

株式掲示板における投稿の信頼度予測

靱 勝彦† Dam Hieu Chit

北陸先端科学技術大学院大学†

1 はじめに

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。定性情報から株価動向を分析する手法が研究されており、抽出した単語の主成分分析から株価を予測するもの [3] や、抽出した単語の感情属性から株価を予測するもの [2] など、様々な手法が一定の成果を上げている。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、情報の信頼性を言及したものは少ないといえる。

本研究では、Yahoo!ファイナンスの textream 掲示板を対象に投稿の信頼度を定義し、投稿の属性（投稿者、銘柄、投稿内容など）から信頼度を推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を目的とする。

2 方法

2.1 信頼度の定義

各投稿に付与されている、投稿を見た人の「そう思う」「そう思わない」という感情を元に、投稿の信頼度を式 1 に定義した。

$$\text{信頼度} \equiv \text{投稿評価値} = \frac{\text{そう思う数} - \text{そう思わない数}}{\text{そう思う数} + \text{そう思わない数}} \quad (1)$$

2.2 データ

掲示板の 2015 年 1 月 1 日から 2016 年 12 月 31 日までを訓練データ、2017 年 1 月 1 日から 2017 年 6 月 30 日までを検証データとし、訓練データ期間中の上位投稿数 30 銘柄に 1,000 以上投稿している投稿者 800 人の投稿を分析対象とした。株式との関係を調べるため、12 ヶ月月次ヒストリカル・ボラティリティ及び月次平均売買代金、月次平均株価収益率を使用した。掲示板の日付は営業日の午前 9 時～翌営業日の午前 8 時 59 分までを当日日付とし、分析に用いた。

Prediction of posting reliability on stock bulletin board
Katsuhiko UTSUBO†, Dam Hieu Chit†
†Japan Advanced Institute of Science and Technology
923-1292, Ishikawa, Japan
{k-utsubo,dam}@jaist.ac.jp

3 銘柄分析

訓練データを用いた銘柄別の投稿評価値の箱ひげ図を、図 1 に示す。丸に示す平均投稿評価値は銘柄ごとに、投稿の信頼度が異なることを示している。

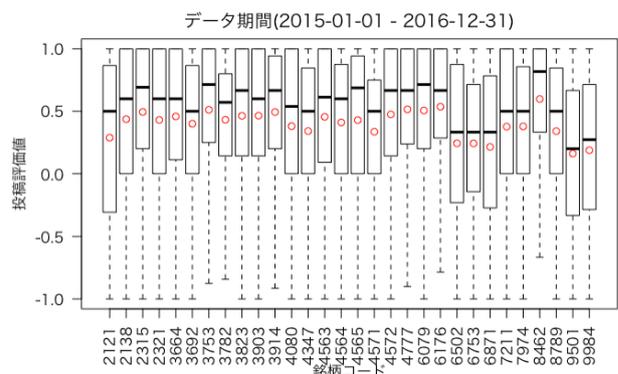


図 1: 銘柄別の投稿評価値。横線は中央値、丸は平均値。

更に、銘柄別の平均投稿評価値と、月次平均株価収益率、12 ヶ月月次ヒストリカル・ボラティリティ及び月次平均売買代金との相関係数を、表 1 に示す。銘柄別の平均投稿評価値は株価の収益率が高いほど、動きが大きくなるほど高く、売買代金が高く売買が活発になるほど低くなる傾向が見られた。

