A-034

語彙極性を獲得するアルゴリズムを用いたニューステキスト分析による株価予測
Predicting Stock Prices Using News Text Analysis With an Algorithm for Acquiring Lexical Polarity

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, 日々発表される情報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おり, 那須川ら[1]のアルゴリズムでは, 初めに用意した辞書単語とその極性を元に, 文中から主節の用言句を辞書単語の候補として抽出し, 一定の閾値を超えた単語を用いて極性辞書を作成した. また, 石垣ら[3]は初めに用意した辞書単語とその極性値を元に, Twitter のツイートを形態素解析して得た単語を辞書単語の候補として抽出し, 共起した単語に極性値を伝搬させて極性辞書を作成して, 為替に特化した極性辞書の構築を行ったが, 文章の否定を考慮していないことや, 為替動向に関係あるとは思えない単語に極性値が伝搬されるという問題があった. そこで本研究では, 金融専門極性辞書[2]を元に, ニュース記事から金融に関連する単語とその極性値を抽出する語彙獲得アルゴリズムを用いた, 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. 金融専門極性辞書の単語を初期の辞書単語とし,辞書に含まれる単語が各見出しに出現した際, その見出し中の"係り受けされる語のみ"と"係り受け含む語"をそれぞれ抽出し, 辞書単語の極性値も終値の変化率に応じて更新する. そして出現した辞書単語が閾値回数を超えた場合, その辞書単語の極性値を, 対応した"係り受けされる語のみ"と"係り受け含む語に伝搬させ辞書単語とし, これを繰り返すことで得た単語を特徴語とする語彙獲得アルゴリズムを作成した. そしてその特徴語が予測前日の見出しに出現した際, TOPIX の株価が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, 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

和泉らの研究[3]では, 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.

2.2 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.

2.3 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. 既存研究では得られた特徴語が多いので, 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 $p_1, \ldots, p_l$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

和泉らの既存研究では, 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, 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.そこで提案手法では目的のニュース記事に対応した, 金融に関連する印象を与える単語を抽出するアルゴリズムを用いて, 文の肯定否定を考慮した特徴量抽出を行った.

3.1 語彙獲得アルゴリズム

はじめに少数の既知表現データを用意し, 初期辞書を作成する. 初期辞書とはあらかじめ極性値が既知である単語群で, 他の単語に極性値を伝搬させる役割を持つ. 今回は金融専門極性辞書に出現する単語の中でも, 絶対値の閾値i を設け極性値のネガティブ・ポジティブ度が高い単語を初期辞書とした. 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・ポジティブ度を極性値として表した辞書であり, ネガティブな単語ほど負の値が大きく, ポジティブな単語ほど正の値が大きい数値データで表されている. そして得た初期辞書に含まれる単語が各見出しに出現した際, 構

†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

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

文解析器の 1 つである Cabocha を用いて, その見出し中の単純エントリと複合エントリを抽出した. 単純エントリとは係り受けされる語のみ, 複合エントリとは係り受けされる語だけでなく, 係り受けする語も同時に抽出する方法である. この時抽出したエントリに対し, Mecab を用いて形態素解析を行い, 以下の余分な品詞や記号を削除する前処理を行うことで, エントリを抽出するか否かを決める辞書単語の出現回数を数えやすくする.

1.符号以外の記号・助詞・助動詞・句読点は削除
2.名詞+動詞の場合, 動詞以降の品詞を削除

1.では株価動向に関連を持つ可能性のある”+”や:”-”といった符号は残し, “～が”や”～の”等の品詞を削除している.
2.では”上昇した”,”値上げしていく”等の用言に対し動詞以降の品詞を削除することで, 名詞のみを抽出している.

