Vol.2022-MPS-141 No.17
2022/12/13

語彙の組み合わせ感性辞書生成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

川﨑拓海[†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．しかし，数値情報ではなく，テキスト情報が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，数値情報では対応が難しいタイプの金融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，ニュース記事から金融に関連する語彙の組み合わせとその極性値を抽出して辞書を生成するアルゴリズムを構築し，その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)の株価予測を行うことによって，その有効性を検証した．

キーワード：テキストマイニング，株価予測，サポートベクターマシン

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[1]や数値情報を用いたテクニカル分析[2]などによる様々な研究が行われている．しかし，数値情報では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，数値情報では説明が難しいタイプの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．このような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，和泉らの研究[3]では，記事に出現した単語を出現パターンごとに抽出し，その中でも出現頻度と株価上昇下落割合がともに閾値以上の単語を特徴語とし，その特徴語が前日に出現した際，TOPIX の株価が予測前日に比べ上昇するか否かをサポートベクターマシン(SVM)に学習させ予測を行っていたが，文章の否定を考慮していない，株価動向に関係あるとは思えない単語が抽出される問題があった．

また，石垣ら[4]は高村ら[5]の単語感情極性対応表の絶対値極性値が高い辞書単語を初期辞書として，Twitter のツイートに辞書単語が出現した際，辞書単語に共起した単語に極性値を伝搬させ抽出し，辞書単語の候補とした．そして全ツイートから獲得した辞書単語候補のうち出現回数が一定以上の単語を辞書に追加し再度抽出した．これを辞書単語が追加されなくなるまで実行し，為替に特化した極性辞書の構築を行っていた．この極性辞書を元に為替の値動きとの比較検証を行っていたが，為替とは関連のない共起した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，新たにニュース記事から金融に関連する単語の組み合わせとその極性値を抽出する語彙獲得アルゴリズムを構築し，獲得した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．これは，金融専門極性辞書[6]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”と”係り受けする語”の組み合わせを全て抽出し，見出し文に出現した辞書単語の極性値を TOPIX 終値の変化率に応じて新しい極性値へと更新する．これを全見出し分実行後，出現頻度が高い辞書単語の極性値を，対応した”係り受けされる語”と”係り受けする語”の組み合わせに伝搬させ，新たに作成した辞書に追加する．2 周目以降は，1 周目で新たに作成した辞書を用いて抽出と更新を行う．作成した辞書に存在する，係り受けの組み合わせが文中に出現した際，見出し文中に存在する残りの”係り受けされる語”と”係り受けする語”の組み合わせを抽出し，出現した辞書の係り受けの組み合わせ極性値を 1 周目と同様に更新する．そして出現頻度が高い辞書の係り受けの組み合わせに対応する極性値を，抽出した”係り受けされる語”と”係り受けする語”の組み合わせに伝搬させ，1 周目で作成した辞書に追加する．

そして作成した辞書がもつ”係り受けされる語”と”係り受けする語”の組み合わせの中でも出現頻度と極性値閾値のともに高い組み合わせを特徴語とし，その特徴語が予測前日の見出しに出現した際，TOPIX の株価が予測前日に比べ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，予測を行った．

2. 提案手法

和泉らの研究では，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， 金融に関連するニュースに対し人が受ける印象を考慮していないことが要因ではないかと考えた． そこでニュース見出し中に， 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

†1 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

える単語が出現すると株価が下落すると考えた．しかし，既に作成されている日本語評価極性辞書や金融専門極性辞書には，時事ニュースに出現する金融に関する重要な単語のうち存在しない単語があるという問題や，“原油価格が上昇”と”売上が上昇”の様に，文章によって単語の極性が反する問題があった．

そこで提案手法では，辞書に含まれる単語がニュース記事の見出しに出現した際，見出しに出現した係り受けする語と係り受けされる語の組み合わせを抽出し，辞書単語の極性値を組み合わせに伝搬することで，金融に関連するニュースの印象を考慮したアルゴリズムを作成し，その中でもニュース見出し中の出現頻度と極性値のともに高い組み合わせを特徴語として抽出し株価の予測を行った．

