Zeszyty Naukowe WWSI, No 14, Vol. 10, 2016, s. 37-48

Identyfikacja wzorców w finansowych szeregach czasowych z wykorzystaniem hierarchicznych metod grupowania na przykładzie kursu BTC/PLN

Kinga Kądziołka*

Prokuratura Okręgowa w Katowicach

Streszczenie

W artykule przedstawiono zastosowanie metody Warda do identyfikacji wzorców w finansowych szeregach czasowych, na przykładzie kursu waluty kryptograficznej bitcoin. Wykorzystując zidentyfikowane wzorce, generowano prognozy zmian kursu w analizowanym szeregu dla danych zbioru testowego, które nie zostały wykorzystane w procesie identyfikacji wzorców. Przeciętny absolutny oraz maksymalny błąd prognozy na danych zbioru testowego był niewielki, natomiast zgodność kierunku zmian kursu BTC/PLN na zbiorze testowym wynosiła tylko 60%.

Słowa kluczowe **–** bitcoin, grupowanie szeregów czasowych, rozpoznawanie wzorców

1 Wprowadzenie

Metody grupowania danych były wykorzystywane do identyfikacji wzorców w finansowych szeregach czasowych na rynku walutowym i na rynku akcji. Stosowano

* E-mail: kinga_kadziolka@onet.pl

Zgłoszono do druku 2 lutego 2016 r.

m.in. metodę k-średnich, algorytmy hierarchicznego grupowania danych, samoorganizujące sieci Kohonena, algorytmy genetyczne, metodę najbliższych sąsiadów [1, 2, 3, 4].

W tym artykule przedmiotem analizy będzie szereg czasowy indeksów łańcuchowych dla cen kryptowaluty bitcoin. Elektroniczny system płatności kryptowalutą bitcoin stwarza możliwości inwestycyjne. W Internecie funkcjonują liczne giełdy umożliwiające handel kryptowalutą. Podejmowane są próby prognozowania kursu bitcoina, implementacji automatycznych systemów transakcyjnych i opracowania strategii inwestycyjnych wykorzystujących rozmaite metody data mining [5, 6, 7, 8].

W niniejszym artykule opisano próbę identyfikacji wzorców w kształtowaniu się kursu kryptowaluty. Analizowane będzie podobieństwo w tygodniowych zmianach kursu BTC/PLN na giełdzie BitMarket.pl.

Do identyfikacji wzorców w analizowanym szeregu wykorzystany zostanie algorytm hierarchicznego grupowania danych metodą Warda. Metodę tą szczegółowo opisuje m.in. Andrzej Stanisz [9]. Wzorce zostaną wygenerowane na podstawie tzw. zbioru treningowego, natomiast ocena grupowania oraz miary jakości prognoz wyznaczone zostaną na podstawie tzw. zbioru testowego, stanowiącego ostatnie obserwacje analizowanego szeregu czasowego (obejmujące okres 5-ciu tygodni), które nie zostały wykorzystane w procesie generowania wzorców. W oparciu o zidentyfikowane wzorce podjęta zostanie próba prognozowania kierunku zmian kursu bitcoina. Prezentowane wyniki uzyskano z wykorzystaniem darmowych programów Gretl i R. Dane dotyczące kursu kryptowaluty bitcoin na wybranych giełdach są ogólnodostępne w Internecie [10].

