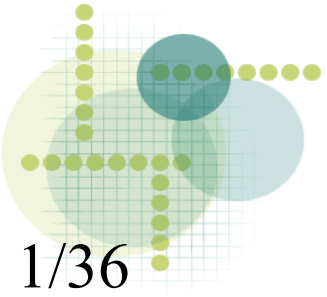
Przewidywanie cen akcji z wykorzystaniem artykułów prasowych

Mateusz Kobos, 21.03.2007
Seminarium Metody Inteligencji Obliczeniowej



Spis treści

Spis treści

Hipotezy Rynku Efektywnego

- Hipotezy Rynku Efektywnego:
  - „Przyszłe ceny akcji nie mogą być przewidziane na podstawie ...”/ „Nie można skonstruować rekomendacji inwestycyjnych umożliwiających zarządzanie portfelem w taki sposób, aby systematycznie osiągać wyniki lepsze niż te, które przynosi strategia >>kup i trzymaj<< na podstawie ...” [Malkiel03]
    - **Słaba (weak)**
      - „... cen historycznych”
    - **Średnia (semi-strong)**
      - „... cen historycznych nawet w połączeniu z publicznie dostępnymi informacjami”
    - **Silna (strong)**
      - „... jakichkolwiek informacji”

Hipotezy Rynku Efektywnego a przewidywanie

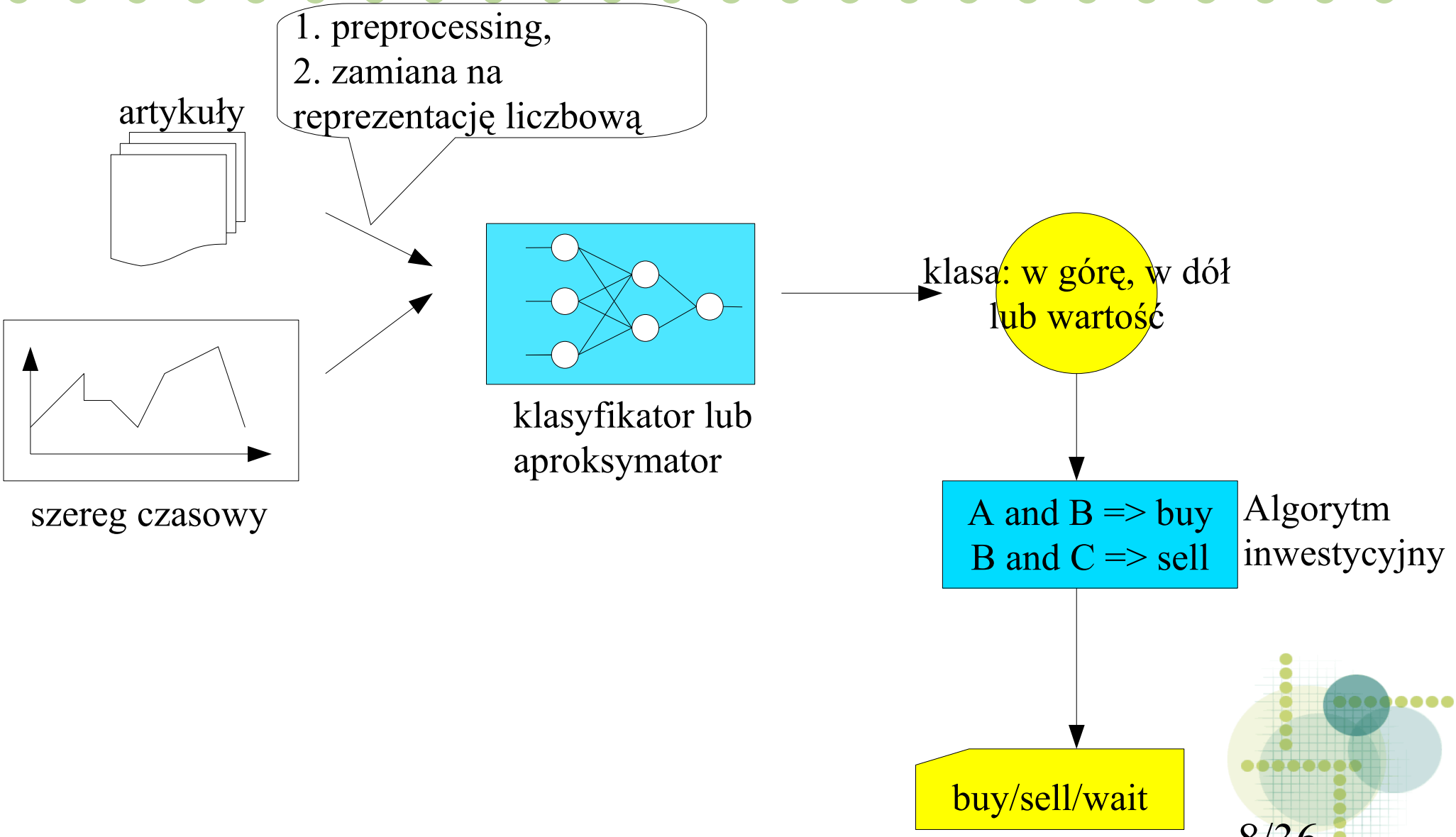
- Hipoteza:
  - Słaba – odrzuca możliwość analizy technicznej
  - Średnia – odrzuca też możliwość analizy fundamentalnej, można wygrywać na prywatnych informacjach
  - Silna – odrzuca możliwość jakiejkolwiek analizy
- Obecnie wśród badaczy najpopularniejszy jest pogląd, że prawda leży między średnią a silną hipotezą
- Przewidywane z wykorzystaniem artykułów odrzuca średnią hipotezę (ale niekoniecznie słabą)

