要 旨

テキストマイニングによる
株価予測に適した機械学習

山口　祐輝

人間のトレーダーが株を売買する際，株価や為替のような定量的データのみでなく，ファンダメンタル分析に分類される，ニュースや会社四季報などを利用した企業の業績や財務状況のような定性的な情報も重要とされている．そこで本研究では仲矢らの研究で利用されたニュース記事を用いたテキスト解析を行い，定性的データを利用した機械学習に最適な手法の比較検討を行う．また，定性的データを十分に活用するため，単語の共起度を利用した単語ベクトルの作成と学習を行う．本実験では比較のため仲矢らの研究で使用されている，NIKKEI NET，Infoseek と MSN 産経ニュースの 2009 年 10 月 1 日から 12 月 22 日までのニュースデータを利用し，日経平均終値の上昇か，下降の 2 値を教師データとして使用する．作成する単語ベクトルは仲矢らの研究で使用されたもの，2 単語の共起性に着目した 2 単語ベクトルの 2 つを作成する．株価予測には k 最近傍法，ニューラルネットワーク，サポートベクターマシンを使用し，ニューラルネットワークではバックプロパゲーションを使用する．実験用データとして 170 記事 121 単語の単語ベクトルと t 検定によって作成した 170 記事 156 単語の単語の共起度を用いた単語ベクトルを使用する．教師データは双方とも翌日の日経平均株価の上昇か下降の二値とし，評価は k-分割交差検証によって検証を行う．その結果サポートベクターマシンによる学習が最も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，株価予測，単語ベクトル

Abstract

Yuki Yamaguchi

In order to predict the stock price, many researches called technical analysis have been studied. Technical analysis mainly uses numerical information, for example, chart or time series of stock prices. Some of these researches use neural network, genetic algorithm, support vector machine are used for some of them. Human stock traders, however, use not only numerical data but also qualitative data, for example, company information, economic, political, or social news. Nakaya et al. have been studied stock price predict using text mining technique and news data on the Internet. In this research comparison of machine learning algorithms for text mining are conducted for the application of stock market predict from market news. Word vector using word co-occurrence is applied for learning data. News data from NIKKEI NET, Infoseek, MSN Sankei from October 2009 to December are used as learning data and the machine learning algorithms predict up or down of the stock price. Therefore training vectors are binary vectors. K-nearest neighbor, neural network, and support vector machine are used to predict. The number of data is 170 news articles and 121 words for conventional word vector and 170 news articles and 156 words for word vector with word co-occurrence. Cross validation with k-division are used for evaluation of algorithms. As a result, support vector machine is the highest accurate algorithm for this application.

key words Machine Learning,Text-Mining,