2 Hierarchiczne metody grupowania danych

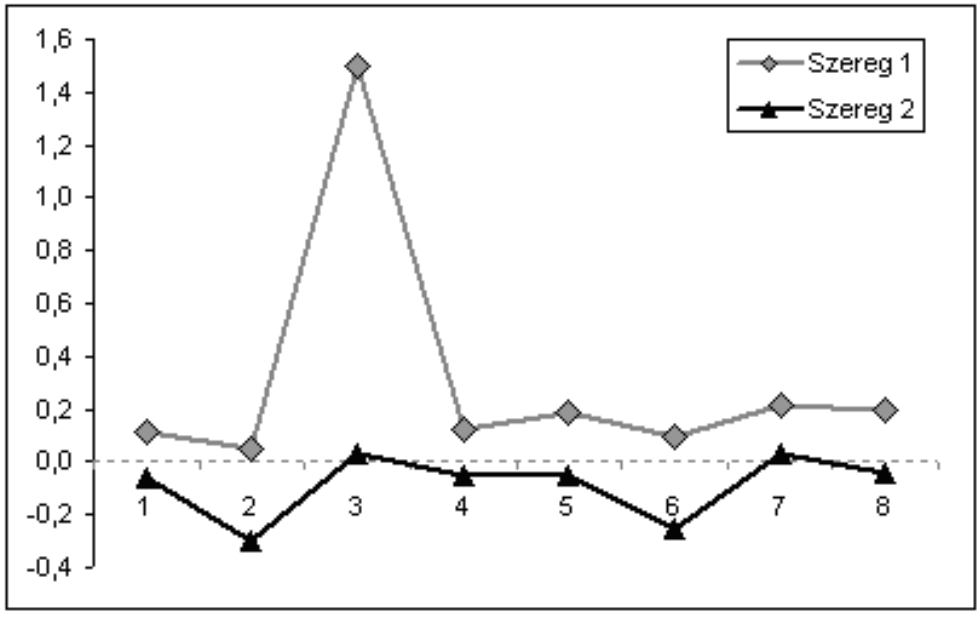
Rezultatem działania algorytmów grupowania hierarchicznego jest drzewo hierarchicznie ułożonych skupień, tzw. dendrogram. W metodzie Warda wykorzystuje się analizę wariancji do szacowania odległości między skupieniami. Na każdym etapie tworzenia dendrogramu, spośród wszystkich możliwych do łączenia par skupień wybiera się tę, która w rezultacie łączenia da skupienie o najmniejszym zróżnicowaniu. Wpływ na uzyskiwany dendrogram ma sposób zdefiniowania miary odległości między obiektami. Tutaj grupowane będą szeregi czasowe.

W literaturze wskazywane są różne propozycje zdefiniowania odległości między szeregami czasowymi, m.in. odległość euklidesowa, DTW (ang. *dynamic time warping*) czy miary oparte na współczynniku korelacji liniowej lub kolejnościowej między porównywanymi szeregami [1, 4, 11, 12, 13, 14, 15].

W tym artykule główna uwaga zostanie skoncentrowana na próbach prognozowania kierunku zmian kursu BTC/PLN, dlatego wykorzystana zostanie miara odległości między szeregami czasowymi oparta na współczynniku korelacji kolejnościowej Spearmana. Odległość ta zdefiniowana jest jako [11]: $m(X,Y)=1-r_s(X,Y)$, gdzie r_s oznacza wartość współczynnika korelacji kolejnościowej dla wartości szeregów X i Y. Wyznaczanie współczynnika korelacji kolejnościowej między dwoma cechami rozpoczyna się od tego, że poszczególnym wariantom obu cech (tutaj poszczególnym elementom analizowanych szeregów) nadaje się rangi, czyli kolejne numery od 1 do n (tutaj n jest liczbą obserwacji analizowanych szeregów), które pozwalają uporządkować ciąg obserwacji (rosnąco lub malejąco). Współczynnik korelacji kolejnościowej wyznaczany jest zgodnie ze wzorem [16, s. 294]:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}, \tag{1}$$

gdzie: n – liczba obserwacji, d_i – różnica między rangami, które są przypisane i-tej obserwacji pierwszej i drugiej cechy.