Hipoteza rynku efektywnego – (nieśmiałe) kontrargumenty

- W latach '90 ukazały się artykuły potwierdzające możliwość sensownej analizy technicznej (za [Thomas03])
- W paru artykułach przedstawia się eksperymenty potwierdzające istnienie krótko- lub długo-terminowego wpływu informacji publicznych na ceny akcji [Pritamani01], [Gidofalvi01], [Radcliffe02]
- Większość badaczy zgadza się, że informacje publiczne mogą generować długoterminowe trendy (ale są też głosy przeciwne) [Chan03]

Spis treści

Struktura i działanie przeciętnego systemu



Spis treści

Klasyfikatory i aproksymatory

- Związane ze sztuczną inteligencją:
  1. Sieć neuronowa (perceptron wielowarstwowy, sieć probabilistyczna)
  2. Support Vector Machine
  3. Zespół algorytmów
  4. Inne: indukcja reguł, algorytm genetyczny Hidden Markov Model, drzewa decyzyjne, Naive Bayes, k-Nearest Neigbours
- Matematyczne:
  1. Modele szeregów czasowych: ARIMA, GARCH
  2. Linear Discriminant Analysis

Spis treści

Dane

- Tekstowe (na wejściu)
  - Źródła
    - Popularne czasopisma finansowe
      - Wall Street Journal, Financial Times
    - Komunikaty prasowe firm i inne komunikaty prasowe
      - Źródła: Dow Jones Newswire, Reutres, Bloomberg
    - Fora dyskusyjne (mało popularne)
      - Niezachęcające wyniki badań (duży szum) [Thomas03]
  - Analiza dokumentów:
    - Zakres dokumentu: tylko nagłówki lub cała treść
    - Sposób analizy: przez człowieka lub automatycznie

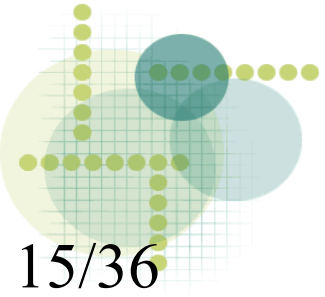
Dane

- Szeregi czasowe (na wejściu i wyjściu)
  - Akcje spółek - najpopularniejsze
  - Akcje spółek z danego sektora (ew. Indeks dla sektora) - niestosowane
  - Indeksy
  - Inne
    - kurs wymiany walut, zwrot z obligacji korporacyjnych
  - Dane makroekonomiczne (jako wejście)
    - Ceny ropy, stopy zwrotu, indeksy, kursy wymiany walut