表 1: 銘柄別の投稿評価平均値と株式との関係

株式指標	相関係数
月次平均株価収益率	0.599
12 ヶ月月次ヒストリカル・ボラティリティ	0.697
月次平均売買代金	-0.497

4 投稿者分析

投稿者別の平均投稿評価値の関係を図 2 に示す。平均投稿評価値が極端に高い人と低い人が存在することがわかる。

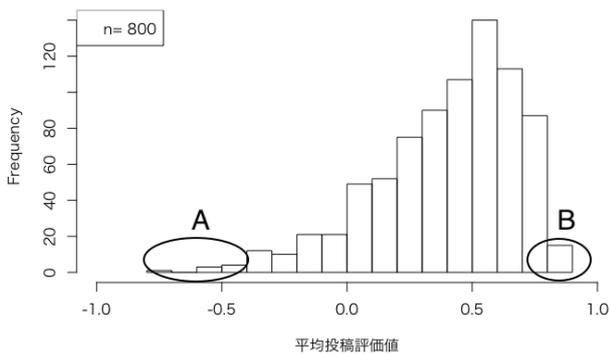


図 2: 投稿者別の平均投稿評価値のヒストグラム。A は” そう思わない”, B は” そう思う” と評価される投稿が多い投稿者。中央値:0.4626

5 投稿内容分析

投稿の形態素解析からネガ・ポジ分析を行い、投稿者別の平均投稿評価値と平均ネガ・ポジ値の関係を調べた。形態素解析の辞書には Neologd0.0.5 を用い、ストップワード¹を取り除いたものを使用した。ネガ・ポジの判定には、投稿に付与されている「強く売りたい」から「強く買いたい」の5段階の感情（以下、投稿感情）から辞書を作成した [1]。その結果、投稿者別の投稿の平均ネガ・ポジ値と投稿評価値の相関係数は 0.333 の正の相関が見られ、ポジティブ投稿ほど投稿評価値が高い傾向が見られた。

6 予測モデル

投稿評価値を予測には、表 2 に示す特徴量を用い、投票評価値が 0 超か未満かの 2 値分類を決定木で行った。訓練データ及び検証データは 2 値が均等になるように無作為に選択したデータを用いた。訓練データで決定木で学習した結果（図 3）から、投稿評価の 2 値分類の値は投稿者の平均投稿評価により決まると言える。予測モデルを検証データで検証した結果、F 値 0.744、正解率 0.756 の精度で投稿評価を予測できた。

7 株価収益率の予測

信頼度の低い投稿者と高い投稿者を 40 名ずつ抽出し、投稿感情の付与された投稿のうち、場中の影響を排除するため営業日 15 時から翌営業日 8 時 59 分までの投稿を用いて翌営業日の株価収益率との関係を調べた結果を表 3 に示す。上位 40 名と下位 40 名の正解率は逆の結果となった。

¹<http://www.dl.kuis.kyoto-u.ac.jp/slothlib/>

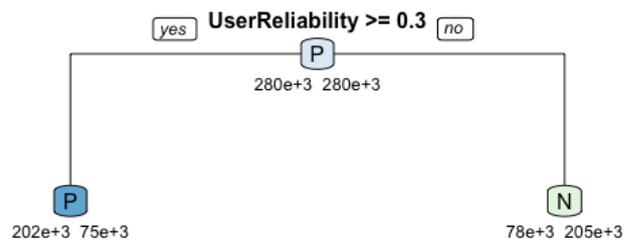


図 3: 決定木。P は 0 超, N は 0 未満の投稿評価値

表 2: 投稿評価 $Evaluation_{jk}$ を予測するための特徴量

変数	項目
$StockReturn_j$	銘柄 j の平均月次株価収益率
$StockVolat_j$	銘柄 j の平均 12 ヶ月月次ヒストリカル・ボラティリティ
$StockTurnover_j$	銘柄 j の平均月次売買代金
$StockReliability_j$	銘柄 j の平均投稿評価値
$UserReliability_p$	投稿 k を投稿した投稿者 p の平均投稿評価値
$Sentiment_k$	投稿 k のネガ・ポジ値

8 まとめ

本研究では、株式掲示板の信頼度を予測するため、信頼度を定義しその予測の 2 値分類モデルを構築した。信頼度は投稿者の平均投稿評価値のみに影響され、翌日株価収益率の予測精度は信頼度の高い投稿者と低い投稿者で逆となることがわかった。

参考文献

- [1] 坪内 孝太, 山下 達雄, 株価掲示板データを用いたファイナンス用ポジネガ辞書の生成, 人工知能学会全国大会 (第 28 回) (2014).
- [2] 山下 達雄, 坪内 孝太, 個人の予測信頼度を加味した株価掲示板情報からの株価予測, 人工知能学会全国大会 (第 30 回) (2016).
- [3] 高橋宏圭, 関和広, 上原邦昭, 株価回帰と web ニュース記事分析を組み合わせた株価動向推定, 電子情報通信学会技術研究報告 111 (2012).

表 3: 訓練データでの株価収益率の予測結果

投稿者信頼度	適合率	再現率	F 値	正解率
上位 40 名	0.575	0.959	0.719	0.566
下位 40 名	0.455	0.850	0.593	0.477