2.1 語彙獲得アルゴリズム

①はじめに少数の単語とそれぞれの単語の極性値を用意し，初期辞書を作成する．初期辞書とはあらかじめ極性値が付与されている単語群で，今回は金融専門極性辞書に存在する単語の中でも，極性値の絶対値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 −1〜1 の極性値として表した辞書である．

②次に得た初期辞書に含まれる単語が各見出しに出現した際，構文解析器の 1 つである Cabocha を用いて，その見出し中の複合エントリを抽出した．複合エントリとは文中に存在する各文節間の係り受け関係をもつ組み合わせを抽出したものである．この時抽出した複合エントリに対し，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，抽出するか否かを決める複合エントリの出現回数を数えやすくする．

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

これは株価動向に関連を持つ可能性のある”+”や”−”といった符号以外を削除し，抽出する際の不必要な品詞を削除している．そして”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．

③この時出現した極性辞書の単語の極性値を更新する．更新式を以下に示す．

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

これは元の極性値に対し，終値の変化率と絶対値の極性値の積を加算していくことで，終値の変化率に応じて極性値を正負に逐次的に更新していく．

見出しに出現した極性単語に否定がかかった場合は次式のように符号を逆転させることで更新後の単語の極性値を更新する．

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

これらの更新式を用いて，各見出しに極性辞書の単語が出現する毎に複合エントリの抽出と極性値の更新を行っていく．

④全見出し実行後，見出しに出現した回数がk_1回以上の極性辞書の単語に対し，その極性辞書が出現した際抽出した複合エントリに極性辞書の極性値を伝搬させる．

⑤新たに複合エントリのみで構成する複合辞書を作成し，抽出した複合エントリを複合辞書に追加することで複合エントリのみの辞書を作成していく．2 周目以降の複合エントリの抽出には，新たに作成した複合辞書を極性辞書として扱っていく．語彙獲得アルゴリズムのフローチャートを図 1 に示す．

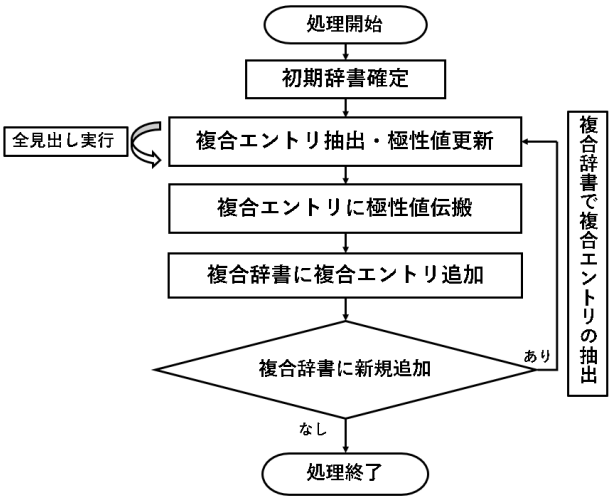

図 1　語彙獲得アルゴリズムのフローチャート

2.2 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を用いて構築した複合辞書を用いたネガティブ・ポジティブ分析による経済動向予測手法を提案する．まず訓練データ内において 1 日に数十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複合エントリを抽出し，特徴語とした．

提案手法では，訓練期間内の各見出しに取り出されたl個の特徴語が生じている場合，入力をその特徴語の極性値，存在しない特徴語に関しては 0 とし，前日のニュースの見出しに出現した特徴量を SVM に学習させ，翌日の株価が上昇するか否かを予測した．

3. 結果

結果の詳細と考察は発表で述べる．

参考文献

[1]萩尾智彦, 佐野睦夫: ローソク足を学習させた畳み込みニューラルネットワークによる仮想通貨価格予測,第 28 回 金融情報学研究会,pp.51-55(2022).
[2]片寄諒介, 吉岡真治: 複数のテクニカル指標を用いた市場動向の予測, 第 34 回人工知能学会, 3Rin4-12(2020).
[3]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).
[4]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).
[5]高村大也, 乾孝司, 奥村学 : スピンモデルによる単語の感情極性抽出, 情報処理学会論文誌, Vol.46, No.2, pp.627-637(2006).
[6]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).
[7]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).