Spis treści

Reprezentacja i etykietowanie dokumentów

- Rodzaje reprezentacji:
  - „Bag-of-words”/”vector space”
  - Wyszukiwanie zdefiniowanych przez eksperta słów kluczowych
  - Słowa kluczowe + „bag-of-words”



Spis treści

Reprezentacja i etykietowanie dokumentów: bag-of-words

- Rodzaje reprezentacji:
  - TF-IDF
  - Inne: binarna
- Etykietowanie (czyli określenie czy dokument pozytywny czy negatywny):
  1. Automatyczne (dość popularne)
     - Jeśli trend szeregu czasowego rosnący => dokument pozytywny
     - Jeśli trend szeregu czasowego malejący => dokument negatywny
  2. Ręczne
     - Człowiek etykietuje zbiór uczący, na podstawie którego uczymy klasyfikator

Reprezentacja i etykietowanie dokumentów: bag-of-words

- Każdemu dokumentowi odpowiada wektor liczb. Jedno słowo, to jeden atrybut wektora
  - oznaczenia: $a(D,w)$ atrybut wekora odpowiadający słowu w w dokumencie D
    T zbiór uczący dokumentów
    -
- Reprezentacja binarna: $a(D,w)=\begin{cases}1 & w \text{ występuje w dokumencie } D\\ 0 & \text{w p.p.}\end{cases}$
  -
- Reprezentacja TF-IDF $a(D,w)=TF(D,w)\,IDF(T,w)$

  $IDF(T,w)=\log\frac{|T|}{|T_w|}$
  - gdzie:

    $TF(D,w)$ częstość występowania słowa w w dokumencie D

    T_w podzbiór T - dokumenty zawierający słowo w

    $IDF(T,w)$ odwrotność częstośći występowania dokumentów ze słowem w wśród dokumentów zbioru T

Reprezentacja i etykietowanie dokumentów: bag-of-words

- Słownik – zbiór słów, którym odpowiadają atrybuty wektora
  - Stosuje się preprocessing:
    - Odrzucenie „stop-words” (przyimki, zaimki, spójniki)
    - Utożsamienie synonimów (za pomocą słownika synonimów WordNet)
    - Stemming – sprowadzenie słów do rdzenia
    - Redukcja wymiarów
      - Odrzucenie słów o najmniejszym współczynniku IDF
      - Odrzucenie słów występujących w takiej samej liczbie w każdej z klas dokumentów (słowa o dużej entropii)

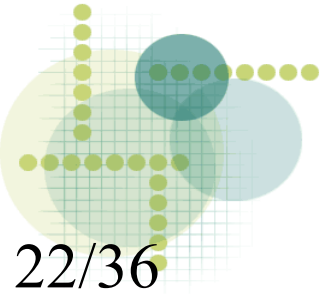
Reprezentacja i etykietowanie dokumentów: bag-of-words

- Pomimo prostoty, reprezentacja bag-of-words nieźle się sprawdza w praktyce
  - Współwystępujące słowa w dokumencie tworzą „kontekst”

