

電子掲示板の投稿を用いた株価リターン予測

桜田 亮太[†] 青野 雅樹[‡]

[†] 豊橋技術科学大学情報工学課程 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

[‡] 豊橋技術科学大学情報・知能工学系 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

E-mail: [†] sakurada@kde.cs.tut.ac.jp, [‡] aono@tut.jp

あらまし 近年インターネットの普及により個人の投資機会は増大している．特に新興市場では個人の及ぼす影響力が極めて大きい．そこで本研究では個人の動向に注目するため，電子掲示板を用いた新興市場の株価の予測を行う．Yahoo! 掲示板の株式カテゴリに寄せられた投稿と，株価時系列データを用い，「株価リターン」を重視した予測手法を提案する．加えて掲示板素性を用いた株価予測についても提案する．これらの予測には「個人の重要度」と定義した，個人の株価上昇の期待値を用いる．結果として株価リターン予測について個人の重要度の上位の人物の投稿を参考に株式を購入することで大きなリターンが得られることがわかった．分類予測については株価を7値に分類し予測を行った結果 34%の分類精度となった．実験により「個人の重要度」は大きな株価リターンが期待できる指標であることが確認された．

キーワード 株価リターン予測, 株価予測

Stock Return Prediction using Messages of the Internet Forum

Ryota SAKURADA[†] Masaki AONO[‡]

[†] [‡] Department of Computer Science and Engineering, Toyohashi University of Technology Hibarigaoka 1-1,

Tempaku-cho, Toyohashi-shi, Aichi 441-8580 Japan

E-mail: [†] sakurada@kde.cs.tut.ac.jp, [‡] aono@tut.jp

Keyword Stock Return Prediction, Stock Prediction

1. はじめに

近年インターネットの普及により個人の投資機会は増大している．個人株主数は 2011 年の延べ人数で約 4600 万人[1]と，10 年前と比べ約 1.37 倍と，個人の動向により株価が変動する機会も日増しに高まり，特に新興市場では個人の及ぼす影響力が極めて大きい．表 1 に示す投資部門別売買状況[2][3]をみると，東証一部などの主力株では個人投資家の占める割合は小さいが，新興市場では大きなウェイトを占めていることがわかる．最近では掲示板に書き込みをすると株価が上昇すると話題になった人物が現れるほど個人の影響力は強くなっている．

本研究では掲示板に投稿された書き込みから個々の人物の重要度を算出，その人物の重要度から株価，特にリターンに特化した予測を行う．個人の動向に注目した株価リターン予測を行うため，Yahoo! 掲示板[4]の株式カテゴリを用いて予測を行う．

Yahoo! 掲示板は Yahoo! JAPAN が運営するインターネット電子掲示板で，話題ごとにカテゴリ分けがなされている．中でも株式カテゴリは，銘柄ごとに約 5000 ものトピックが用意されている．Yahoo! 掲示板の株式

カテゴリは株式を扱う掲示板で規模が最大級で，人気の銘柄では一日に約 500 件の書き込みがある．書き込みの際にユーザ登録が必要であるため，ユーザの動向を調査することができる．投稿時に自分の意志を選択する欄(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 段階)があり，投稿を分析する際に活用できるのだが，実際に投稿時に自分の意志を表示するのは全体の 2 割程である．投稿には買いを薦める投稿と売りを進める投稿が混在し，意志を表示していない投稿については，どちらの意志であるかを文脈により推測しなければならない．掲示板には API は搭載されておらず，通常閲覧時と同様の HTML データから情報を取得する必要がある．また，一度に閲覧できる最大投稿数は 5 件であり，多くの記事を取得するには大量のアクセスを要することとなる．

本論文の構成を以下に示す．2 節で関連研究を紹介し，3 節で Yahoo! 掲示板について述べ，4 節において株価リターン予測における提案手法を述べる．5 節でシステム設計について述べ，6 節で提案手法の有用性を調査する評価実験を行う．7 節でまとめと今後の課題について述べる．

表 1 投資部門別売買状況 (2011 年)

	東証一部	東証マザーズ	JASDAQ
法人	9.60%	3.40%	7.10%
海外投資家	59.50%	13.50%	7.60%
証券会社	2.60%	8.40%	20.50%
個人投資家	28.30%	74.70%	64.80%

2. 関連研究

本節では、掲示板など個人の投稿と株式に関係に関する研究及び今回採用する Yahoo! 掲示板に関連する代表的な研究を述べる。

2.1. 掲示板の投稿と株式の関連性の研究

Antweiler(2004)[5]らは、ダウ・ジョーンズ工業株指数とダウ・ジョーンズインターネット指数に採用されている 45 社について Yahoo! と Raging Bull の投稿の投稿数および内容を、自然言語処理により分析し、株式指標との関係を実証的に検証している。結果として、掲示板は株式リターンを予測できないこと、投稿は出来高を予測できること、投稿はボラティリティを予測できることなどを指摘している。