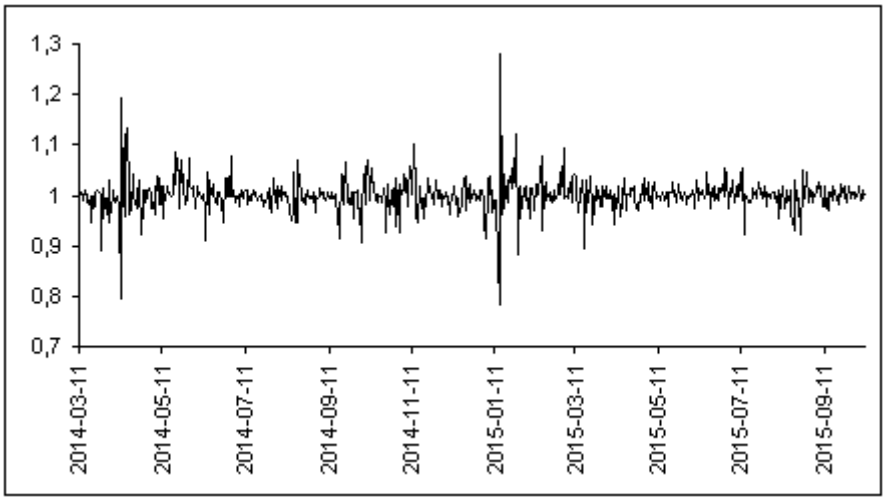


Rysunek 1. Przykładowe szeregi czasowe

Na możliwość wykorzystania współczynnika korelacji kolejnościowej do grupowania podszeregów wyróżnionych z szeregów czasowych zwrócono uwagę na przykład w pracach [4, 11]. W odróżnieniu od współczynnika korelacji liniowej, współczynnik korelacji kolejnościowej mierzy również monotoniczne zależności nieliniowe. Przykładowo, w przypadku szeregów przedstawionych na rys. 1 wartość współczynnika korelacji kolejnościowej dla wartości tych szeregów wynosi 0,994. Zależność ta jest istotna statystycznie na poziomie istotności 1%. Odległość między analizowanymi szeregami wynosi $m(X,Y)$=1-0,994=0,006. Z kolei współczynnik korelacji liniowej dla wartości analizowanych szeregów wynosi 0,4749 i nie różni się statystycznie od zera na poziomie istotności 10%.

3 Identyfikacja wzorców na przykładzie kursu BTC/PLN na giełdzie BitMarket.pl

Przedmiotem analiz jest szereg indeksów łańcuchowych wyznaczonych na podstawie cen kryptowaluty bitcoin z zamknięcia sesji na giełdzie BitMarket.pl w okresie 11 marca 2014 r. – 11 października 2015 r. (rysunek 2).



Rysunek 2. Szereg indeksów łańcuchowych dla cen kryptowaluty bitcoin

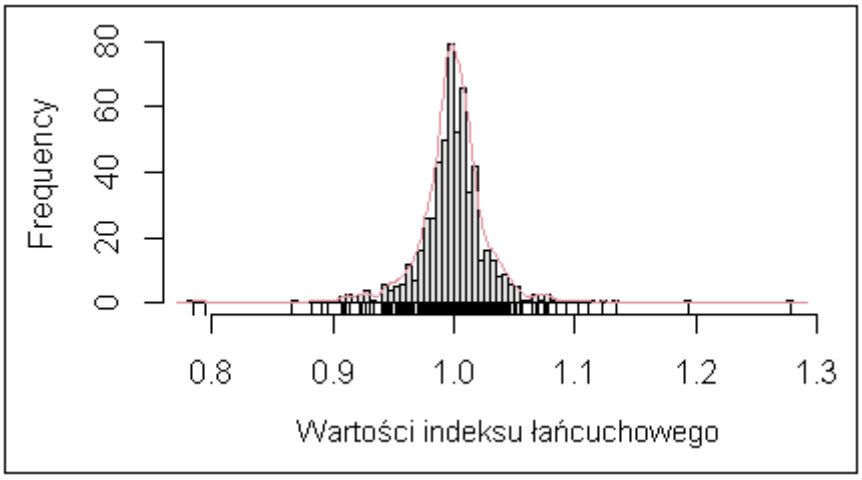
Indeksy łańcuchowe określone były następująco:

$$y(t)=p(t)/p(t-1), \tag{2}$$

gdzie p(t) – cena kryptowaluty bitcoin na giełdzie BitMarket.pl z zamknięcia sesji w chwili t.

Analizowany szereg czasowy indeksów łańcuchowych był stacjonarny (na co wskazywały wyniki testów KPSS i ADF). Wartość wykładnika Hursta dla analizowanego szeregu wynosiła 0,5367 i przekraczała poziom graniczny 0,5, co sugerowało, że w analizowanym szeregu zachodzi długoterminowa zależność danych i istnieje możliwość przewidywania przyszłych kierunków zmian kursu BTC/PLN na podstawie wcześniejszych wartości.

Rys. 3 przedstawia histogram uzyskany na podstawie wartości analizowanego szeregu.