Spis treści

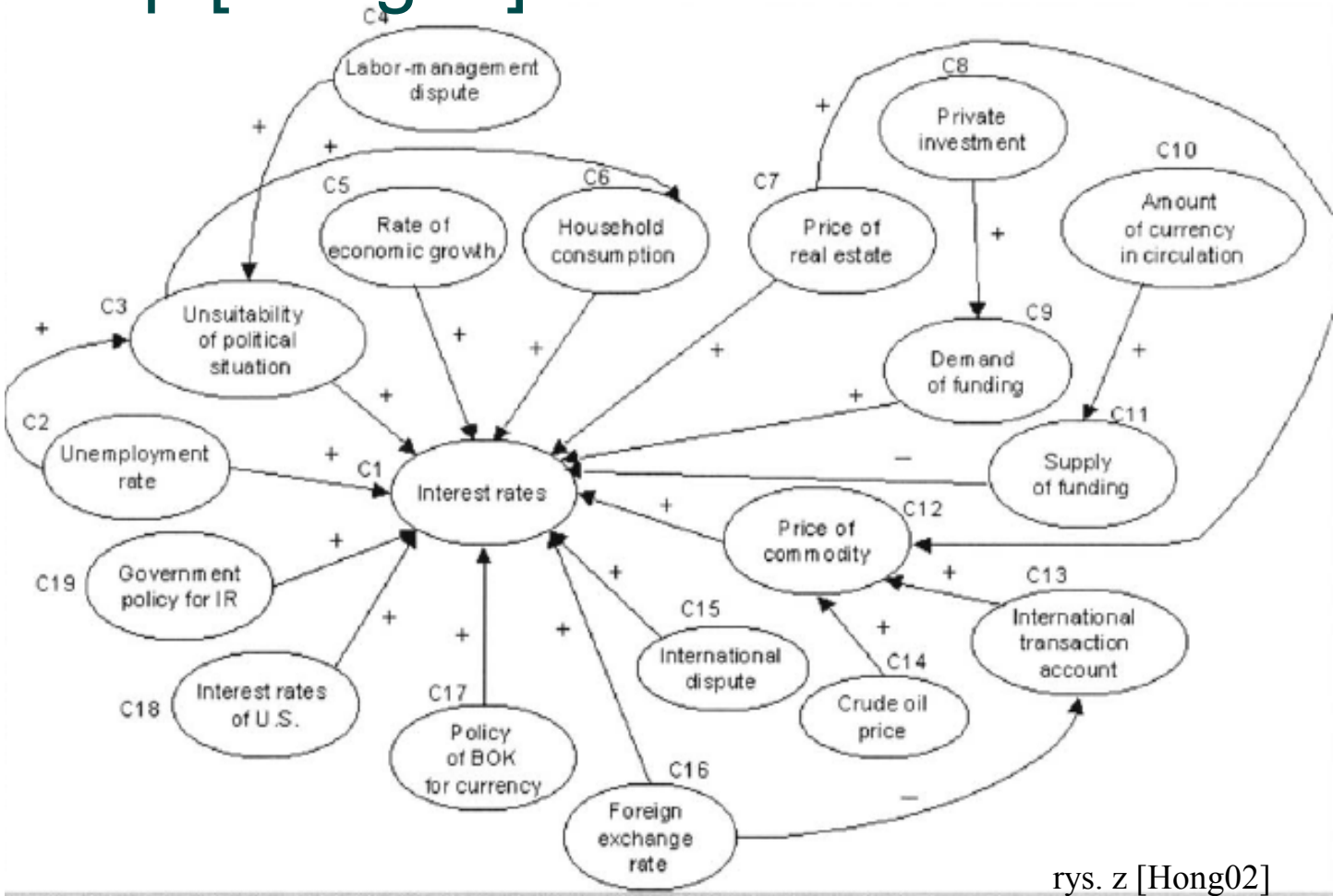
Wiedza ekspercka

- Analiza artykułu przez człowieka i zastosowanie reguł eksperckich np:
  - „Jeśli sytuacja międzynarodowa się pogarsza, to ceny zazwyczaj maleją (i odwrotnie)” [Kohara97]
- Tworzenie wektora słów tylko z wyróżnionych słów kluczowych (dość popularne)
  - Można rozróżniać też wyrażenia pozytywne i negatywne
  - Generowanie w trakcie nauki reguł decyzyjnych opartych na tych słowach [Permuentilleke02]



Wiedza ekspercka: Cognitive Map

- Wiedza ekspercka o zależnościach zapisana w Cognitive Map [Hong02]



rys. z [Hong02]

Wiedza ekspercka: Cognitive Map cd.

