Vol.2020-MPS-131 No.28
2020/12/18

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

川崎　拓海[1]　　穴田　一[1]

概要：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マーケティングの予測を精度高く行える可能性があると考えられる. そこで本研究では金融に関係する単語を分析する金融専門極性辞書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.

キーワード：株価予測, サポートベクターマシン, テキストマイニング

Economic Trend Prediction by News Analysis using Economic Polarity Dictionary

TAKUMI KAWASAKI[1]　HAJIME ANADA[1]

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 the economic trend prediction using the economic polarity dictionary.

Keywords: stock prediction, support vector machine, text mining

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マーケティング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. 本研究では予測前営業日の見出しのテキストデータから, 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 SVM に学習させる. 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　時系列テキスト分析

既存手法では予測前営業日と予測当日のテキストを用いる時系列性に着目し, m 期間のテキストの特徴語ベクトルから二値分類した値y_tを求める.

$$y_t = f(x_{t-m+1,\ldots,x_t})$$

ここで, fは手法を表し, xは時刻$t-1$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その単語が出てきた時, 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く.

抽出したl個の特徴語の出現パターンを$p_1, \ldots, p_l$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次元の専門用語と株価の利益率に関する特徴量ベクトルを SVM に学習させた.

3 提案

和泉らの抽出した専門単語には"リビア"や"写真"などが挙げられているが, いずれも株価の上昇・下落と関係性があるとは思えない. その影響か, 全体の正解率は約 70%であるが, 悪い年は約 55%と不安定である. これは 単語の出現数のみ考慮していて, 単語の印象を考慮していないことが起因し, 人へ良い印象を与える単語は株価が上昇すると考えた. 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・どちらでもない・ポジティブ の範囲に単語を分類する辞書であり, それぞれ−1, 0, 1 の数値データとして扱う.

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 1 日ごとにまとめ, 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以上かつ極性値閾値以上
II. 株価上昇割合 1−θ以下かつ極性値閾値以下

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 に属する単語が生じている場合, 特徴量を 1, 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 TOPIX-連動型上昇投資信託(ETF)の上昇下落の予測を行った. 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, 訓練データを 2018 年 3 月~2020 年 2 月までの 2 年間, テストデータを 2020 年 3 月~2020 年 8 月の半年間とした. 金融専門極性単語の出現数をカウントし 20 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が 0.6 以上 極性値 0.002 以上と株価の上昇割合 0.45 以下 極性値−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. 図 1 に混同行列を用いた提案手法の結果を示す.また表 1 を元に Accuracy(正解率)や Precision (適合率) , Recall (再現率) , Matthews Correlation Cofficient : MCC (マシューズ相関係数) を求め, 表 2 に示した.

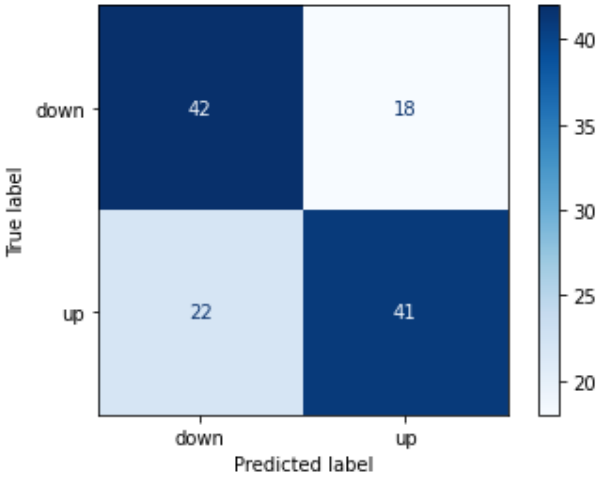

図 1 混同行列の結果

表 2 混同行列を用いた結果

Accuracy	0.67
Precision	0.70
Recall	0.65
MCC	0.35

結果の詳細と考察は発表で述べる

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)