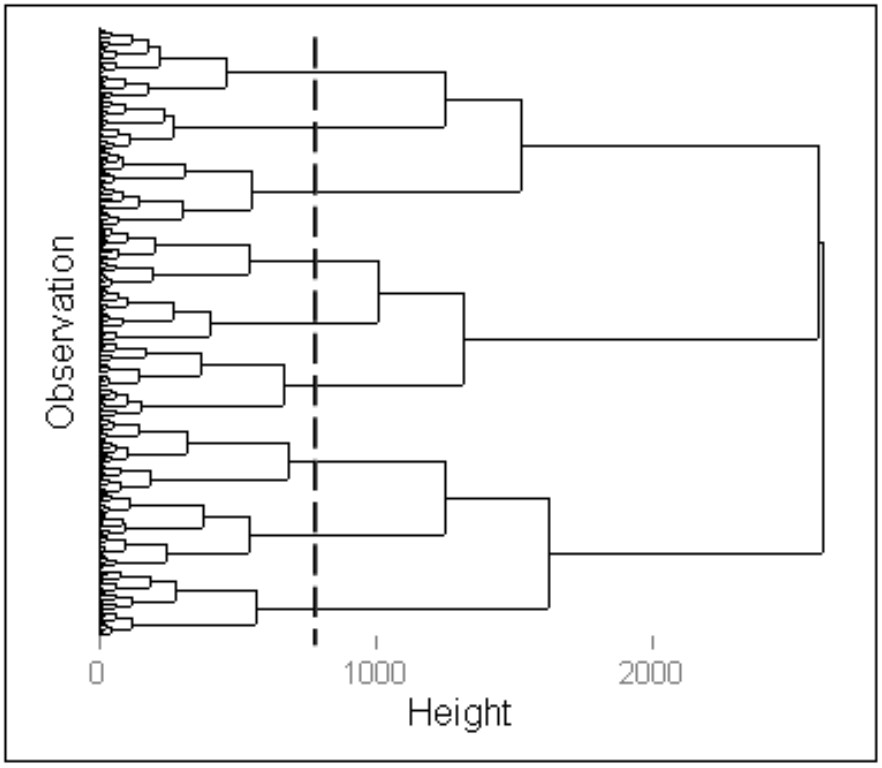


Rysunek 3. Histogram dla wartości indeksu łańcuchowego kursu BTC/PLN

Analizowany zbiór danych został podzielony na dwa podzbiory. Dane za okres 11 marca 2014 r. – 6 września 2015 r. stanowiły tzw. zbiór treningowy, na podstawie którego poszukiwane były wzorce w analizowanym szeregu, natomiast dane za okres 7 września 2015 r. – 11 października 2015 r. stanowiły tzw. zbiór testowy. Dane zbioru testowego, które nie były wykorzystane w procesie generowania wzorców, posłużyły do oceny dokładności prognoz uzyskiwanych w oparciu o zidentyfikowane schematy w szeregu. Dokonano identyfikacji wzorców na podstawie wektorów generowanych przez 7 kolejnych wartości indeksów łańcuchowych wyznaczonych dla cen kryptowaluty bitcoin z zamknięcia sesji.

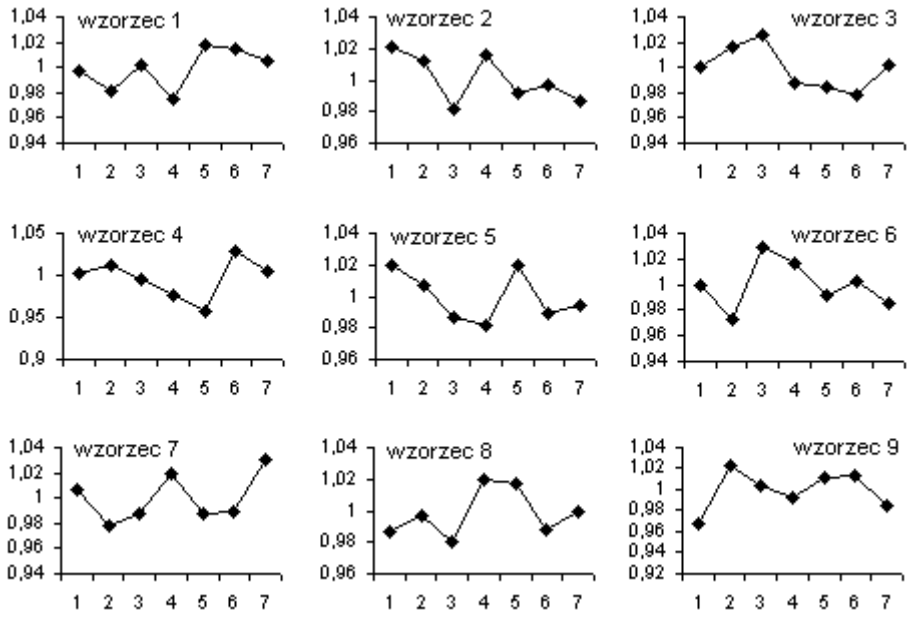
W zaprezentowanym przykładzie liczbę uwzględnionych opóźnień przyjęto arbitralnie. Zbiór treningowy stanowiło 539 obiektów 7-wymiarowych postaci: (y(t-6),…, y(t-1), y(t)), gdzie y(t) – wartość indeksu łańcuchowego dla cen bitcoina z zamknięcia sesji w chwili t). Tak zdefiniowany zbiór obiektów poddano grupowaniu metodą Warda, przyjmując miarę odległości zdefiniowaną jako 1- r_s, gdzie r_s – współczynnik korelacji kolejnościowej Spearmana.

