為替ニュース記事を用いたSVMによる株価予測

Using news articles of foreign exchange to predict stock prices by SVMs

石黒 祐輔[1*]　ダヌシカ ボレガラ[2]　伊庭 斉志[1]
Yusuke Ishiguro[1], Danushka Bollegala[2], Hitoshi Iba[1]

[1] 東京大学大学院 情報理工学系研究科
[1] School of Information Science and Technology, The University of Tokyo
[2] リヴァプール大学 電気工学・電子工学・コンピュータ科学科
[2] Department of Computer Science, The University of Liverpool

Abstract: We propose a method to predict stock prices by SVMs using news on foreign exchange rates on the Web. Our method targets Japanese news and stocks. We compare several parameters for predicting the span, and fixed span to 50 minutes. We then apply the proposed method to 15 different stock issues from Nikkei 225. Although our preliminary results are encouraging, we plan to further improve the accuracy of our approach in future.

1　はじめに

近年高度な情報社会化が進み, ウェブ上に大量の情報が蓄えられるようになった. そこでそれらビッグデータを活用した研究が盛んに行われており, ビッグデータから何らかの知識を抽出するデータマイニングが特に盛り上がりを見せている. それは金融分野でも例外ではなく, その中の 1 つに金融テキストマイニングがある. もしこの分野の研究が上手くいけば, 金融モデルの解明に繋がる可能性があるので, 重要性は増している.

1.1　関連研究

関連研究として, ウェブ上のニュースをテキストマイニングし, 株価予測を行なっている国内外の研究を紹介する.

1.1.1　英語のウェブニュースによる株価株価予測

Shumaker ら [1, 2] は, 英語版の Yahoo! Finance のニュースをテキストマイニングし, SVR で学習したモデルの評価をした. 株価の予測は, ニュースが発表された 20 分後の株価と短期の予測を行なっている.

テキストマイニングの手法としては, 記事内の固有名詞の出現回数, 主観性判断, 極性判断を組み合わせた.

評価方法としてはトレンド予測, 変動率予測, 仮想取引の 3 つの視点から評価している. 結果としては複雑なものとなっているのでここでは詳細を省略する.

1.1.2　日本語のウェブニュースによる株価予測

辻ら [3] はウェブ上にある日本の全国紙のニュースをもとに SVM と SVR によって株価を予測する研究を行った. ただし株価変化の予測単位は 1 日と比較的長いスパンとなっている.

テキストマイニングの手法は, ニュースより企業名と経済用語を含む辞書の単語を抽出し, どちらも入っていないニュースは排除している. また文章の構文解析も行い, 企業名と係り受け語なども活用している.

株価のトレンドと変動率を予測する実験を別々に行い, ベースラインとして記事の単語の生起情報のみで学習した場合と比較した. トレンド予測に関しては企業名と単語の生起情報の組み合わせの時最大の 58.89%となり, ベースラインと比べて正答率が 5%程度上昇したが, 係り受け語などを付加した場合は逆に正答率は減少している. また変動率の予測に関しては, ベースラインと比較して僅かに性能が改善された程度だったため, 考察は行われていない.

1.2　研究の目的

まず輸出入を行う企業にとっては為替が非常に大事である. それはよく知られている通り, 円安の場面では

*連絡先： 東京大学大学院 情報理工学系研究科 伊庭研究室
東京都文京区本郷 7-3-1 東京大学本郷キャンパス 工学部 2 号館 122B2
E-mail: ishiguro@iba.t.u-tokyo.ac.jp

輸出企業にとってプラス, 輸入企業にとってマイナスに働き, 逆に円高の場合は輸出企業にとってマイナス, 輸入企業にとってプラスに働くためである. もちろん企業によってはヘッジ取引などを用いて為替変動に対策を取ることもあり, その影響も小さくなったり, もしくは逆に働く可能性もある.

そこで本研究では, 為替のニュースと為替市況の相場観の関係を学習することにより, 為替相場と株価の関係を分析することを目的とする. 手法としては Shumaker らの手法をベースとするが, 論文の中では細かく言及されていない部分が多いことや, そもそも対象としている言語が違うため, 改良を試みるために様々な独自の手法を追加した.