Bollen(2010)[6]らは Twitter の投稿を感情分析することで、3 日後の「ダウ・ジョーンズ工業株価平均」の変化の方向性を 86.7% の精度で予測できるという研究成果を報告した。彼らは投稿を、気分追跡ツール (OpinionFinder) 及び気分プロフィール検査 (POMS) を用い分析する心理学的なアプローチにより予測を行っている。

2.2. Yahoo! 掲示板に関する研究

Yahoo! 掲示板に関する研究として、丸山ら(2008)[7]は、Yahoo! 掲示板の投稿と株式指標との関連性を実証的に指摘している。彼らは、Yahoo! 掲示板より取得した投稿データを自然言語処理及び機械学習により投稿を「弱気」「中立」「強気」の三種類に分類し調査を行った。結果として、掲示板の投稿内容分析を考慮することで株価リターン予測が不可能ではないこと、投稿数と出来高やボラティリティが関連していることなどを報告している。

3. 提案手法

本節では、Yahoo! 掲示板の株式カテゴリに寄せられた投稿と、株価時系列データを用い、「株価リターン」を重視した予測手法を提案する。また、「総投稿数」「個

人の重要度」など 5 素性を掲示板の投稿から算出、一方、「テクニカル指標」11 素性（移動平均、方向性指数など）を株価時系列データから算出し、これら 16 素性を用いた株価予測手法についても提案する。従来研究に対してもっともユニークな素性である「個人の重要度」に関しては、ある個人が投稿をしたのち、3 日後の株価の変動率を人物ごとにまとめ、その平均変動率から算出する。以下では「個人の重要度」の算出、「株価リターンを重視した予測手法」、「掲示板素性を用いた株価予測」を述べる。

3.1. 個人の重要度

個人の重要度は Yahoo! 掲示板に投稿のあるユーザーから、設定した条件を満たしたユーザー全てについて算出する。数値としては理論上 $+\infty \sim -\infty$ まで取り得る。算出方法は、ユーザーの投稿を探索し、書き込みがあった場合 3 日後の株価の変動率を求め、その操作を全ての銘柄、全ての期間において行い変動率の総和から求められる。そこで、 i 番目のユーザー U_i の重要度 $Imp(U_i)$ を、

$$Imp(U_i) = \sum_{i=1}^{\text{ユーザー数}} \sum_{j=1}^{\text{銘柄数}} \sum_{t=1}^{\text{日付}} \left(\frac{x_{j,t+3}}{x_{j,t}} - 1 \right) exist(P_{i,j,t})$$

ここで、

i : i 番目のユーザー

$x(j,t)$: 銘柄 j , 日付 t における株価

$P_{i,j,t}$: 銘柄 j の t 日のユーザー i の書き込み

$$exist(P_{i,j,t}) = \begin{cases} 0 & \text{ユーザーの書き込みが存在しない} \\ 1 & \text{存在する} \end{cases}$$

という式で定義する。この式で与える重要度が大きいほど重要な人物であると想定する。なお、重要度はノイズを除去するため、ある条件を満たしたユーザーのみに適用する。

3.2. 株価リターンを重視した株価リターン予測

4.1 節で算出した個人の重要度を使用し、株価リターンを重視した予測を行う。株価リターンとは株式による収益を上げることが目的とした予測である。予測手法は、個人の重要度のランキングに基づいて、ランキング上位の人物が掲示板に投稿を行った銘柄が株価リターンの大きい銘柄とする。

3.3. 掲示板素性を用いた株価予測

掲示板素性及びテクニカル指標を素性にし、3 日後の株価を予測する。Yahoo! 掲示板から観測される素性として 5 素性を算出、一方、テクニカル指標 11 素性を株価時系列データから算出し、これら 16 素性を用いた株価予測手法について提案する。分類手法には

SVM(libSVM)[8]を使用し、分類は 7 値(大幅上昇, 中幅上昇, 小幅上昇, 微小変動, 小幅下落, 中幅下落, 大幅下落)とする。

Yahoo! 掲示板から観測される素性は以下のとおりである。

- ・当日書き込み数 5 日乖離率
 - ・当日「強く買いたい」数 5 日乖離率
 - ・当日「強く売りたい」数 5 日乖離率
 - ・ID の重要度から算出した今後の期待値
 - ・当日の投稿の最も重要度の高い人物の重要度
- 一方、テクニカル指標からは、
- ・当日出来高 5 日乖離率
 - ・当日売買代金 5 日乖離率
 - ・直近 25 日移動平均標準偏差
 - ・5, 25, 50 日株価移動平均線乖離率
 - ・サイコロジカルライン
 - ・ADX
 - ・MFI
 - ・CCI
 - ・ATR