Na podstawie uzyskanego dendrogramu dokonano podziału obiektów zbioru treningowego na 9 grup. Przerywaną linią zaznaczono przyjęte (arbitralnie) miejsce podziału dendrogramu (rys. 4).



Rysunek 4. Dendrogram uzyskany dla zbioru treningowego.

Środki poszczególnych grup były zidentyfikowanymi wzorcami. Uzyskane w ten sposób wzorce przedstawia rys. 5.



Rysunek 5. Wzorce dynamiki indeksów łańcuchowych dla cen bitcoin

4 Prognozowanie przyszłych zmian kursu BTC/PLN na podstawie zidentyfikowanych wzorców

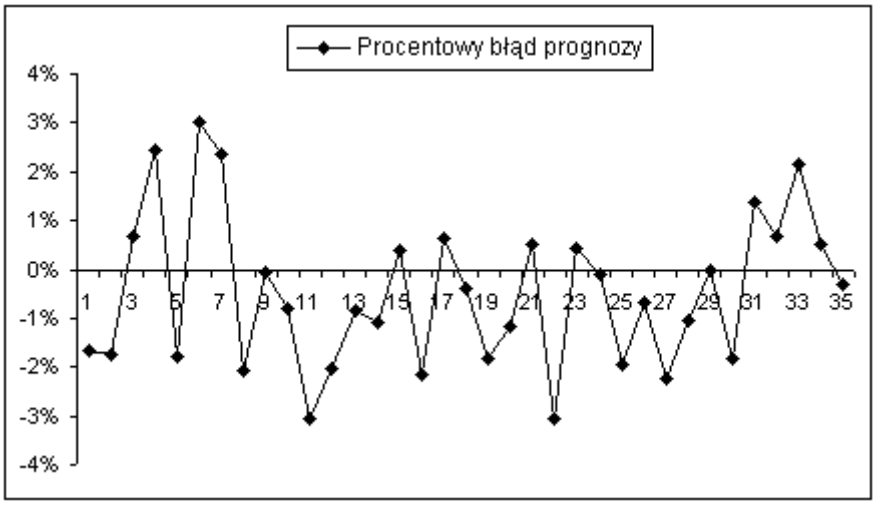
Dla każdego elementu zbioru testowego wyznaczono prognozę indeksu łańcuchowego cen kryptowaluty bitcoin z zamknięcia sesji w oparciu o wartości tego indeksu z 6-ciu poprzednich sesji. W tym celu dla każdej obserwacji zbioru testowego wyznaczono najbliższy (według przyjętej miary odległości) jej wzorzec w 6-wymiarowej przestrzeni, tzn. uwzględniając współrzędne (y(t-6), …, y(t-1)). Jako poszukiwaną prognozę indeksu łańcuchowego w chwili t przyjęto współrzędną y*(t) wzorca, który był najbliższy danej obserwacji.

