A-016

金融極性辞書を用いたニューステキスト分析による経済動向予測
Economic Trend Prediction by News Analysis Using the Economic Polarity Dictionary

川崎　拓海‡　　穴田　一‡
Takumi Kawasaki　Hajime Anada

1. 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, 金融に関係する単語を分析する金融専門極性辞書[2]を用いたニューステキスト分析による東証株価指数(TOPIX)の株価予測を提案する. 本研究では予測当日 3 日前から予測前営業日ごとに見出しのテキストデータをまとめ, 形態素解析ツール janome を用いて金融極性単語を抽出し, 一定割合以上出現した単語の中で株価上昇確率と極性値の条件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サポートベクターマシン(SVM)に学習させる. そして, その学習モデルを用いて別の期間のテキストを用いて予測当日の TOPIX の日中の上昇・下落を予測し, 本研究の有意性を確認した.

2. 既存研究

2.1 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.

2.2 テキストの時系列出現パターン

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する. 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”と定義する.

2.3 特徴語の抽出

既存研究で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 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 $(\theta > 0.5)$を取り出す.

2.4 SVM を用いた株価予測

既存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 を用いる. 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.

3. 提案

全体の正解率は 71.4%であり, 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, 人に良い印象を与える単語は株価が上昇し, 人に悪い印象を与える単語は株価が下落すると考えた. そこで提案手法では金融専門極性辞書を用いて経済に関する単語の印象を考慮した. 金融専門極性辞書とは 金融専門単語についてネガティブ・ポジティブ度を数値化した辞書であり, -1以上 1 以下の数値データで表されている.

本研究では IT・経済ニュースの記事に対して金融専門極性辞書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を予測当日 3 日前から予測前営業日までの 3 日間ごとにまとめ, 形態素解析ツール janome を用いて, 金融専門極性辞書の単語が k 回以上出現した中から

I. 株価上昇割合 θ_1以上かつ極性値閾値 η_1以上
II. 株価上昇割合 θ_2以下かつ極性値閾値 η_2以下

†東京都市大学大学院　総合理工学研究科
Graduate School of Integrative Science and Engineering, Tokyo City University

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Ⅰ,Ⅱに属する単語が生じている場合, 特徴量を 1,いずれにも属さない場合は 0 とする. 出力を予測対象日の株価が上昇した場合 1, 下落した場合は 0 とし, SVM に学習させる.

4. 結果

提案手法の有意性を確認するため Livedoor 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 3 月~2020 年 8 月までの TOPIX-連動型上昇投資信託(ETF)とし, 予測当日 3 日前から予測前営業日までの 3 日間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 2 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.

金融専門極性単語の出現数をカウントし 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 θ_1が 0. 7 以上かつ極性値閾値 η_1が 0. 002 以上と株価の上昇割合 θ_2が 0. 4 以下かつ極性値閾値 η_2が-0.002 以下のパターンを抽出し, 特徴語として用いた. 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, Matthews Correlation Cofficient : MCC (マシューズ相関係数) も求め, それぞれの結果を表 2 に示した.MCC とは-1 と +1 の値の間をとり, 実際のクラスと予測したクラスが一致しているほど+1 に近い値をとる.

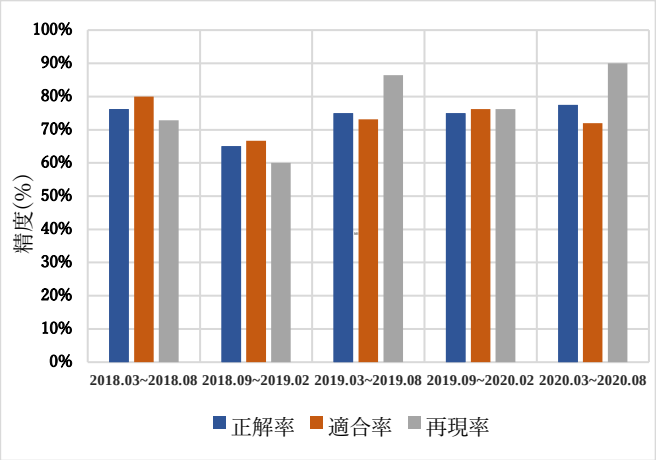

図 1　混同行列を用いた結果

表 2　混同行列を用いた結果

テスト期間	MCC	正解率	適合率	再現率
2018-03~2018-08	0.53	76.2%	80.0%	72.8%
2018-09~2019-02	0.30	65.0%	66.7%	60.0%
2019-03~2019-08	0.50	75.0%	73.1%	86.4%
2019-09~2020-02	0.50	75.0%	76.2%	76.2%
2020-03~2020-08	0.57	77.5%	72.0%	90.0%
全体の平均	0.48	73.7%	73.6%	77.1%

表 2 より, 精度の低い年でも正解率が 65%を超え, 全体の平均も約 74%と既存手法を上回る結果となった. また, 株価の急落が起こった 2020 年 2 月~3 月でも 75%以上の精度で予測が出来ていた. しかし, 2018 年 9 月~2019 年 2 月の全体の精度がほかの年と比べ, 正解率が低くなった. しかし, 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.

特徴語として用いた金融極性単語は発表で述べる.

5. 考察

2018 年 9 月~2019 年 2 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だと考えられる. しかし, 急落・急騰の激しい時期であった 2019 年 9 月から 2020 年 8 月にかけての正解率は 70%以上を超えていた. これは直近 2 年間を訓練データとしているので急落が激しかった 2018 年 9 月~2019 年 2 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.

また, 2018 年 9 月~2019 年 2 月と 2019 年 9 月~2020 年 2 月のテスト期間を株価の推移と比較すると, 急騰急落の激しい期間にもかかわらず, 比較的高い精度が出力できることが分かった. また, 既存研究でも急騰・急騰の期間で高い正解率をもつことが確認されている. したがって, テキストを用いた急騰・急落の予測は有用であると考えられる.

6. 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・下落としていたので, 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.

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] 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).

[3] 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).

[4] 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).