2 提案手法

本稿で提案する手法は, 為替市況に関するニュースをテキストマイニングにすることによって, 輸出入企業の株価の方向性を予測することを目標とする.

提案手法の全体像は図 1 のようになる. 銘柄ごとにこのシステムを適用することによって, その銘柄専用の SVM のモデルが作られる. 以降, 各部分を説明する.

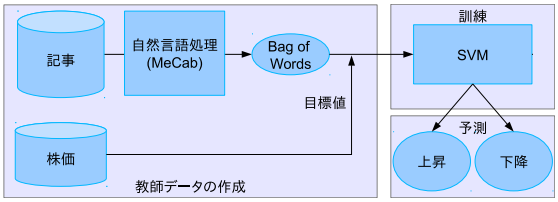

図 1: 提案手法の全体像

2.1 サポートベクターマシン

一番メインとなる機械学習のアルゴリズムとして, 本研究ではサポートベクターマシン (SVM)[4] を用いることにし, またそのライブラリとして libsvm[5] を用いた. SVM は教師有り機械学習のアルゴリズムとして非常にメジャーなもので, その汎用性の高さと結果の良さから採用されることも多いが, デメリットとしては計算量が他のアルゴリズムと比較して大きいことが挙げられる.

2.2 テキスト前処理

本研究では, 日本語の分かち書きを行うために MeCab[6] を用いた. まず入力テキストを MeCab によって分かち書きに分解し, 単語ごとの出現回数を数えたものを記事の表現ベクトルの一種である Bag-of-Words (BoW)[7] と呼ばれるものに変換する. この際, 数字, 記号, 助詞など単語単体で意味をなさないものは破棄する. こうしてできた BoW をその記事の特徴ベクトルとする.

ここで MeCab に入力するテキストは通常記事本文とすることが自然であるが, 記事本文をそのまま BoW に変換すると, 本文中の重要ではない単語や, 単語単位で見た時に本来の文脈と異なる特徴量に入ってしまう可能性がある. そこで入力するテキストを記事本文ではなく, 記事のタイトルとすることによって, そのニュースの要約をノイズ無く得ることができると考えられるので, 本研究では記事のタイトルを用いた.

最後に, 全記事における個々の単語の出現確率を調べ, それが 2.5%以上 80%以下の単語は BoW から取り除く. これは出現確率が低いものは学習しにくく, 出現確率が高いものは予測結果に影響しないと考えるためである.

2.3 教師データ

SVM に入力する教師データとして, 記事が配信された時刻から見て x 分後に株価が上がった場合は”+”を, 下がった場合は”-”を, 変化しなかった場合は”0”を与えることを考えるのが自然であるが, この場合 3 クラスの分類問題と 2 クラスに比べて予測が難しくなってしまう. また大なり小なり株価が変動しているのが自然であるので”0”のクラスが少なくなり, 不均衡データとなってしまい SVM の学習が難しくなる.

そこで学習を効率よく行わせるために, 株価が変動しなかった場合のデータは採用しないことにする. またこの方法のメリットとしては, 結果が見やすくなることも挙げられる.

3 実験

実験では, まず予測期間を固定するために, 銘柄を 1 社に固定して様々な予測期間を試し, 最も成績が良かったものを採用する. そして次に, 様々な銘柄に対して提案手法を適用し, その結果を評価する.

3.1 データ

ニュースデータとして, 日経 QUICK ニュース社より平日の 8 時半, 10 時, 12 時, 17 時に配信されている為替市況のニュース [1] を用いた. ただし 17 時のニュースは東証の立会時間外なので用いない. [2]

[1] 必ずしもその時間に配信されるわけではなく, 通常 10〜20 分程度遅れる

[2] 厳密には 8 時半も時間外であるが, 30 分後より開くので採用した

株価データは, 日経平均を構成する 225 銘柄の 1 分足のものを取得して用いた.

