要 旨

株価の時系列変化の予測のための特徴選択

中井　淳人

ビックデータ時代の到来によりニュースなどのテキストデータを用いた株価の予測が注目されている．そこで本研究では機械学習による株価予測に用いる新聞記事のテキストデータの特徴選択について検討を行う．本研究では, 特徴に加え, 先行研究で精度の高いことが報告されているサポートベクターマシンと, テキストデータを学習するという点で類似しているスパムメールの判別で精度の高いことが報告されているランダムフォレストで学習, 株価の予測を行い比較する．2013 年 1 月 1 日から 12 月 31 日までの日本経済新聞の朝刊の記事を利用し, 日経平均株価の上昇か中間 (値動きなし) か下降かの 3 クラスを教師信号として使用する．まず, 単語の出現頻度と株価の動きとの相関から単語の選出を行う．予測の際に当日の記事だけでなく前日と前々日の記事も用いるため, 過去 3 日間の時系列単語ベクトルを提案する．これに加え, 時系列単語ベクトルを特異値分解により次元を削減し機械学習に用いる．実験データは, 記事より得られた単語から 500 単語を相関係数を用いて選出し, 233 日分の時系列単語頻度ベクトルを作成し使用する．その結果, 相関による単語の選出, 過去 3 日間の時系列単語頻度ベクトル, 特異値分解を行った場合, それらを行わなかった場合に比べて精度が高くなることを示す．

キーワード　テキストマイニング, 機械学習, 株価予測

Abstract

Feature Selection for Prediction of Stock Price Time Series

Atsuhito NAKAI

Prediction of stock price using the text data of the news in the Internet attracts many researchers. This thesis proposes a feature selection of text data obtained from newspaper articles in the Internet for the stock price prediction by machine learning. In addition, the random forest algorithm is used to improve the accuracy of prediction. The random forest has been achieved high accuracy for text data learning in previous research. The result using random forest is compared with that of support vector machine, which is a popular algorithm in machine learning. The articles are retrieved from Nihon Keizai Shimbun Morning Newspaper of January 1 to December 31, 2013, and the training data are given as price up, price down, and stationary price. For feature selection, words extracted from articles are selected by correlation between prices and the frequency of the occurrence of the word. Furthermore three word occurrence vectors for three days latest to the day for prediction are combined and used for better prediction. We propose the vector as time-series word frequency vector. In addition, singular value decomposition is employed to time-series word frequency vector in order to reduce the dimension of word frequency vector. The experimental data, which, consists of 500 words selected by the correlation are obtained from the articles for 233 days. As a result, the accuracy of prediction improves in the case of applying proposed methods, i.e., word feature selection using correlation, time-series word frequency vector, and

singular value decomposition.

key words Machine Learning, Text Mining, Prediction of Stock Price