3L3-OS-16a-3in2

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

Stock Price Prediction via Stock Bulletin Board with considering users' prediction ability

山下 達雄[*1] 坪内 孝太[*1]
Tatsuo Yamashita Kota Tsubouchi

[*1] ヤフー株式会社
Yahoo Japan Corporation

This paper describes the stock price prediction with stock bulletin board data with considering users' prediction ability. Users submit their comment to stock bulletin board after adding the tag (I want to buy / sell this stock!!), and the tag shows users' prediction for the stock price. We focus on the users' ability and research the relationship between each users' prediction and real stock price. As the result, some users with high-skill are found and it is confirmed that this approach is really useful for the stock price prediction.

1. 序論

株価掲示板に投稿された情報を解析し、株の値動きを予測する事を目指す。

株価掲示板情報とは、特定の株式銘柄について議論されている情報である。対象銘柄に対するコメントに加え、「買いたい」や「売りたい」といった投資行動がメタデータとしてひもづいている情報をさし、本稿ではそれらのメタデータを銘柄に対する「感情」と呼称する。

坪内ら [坪内 14] は、株価掲示板情報の投稿コメントと、感情スコアがセットになっているデータを解析することで、株価掲示板情報を解析するために基礎となる感情辞書を生成し、感情スコアがついていない記事に対しても感情スコアを予測できる手法の開発に成功した。また、坪内ら[坪内 15] は感情スコアと、実際の株価の動きとの相関を調べる事を目的とし、株価掲示板の情報や株の値動きを時間軸でいくつかのパタンに分け、両者の関連を調べた。

本稿は、後者の結果を詳細に調べ、掲示板に投稿する「ユーザ」に着目した。ユーザが入力する感情スコアの動きと実際の株価の動きとの相関を調べ、ユーザ行動を考慮することが株価予測にどのような影響をあたえるか検証することを目的とする。

2. 分析対象データ

本研究では Yahoo! 株価掲示板情報を対象とする。Yahoo! 株価掲示板は、以下の情報からなる。

- 対象銘柄: 何の銘柄についての情報か
- タグ情報: 強く買いたい、買いたい、様子見、売りたい、強く売りたいの 5 種
- タイトル: 投稿のタイトル
- テキスト: 投稿の本文
- 投稿日時: 記事を投稿した日時
- 投稿者: 記事を投稿したユーザの ID（解析には匿名化）

対象銘柄についての個別の議論がなされている事と、対象銘柄に対する投資意向を示すタグがセットで投稿されている点が他の一般的な掲示板とは異なっているといえる。本研究では過去の研究に期間をあわせ、約 2 年間の投稿を使用した。

連絡先：山下達雄, ヤフー株式会社,
tayamash@yahoo-corp.jp

- 期間: 2012-11-21 から 2014-11-17 (727days)
- 投稿数: 14,015,844
- 銘柄数: 3,584
- ユーザ数: 199,328

以降、5 種のタグ情報「強く買いたい、買いたい、様子見、売りたい、強く売りたい」をそれぞれ「5,4,3,2,1」の数値に置き換え「感情スコア」と呼称する。

3. 調査対象ユーザの選定

株価との相関を調べるためにはある程度の量を投稿しているユーザ、つまり「常連ユーザ」を調査対象とする必要がある。

長期にわたって活動していることに加え、一つの銘柄について投稿活動を継続していることも重要である。つまり、彼らの投稿を銘柄ごとに分けたときに十分な量が確保できることが必須である。

本研究では、常連ユーザの条件の一つとして、「株式市場の営業日に特定の銘柄の掲示板で多くの日数で投稿している」ことを考慮する。

今回の調査では、特定の銘柄の掲示板で市場営業日のうち 50 日分以上投稿しているユーザとした。さらに、感情スコアを利用する前提であるため、投稿のうち 50%以上に感情スコアが付与され、かつ、全てが同じスコアでない（一様ではない）という条件を課した。

以上の条件で抽出した常連ユーザ数は 359、彼らの投稿先掲示板数（銘柄数）は 102、ユーザと銘柄の組み合わせ数は 420 であった。

4. 相関をみる対象

対象銘柄の特定の「1 日」に付与されている感情スコアと、株価との相関を調査する方法について述べる。

感情スコアにおける 1 日の時間の区切り方の定義は、坪内ら[坪内 15]と同様のものを用いた。今回は日本の株式市場を対象とするが、当該市場は、9 時から 15 時まで市場が開催している。そこで、株式市場の開催時間に応じたデータの区切り方を目指す。また、それと比較する株価変化は終値と始値の差を始値で割った変化率を用いた。

