3L3-OS-16a-4in1

ファイナンス掲示板情報からの株価予測

Daily Stock Price Prediction via Stock Bulletin Boards Data

坪内 孝太[*1] 伊藤 友貴[*2] 山下 達雄[*1] 和泉 潔[*2]
Kota Tsubouchi Tomoki Ito Tatsuo Yamashita Kiyoshi Izumi

[*1] ヤフー株式会社
Yahoo Japan Corporation

[*2] 東京大学大学院
The University of Tokyo

This paper describes the research on daily stock price prediction via Yahoo! Stock bulletin boards data. Though each words in bulletin board are treated as a feature for prediction in previous research, the words are often over-fitted because the same words are used in different cases. So we focus on the relationship of the word in this research. Concretely, we vectorize the words in bulletin boards with word2vec and we developed the estimation model with the vectorized words. As a result, it is confirmed that out proposed method performs well.

1. はじめに

本論文では Yahoo! 株価掲示板情報を用いて株価を予測する事を目的とする。これまでの研究では、掲示板情報に現れる単語を特徴量として株価の上下を推測するものが多かった[坪内 15]が、本稿では単語同士の関係について着目する。word2vec [Mikolov 2013]を用いて単語の特徴量を単語同士のベクトルで表現し、掲示板で発言された数日後の株価の上下を推測できるかどうか予測した。

2. 手法の概要

本手法の概要を図 1 に示す。本手法はヤフーが提供する株価掲示板の投稿情報を解析し、株価の変動を予想するというものである。提案手法は、前処理で word2vec を用いて単語を選定およびグループ化する以下の 4 つのステップからなる。

1) 掲示板に出現する単語のベクトル表現
2) 単語の選定
3) 単語のグループ化
4) 機械学習

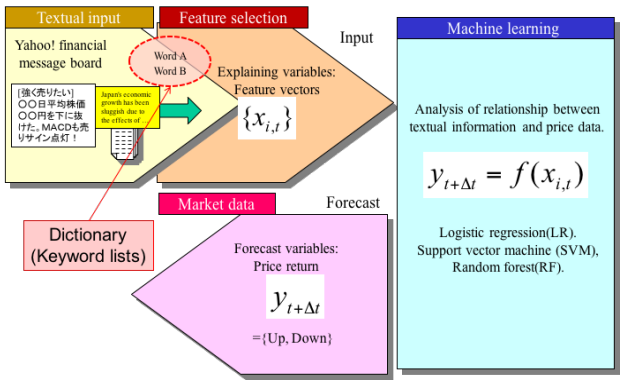

図 1. 掲示板情報を用いた株価予測手法の概要

連絡先：坪内孝太，ヤフー株式会社 Yahoo! JAPAN 研究所，
ktsubouc@yahoo-corp.jp

2.1 本手法に必要なデータセット

本手法を行うにあたり、2 つのデータセットが必要になる。株価掲示板情報を情報、金融の専門家がスコアリングした極性辞書である。

後者は金融実務者が株価予測において重要と考えピックアップした単語 1,319 語である。単語ごとにスコアが付与されている。このスコアは、複数の金融実務者がセンチメントスコアを自分の判断で回答しその平均をとったスコアである。

2.2 掲示板に出現する単語のベクトル表現

まず、前処理として株価掲示板に出現するすべての単語をベクトルで表現する。同じ投稿に共起する単語かを入力とし、word2vec により 200 次元のベクトルによりスコア化した。

結果、たとえば「ストップ高」と似たようなベクトル表現を持つ単語として「騰がり」、「天井」という単語のベクトルに「買い」のベクトルを足したスコアに近い単語として「全力」が、導出され、株式取引に近い感覚の結果となった。

2.3 有意な単語の選定

掲示板の単語を機械学習の特徴量として採用し予測することを試みるが、掲示板情報で用いられている単語はバラエティに富んでおり、そのまま採用した場合は機械学習の精度に影響がでることが考えられる。そのため、まずは単語の選定を行う。

本手法では、掲示板に出てくる多くの単語のうち、専門家のインタビューにより作られた極性辞書に含まれる語のみを対象とした。用いた極性辞書（1,319 単語）との突き合わせの結果、103 語のみを対象として選定された。

2.4 単語のグループ化

選定された 103 語のグループ化を行う。対象の 103 語について 200 次元のベクトルのコサイン類似度を計算し、K-means 法によるクラスタリングを行う。本稿では様々なパターンを試し、K=25 を採用した。これにより、103 語の単語が 25 個にグループ化された。

2.5 機械学習による予測

掲示板(2014 年 9 月)の各記事テキストに対象となる 103 語が含まれていたら、その単語が何番目のグループに属するかを調べ、極性辞書に付与されている単語の重み分を加算する。結果、1 つの投稿の記事が 25 つの特徴量で表現される。

得られた 25 次元のベクトルを説明変数として、その記事の銘柄の株価の記事が出てから 1,5,10 分後、引けまでの変動率(リターン)を非説明変数で外挿予測(2 値分類)を行う。2 値分類には、ロジスティック回帰、サポートベクターマシン、ランダムフォレストの 3 つの方法を用い、性能を比較した。

3. 株価掲示板情報を用いた実験

実際の株価掲示板情報を用いて提案手法の評価を行う。予測性能の評価に加え、途中成果物として得られるクラスタリング結果の妥当性も定性的に評価する。

3.1 株価掲示板情報

本研究では、ヤフーが提供する株価掲示板情報を用いる。2014 年 9 月の 1 ヶ月分のデータを用いて株価予測モデルを構築する。

9 月 1 日～20 日の 20 日間の投稿データにより学習を行い、そこで得られたモデルを 9 月 21 日～30 日の 10 日間のテストデータに対して推測を行い、性能評価とした。