- Wyszukujemy wyrażenia odpowiadających poszczególnym węzłom (negatywne i pozytywne)
  - Każdy dzień jest reprezentowany przez wektor pozytywnych i negatywnych węzłów. Te wartości negatywne i pozytywne są „propagowane” w Cognitive Map i otrzymujemy sumaryczny wpływ na przewidywaną wartość (tu: zysk z obligacji korporacyjnych). Tą wartość używamy jako wejście w aproksymatorze (tu: MLP ANN).

Wiedza ekspercka: Przypisywanie artykułów do kategorii

- Rozpoznawanie, czy dokument należy do jednej z wielu predefiniowanych kategorii (za pomocą wyrażeń regularnych) [Thomas03]
  - Przykład części wyrażeń rozpoznających jedną z kategorii:
    - New Financing: Discussion of issues of new debt, shares, other securities, or loans. ("Charter Communications files $4 billion shelf", CHTR, 3/9/2001)

      1. "offer(s)?|sell(s)?|issue(s)?|sale(s)?|pricing|price(s)?|launch(es)?|placement|raise(s)?|tap(s)?" AND "note(s)?|bonds|debt|debenture(s)?" NAND "pay down|research"
      2. "offer(s)?|issue(s)?|sale(s)?|pricing|launch(es)?|placement|raise(s)?|tap(s)?" AND "shares"

rys. z [Thomas03]

Spis treści

Reprezentacja dokumentów: Bag-of-words + wiedza ekspercka

- Wymuszenie występowania predefiniowanych słów w słowniku reprezentacji bag-of-words [Mittermayer06]
- Predefiniowane frazy (słowa znajdujące się blisko siebie) występują jako oddzielne wymiary w wektorze [Seo04]

Spis treści

- Przewidywanie cen akcji a Hipoteza Rynku Efektywnego
- Struktura i działanie przeciętnego systemu
  - Klasyfikatory i aproksymatory
  - Dane
  - Reprezentacja i etykietowanie dokumentów
    - Bag-of-words
    - Wiedza ekspercka
    - Bag-of-words + wiedza ekspercka
- Otrzymywane wyniki

Otrzymywane wyniki: parametry

- Dane uczące:
  - Liczba analizowanych artykułów:
    - od paruset do paru tysięcy (w sumie)
  - Zakres czasowy:
    - od 3 miesięcy do 3-4 lat
- Horyzont czasowy dokonywanych inwestycji/horyzont predykcji:
  - 15 min [Mittermayer06], 1h [Mittermayer04], 3-5 dni [Fung05], 20 dni [Pritamani 01], 1 miesiąc [Hong02], 60 dni [Wuthrich98]

Otrzymywane wyniki: uwagi

- Prawie nigdy nie uwzględnia się kosztów transakcji
  - Często otrzymuje się system, który dokonuje przewidywania, ale w przypadku uwzględnienia kosztów transakcji nie jest on w stanie wygrać z rynkiem
- Często proponowany system porównuje się z modelem błądzenia losowego (Random Walk)
  - Jest parę wersji tego modelu:
    - „jutrzejsza cena będzie taka jak dzisiejsza”
    - „jutrzejszy wzrost będzie taki sam jak historyczny średni wzrost”

Otrzymywane wyniki: dane liczbowe

- Przewidywanie kursu wymiany walut (up/down/steady) 1,2,3 godziny naprzód [Permuentilleke02]
  - dokładność: 28-53% (człowiek: ok.50%)
- Przewidywanie zwrotu z obligacji miesiąc naprzód (aproksymacja) [Hong02]
  - Średni moduł błędu procentowego (MAPE): 3,34%-4,65% (model Random Walk: 11%-15,31%)
- Przewidywanie zwrotu z akcji 20 dni naprzód (za pomocą prostych reguł) [Pritamani01]
  - Średnia roczna stopa zwrotu względem losowego portfela akcji (z pewnymi sensownymi warunkami) uwzględniająca koszty transakcji: 12-18%

