文章極性を考慮したニュースステキスト分析による経済動向予測

Economic Trend Prediction by News Analysis Considering Text Polarity

川崎　拓海 [1*]　　穴田　一 [1]
Takumi Kawasaki　　Hajime Anada

[1] 東京都市大学大学院
[1] Tokyo City University Graduate School

Abstract: In recent years, various researches in the field of economic prediction have been carried out using fundamental analysis and technical analysis with numerical information. Considering news articles containing not only numerical information but also textual information means that we can pay attention to public opinion, and thus we can make more accurate economic trend prediction which are difficult to predict only by numerical information. In this study, we propose a news text analysis for economic trend prediction using polarity dictionaries.

はじめに

近年，金融予測の分野ではローソク足の画像を用いた分析やファンダメンタル分析，数値情報を用いたテクニカル分析などによる様々な研究が行われている．その中でも数値情報だけでなくテキスト情報も含まれているニュース記事を考慮することは，世論に目を向けることを意味し，数値情報だけでは説明が難しい市場の予測を精度高く行える可能性があると考えられる．そこで本研究では，テキストマイニング手法を用いてニュース記事から株価の上昇・下落の予測を行った．テキストマイニング手法を用いた金融予測についても様々な研究が行われているが，本研究では，新聞記事の予測前営業日と予測当日のテキストを用いて株価の上昇下落を予測した和泉らの研究[1]を基に，日本語評価極性辞書[2][3]と金融に関係する単語を分析する金融専門極性辞書[4]を用いたニュースステキスト分析による東証株価指数(TOPIX)の株価予測を提案し，その有効性を確認した．本研究では予測前営業日の見出しのテキストデータをまとめ，文中の否定文を考慮した極性単語を抽出し，一定割合以上出現した単語の中で株価上昇確率を共に満たすものを特徴語とする．そしてテキストにその特徴語が出現した際，TOPIXの株価は上昇するか否かを勾配ブースティング決定木(GBDT)に学習させて，その有効性を検証した．

*連絡先：東京都市大学大学院　総合理工学研究科
〒158-8557 東京都世田谷区玉堤１丁目２８番１号
E-mail: g2181429@tcu.ac.jp

既存研究

テキストの時系列出現パターン

従来のテキスト分類を用いた市場予測では，テキストの時系列性に着目し，直近m個のテキストの特徴ベクトルから二値分類した値y_{t+1}を求める．

$$y_{t+1} = f(x_{t,\dots,}x_{t-m-1})$$

ここで，fは手法を表し，x_tは時刻tにおけるテキストの特徴語ベクトルを表す．

テキストの時系列出現パターン

和泉ら[1]は新聞記事の予測前営業日x_{t-1}と予測当日x_tのテキストで，単語の出現パターンを作成した．予測前営業日のテキストx_{t-1}では出現していないが予測当日x_tでは出現している場合　”新出”．予測前営業日のテキストx_{t-1}に出現している，かつ予測当日のテキストx_tにも出現している場合　”続出”．予測前営業日のテキストx_{t-1}には出現してい

るが予測当日のテキストx_tには出現していない場合 ”消滅”と定義した.

特徴語の抽出

和泉らは，日本経済新聞の予測前営業日と予測当日の記事のリード(第一段落)と見出しを結合してMecab を用いて形態素解析を行い，TeamExtract で専門用語を抽出し，特徴語とした. TeamExtract は形態素解析で分割された専門用語を再度組み合わせ，専門用語として抽出するものである．これを訓練期間内に出現した記事のテキストデータに用いた．出現パターンを考慮した専門用語の出現数を調べ，k回以上出現したものの中から，テキストに出現パターンを考慮した単語が出てきた時，株価が上昇した確率が θ以上のものと$1-\theta$以下のもの $(\theta > 0.55)$を取り出した.

SVM を用いた株価予測

和泉らの研究では抽出した特徴語で株価の上昇・下落を予測するために SVM を用いた. SVM とは互いに一番近いベクトルの距離を最大化することで未知データを 2 クラスのどちらかに分類する手法である．既存研究では単語の特徴量が多いので，カーネルトリック法という非線形分離型の分類器を用いて実験を行っている.

抽出した l個の特徴語の出現パターンを $p_1, \dots, p_l$とし，訓練期間内のテキストに出現パターン p_iの単語 iが生じている場合，i次元の特徴量を1そうでない時は0とした．出力を当日の株価の利益率が 0または正のとき1，負の場合は−1とし， 作られた l次元の専門用語に関する特徴ベクトルと株価の出力の関係を SVM に学習させた.