3.2 単語のグループ化の定性評価

本手法の有用性を定性的に検証する目的で、word2vec により得られた記事ごとの 200 次元のベクトルの類似性から k-means 法によりクラスタリングした結果を表 1 に示す。

結果、例えば 5 番のクラスタには上向きの単語が揃っている、17番には協業系の単語が揃っているなど、同じような意味合いの単語が同じクラスタに入っている。実際に専門家のインタビューにより、つけた極性辞書の weight も、同じクラスタ番号の中には似通った極性となっているケースが多いことが分かり、本手法によりもクラスタリングは納得の行く結果となっている。

一方で、8 番や 10 番のクラスタのように、クラスタに属する単語が少ないケースでは極性辞書の weight の正負が分かれる事例も見られた。

表1. 極性辞書の単語のクラスタリングの結果

	cluster		cluster		cluster
赤字	1	視野	9	受託	17
注意	2	案件	9	縮小	18
参入	3	遅れ	9	納入	18
リニューアル	3	営業益	10	採算	18
下支え	3	無配	10	加速	18
復配	3	可能性	11	増産	18
上乗せ	3	低迷	12	特需	18
子会社化	3	継続	12	消費増税	18
上場廃止	3	急増	12	出店	18
値上げ	3	下落	12	悪化	18
引き下げ	3	導入	13	太陽光発電	18
高騰	3	開発	13	海外展開	18
上方修正	4	増益	14	自動車	19
追い風	5	下方修正	14	新製品	20
寄与	5	設備投資	14	下降	20
恩恵	5	増収	14	合併	20
拡充	5	進捗	14	買収	20
増強	5	減益	14	懸念	20
増税	6	配当性向	14	保守	20
低下	6	倍増	14	リスク	20
続伸	6	黒字	14	貢献	21
人件費	6	驚き	15	提携	21
横ばい	6	減配	15	増加	21
為替差益	6	上昇	16	改善	21
反動	6	連携	17	採用	21
自社株買い	7	共同開発	17	前期	22
増額	7	業務提携	17	増配	23
自動車部品	7	受注	17	需要	24
推進	7	拡販	17	研究開発	24
伸び	8	協業	17	稼働	24
減少	8	効率化	17	拡大	24
新設	9	主力	17	設立	24
回復	9	強み	17	ロボット	25
反発	9	強化	17		
撤退	9	北米	17		

3.3 株価掲示板情報を用いた株価予測の結果

提案手法による株価の動向予測のシミュレーション結果を図 2 に示す。precision と recall、それらから計算される F 値の 3 つの指標を用いて評価した。ロジスティック回帰（LR）、サポートベクターマシン（SVM）、ランダムフォレスト（RF）の 3 手法で比較を行った結果で、図の左軸が precision と recall を、F 値について右軸を示している。。

結果を見ると、LR, SVM, RF の順で結果は良くなっているが、どれもチャンスレベルの予測結果にとどまっている。ただ、RF においては他の手法と比べ有意な結果となった。

Precision は株価予測において有用な結果となっているが、課題は再現率である。これは、ワードを 103 語に限定したことが原因であるとかんがえられる。対象となる 103 語が投稿に含まれない場合は、予測不可能となるためである。

数万語ある単語が 103 語と削られてしまったのは、専門家のインタビューによって作成された極性辞書を用いた事が原因として考えられる。株価掲示板の投稿者のような一般大衆と専門家とでは、用いる言葉や表現に差が生じることは必至である。株価掲示板と相性の良い辞書により単語の選定を行うことで、再現率に改善が期待できる。

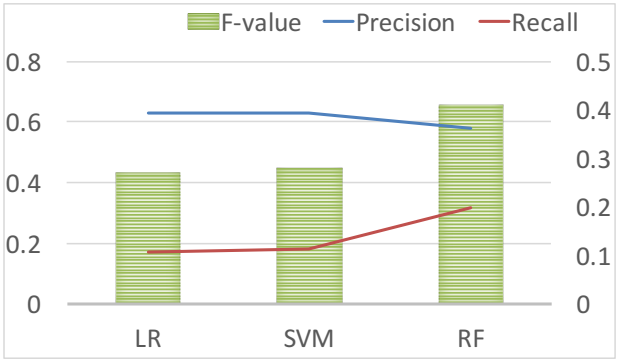

図 2. 提案手法による株価予測の結果

4. 結論

株価掲示板の投稿情報を用い、短時間での株価の動向に対する予測を行った。提案手法は、まず word2vec によりベクトル化したうえで掲示板に出現する単語を選定およびグループ化し、機械学習により株価予測を行うというものである。

シミュレーション実験の結果、word2vec を用いたクラスタリングは定性的にみて有用なクラスタリング結果となった。それを用いた株価の予測性能は極性辞書による専門辞書により単語を限定した事が原因で、チャンスレベルとなっている、ランダムフォレストでは有意な性能向上が見られた。

今後は、極性辞書に置き換わるウェブの掲示板情報に見合う辞書の作成などが課題となる。

参考文献

[Mikolov 2013] T. Mikolov, et. al,: Distributed Representations of Words and Phrases and their Compositionality”, NIPS 2013, pp. 3111-3119, 2013.

[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.