Prediction of posting reliability on stock bulletin board

Katsuhiko Utsubo

Graduate School of Advanced Science and Technology,

Japan Advanced Institute of Science and Technology

August 2019

keywords:stock bulletin board, reliability, machine learning, decision tree

Although it is difficult to predict trends in stock prices and indices, but if these trends can be predicted, they can be used as a basis for investment management. Traditionally, stock price is predicted using quantitative information, but there is also a phenomenon called "material exhaustion" that the stock price drops even if the settlement information is good. In many cases it is not possible to explain by quantitative information alone. Here, "material expenditure" means that information affecting the market is exhausted and that future stock price increases will not be visible.

In recent years, methods for analyzing stock price trends from information such as news have also been studied. This research predicts stock prices from extracted words, or predicts stock prices from emotional attributes of extracted words. These have achieved some success. However, these methods do not verify the authenticity of the information. All information is handled uniformly. There are few researches that mention the credibility of information and analyze stock market trends.

In this research, we propose a method to predict the reliability of the contents posted on the stock bulletin board, which is qualitative information. The purpose of this research is to construct a model for analysis of stock price trends using the reliability of information.

The reliability of a post is quantified by the post evaluation value given to the post, which is used as a target variable. The forecasting model used the explanatory variable as the index of stock price, price return, stock price historical volatility, and turnover. Furthermore, the evaluation value of stock brand and the post evaluation value of the poster and the negative / positive value of the post contents of the bulletin board were used as explanatory variables. We constructed a model to predict objective variables from explanatory variables using a bulletin board and stock price data. The data is from January 2015 to December 2016 as the training data, and from January 2017 to June 2017 as the validation data.

As a result of examining the relationship between each explanatory variable and the post evaluation value, the following was found. When the price return or the stock price historical volatility becomes high, the posting evaluation value of the bulletin board rises. When the turnover goes up, posting evaluation value of the bulletin board decreases. This is a situation in which the price return is high and the fluctuation is strong, that is, a situation in which it is easy for investors to obtain a profit. It is speculated that this would

mean that posts with high post evaluation values will increase.

Also, it was found that the contributors are classified into groups with high and low post evaluation values. A positive correlation was found between the posting evaluation value of the bulletin board and the posting evaluation value of the poster. Furthermore, correspondence analysis was performed using the posting evaluation value of the bulletin board and the posting evaluation value of the poster. As a result, it was found that contributors with high post evaluation value gather on a bulletin board with high post evaluation value, and contributors with low post evaluation value gather in a bulletin board with low post evaluation value.

Furthermore, the negative-positive analysis of posts by natural language processing showed a positive correlation between the reliability of posts and the negative-positive value of posts, and it was found that the post evaluation value was higher for posts with positive emotions. In order to confirm this result, the actual posting contents were extracted 5 posts in the descending order of the post evaluation value and 5 posts in the low order, and each post was visually confirmed. The posts with low post evaluation value have many dirty words and symbols, and many posts do not receive a good impression. Conversely, posts with high post evaluation value are polite sentences, and they are post contents that have a good feeling of favor. This result is consistent with negative-positive analysis.

Next, we created a model that predicts post evaluation values using a binary classification of positive or negative post evaluation values. The model was constructed using a decision tree with the contribution evaluation value as the objective variable, the stock return, stock price historical volatility, turnover , the evaluation value of stock brand, the post evaluation value of the poster, and the negative value of the content of the contribution as explanatory variables. The correct answer rate of the model is 0.756, and the F value is 0.744, which makes it possible to predict the post evaluation value. Furthermore, when the decision tree model was visualized and details of the model were confirmed, it was found that the post evaluation value is determined only by the post evaluation value of the poster. That is, a post with a high post evaluation value is predicted to have a high post evaluation value, and conversely, a post with a low post evaluation value is predicted to have a low post evaluation value.