提案手法

和泉らの研究での全体の平均正解率は 71.4%であるが，悪い年は 56.3%と不安定である．これは 単語の出現数や出現パターンのみ考慮していて，単語の印象を考慮していないことが要因であると考えられる．なぜなら人に良い印象を与える単語が出現すると株価が上昇し，人に悪い印象を与える単語が出現すると株価が下落すると考えたからである．そこで提案手法では肯定文中極性単語と否定文中極性単語を考慮した特徴量抽出を行った.

今回極性単語を扱うにあたり，日本語評価極性辞書と金融専門極性辞書を利用した．日本語評価極性辞書とは様々な用法や名詞に対し，Negative・Positive と極性が振られている辞書である．また，金融極性辞書とは金融専門単語についてネガティブ・ポジティブ度を極性値として表した辞書であり，−1以上 1 以下の数値データで表されている.

この２つの極性辞書から感情分析ツールの１つである Oseti を用いて否定文を考慮した極性単語の抽出を行った. Oseti とは形態素解析ツール Mecab を用いて文章極性スコアを算出するものである．単語に”せず”や”ない”等の否定が掛かっている場合，その単語の極性を反転させスコアを求める．よって Positive(Negative)の極性単語に否定が掛かっている場合 Negative(Positive)とし，肯定文中極性単語と否定文中極性単語をそれぞれ抽出した。

本研究では IT・経済ニュースの記事に対して２つの辞書から得られる極性単語を用いたネガティブ・ポジティブ分析(以下ネガポジ分析とする)による経済動向予測を提案する．まず訓練データ内において 1 日に数件ずつ掲載されている IT・経済ニュースの見出しから, Oseti を用いて否定文中か肯定文中かを考慮した極性単語を抽出した. Oseti に用いる金融専門極性辞書の極性値ηは$\eta > \eta_{th}$ の場合 Positive, $\eta < -\eta_{th}$ の場合 Negative と分類された極性単語を抽出した．得られた極性辞書の単語がk回以上出現した中から株価上昇割合 θ_1以上， 株価下落割合 θ_2 以上の単語を取り出し特徴語とした.

取り出されたl個の特徴語に対し，訓練期間内のテキストに特徴語が生じている場合，特徴量を 1，存在しない特徴語に関しては 0 とし，勾配ブースティング決定木に学習させた.

結果

提案手法の有効性を確認するためロイターニュース IT・経済ニュースの見出しを用いて，予測対象を半年ごとに分けた 2018 年 7 月~2020 年 12 月までの TOPIX-連動型上昇投資信託(ETF)とし,予測前営業日のニュースの見出しで上昇下落の予測を行った．訓練データの期間は直近の過去 3 年間を用いた．また，予測前日の終値と予測対象日の終値の差分を TOPIX -ETF の上昇・下落の基準とした.

極性値の閾値をη_{th} = 0.03 とし， 得られた極性単語で 10 回以上出現した単語の中から予測当日の株価の上昇割合 η が 0. 75 以上と株価の下落割合 η が 0. 65 以上のパターンを抽出し，特徴語として用いた．モデルのパラメータはグリットサーチを行い，

最適なパラメータを選択した.

予測結果は表 1 の混同行列を用いて評価する.

表 1　混同行列の例

実際のクラス	Negative	TN(True Negative)	FN(False Positive)
	Positive	FP(False Negative)	TP(True Positive)
		Negative	Positive
		機械学習モデルの予測	

True は予測が正しく False は予測が正解のクラスと異なったことを表す. 表 1 を元に Accuracy(正解率)や Precision (適合率) , Recall (再現率) を求め, グラフ化した結果を図 1 に示し, F 値も求め, それぞれの結果を表 2 に示した.