なお上述の為替市況のニュースが配信開始されたのが比較的新しいため, データ量としては十分ではないが以降の実験では訓練セットとして 2012 年 11 月から 2013 年 11 月まで, テストセットとして 2013 年 12 月のデータを用いることにする.

3.2 評価方法

各実験での評価方法としては, 以下の 3 つを用いる.

- 訓練セット正答率
- テストセット正答率
- 平均 F 値

このうち平均 F 値は各クラスのマクロ平均, すなわち正例と負例のそれぞれで F 値を計算し, その平均をとることにする. これは正例の F 値のみの場合, 正例と負例が同確率だと仮定するならば予測を全て正例とする分類器でも高い F 値となるためである.

株価が上昇する確率と下降する確率がそれぞれ等しいとし, それをランダムに予測したのならば, テストセット正答率と平均 F 値はそれぞれ 0.5 となるので, これがベースラインとなる.

3.3 予測期間の実験

まず, ニュースが配信されてから何分後の株価を予測するかを決めるために, トヨタ自動車株式会社の株価を用いて実験した. 結果としては, 図 2 のようになった.

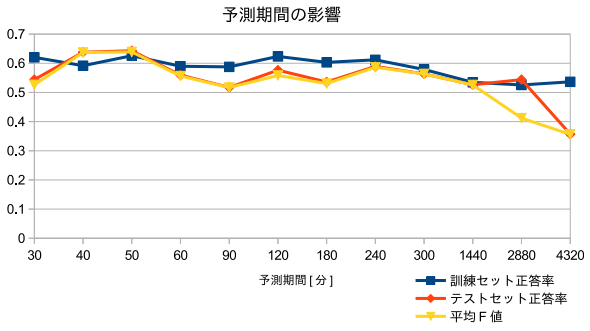

図 2: 予測期間の違いによる影響

この結果では 50 分後の株価を予測したときが最も成績が良く, テストセットの正答率と平均 F 値がそれぞれ 64.28%と 0.6387 となった. また 40 分の場合でもそれぞれ 63.79%と 0.6370 と, ほぼ 50 分と同等の成績であった. これらのテストセット正答率は, 共に有意水準 5%の二項検定で有意であった. 全体的な傾向としては, 50 分ピークに短期間の予測は他より高く, 特に予測期間が 1 日を超えると成績は悪化し, 0.5 を下回る成績となっている. これより, 提案手法は短期間の予測に適していると考えられる.

3.4 企業の違いによる予測の影響

前の実験の結果, 予測期間を 50 分に固定できたので, 今度は一般的に言われている輸出企業, 輸入企業, 内需企業のそれぞれを日経平均の中から適当に 5 銘柄ほど抽出し, 提案手法を適用したときの結果の差を見ることにする.

用いた銘柄は以下の通りである. (株式会社略)

- 輸出企業
  トヨタ自動車, ソニー, ニコン, ファナック, ブリヂストン
- 輸入企業
  丸紅, 東京ガス, 東京電力, JX 日鉱日石エネルギー, 日清製粉グループ
- 内需企業
  鹿島建設, 三菱 UFJ フィナンシャル・グループ, 東日本旅客鉄道, セブン&アイ・ホールディングス, 東京ドーム

結果は表 1 のようになった.

テストセット正答率や平均 F 値が 0.55 を超えてるような特に成績が良かったものは, トヨタ自動車, ブリヂストン, 東京ガス, 鹿島建設の 4 社であった. このうちテストセット正答率が有意水準 5%の二項検定で有意であったものは, トヨタ自動車のみであった. 東京ガスの方が正答率は高いのに有意で無かったのは, 総度数 N が小さかったためである. 逆に成績が悪かったものは, 東京電力, 東日本旅客鉄道, セブン&アイ・ホールディングスの 3 社であるが, それでも 0.4 を下回るほど極端に悪くは無かったことや, 有意水準 5%で有意と言えないことを考えると, 誤差の範囲と考えられる.

