E-016

クラスタリングと単語分散表現を用いたニュース記事からの株価動向予測

Prediction of Stock Price Trend from News Articles Using Clustering and Distributed Representation of Words

易 迪† 杉本 徹†
Yi Di Sugimoto Toru

1. はじめに

近年社会の情報化に伴い大量の企業経営状況や決算発表，事件などの金融情報がウェブ上で入手可能となっており，データマイニングの手法によって市場をモデル化し，市場の行動を予測する研究が様々な側面から行われている。本研究では，日本株の銘柄ニュースから株価動向を予測することを目的とする。ニュース記事に書かれている話題は様々なものがあり，話題によって記事と株価動向の相関性が異なる。また，多くの種類の単語が含まれており，単語の意味を考慮した処理が必要である。そこで本研究では，ニュース記事を話題によりクラスタリングし，各クラスタのニュースを単語の分散表現を用いてベクトルに変換することで株価動向予測の精度向上を目指す。

2. 背景

2.1 関連研究

テキスト情報を用いた株価動向予測に関する研究として Schumaker ら[1]は，Yahoo! Finance から収集したニュースを分析し，会社名や人名などの固有名詞とあらかじめ決めておいた単語のみに注目し，そのニュースが配信されてから 20 分後の株価との関係を SVR を用いてモデル化する方法を提案した。また，高橋ら[2]はヘッドラインニュースを対象とし， Naive Bayes を用いた Good/Neutral/Bad の記事分類結果と株価のリターンとの間に有意な関連があることを示した。Hagenau ら[3]はニュースに出現する 2 単語の組み合わせをカイ二乗検定を用いてフィルタリングしたものを素性として株価動向を予測する方法を提案している。

2.2 問題設定

日本語版 Yahoo! ファイナンスのサイトから日本上場企業に関連するニュースを収集し，ニュースの話題，内容による株価動向への影響について検討する。

ニュースとして，Yahoo! ファイナンスに掲載された日経 225 企業に関するニュース記事本文を収集する。株価データとして，Yahoo! Finance API から各個別銘柄の各取引日の株価始値と終値データを取得し（注：2017 年 6 月現在この API で日本銘柄の株価データは取得できなくなったようである），各ニュース記事に対し，その記事に関連する銘柄がそのニュースの発表日（15 時以降発表の場合は翌取引日）において始値<終値であれば「Up」，始値>終値であれば「Down」というラベルデータを与える。なお，始値=終値となるニュースは対象外とした。

3. 提案手法

3.1 処理の流れ

システムの全体像を図 1 に示す。

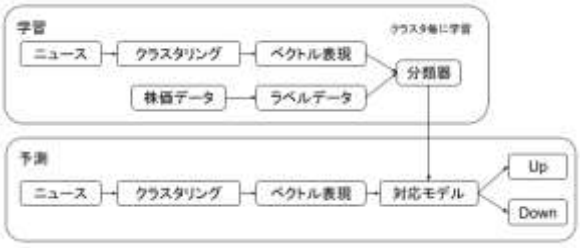

図 1 処理の流れ

学習段階では，まずウェブからニュースを収集し，ニュースの内容，話題に基づいてクラスタリングを行う。次に，ニュースの配信日付，時間と企業名，銘柄コードからニュースの対応ラベルデータを取得する。最後に，ニュースのベクトル表現とラベルデータを分類器の入力とし，クラスタ毎のモデルを作成する。予測段階では，与えられたニュースが属すクラスタを求めて，ベクトル表現に変換した後，対応するモデルに入力し，予測値を出力する。

3.2 ニュースのクラスタリング

吉田ら[4]は，新聞記事のクラスタリングによって銘柄の取引高に影響を与える可能性が大きい記事を選別できることを示している。本研究でも，ニュースの話題に着目する。ウェブ上の金融ニュースは企業経営状況や決算発表，事件など様々な情報を含んでおり，話題によって株価動向への影響も異なると考えられる。

この問題に対処するため，本研究ではトピックモデルを用い，ニュースを話題ごとにクラスタリングする。トピックモデルとしては，Latent Dirichlet Allocation を用いる。ニュース記事の各単語は，潜在的なトピックを持ち，あるニュース中に出現する単語の潜在トピック分布から，ニュースが属すトピックを推定できる。この推定結果に基づき，ニュースのクラスタリングを行う。

3.3 ニュースのベクトル表現

3.3.1 カイ二乗値を用いた単語選択

ニュースからベクトル表現を生成するため，まず記事本文を形態素解析し，一般名詞，サ変接続名詞，自立動詞，自立形容詞，形容動詞語幹の単語を抽出する。これらの単語の中には分類に有用な単語とそうでない単語が含まれているので，カイ二乗値を用いて分類に有用な単語を選択する。カイ二乗値 χ^2 はカテゴリと単語がどれくらい依存しているかを表す尺度である。カテゴリ「Up」，「Down」のラベルがそれぞれ付与された単語 w が出現するニュースの数と単語 w が出現しないニュースの数を用いて，単語 w の各カテゴリにおける出現頻度の期待値を求める。カイ二乗値 χ^2 は単語 w の各カテゴリにおける出現頻度の期待値と実際の出現頻度がどれだけ乖離しているかを表す。この乖離が大きいほど単語とカテゴリの依存度合いが強いと解釈でき，分類に有用な単語であると考えられる。