以上の素性を採用, その中から有効と思われる素性のみに用いて実験を行う。

4. システム設計

提案手法の実装にあたり Java 言語を用い, Yahoo! 掲示板の個人の人物の重要度に基づいた株価リターン予測システム及び掲示板素性を用いた株価予測システムを設計した。

4.1. Yahoo! 掲示板の投稿データ収集

Yahoo! 掲示板には API が存在しないため、投稿収集はユーザが閲覧するページにアクセスすることで行う。1 度のアクセスで最大 5 件のメッセージが取得できる。本研究で株価リターン予測に用いるデータは、個人の影響が受けやすい東証マザーズと JASDAQ の新興 2 市場の銘柄に限定する。

また、株価時系列データは「株価データ ダウンロードサイト」[9]より取得する。本 Web サイトは CSV 形式で東京証券取引所及び大阪証券取引所の取扱い銘柄の株価時系列データ(始値, 高値, 安値, 終値, 出来高, 売買代金)が取得可能である。取得可能期間は 2007 年 1 月 1 日から当日までの株価となっている。取得された掲示板の投稿データは HTML 形式のため、必要な情報のみを抽出する必要がある。HTML から必要情報の抽出には文字列処理操作を用いた。

4.2. 個人の重要度の算出

個人(各ユーザ)の重要度を算出する。個人の重要度

は Yahoo! 掲示板に投稿しているユーザごとに与え、各ユーザの期待値を計る指標と定義する。重要度の算出には 5.1 で取得した掲示板投稿データ及び株価時系列データを使用する。なお、今回算出に使用する投稿は、投稿時の気持ちに「強く買いたい」または「買いたい」が付加されている投稿に限定する。投稿には売りを薦めている投稿も見受けられ、そのような投稿を含めて算出すると正確な個人の重要度が算出できない。そこで、気持ちが付加されている投稿に限定することで、正確な算出が出来るようになる。

4.3. 個人の重要度に基づいた株価リターン予測

個人の重要度ランキング上位の人物が投稿をした銘柄を高株価リターン銘柄と判定する。また、評価実験を行うため、個人の重要度に基づき株式売買シミュレーションシステムを作成した。シミュレーションは重要度上位 N 人(N は任意の数字)までの人物が投稿した銘柄を、投稿の翌日の始値で購入、2 日後の終値で売却するシミュレーションを行う。

4.4. 掲示板素性を用いた株価予測

4.3 節で記述した掲示板素性及びテクニカル指標の素性を用いて SVM 分類器により 3 日後の株価を 7 値に分類するシステムを作成する。掲示板素性については株価時系列データより、テクニカル指標の素性については株価時系列データより算出する。

5. 評価実験

提案手法の有効性を検証するために、二つの実験を行った。

一つ目は個人の重要度の有用性を調査するため、株価リターン予測した結果を株式売買シミュレーションを行うことで合計損益率を求める実験である。実験方法としては、まず個人の重要度を算出、その後各人物について投稿検知した場合、翌日の始値で購入、その 2 日後の終値で売却すると仮定したときの損益率を加算することで各人物についての合計損益率を求める。個人の重要度のランキング上位 N 人の投稿を考慮した場合の合計損益率を算出し、個人の重要度の有用性を調査する。データセットは 2010 年, 2011 年, 2012 年と 3 分割し、二つを訓練データ、一つをテストデータとして評価を行う。

二つ目として個人の重要度やその他掲示板及びテクニカル指標を素性として株価予測を行う実験である。提案する 16 素性について有効と思われる素性を採用し SVM 分類器により予測を行う。

なお、いずれの実験も期間は 2010 年 1 月 1 日から 2012 年 12 月 10 日、採用銘柄は東証マザーズ及び

JASDAQ の新興 2 市場とする。データセットとして Yahoo! 掲示板の投稿を約 253 万件、株式時系列データの 1103 銘柄の日毎のデータを取得した。評価方法は、一つ目の実験については合計損益率及び DCG、二つ目については適合率とする。なお、DCG グラフのベースラインは ID をランダムに抽出し算出した結果の 10 回の平均とする。