もし輸出入企業は為替相場の影響を受け, 内需企業は影響を受けないという仮説が正しいとするならば, 輸出入企業の結果は 0.5 を超え, 内需企業の結果は 0.5 前後になるはずであるが, 結果としては分類に関係なく銘柄によって結果がバラバラである.

これはデータ不足などの理由により学習が上手くいっていない可能性があるが, そもそも例え内需企業といえども為替相場に影響を受ける場合や, 輸出入企業といえども為替相場に影響を受けない場合もあるので, 実の為替レートとの相関性も含めて検討する必要がある.

銘柄	訓練セット 正答率	テストセット 正答率	平均 F 値
トヨタ自動車	62.58%	64.29%*	0.6387
ソニー	58.66%	53.57%	0.5333
ニコン	59.19%	51.72%	0.4874
ファナック	52.43%	49.15%	0.4878
ブリヂストン	55.04%	58.93%	0.5860
丸紅	56.88%	51.16%	0.5019
東京ガス	52.05%	64.44%	0.6000
東京電力	56.52%	46.94%	0.4292
JX 日鉱日石エネルギー	60.5%	54.35%	0.5158
日清製粉グループ	57.21%	48.21%	0.4807
鹿島建設	60.34%	58.33%	0.5804
三菱 UFJ ファイナンシャル・グループ	58.68%	52.08%	0.4829
東日本旅客鉄道	55.61%	46.43%	0.4531
セブン&アイ・ホールディングス	54.73%	45.76%	0.4499
東京ドーム	54.28%	49.06%	0.4787
平均	56.98%	52.96%	0.5137

*: 有意水準 5%の二項検定で有意

表 1: 輸出・輸入・内需企業による予測結果の違い

4 おわりに

4.1 まとめ

本研究では, 為替に関するニュースから為替市況の相場を SVM によって機械学習し, 輸出入企業を含む様々な企業の株価の方向性を予測した. 予測期間の調整をする実験で有意と思われる傾向は見られたが, その後の銘柄間の差を調べる実験で事前予想と異なる結果が得られた, 今後データの更なる追加や手法の工夫によって, 為替に関するニュースから株価を予測する可能性を残すことができた.

4.2 今後の課題

この研究の今後の課題や展望としては, 以下の点が挙げられる.

- 為替相場と株価の相関性が結果と一致するか
  もし単純な「輸出入企業は為替相場の影響を受け, 内需企業は影響を受けない」という仮定と矛盾する実験結果が得られた. 自然に考えるのならば, 為替相場の影響の受け方は分類によって一定ではない可能性が考えられ, これを検証する必要がある.
- 仮想トレードによる成績の評価
  方向性の予測がもし上手くいけば, それを利用した取引の成績も良くなるはずである. そこで提案手法を用いた仮想トレードの実験を行い, その評価する.
- 事後確率推定オプションの追加
  libsvm には事後確率を推定するオプションがある. これを利用することによって, 高い確率なのに分類を間違えてしまったのか, 低い確率だから学習データが足りてないのかなどを区別することができる.

参考文献

[1] Robert P. Schumaker and Hsinchun Chen. A discrete stock price prediction engine based on financial news. *Computer*, Vol. 43, No. 1, pp. 51–56, January 2010.

[2] Robert P. Schumaker, Yulei Zhang, Chun-Neng Huang, and Hsinchun Chen. Evaluating sentiment in financial news articles. *Decis. Support Syst.*, Vol. 53, No. 3, pp. 458–464, June 2012.

[3] 辻洋平, 古宮嘉那子, 小谷善行. Web ニュース中の複数企業に対応した株価予測. 電子情報通信学会技

術研究報告. IBISML, 情報論的学習理論と機械学習, Vol. 110, No. 476, pp. 109–113, mar 2011.

[4] V. Vapnik. *Statistical Learning Theory*. Wiley, Chichester, GB, 1998.

[5] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp. 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[6] 工藤拓, 山本薫, 松本裕治. Conditional random fields を用いた日本語形態素解析 (解析). 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2004, No. 47, pp. 89–96, may 2004.

[7] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.