† 芝浦工業大学, Shibaura Institute of Technology

3.3.2 単語分散表現の利用

文章からベクトルを生成する手法として，文章中の単語を種類ごとに次元に割り当てて単語の出現頻度情報からベクトルを作る Bag-of-Words 法が多く用いられている。しかし，文章中の各単語は独立ではなく，類似単語も存在するので，Bag-of-Words 法だと文章の意味を十分に捉えることができないという問題がある。一方，Mikolov ら[5]が提案した単語の分散表現は大規模コーパスを用いた学習により単語の意味的類似性を反映したベクトル表現である。本研究では，ニュースから抽出した単語に対してまずカイ二乗値を用いた選択を行い，選ばれた単語に対して word2vec [5]を利用して作成したベクトル表現を求め，それらの平均をニュースのベクトル表現とする。

4. 実験

4.1 実験方法

実験では，2015 年 5 月 1 日から 2016 年 12 月 31 日までに配信された日経 225 企業に関する 7485 件のニュースを対象とする。単語の分散表現は，日本語 Wikipedia と毎日新聞の本文データから word2vec の CBOW モデルを用いて 200 次元のベクトルデータを作成した。前節で述べた方法により作成したニュースのベクトル表現を用いて Support Vector Machine により「Up」，「Down」の 2 値分類の学習を行った。SVM のカーネルは RBF カーネルを用いた。評価実験は，本研究で提案したニュースのクラスタリング，カイ二乗値による単語選択，単語分散表現の利用のそれぞれについて，行う場合と行わない場合の全組み合わせ（8 パターン）に対して 10 分割交差検定を行い，求めた平均分類精度を比較する。

4.2 実験結果

それぞれの方法で株価動向予測を行った場合の平均精度を表 1 に示す。

表 1 それぞれの方法による予測精度

	BoW	
	全単語を使用	$\chi^2 > 1$ の単語のみ使用
クラスタリング無し	53.11%	53.00%
クラスタリング有り	54.07%	55.69%
	単語分散表現を利用	
	全単語を使用	$\chi^2 > 1$ の単語のみ使用
クラスタリング無し	54.68%	57.17%
クラスタリング有り	55.38%	58.35%

この結果から，ニュースを話題によりクラスタリングし，各クラスタに対して適切なモデルを構築することで，全ニュースを 1 つのモデルにまとめるよりある程度良い精度が得られることが分かった。また，カイ二乗値による単語選択と単語分散表現の利用により，それぞれ予測精度がやや向上することが分かった。

次に，クラスタごとの動向予測精度を図 2 に示す。

クラスタ 8 のニュースは，ベクトル化の手法にかかわらず予測精度が 50%程度と低かった。クラスタ 8 のニュースの内容を確認したところ，株価，営業利益，為替レートなどが話題となっていて，使われる単語は大体決まっており，情報は数字で伝えている場合が多かった。ニュースをベクトル化する過程で数字が表す情報が無くなったため，うまく予測できなかったと考えられる。

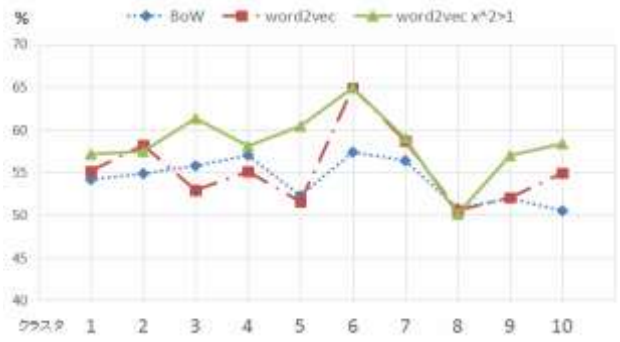

図 2　クラスタごとの予測精度の比較

4.3 出力の抑制

システムの利用者に動向予測の確信度が高いニュースのみを提示する方法を検討する。libsvm 分類器は入力が各カテゴリに属する確率を出力することができるので，この確率に閾値を設けて，閾値を超えたニュースのみを利用者に提示する。また，クラスタ 8 に分類されるニュースは予測精度が低いので利用者に提示しないことにする。この方法を実験した結果，各カテゴリに属する確率に対して設ける閾値の大きさとニュースからの予測精度はある程度の正の相関関係があることを確認できた。

5. おわりに

本研究では，ウェブ上のニュース記事と株価変動との関連性について分析を行った。その結果，ニュースの内容，話題によって株価動向をある程度予測できるニュースと予測が難しいニュースとに分別できることと，有用性が高い単語のみに注目することで予測精度を向上できることが分かった。また，分類器が出力する確率を確信度と見なし，確信度が高いニュースのみを利用者に提示することでより良い精度が得られることが分かった。

今後の課題として，学習データの拡張と株価の変動量の予測が挙げられる。企業の株価動向に影響を与える要素はニュースのみではなく，様々な情報が株価動向と相関関係にあると考えられる。今後はニュースと企業に関するツイートデータを組み合わせることにより，動向予測の精度向上と株価変動量の予測を試みたい。

参考文献

[1] Robert P.Schumaker, Hsinchun Chen, “A Discrete Stock Price Prediction Engine Based on Financial News”, Computer, Vol.43, No.1 (2010).
[2] 高橋悟, 高橋大志, 津田和彦, “ヘッドラインニュースと金融市場の関連性の分析”, 経営情報学会全国研究発表大会要旨集, (2008).
[3] Michael Hagenau, Michael Liebmann, Markus Hedwig, Dirk Neumann, “Automated news reading:Stock Price Prediction Based on Financial News Using Context-Specific Features”, HICSS, (2012).
[4] 吉田稔, 中川裕志, 石田智也, 中嶋啓浩, 松井藤五郎, 和泉潔, 池田翔, 本多隆虎, “ニュース記事クラスタリングによる取引高予測の試み”, 人工知能学会全国大会論文集, (2011).
[5] Mikolov Tomas, Kai Chen , Greg Corrado, Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space”, ICLR2013, (2013).