Tabela 1 przedstawia odległości poszczególnych elementów zbioru testowego od wzorców w sensie przyjętej miary odległości w przestrzeni 6-wymiarowej. Najbliższy danej obserwacji będzie ten wzorzec, dla którego wartość 1- r_s będzie najbliższą zeru liczbą, czyli, im współczynnik korelacji kolejnościowej Spearmana będzie bliższy 1. Jeśli dla większej niż 1 liczby wzorców miara odległości między daną obserwacją, a tymi wzorcami przyjmowała minimalną wartość, wówczas jako prognozę przyjęto średnią arytmetyczną wartości ostatnich współrzędnych tych wzorców.

Tabela 1. Odległości obserwacji zbioru testowego od wzorców

wz1	wz2	wz3	wz4	wz5	wz6	wz7	wz8	wz9	prognoza
0,5143	1,3714	1,7143	0,8000	1,0857	1,3143	1,3714	0,5714	0,2286	0,9840
0,8000	0,9714	1,3143	1,8857	0,9143	0,8571	0,6286	0,5714	1,6000	0,9994
1,3143	1,1429	1,2000	0,9143	1,9429	0,4000	0,6286	0,6857	0,8571	0,9858
0,7429	1,6000	0,3429	1,4857	0,8571	1,0286	1,6000	1,3143	0,9714	1,0013
1,8857	0,4571	0,4000	1,0286	1,0857	1,2571	1,0286	0,7429	1,0857	1,0013
1,0286	0,6857	0,8571	0,2857	0,7429	0,8000	0,6857	1,7714	1,2571	1,0052
0,8000	1,0857	0,7429	1,4286	0,3429	1,7143	1,6571	0,8571	0,8571	0,9946
1,8857	0,2857	0,5714	1,2571	1,1429	0,9143	0,5143	0,7429	1,6000	0,9861
0,9143	1,2571	0,6286	0,4000	1,2000	0,4571	0,9143	1,8857	1,0286	1,0052
0,6286	1,1429	1,1429	0,6857	0,4000	1,8857	1,8286	0,9143	0,2286	0,9840
1,2000	0,2286	1,4286	1,2571	0,6286	1,1429	0,3429	0,6286	1,6571	0,9861
0,6286	1,4857	1,3714	0,9143	1,6000	0,2286	0,6286	1,2000	1,0286	0,9858
0,4571	1,6000	1,2571	1,0857	0,8571	1,5429	1,8286	0,6857	0,1714	0,9840
1,4857	0,2286	1,2571	1,4286	0,8571	1,2000	0,4571	0,3429	1,5429	0,9994
1,2000	1,0857	0,6857	1,2571	1,5429	0,0571	0,3429	1,3714	1,7143	0,9858
1,2000	1,3714	0,2857	1,2571	1,0857	1,2571	1,7143	0,9714	0,6286	1,0013
1,2000	0,4571	1,2571	0,1143	0,8571	1,0857	0,6857	1,2571	0,9143	1,0052
0,1714	1,2000	1,3714	1,1429	0,2286	1,3143	1,2000	1,2571	1,0857	1,0043
1,1429	0,7429	1,7714	1,0857	1,3714	1,0286	0,6857	0,1714	0,8571	0,9994
0,4000	1,7714	1,1429	1,1429	1,4286	0,2857	0,9143	1,3714	1,0286	0,9858
1,1429	1,2571	0,8571	1,3143	1,1429	1,4857	1,6571	0,4000	0,4000	0,9917
1,6571	0,4000	0,4571	1,2571	0,9714	0,7429	0,4571	1,1429	1,8286	0,9861
0,9714	1,3714	0,5714	0,3429	1,3143	0,5714	1,1429	1,7714	0,7429	1,0052
0,7429	0,9714	0,9714	0,8000	0,2286	1,9429	1,7714	0,9714	0,4571	0,9946
1,3714	0,1714	1,2000	1,3143	0,5714	1,3143	0,5143	0,5714	1,6000	0,9861
0,8571	1,3714	0,9714	0,9143	1,6000	0,0571	0,5143	1,4857	1,3143	0,9858
0,9714	1,4857	0,6857	1,2571	1,0857	1,4286	1,8286	0,6857	0,3429	0,9840
1,3714	0,2286	1,2000	0,4000	0,8571	0,9143	0,3429	1,2000	1,3714	0,9861
0,1714	1,6000	0,9714	1,3714	0,5143	1,0286	1,3714	1,4286	1,0857	1,0043
1,5429	0,5714	1,3714	0,9143	1,3714	1,2571	0,8571	0,1714	0,6857	0,9994
0,4000	0,5714	1,2571	0,6857	1,0286	0,4571	0,7429	1,6571	1,1429	1,0043
1,0286	1,2000	0,8571	1,6571	0,9143	1,4857	1,5429	0,4000	0,7429	0,9994
1,8286	0,4000	0,5714	1,0857	1,3714	0,5143	0,2857	1,0286	1,7143	1,0304
0,6857	1,5429	0,6286	0,4000	1,0857	0,7429	1,3714	1,8286	0,6286	1,0052
0,9143	0,7429	1,1429	0,8571	0,2857	2,0000	1,6000	0,6857	0,5143	0,9946