Using this model, we evaluated the forecasting performance of the price-earnings ratio on the next day with 40 highly reliable and 40 low unreliable contributors among regular contributors. When the post sentiment attached to the post is "want to buy" or "want to buy strongly", the post predicts that the price / earnings ratio will rise the next day. Conversely, when "I want to sell" or "I want to sell strongly", the post predicts that the price / earnings ratio will decline the next day. The accuracy rate of the prediction at this time was analyzed by binary classification. As a result, it was found that the accuracy rate of prediction of a highly reliable poster is 0.566, and the accuracy rate of a low reliability poster is 0.477, and the prediction accuracy rate of a highly reliable poster is high. Furthermore, as a result of conducting a chi-square test, it was shown that this accuracy rate difference has an advantage and the accuracy rate of the prediction of a highly reliable poster is high. From this, it can be said that highly reliable information can be obtained from a highly reliable person, and the prediction performance of the stock price of the highly reliable information is high.

From the above results, the model proposed by this study shows that it is possible to predict the reliability of posts by examining the reliability of posters. Furthermore, it is

possible to predict the price return of the next day from highly reliable posts.

By using the model proposed by this research, it is possible to extract highly reliable posts as a preliminary step of analysis of qualitative data. By extracting reliable information, it is possible to contribute to investors' judgments on stock investment.

株式掲示板における投稿の信頼度予測

靱　勝彦

北陸先端科学技術大学院大学

先端科学技術研究科

令和 元年 8 月

キーワード:株式掲示板, 信頼度, 機械学習, 決定木

株価や指数の動向を予測することは困難であるが、この動向を予測できれば投資家への運用の判断材料になる。従来、株価の予測には定量的な情報を用いて行われているが、決算情報が良くても株価が下がる「材料出尽くし」という現象もあり、定量的な情報だけでは説明がつかない場合も多い。ここで「材料出尽くし」とは、相場に影響する情報が出尽くしてしまい、今後の株価上昇が見えないことを言う。

近年では、ニュースなどの定性情報から株価動向を分析する手法も研究されており、抽出した単語から株価を予測するものや、抽出した単語の感情属性から株価を予測するものなど、様々な手法が一定の成果を上げている。ただし、これらの手法では、定性情報の真偽を確かめることなくすべて一律に扱っており、定性情報の信頼性に言及し、株価動向を分析したものは少ないといえる。

定性情報の信頼性の研究では、フェイクニュースの信頼度を分類する研究が行われており、ニューラルネットワークを用いて情報を分類するものや、定性情報の伝播状況から情報の信頼性を分類するものや、情報の発信者の信用履歴を用いて分類するものなど様々なものがある。しかし、フェイクニュースは様々な種類があり、それぞれが異なるテキストの指標を持っていると報告するものもあり、単一のアプローチでは難しいと言える。

SNS など、コミュニケーションツールの重要性はますます高まっている。特に、個人投資家にとっては、機関投資家に比較し、情報の取得量の格差は依然として大きい。また、個人投資家は情報を得るために、知識の交換の場として掲示板などの SNS を利用することが多い。そのため、株式掲示板を分析することにより、投資家の発言としての形式知と、実際の行動としての暗黙知を、掲示板の信頼度の分析という形で、信頼度を定量的に評価することが可能となり、知識科学的に意味があると言える。

本研究では、定性情報としての株式掲示板における投稿内容の信頼性を予測する手法を提案し、情報の信頼性を踏まえた株価動向の分析への手かがりとするモデルの構築を行うことを目的とする。

投稿の信頼度は、投稿に付与された投稿評価値で定量化し、これを目的変数とし、説明変数を株価の指標である、株価収益率、株価ヒストリカル・ボラティリティ、売買代金と、 銘柄の投稿評価値、投稿者の投稿評価値、掲示板の投稿内容のネガポジ値を説明変数として予測モデルの構築を、掲示板と株価データを用いて行った。データは 2015 年 1 月から 2016 年 12 月までを学習データ、2017 年 1 月から 2017 年 6 月までを検証データとした。