5.1. 実験結果

5.1.1. 株価リターン予測の合計損益率

各個人の損益率を算出し、上位 N 人までを考慮した場合の損益率を以下の図 1 に示す。なお、2012 年のグラフは 2011 年と 2010 年を訓練データとし 2012 年をテストデータとしている。同様に 2011 年のグラフは 2010 年と 2012 年を訓練データ、2011 年をテストデータ、2010 年のグラフは 2011 年と 2012 年を訓練データとし、2010 年をテストデータとしている。

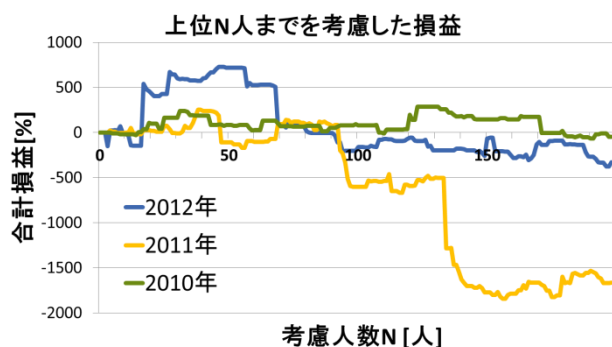


図 1 上位 N 人までを考慮した損益率

結果を見ると、全体としてのグラフとなっていることがわかる。その中で上位に目を向けると利益が多く出ている部分が多く見受けられる。つまり個人の重要度上位と判定された人物は利益が多く出ていることがわかる。図 2 に DCG 値変換したグラフを示す。DCG はランキング上位ほど強く効く指標となっている。

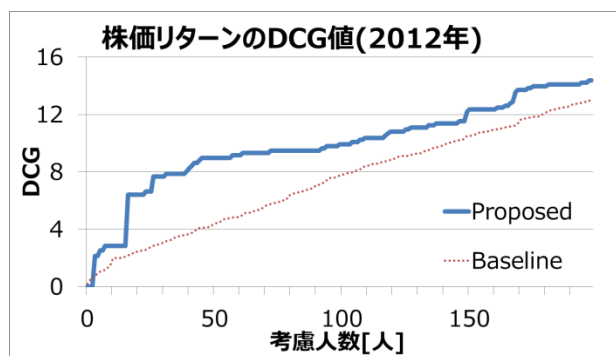


図 2 2012 年の株価リターンの DCG 値

結果のグラフを見ると提案手法がベースラインを

大幅に上回っていることがわかる。以上の結果より、個人の重要度に基づいた株価リターン予測は大きな株価リターンが得られることがわかった。

5.1.2. 掲示板素性を用いた株価予測

二つ目の評価実験として SVM による株価 7 値分類を行った。素性は提案する 16 素性の中から、試行錯誤により有効な素性のみ限定し予測を行う。

以上素性を使用し SVM 分類器により 7 値分類を行った結果、分類精度は 34% となった。現状として微小変動やその前後に正解ラベルに訓練データが多く振られていて、結果として予測の方も微小変動とその前後に分類されることが多くなっている。分類精度向上のために、分類の閾値を全ての正解ラベルで均等になるように調整する必要がある。

6. 結果・今後の課題

実験結果から、重要度の高いユーザの動向に着目した株価予測は、モデルの選定方法にもよるが極めてリターンの大きい株価予測となることがわかった。掲示板素性を用いた株価予測については、34% の結果となった。

今後の課題として、予測の精度を向上させるため、重要度算出モデルの改善、調査期間の拡大、投稿時の意思の予測の精度の向上、予測に考慮すべきテクニカル指標の調査、調査銘柄の拡大、掲示板素性を用いた株価予測については正解ラベルを均等になるように振り分ける等で高精度の株価予測を行うことが挙げられる。

参 考 文 献

- [1] 東証：株式分布状況調査,
www.tse.or.jp/market/data/examination/distribute/.
- [2] 東証：投資部門別売買状況,
www.tse.or.jp/market/data/sector/index.html.
- [3] JASDAQ: 投資部門別売買状況,
www.ose.or.jp/market/statistics/investment_jasdaq.
- [4] Yahoo! 掲示板, /messages.yahoo.co.jp/.
- [5] Antweiler W, Frank MZ, "Is all that talk just noise? the information content of internet stock message boards.", The J of Finance 59 2004.
- [6] Johan Bollen, Huina Mao, Xiao-Jun Zeng, "Twitter mood predicts the stock market", arXiv.org 2010.
- [7] 丸山健, 梅原英一, 諏訪博彦, 太田敏澄, "インターネット株式掲示板の投稿内容と株式市場の関係", 証券アナリストジャーナル Vol.46 2008.
- [8] LIBSVM -- A Library for Support Vector Machines,
www.csie.ntu.edu.tw/~cjlin/libsvm/
- [9] 株価データ ダウンロードサイト,
k-db.com/site/default.aspx.