Na przecięciu *i*-tego wiersza i *j*-tej kolumny podana jest odległość *i*-tej obserwacji od *j*-tego wzorca. Kolorem szarym wyróżniono najmniejszą z odległości dla poszczególnych obserwacji. Dodatkowo w tabeli znajduje się kolumna o nazwie prognoza zawierająca wygenerowaną prognozę y*(t).Przeciętny absolutny procentowy błąd prognozy indeksów łańcuchowych dla cen bitcoina na zbiorze testowym wynosił 1,34% a maksymalny błąd procentowy wynosił 3,06% (rys. 6).



Rysunek 6. Procentowy błąd prognozy na zbiorze testowym

Niestety niskim błędom prognoz nie towarzyszyła wysoka zgodność kierunku zmian kursu. Dla danych zbioru testowego zgodność kierunku zmian wynosiła 60%. Opracowanie metody, która pozwalałaby z dużą dokładnością prognozować kierunek zmian kursu kryptowaluty bitcoin mogłoby zwiększyć szanse osiągnięcia zysku w przypadku inwestycji z wykorzystaniem opcji binarnych. Opcje binarne umożliwiają inwestowanie w wirtualne monety bez konieczności ich posiadania. Inwestor ma za zadanie przewidzieć czy kurs bitcoina wzrośnie czy spadnie, czyli czy w momencie zakończenia opcji cena kryptowaluty będzie wyższa czy niższa od ceny początkowej [17].

Dokonano również oceny dokładności przyporządkowania obserwacji zbioru testowego do wzorców za pomocą miernika określonego wzorem [4]:

$$Acc = \frac{\sum_{i=1}^{n} u_i}{n} \tag{3}$$

gdzie $u_i = \begin{cases} 1 & dla \quad m(y_i, w_i) \le p \\ 0 & dla \quad m(y_i, w_i) > p \end{cases}$,

n – liczba elementów (tutaj liczba obserwacji zbioru testowego),
y_i – i-ty element zbioru testowego,
w_i – wzorzec najbliższy i-temu obiektowi,
m – miara odległości między obiektami,
p – ustalona wartość.

Miernik *Acc* przyjmuje wartość z przedziału [0,1]. Im wartość ta jest bliższa 1 tym lepsza dokładność przyporządkowania obiektów do wzorców. Przyjęto arbitralnie, że porównywane 6-elementowe szeregi czasowe są podobne jeśli współczynnik korelacji kolejnościowej Spearmana między nimi wynosi co najmniej 0,7 i dlatego określono parametr p=0,3. Wówczas dla danych zbioru testowego *Acc*=0,611.

Podsumowanie

Prognozy uzyskane na podstawie zidentyfikowanych wzorców obarczone były niskimi błędami. Jednakże zgodność kierunku zmian kursu na zbiorze testowym wynosiła tylko 60%. W przypadku stosowania hierarchicznych metod grupowania danych występuje pewien element subiektywizmu związany m.in. z ustaleniem miejsca podziału dendrogramu.

Do wyznaczenia liczby skupień można wykorzystać m. in. wykres przebiegu aglomeracji, różnice odległości między kolejnymi węzłami czy regułę Mojeny [9]. Nie ma jednak „uniwersalnej" metody ustalenia punktu, w którym należy dokonać podziału dendrogramu (a tym samym ustalenia liczby skupień, na które należy podzielić zbiór danych), która dawałaby najlepsze rezultaty w każdym przypadku. Ponadto rezultat grupowania zależy też od sposobu zdefiniowania miary odległości oraz metody wiązania skupień. Istotny wpływ na uzyskiwane wyniki ma też sposób transformacji zmiennych wejściowych.

