ニュース記事の見出しを利用した取引高予測の試み

Towards Prediction of Volume of Transactions Using the Related News Articles

吉田 稔 [1*]　中川 裕志 [1]　石田 智也 [2]　中嶋 啓浩 [2]
松井 藤五郎 [3]　和泉 潔 [4,5]　池田 翔 [4]　本多隆虎 [6]
Minoru Yoshida[1]　Hiroshi Nakagawa[1]　Tomonari Ishida[2]　Akihiro Nakashima[2]
Tohgoroh Matsui[3]　Kiyoshi Izumi[4,5]　Sho Ikeda[6]　Takatora Honda[7]

[1] 東京大学情報基盤センター
[1] Information Technology Center, The University of Tokyo
[2] 野村證券株式会社
[2] Nomura Securities Co.,Ltd.
[3] 中部大学生命健康科学部, 工学部
[3] College of Life and Health Science, and College of Engineering, Chubu University
[4] 東京大学大学院工学系研究科
[4] School of Engineering, The University of Tokyo
[5] JST さきがけ
[5] PRESTO, JST
[6] 東京大学 工学部
[6] School of Engineering, The University of Tokyo
[7] 早稲田大学大学院基幹理工学研究科
[7] Graduate School of Fundamental Science and Engineering, Waseda University

Abstract: We report our on-going research to develop a system to predict volume of transactions of a brand using news articles related to the brand. We investigated the effect of keyword-based article selection and online learning algorithms on prediction accuracies.

1　はじめに

現在我々は，ニュース記事を用いて株の取引高を予測する手法について研究を進めている．

「テキストと株価の関係」に関する従来研究としては，例えば，小川ら [1] は，新聞記事をルールベースでテーマ分類し，テーマが株価動向にどのような影響を及ぼすかを解析した．高橋 [2] らは，ヘッドラインニュースを情報源とし，Naive Bayes 法により分類されたニュースの Good/Bad のラベルと，ニュース配信時の株価リターンとの関連を調査し，有意な関連があったと報告している．また，和泉 [3] らは，日本銀行の金融経済月報を題材として経済市場分析を試みている．単語共起関係抽出ツール KeyGraph[4] を用い，抽出された共起パターンと月末における金利の関係を主成分分析によって解析し，金融経済月報が市場金利に対して，一定の説明力を持つ可能性が高いことを示した．また，[5] においては，国際金融情報センターの発行する市場解説記事を自動分類した結果を，人工市場の分析に利用する試みを行っている．張 [6] らは，株価の変動を記事や語句の評価値の推定に用い，係り受け関係を使うことで良好な結果を得ている．

しかしながら，テキストからの個別銘柄の株価予測に関しては，高い精度で実現することの難しさが既存の研究でも指摘されている [7] ため，本研究では，最初の目標として，比較的予測が容易と考えられる,「取引高」に着目し，その予測手法について研究を行っている．

[8] において，我々は，ニュース記事のトピック推定とそれに基づく記事クラスタリングを行い，取引高予測に有用な話題とそうでない話題があることを示唆す

*連絡先：東京大学情報基盤センター
〒 113-0033 文京区本郷 7-3-1
E-mail: mino@r.dl.itc.u-tokyo.ac.jp

る結果を得たことを報告した．この「話題による取引高の偏り」について調査するため，得られた話題の中でも，特に，高い有用性を示唆していた「金銭関係の話題」に着目し，それを扱うニュース記事をもとに取引高の予測を試みる．

2 問題設定

入力として，

D_t: ある銘柄に関する，日付 t の記事集合

v_t: ある銘柄の，日付 t の取引高

が与えられるとする．ここで，取引の無い日付については，D も v も定義されず，t は，日付そのものではなく，取引の有った日付を古い順に並べた順番を表すものとする．[1]