また, この時出現した極性辞書の単語の極性値も更新する. 更新式を以下に示す.

$$新しい極性値 = 元の極性値 \times \left(\frac{翌日の終値 - 見出し日の終値}{見出し日の終値} + 1 \right)$$

その見出しに出現した極性単語に否定がかかった場合は符号を逆転させることで更新後の単語の極性値を確定させる.

$$新しい極性値 = 元の極性値 \times \left(-\frac{(翌日の終値 - 見出し日の終値)}{見出し日の終値} + 1 \right)$$

このように, 各見出しに極性辞書の単語が出現する毎にエントリの抽出と極性値の更新を行っていく.

全見出し実行後, 出現したエントリに対応した極性辞書の出現見出し数がk %以上の場合, その更新された辞書単語の極性値を対応したエントリに伝搬させ, 新たに極性辞書として追加する. また, 同じエントリに対応した極性辞書の単語が複数存在し, いずれの極性辞書の出現見出し数がk %以上を超えている場合, 複数ある単語辞書の極性値の平均値を伝搬させる.

以上のアルゴリズムを新しい辞書単語が存在するまで行う.この語彙獲得アルゴリズムのフローチャートを次に表す.

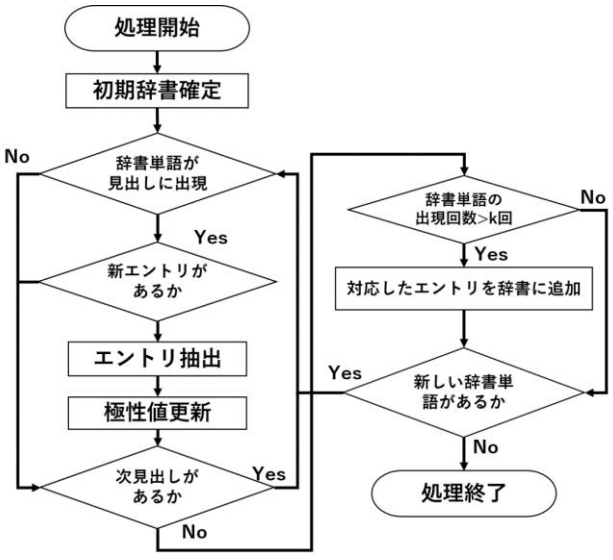

図 1　語彙獲得アルゴリズムのフローチャート

3.2 極性辞書を用いた株価予測

本研究では IT・経済ニュースの記事に対して語彙獲得アルゴリズム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語彙獲得アルゴリズムを用いて極性単語を抽出した. そして得られた極性辞書の単語を特徴語とした.

提案手法では, 見出しに出現した極性辞書の単語に否定が掛かっている場合”単語 Negative”として極性を反転させ, 文の肯定否定を考慮して同一な極性辞書の単語でも否定が掛かった単語と掛かっていない単語を区別して特徴語として抽出した. 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, 特徴量をその特徴語に対応する極性値, 存在しない特徴語に関しては 0 とし, 各見出しに出現した極性辞書の単語の特徴語を SVM に学習させた. この時, 見出しに単純エントリと複合エントリの係り受けされる単語が重なった場合, 複合エントリの極性値のみを特徴量とした. 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 Accuracy(正解率), Precision (適合率) , Recall (再現率), F 値を求めた.

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	

4. 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それぞれの予測日の直近の過去 3 年間を用いた. また, TOPIX-ETF の予測前日の終値を予測対象日の終値の上昇・下落の基準とした. 語彙獲得アルゴリズムの初期辞書として扱う金融専門極性辞書は, 極性値の絶対値iが 1 以上の単語を初期単語とし,出現見出し数が 5%以上を超える極性辞書の単語の極性値をエントリに伝搬させ, 極性辞書に追加した.

特徴語として得られた極性辞書の単語や結果と考察, 今後の課題は発表で述べる.

参考文献

[1] 那須川哲哉, 金山博：文脈一貫性を利用した極性付評価表現の語彙獲得, 情報処理学会研究報告, pp.109-116(2004).
[2] 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).
[3] 石垣藍睦, 沼尾雅之：Twitter からの為替予測に特化したドメイン辞書構成法の提案, 第 13 回情報科学技術フォーラム, RO-001(2014)
[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).

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.