それぞれの説明変数と投稿評価値の関係を調べた結果、次のことがわかった。株価収益率または株価ヒストリカル・ボラティリティが高くなると、掲示板の投稿評価値は上昇し、売買代金が高くなると掲示板の投稿評価値は減少した。これは株価収益率が高く、変動が激しい状況、すなわち投資家の利益の得やすい状況になると、投稿評価値が高い投稿が増えるということになるのではなかと推測される。

また、投稿者は投稿評価値の高いグループと低いグループに分類されることがわかった。掲示板の投稿評価値と投稿者の投稿評価値には正の相関関係が見られ、さらに、掲示板の投稿評価値と投稿者の投稿評価値でコレスポンデンス分析を行った結果、投稿評価値の高い掲示板には投稿評価値の高い投稿者が集まり、投稿評価値の低い掲示板には投稿評価値の低い投稿者が集まることがわかった。

さらに、自然言語処理による投稿のネガポジ分析から、投稿の信頼度と投稿ネガポジ値には正の相関がみられ、ポジティブな感情の投稿ほど投稿評価値が高いことがわかった。この結果を確認するために、実際の投稿内容を投稿評価値が高い順に 5 投稿、低い順に 5 投稿抽出し、それぞれの投稿を目視にて確認したところ、投稿評価値の低い投稿は、汚い単語や記号を多用しておりあまりいい印象を受けない投稿が多く、逆に投稿評価値の高い投稿は、丁寧な文章であり、好感の持てる投稿内容であり、ネガポジ分析と一致するような結果となった。

次に、投稿評価値を投稿評価値が正か負かの 2 値分類で予測するモデルを作成した。モデルは投稿評価値を目的変数とし、株価収益率、株価ヒストリカル・ボラティリティ、売買代金、投稿者の投稿評価値、投稿内容のネガポジ値を説明変数として、決定木によるモデルを構築した。作成したモデルは正解率が 0.756、F 値が 0.744 となり、投稿評価値を予測することができるモデルとなった。さらに、決定木のモデルの可視化を行い、モデルの詳細を確認したところ、投稿評価値は投稿者の投稿評価値のみによって決定されることがわかった。つまり、投稿評価値の高い投稿者の投稿は投稿評価値が高いと予測され、逆に投稿評価値の低い投稿者の投稿は投稿評価値が低いと予測される結果となった。

このモデルを用い、常連投稿者のうち、信頼度の高い 40 名と信頼度の低い 40 名で、翌日株価収益率の予測性能を検証した。投稿に付与されている投稿感情が「買いたい」「強く買いたい」のときにその投稿は翌日株価収益率が上昇すると予測しているとし、「売りたい」「強く売りたい」のときにその投稿は翌日株価収益率が下降すると予測するとした時の、予測の正解率を、2 値分類により分析した。その結果、信頼度の高い投稿者の予測の正解率が 0.566、信頼度の低い投稿者予測の正解率が 0.477 となり、信頼度の高い投稿者の予測正解率が高いことがわかった。さらに、カイ二乗検定を行った結果、この正解率差には優位性があり、信頼度の高い投稿者の予測の正解率が高いことが示された。 このことから、信頼度

の高い情報は信頼度の高い人から得ることができ、その信頼度の高い情報の株価の予測性能は高いということが言える。

以上の結果から、本研究により提案するモデルは、投稿者の投稿評価値、すなわち投稿者の信頼度を調べることにより、将来に投稿された投稿の投稿評価値、すなわち掲示板に投稿された投稿の信頼度の予測に対して有効であるといえる。さらに、その信頼度から翌日株価収益率が予測可能であることを示した。

本研究が提案するモデルを用いることにより、定性データの分析の前段階として、信頼度の高い投稿を抽出することが可能である。信頼性の高い情報を抽出することで、投資家の株式投資への判断材料に貢献することができるといえる。