ここで，取引高 v_t に対し，前日 N 日と比較しての増加傾向，減少傾向を表す値 y_t を，以下で定義する．

$$y_t = \frac{v_t}{a_t}$$

ただし，

$$a_t = \frac{\sum_{t-N \leq t' \leq t-1} v'_t}{N}$$

（現在は，$N = 5$ を設定している．）本研究の目的は，D_t が与えられたときに y_t に関する予測を行うことであるが，問題設定としては，値の大小を 2 値分類することを考える．y_t は比率のため，1.0 より大きければ取引高の「増加傾向」，小さければ取引高の「減少傾向」を表していると考えられる．すなわち，

$$z_t = sign(y_t - 1.0)$$

を定義し（ここで $sign(x)$ は，$x \geq 1.0$ ならば $+1$，さもなくば -1 を返す関数とする．[2]），z_t の予測を行う．

すなわち，本稿の提案システムのタスクは，入力 $\mathcal{D} = (D_1, D_2, \ldots)$ と $\mathcal{V} = (v_1, v_2, \ldots)$ が与えられたときに，$\mathcal{Z} = (z_1, z_2, \ldots)$ を返すことである．

3 予測手法

本稿では，「記事の絞り込み」と「オンライン学習」の 2 つの手法について，取引高上昇の予測に与える影響を調査する．以下，両者について順に解説する．

[1] 取引の無い日の記事は，翌取引日の t に対応づけられるものとする．

[2] 定義から，y_t が厳密に 1.0 となる可能性は低いため，ここでは $z_t = 0$ となるケースは考えていない．

3.1 キーワードによる記事の絞り込み

記事の絞り込みは，単純なキーワードマッチングにより行う．具体的には，文字列「円」を含む記事を，金銭関係の記事 (以下，“記事集合 A” と呼ぶ) として用いる．

ここで，金銭関係の記事に「決算発表」関連の記事が多く含まれていることを考慮し，決算関連の記事を除いた “記事集合 B” を作成する．これは，決算発表は，予め実施日が確定していることが普通であり，ニュース記事による取引高予測の対象としては不適切である可能性がある[3]ためである．記事集合 B の生成も，キーワードを用いて行う．具体的には，**禁止文字列**の集合を用意し，記事集合 A の中で禁止文字列を**含まない記事**を集め，記事集合 B とする．禁止文字列集合は，現在は {”益”,”字”,”売上”} となっている．また，参考のため，逆にこれらの禁止文字列を**含む**記事を集めた “記事集合 C” も作成する．

3.2 オンライン学習による取引高上昇の予測

本研究が目標とするのは,「ある銘柄に関し，その銘柄に関するニュースを見て，当日の取引高が上昇する可能性が高いときに，それを教えてくれるシステム」である．このとき，入力テキスト（ニュース記事）は時系列に並んでおり，システムは時系列順にテキストを受け取り，取引高予測を行うことになる．

我々は，この目的に合致した取引高予測の学習・分類の枠組みとして，**オンライン学習**を採用する．**オンライン学習**では，一般に，データを順次受け取り，そのたびにパラメータの更新を行う．パラメータ更新の際にデータ全体を見ることのできる**バッチ学習**に比べ，学習速度，必要記憶量の少なさの点で優れており，また，直近のデータへの適応性という点でも，本研究のような，**時系列的データを利用した学習**に適している．本研究では，実装の容易性と性能の点を考慮し，以下で説明する Online Bayes Point Machine (OBPM) [9] を採用した．

3.2.1 Online Bayes Point Machine

Online Bayes Point Machine (OBPM) は，Bayes Point Machine (BPM) [10] を応用したオンライン学習アルゴリズムである．

Bayes Point Machine は，学習データが与えられた時に，それと矛盾しない**可能なパラメータ**空間内の**重心**を求めるアルゴリズムである．Herbrich ら [10] は，複数のパーセプトロン学習器を並列に実行し，それぞ