「予測」と「観測」の 2 種類を用意した。

4.1 予測

前日の市場がクローズしてから、当日の市場が開くまでの期間の反応のみを集めて、スコア化したもので、当日の株価変化を予測する。

- 感情値スコア計算：前日 15 時～当日 9 時
- 株価の変化：終値と始値の差を始値で割ったもの

4.2 観測

取引中のみにしぼった反応をスコア化したもので、期間内の株価変化と感情値の関係を観測する。

- 感情値スコア計算：当日 9 時～15 時
- 株価の変化：終値と始値の差を始値で割ったもの

5. 比較実験

前述の「予測」「観測」のそれぞれの設定で、常連ユーザのある銘柄の掲示板での 1 日ごとの投稿の感情値スコア平均と株価変化率の相関係数を算出した。

比較対象として、各銘柄の前日の株価変化率と当日の株価変化率の相関係数（A）と各銘柄の全ユーザの投稿による感情スコアと株価変化率の相関係数（B）を計算した。後者の感情スコアは「補正平均感情スコア」（坪内ら[坪内 15]）、つまり対象銘柄の 1 日における全ユーザの感情スコアの平均を用いている。

Aにおける相関係数の最高値は 0.1514（銘柄 6740、以下括弧内は同様）、最低値-0.2261（6862）であった。Bにおいては、「予測」の相関係数の最高・最低値は 0.2483（1928）と-0.1839（9843）で、「観測」では 0.2948（3668）と-0.2744（7912）であった。常連ユーザでの相関係数の最高・最低値は「予測」では 0.4835（9501）、-0.3005（4587）、「観測」では 0.4338（9501）、-0.4500（2931）で、これらはすべて異なるユーザである。

相関の分析にあたって、常連ユーザの相関係数と全ユーザの相関係数の絶対値の差に着目した。

「予測」においては、絶対値の差が大きいほど予測能力が高いと示唆される。表 1 に差が大きい上位 5 件をあげた。たとえば、”3A” のユーザは銘柄 9501 に対しては、非常に高い予測性能を出している事が分かる。”ia”における銘柄 3765, “jS”における銘柄 2138 も同様である。

銘柄	相関係数	投稿数	差	ユーザ ID
9501	0.4835	60	0.4715	3A
3765	0.3457	57	0.3114	ia
2138	0.3110	53	0.2804	jS
6871	0.3041	57	0.2804	cp
4576	0.2932	148	0.2741	4s

表 1:「予測」における上位 5 件

しかし、複数の銘柄に渡って相関の高いユーザは見受けられなかった。比較的高いユーザを表 2 のユーザ"9s","8E"としてあげた。ほとんどはユーザ"FS"のようにほぼ予測ができていない。

一方、「観測」においては、絶対値の差が大きいほど市場の変化に対して激しく反応していると示唆される。表 3 に差が大きい上位 5 件をあげた。市場の変化に抗った感情投稿を行うユーザ（"XS", "LE"）も上位に入っている。

「予測」と異なり「観測」では、複数の銘柄に渡ってほぼ一貫して相関が比較的高いユーザが存在する。例として、表 4 に常連ユーザのデータを上げた。単に市場の動きにそった感情を投稿しているにすぎないともいえるが、予測に際してはこのようなユーザを取り除くことが有用であり、その判別に役立つであろう。また、このユーザは予測では 1 件だけしか現れておらず、これは市場の開いている時間だけ活動している「流れを楽しむ」ユーザであることを示唆している。

ユーザ ID	銘柄	相関係数	投稿数	差
9s	3653	0.0228	80	0.0106
9s	3765	0.0707	54	0.0364
9s	6666	0.1114	69	0.109
9s	2489	0.2254	153	0.1991
8E	3843	0.1094	140	0.0012
8E	3782	0.1372	68	0.0519
8E	6871	0.1603	84	0.1366
FS	9501	-0.0019	83	-0.0101
FS	7211	-0.0084	93	-0.0059
FS	998407	-0.0984	187	0.0478
FS	4755	0.0149	55	-0.0922

表 2: 常連ユーザの相関（予測）

銘柄	相関係数	投稿数	差	ユーザ ID
2121	0.4255	59	0.3499	el
3661	-0.3650	52	0.3449	XS
2931	-0.4500	64	0.2889	LE
2121	0.3567	63	0.2811	dE
3665	0.2985	76	0.2433	K8

表 3:「観測」における上位 5 件

ユーザ ID	銘柄	相関係数	投稿数	差
K8	4587	0.2240	50	0.0291
K8	2497	0.2504	61	0.0043
K8	3668	0.2754	91	-0.0194
K8	7779	0.2906	50	0.1186
K8	3665	0.2985	76	0.2433
K8	2121	0.3175	56	0.2419

表 4: 常連ユーザの相関（観測）

6. 結論

株価掲示板の情報を用いた株価予測においてユーザごとの予測性能を考慮する事の有用性を検証した。

ユーザを常連のみに限定し、考察をした結果、実際の株価に対し予測性能が全般的に高いユーザが存在することが明らかになった。また、全体的に高くはなくても、特定の銘柄においては突出して予測能力の高いユーザや、真逆の行動が株価と相関が出るようなユーザなども確認できた。

このような様々な信頼度をもったユーザの行動を集合知として統合的に加味し、頑強かつ汎用性の高い株価予測システムが構築できる。

参考文献

[坪内 15] 坪内孝太, 山下達雄 ”株価掲示板情報の感情解析と株価との相関の研究”，2015 年度人工知能学会全国大会講演集, 1J5-OS-13b-2in.

[坪内 14] 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”，2014 年度人工知能学会全国大会講演集, 3L4-OS-26b-1in.