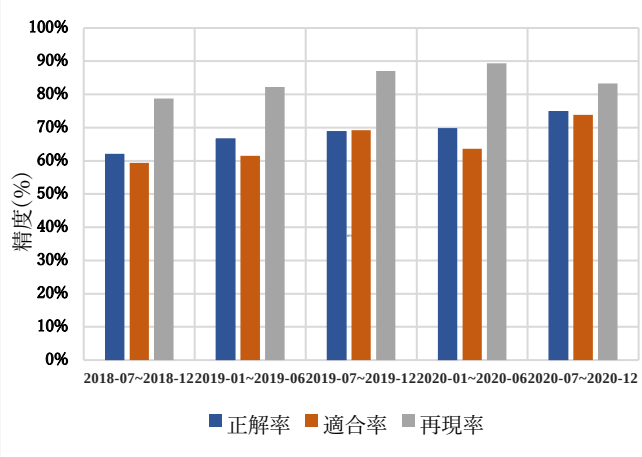

図 1　混同行列を用いた結果

表 2　混同行列を用いた結果

テスト期間	F 値	正解率	適合率	再現率
2018-07~2018-12	**0.68**	**62.1%**	**59.4%**	**78.8%**
2019-01~2019-06	**0.67**	**66.7%**	**61.5%**	**82.2%**
2019-07~2019-12	**0.77**	**68.9%**	**69.2%**	**87.1%**
2020-01~2020-06	**0.74**	**69.8%**	**63.6%**	**89.4%**
2020-07~2020-12	**0.78**	**75.0%**	**73.8%**	**83.3%**
全体の平均	**0.73**	**67.5%**	**64.4%**	**84.2%**

表 2 より, 先行研究の全体の正解率が約 70%に対し, 提案手法の精度は低く, 全体の正解率が 67.5%という結果となった. しかし, 悪い年の正解率でも約 62%と既存手法より安定した結果を得ることができた. 実際に上昇すると予測するデータのうち, 上昇すると予測できた割合を示す再現率の平均が約 84%と高い精度で予測できていた.

また, 株価の急落が起こった 2020 年 1 月~6 月の予測と 2020 年 7 月以降の期間においての予測が精度高く出来ていた. 金融業界では正解率が常に 55%以上あれば有用であると評価されていて, 今回それを越えることができていた.

特徴語として得られた極性単語は発表で述べる.

考察

2018 年 7 月~2019 年 6 月の精度が低かったが, これは米国の業績悪化の影響で起こった大幅な急落が主な原因で, 訓練期間のデータではうまく学習できなかったと考えられる. この期間の精度は, 異なるニュース記事で予測を行った場合にも同じように悪くなった.逆に, 急落・急騰の激しい時期であった 2020 年 1 月から 2020 年 12 月にかけての正解率が, ほかの年と比べ上がっていた. これは直近 3 年間を訓練データとしているので急落が激しかった 2017 年 1 月~2019 年 12 月にかけての暴落をうまく学習したからだと考えられる.

今後の課題

前日と予測当日の終値の差が 1 円以上あれば上昇・下落と株価の変化がほとんど無い横ばいの変動時でも二値で分類してしまうため, うまく分析できていない. 今後は一定の閾値を設けて上昇・横ばい・下落の多値分類を行うことで精度を高めていきたい.

また, 予測前営業日のみで当日の株価予測を行っているため時系列性を考慮できていない. よって予測前営業日だけでなく, 2 日間や 3 日間の出現パターンを考慮して予測行っていきたい.

参考文献

[1] 和泉潔, 松井藤五郎 : 新聞記事の時系列テキスト分析による株式市場の動向予測, 第 30 回人工知能学会, 3L3-OS-16a-6 (2016).

[2] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一, 意見抽出のための評価表現の収集, 自然言語処理, Vol.12, No.3, pp.203-222, 2005. / Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi. Collecting Evaluative Expressions for Opinion Extraction, Journal of Natural Language Processing 12(3), 203-222 (2005)

[3] 東山昌彦, 乾健太郎, 松本裕治, 述語の選択選好性に着目した名刺評価極性の獲得, 言語処理学会第 14 回年次大会論文集, pp.584-587, 2008. / Masahiko Higashiyama, Kentaro Inui, Yuji Matsumoto. Learning Sentiment of Nouns from Selectional Preferences of Verbs and Adjectives, Proceedings of the 14th Annual Meeting of

the Association for Natural Language Processing, pp.584-587 (2008)

[4] Ito T., Sakaji H., Tsubouchi K., Izumi K., Yamashita T. Text-Visualizing Neural Network Model: Understanding Online Financial Textual Data. In: Phung D., Tseng V., Webb G., Ho B., Ganji M., Rashidi L. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2018. Lecture Notes in Computer Science, Springer, vol 10939, pp 247-259 (2018).

[5] 中川裕志, 森辰則, 湯本紘彰:出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, Vol. 10, No. 1, pp. 27-45 (2003).

[6] 東山昌彦, 乾健太郎, 松本裕治:述語の選択選好性に着目した名刺評価極性の獲得, 言語処理学会第 14 回年次大会論文集, pp.584-587 (2008).