Otrzymywane wyniki: dane liczbowe

- Przewidywanie zwrotu z akcji 15 min. naprzód [Mittermayer06]
  - Dokładność: 82%, sumaryczny zwrot z przeprowadzonych 329 transakcji: 89% (algorytm losowy: dokładność: 33%, sumaryczny zwrot:0%)
    - Autorzy twierdzą, że po uwzględnieniu kosztów transakcji prawdopodobnie zwrot byłby =0
- Przewidywanie 5 indeksów 60 dni naprzód (up/down/steady) [Wutrhich98]
  - Dokładność 47%, zwrot: 2,4% w ciągu 3 miesięcy względem indeksu (9,6% w ciągu roku)

Otrzymywane wyniki: dane liczbowe

- Przewidywanie zwrotu z akcji parę dni naprzód (klasyfikacja: up/down) [Fung05]
  - Dokładność: 51% (horyzont inwestycji:1 dzień)-65% (horyzont: 5 dni),
  - sumaryczna stopa zwrotu dla inwestycji z przeciągu 5 miesięcy: 18% (strategia Buy&Hold: -20,6%)

Bibliografia

- [Chan03] Wesley S. Chan, *Stock price reaction to news and no-news: drift and reversal after headlines*, Journal of Financial Economics 70, 2003
- [Fung05] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Hongjun Lu, *The Predicting Power of Textual Information on Financial Markets,* IEEE Intelligent Informatics Bulletin vol.5 No.1, 2005
- [Gidofalvi01] Győző Gidófalvi, *Using News Articles to Predict Stock Price Movements*, Department of Computer Science and Engineering University of California, San Diego, http://citeseer.ist.psu.edu/gidofalvi01using.html, 2001
- [Hong02] Taeho Hong, Ingoo Han, *Knowledge-based data mining of news information on the Internet using cognitive maps and neural networks*, Expert systems with applications 23, 1-8, 2002
- [Kohara97] Kazuhiro Kohara, Tsutomu Ishikawa, Yoshimi Fukuhara, Yukihiro Nakamura, *Stock price prediction using prior knowledge and Neural Networks*, Intelligent systems in accounting. finance and management vol, 6, 1997
- [Malkiel03] Burton G. Malkiel, *Błądząc po Wall Street*, WIG-Press Warszawa, 2003
- [Mittermayer04] Marc-André Mittermayer, *Forecasting Intraday Stock Price Trends with Text Mining Techniques*, Proceedings of the 37th Hawaii International Conference on System Sciences, 2004
- [Mittermayer06] Marc-André Mittermayer, Gerhard F. Knolmayer, *NewsCATS: A News Categorization And Trading System,* Proceedings of the Sixth International Conference on Data Mining (ICDM'06), 2006

Bibliografia cd.

- [Peramunetilleke02] Desh Peramunetilleke, Raymond K. WongCurrency, *Exchange Rate Forecasting from News Headlines,* Proceedings of the 13th Australasian database conference - Volume 5, 131-139, 2002
- [Pritamani01] Mahesh Pritamani, Vijay Singal, *Return predictability following large price changes and information releases*, Journal of Banking & Finance 25, 2001
- [Radcliffe02] Radcliffe G. Edmonds Jr, Ali M. Kutan, *Is public information really irrelevant in explaining asset returns?*, Economics Letters vol. 76, issue 2 , 2002
- [Seo04] Young-Woo Seo, Joseph Giampapa, Katia Sycara, *Financial News Analysis for Intelligent Portfolio Management*, Tech. Report CMU-RI-TR-04-04, Robotics Institute, Carnegie Mellon University, 2004
- [Thomas03] James D Thomas, *News and Trading Rules,* Carnegie Mellon University Ph.D. Thesis, 2003
- [Wuthrich98] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, W. Lam, *Daily Stock Market Forecast from Textual Web Data*, Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on, 1998

Dziękuję za uwagę!