[3] ユーザーは，わざわざ新聞記事をチェックせずとも，決算発表の日付のみをチェックすればよいことになるため．

れにおいて学習された異なるパラメータの平均を取ることで，重心を近似的に求めるアルゴリズムを提案した．このとき，学習器毎に異なるパラメータを得るために，各学習器に異なる順序のデータ列を与える．しかし，オンライン学習の枠組みにおいては，データの順序を自由に入れ替えることができないため，この手法が適用できない．

これに対し，OBPM では,「各データを学習に用いるか否かをランダムに決定する」という仕組みを導入することによって，データの順序が一定の場合でも，各学習器毎に異なるパラメータの学習を実現した．具体的には，N 個の学習器毎に,「確率 p で，$\mathbf{x}_i$ を用いて学習を行う（確率 $1-p$ で，$\mathbf{x}_i$ を無視する）」という操作でパラメータ学習を行い，学習器毎に異なるパラメータを得る．その後，全体の学習器のパラメータを，学習された N 個のパラメータの平均として求める．

各学習器としては，単純パーセプトロンを採用する．単純パーセプトロンは，線形分離器 $\mathbf{w}$ を学習するアルゴリズムである．具体的には，ベクトルで表現された入力データ $\mathbf{x}$ に対し[4]，出力ラベル $z \in \{-1, 1\}$ を，線形分離式 $z = sign(\mathbf{w} \cdot \mathbf{x})$ により予測する学習器であり，学習結果はベクトル $\mathbf{w}$ で表現される．学習は，新しい学習データ $\mathbf{x}_i$ が与えられる度に，

1. 現在の学習器で正しく $\mathbf{x}_i$ がラベル付けされる場合，何もしない．
2. 正しくラベル付けされない場合，$\mathbf{w}$ に $z \cdot \mathbf{x}$ を加える．

という手順で $\mathbf{w}$ を更新することにより行う．

4 実験

4.1 実験設定

対象となる銘柄の取引高と，各銘柄名を見出しに持つ新聞記事（2005-2007 年日経新聞の記事）を，日付で対応付け，記事の存在する日付について，z_t の予測を行った．2005 年，2006 年の記事を学習データ，2007 年の記事をテストデータとして，正解率 (Accuracy)・精度 (Precision)・再現率 (Recall)・F-value(精度と再現率の調和平均) を測定した．各記事のタイトルを CaboCha[11] の解析結果を利用し単語に分割し，その中から，予備実験の結果有用性の高かった「動詞」（ただし,「サ変接続名詞」を含む）を素性として用いた．

使用した銘柄は，トヨタ自動車，本田技研工業，ソニー，東芝，三菱電機，三菱商事，三菱重工業の 7 銘柄である．テストデータの 1 年間 7 銘柄のうち，記事の抽出できた日付は延べ 1011 日，このうち,「円」を含む記事 (記事集合 A) を持つ日付は延べ 148 日あり，決算関係の記事 (記事集合 C) に限ると 33 日，決算関係の記事を除いた場合 (記事集合 B) は延べ 124 日となった[5]．記事を絞り込むことにより，本来予測してほしい「取引高が上昇する日」を大幅に見逃してしまうことになり，再現率が下がることに注意されたい．本研究の目的は，記事の絞り込みや，機械学習による予測を用い，精度を上昇させることができるか否かを確認することにある．

OBPM の学習器数 N は 100 とし，データ使用確率 p は 0.5 とした．同一の設定での実験は 10 回行い，その平均をとった．

4.2 結果

結果を表 1 に示す．記事を絞り込むことにより，記事全体を利用した場合と比べ，予測精度が上昇することを確認した．

しかしながら，これは主に，絞り込まれた記事集合に「取引高上昇日」が多く含まれているためと考えられる．このため,「選択された記事集合すべてについて取引高上昇と判定する」ベースラインアルゴリズムが考えられる．このベースラインに従うと，精度は，全記事について 0.45，記事集合 A について 0.60，記事集合 B について 0.59，記事集合 C について 0.67 となり，いずれも，OBPM を用いた場合と比べてそれほど遜色のない精度となる．すなわち，精度上昇には主に「キーワードによる記事絞り込み」が貢献しており，OBPM については，それほどの貢献は無かったと考えられる．

