenia własne

PODSUMOWANIE

Klasyfikator liniowy typu CPL może zostać zbudowany tak, aby maksymalizować potencjalny zysk z dobrych decyzji, a nie jedynie ich ilość. Potwierdzają to wyniki uzyskane na zbiorach uczących. Taka cecha klasyfikatora wydaje się być bardzo interesująca do zastosowań giełdowych. Jednak w przeprowadzonym eksperymencie zyski na zbiorach testowych były mniejsze (rysunek 3). Standaryzacja danych wejściowych nie zmieniła tego zachowania. Natomiast liniowe przeskalowanie danych do przedziały <-1,1> zaowocowało wzrostem zysku. Wzrost jest znaczący, jednak osiągnięty wynik nie jest lepszy od tego dla danych nie przekształconych i klasyfikatora nie uwzględniającego kosztów.

BIBLIOGRAFIA

Bobrowski L. (2005) Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych, Wydawnictwa Politechniki Białostockiej.

Bobrowski L., Łukaszuk T. (2009) Feature selection based on linear separability, Biocybernetics and Biomedical Engineering, Volume 29, Number 2, str. 43-59.

Bollerslev T. (1986) Generalized Autoregressive Conditional Heteroskedasticity, Journal of Econometrics 31, 307-327.

Box G.E.P, Jenkins G.M. (1983) Analiza szeregów czasowych, Państwowe Wydawnictwo Naukowe.

Cortes C. Vapnik V. (1995) Support-Vector Networks, Machine Learning 20

Duda O.R. Hart P.E., Stork D.G. (2001) Pattern Classification, J. Wiley, New York.

Edwards R.D. Magee J. (1997) Technical Analysis of Stock Trends, 7th edition, Amacom

Engle R.F (1982) Autoregressive Conditional Heteroskedasticity with the Estimates of the Variance of U.K. Inflation, Econometrica 50, No. 4, 987-1007.

Hamiliton J.D. (1994) Time Series Analysis, Princeton University Press.

Huang W. Nakamori Y. Wang, S.Y. (2005) Forecasting Stock Market Movement Direction with Support Vector Machine, Computers & Operations Research 32 str. 2513-2522.

Kim KJ. (2003) Financial time series forecasting using support vector machines, Neurocomputing Volume 55, Issues 1-2, str. 307-319.

Krawczuk J., Bobrowski L. (2010) Short term prediction of stock indexes changes based on a linear classifier, Symulacja w badaniach i rozwoju, Vol.1 nr 4/2010.

Nison S. (2001) Japanese Candlestick Charting Techniques, Second Edition Prentice Hall Press.

Vapnik, V.N (1995) The Nature of Statistical Learning Theory. New York, Springer.

Zhang, G., Patuwo, B.E., Hu, M.Y. (1998) Forecasting with Artificial Neural Networks: the State of the Art, International Journal of Forecasting 14 str. 35-62.

COST-SENSITIVE CPL LINEAR CLASSIFIER AS A MARKET PREDICTION TOOL

Abstract: One kind of data mining – classification – can be used for purpose of predicting changes in market prices. In the simplest scenario we can classify every daily market move as one of two classes: increases or decreases. The standard approach to building a classifier is to optimize correctly classified instances (market moves). However, in the case of predicting the stock market, a better measure of model quality could be a potential profit. This article describes such an approach (cost-sensitive classification) for a linear classifier based on a convex and piecewise-linear penalty function (CPL).

Key words: linear classifier, market prediction, CPL function