W zaprezentowanym przykładzie uwaga została skoncentrowana na prognozowaniu kierunku zmian kursu BTC/PLN i wykorzystano jako zmienne wejściowe indeksy łańcuchowe cen kryptowaluty z zamknięcia sesji. Identyfikując wzorce w finansowych szeregach czasowych można również zastosować inne transformacje danych wejściowych, m.in. stopy zwrotu czy normalizację lub wykorzystać dane w postaci nieprzekształconej [3, 4]. Również w tej kwestii brak jest uniwersalnego podejścia, które umożliwiałoby uzyskiwanie najlepszych rezultatów dla każdego zbioru danych.

Bibliografia

[1] Braak P., Nayak R., *Temporal Pattern Matching for the Prediction of Stock Prices*, http://www.eprints.qut.edu.au/14267/ [dostęp: 18.12.2015]

[2] Lee H.M. et al., *Stock Trend Prediction by Using K-Means and AprioriAII Algorithm for Sequential Chart Pattern Mining*, http://www.iis.sinica.edu.tw/page/jise/2014/201405_07.pdf [dostęp: 18.12.2015].

[3] Lula P., Morajda J., *Klasyfikacja wzorców występujących w finansowych szeregach czasowych przy użyciu sieci neuronowych Kohonena*, „Zeszyty Naukowe Akademii Ekonomicznej w Krakowie" 2002, nr 604

[4] Pichura M., *Podobieństwa w szeregach czasowych na przykładzie polskiego rynku akcji*, „Zeszyt Naukowy Śląskiej Wyższej Szkoły Zarządzania w Katowicach" 2011, nr 24

[5] Kądziołka K., *Wielowymiarowy obraz ryzyka na giełdach walut kryptograficznych*, „Kwartalnik Prawo – Społeczeństwo – Ekonomia" 2015, nr 4

[6] Li X. et al., *Exploring the Determinants of Bitcoin Exchange Rate*, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2515233 [dostęp: 11.03.2015].

[7] Madan I., Saluja S., Zhao A., *Automated Bitcoin Trading via Machine Learning Algorithms*, http://cs229.stanford.edu/projects2014.html [dostęp: 11.03.2015]

[8] Moser A. et al., *Bitcoin Stock Prediction Using Artificial Neural Networks*, http://academia.edu/17464066/Bitcoin_Stock_Prediction_Using_Artificial_Neural_Networks [dostęp: 6.11.2015].

[9] Stanisz A., *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 3. Analizy wielowymiarowe*, StatSoft, Kraków 2007

[10] http://bitcoincharts.com/markets/list/ [dostęp: 18.10.2015].

[11] Baek J., Young S.S., *A modified correlation coefficient based similarity measure for clustering time-course gene expression data*, "Pattern Recognition Letters" 2008, No. 29, Vol. 3

[12] Borisov A., Grabusts P., *Clustering methodology for time series mining*, "Scientific Journal of Riga Technical University" 2009

[13] Iglesias F., Kastner W., *Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns*, "Energies" 2013, No. 6, Vol. 2

[14] Liao T.E., *Clustering of time series data – a survey*, "Pattern Recognition" 2005, No. 38, Vol. 11

[15] Sangetta R., Sikka G., *Recent Techniques of Clustering of Time Series Data: A Survey*, "International Journal of Computer Applications" 2012, No. 52, Vol. 15

[16] Wasilewska E., *Statystyka opisowa od podstaw. Podręcznik z zadaniami*, Wydawnictwo SGGW, Warszawa 2011
[17] http://binarneopcje.pl/bitcoin-a-opcje-binarne-czyli-nowosc/, [dostęp: 4.01.2016]

Pattern recognition in financial time series using hierarchical clustering. The case of BTC/PLN exchange rate prediction

Abstract

The aim of this article was to present the use of Ward's method to identify patterns in BTC/PLN exchange rate. Identified patterns were used to predict BTC/PLN movement direction. Mean absolute percentage error and maximal percentage error on the test set were small, but the movement direction was correctly predicted only in 60% of cases.

Keywords – Bitcoin, time series clustering, pattern recognition