要 旨

潜在意味解析を用いた
トピック抽出による株価予測

塩見 侑司

テキストマイニングを用いた株価予測が研究されている．それらの研究は人間が新聞記事やニュース記事などのテキストデータを投資材料としていることから，コンピュータにおいてその行動を実現することを目的としている．しかし，テキストデータは単語ベクトルと呼ばれる形で表現されるが，一般にその次元は数万を超えそのままでは機械学習に適さない．自然言語処理分野ではテキストデータから特徴的な情報の抽出を行う潜在意味解析が提案されている．潜在意味解析は本質的な特徴量を残しつつ必要のない情報量を削減することで次元削減を行う．そこで本研究ではこの潜在意味解析を用いて経済新聞記事からトピックを抽出し，そのトピックにより SVM(サポートベクターマシン) を用いて株価の予測を行う．株価予測精度の向上と株価予測に影響されるトピックの検証を目的とする．テキストデータとして 2014 年 のニュース記事 (全 64564 記事) を使用する．各記事に形態素解析を行い，名詞を抽出し，各単語の出現回数を日付毎に求た単語ベクトルを用いる．教師データとして，日経平均終値の上昇が前日比 +0.5%より増加すれば上昇，前日比 −0.5%より減少すれば下降，その間を値動きなしとした 3 クラスのデータを用いる．結果は，4，5 個のトピックを選んだ時の 66.7%が最大の識別率であり，先行研究の奥村 [1] の相関を用いた識別結果よりも 11.6 ポイント向上したことを示す．また選出したトピックも直感的に株価の影響のあるトピックが選ばれていることから株価に影響するトピックが存在することを示す．

キーワード　株価予測，潜在意味解析，テキストマイニング，トピック，TF-IDF，サポートベクターマシン

Abstract

Prediction of Stock Market Price by Topic Extraction using Latent Semantic Analysis

Stock price prediction using text mining has been studied. For human stock dealers, news articles are used to make decision for dealing. Therefore stock market price prediction using text mining has been studied in the area of artificial intelligence. Text data is generally expressed by word vector. The dimension is over ten thousands and it is not suitable for machine learning. Latent semantic analysis(LSA) has been proposed in natural language processing and it can extract the essential feature from text data. LSA reduces the dimension by keeping essential feature while unnecessary information reduced. This study propose the prediction method of stock prices by topic extraction from news articles using LSA. The purpose of this study is to improve the prediction accuracy and to clarify whether stock market price is influenced by topic. As a result, choosing 4 or 5 topics achieves the highest accuracy. The prediction accuracy is 66.7%. Also this thesis shows that the results of selected topics affect the stock market prices.

key words stock market prices, latent semantic analysis, text mainig, TF-IDF, support vector machine