5 おわりに

現在我々が進めている，取引高の上昇予測に新聞記事を用いる研究について報告した．記事集合を特定のキーワードで絞り込むことにより，予測の精度（precision）を上昇させる効果があることを確認し，また，OBPM による，オンライン学習の応用についても紹介を行った．記事の絞り込み，オンライン学習のいずれにおいても，まだ多くの課題が残されており，より適切な記事絞り込み手法の研究，オンライン学習の精度向上の両者について，今後も研究を進めていく予定である．

[4] 本研究においては，$\mathbf{x}$ は各文書 D を bag of words で表現したベクトル

[5] データは「各日付」を「その日付の，関連記事の集合」で表現しているため，排他的に分割された記事をもとにデータ（日付集合）を作成しても，両データ間で共通する日付が存在することに注意されたい．

表 1: 実験結果

キーワード	データ数	Accuracy	Precision	Recall	F-measure
全記事	1011	0.50 (504.7/1011)	0.44 (179.0/405.3)	0.39 (179/459)	0.41
記事集合 A	148	0.46 (68.0/148)	0.57 (38.1/67.2)	0.43 (38.1/89)	0.49
記事集合 B	124	0.54 (66.8/124)	0.65 (34.6/53.4)	0.47 (34.6/73)	0.55
記事集合 C	33	0.54 (17.8/33)	0.66 (13.8/20.8)	0.63 (13.8/22)	0.64

参考文献

[1] 小川 知也，渡部 勇．株価データと新聞記事からのマイニング．情報処理学会 自然言語処理研究会研究報告 2000-NL-142-19 (2000)

[2] 高橋悟，高橋大志，津田和彦．ヘッドラインニュースに対する株価の反応について．第 6 回行動経済学ワークショップ. (2007)

[3] 和泉潔，後藤卓，松井藤五郎. テキスト情報を用いた金融市場分析の試み. 人工知能学会第 22 回全国大会 (2008)

[4] 大澤幸生, ネルス E. ベンソン, 谷内田正彦．KeyGraph. ：語の共起グラフの分割・統合によるキーワード抽出．電子情報通信学会論文誌 D-1, Vol.J82-D-1, No.2, pp.391-400, (1992)

[5] 和泉潔，松井宏樹，松尾豊．人工市場とテキストマイニングの融合による市場分析．人工知能学会誌, Vol. 22, No. 4, pp. 397-404 (2007)

[6] 張 へい，松原 茂樹，株価データに基づく新聞記事の評価，第 22 回人工知能学会全国大会論文集, (2008)

[7] Moshe Koppel and Itai Shtrimberg. Good News or Bad News? Let the Market Decide. Computing Attitude and Affect in Text: Theory and Applications, 297-301 (2006)

[8] 取引高とニュース記事の関連性の分析. 吉田稔, 中川裕志, 松井藤五郎, 和泉潔, 石田智也, 中嶋啓浩. 第 4 回ファイナンスにおける人工知能応用研究会（SIG-FIN）, pp. 60–64 (2010).

[9] Edward Harrington, Ralf Herbrich, Jyrki Kivinen, John C. Platt, Robert C. Williamson. Online Bayes Point Machines. PAKDD 2003, pp. 241–252 (2003)

[10] Ralf Herbrich, Thore Graepel, Colin Campbell. Bayes Point Machines. Journal of Machine Learning Research 1, pp. 245–279 (2001)

[11] Taku Kudo and Yuji Matsumoto, Japanese Dependency Analyisis using Cascaded Chunking, Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002), pp.63–69 (2002)