The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

深層学習を用いた新聞記事分析による市場動向予測における予測精度の改善

Improvement of Prediction Accuracy in Predicting Market Trends by Newspaper Article Analysis Using Deep Learning

松本 一樹 *1　　松井 藤五郎 *2*3
Kazuki Matsumoto　　Tohgoroh Matsui

*1中部大学 大学院 工学研究科 情報工学専攻
Department of Computer Science, Graduate School of Engineering, Chubu University

*2中部大学 生命健康科学部 臨床工学科
Department of Clinical Engineering, College of Life and Health Sciences, Chubu University

*3中部大学 工学部 情報工学科
Department of Computer Science, College of Engineering, Chubu University

In this paper, we analyze newspaper articles using deep learning to forecast market trends. We have proposed a method to forecast market trends based on time-series text analysis using deep learning. This method works very well for forecasting TOPIX from The Nikkei (Nihon Keizai Shinbun) between 2008 and 2014, but the prediction accuracy falls after 2015. In this paper, we propose to reduce the duration of the training data in order to improve the prediction accuracy after 2015. As a result of the period of training data over the past three years, the prediction accuracy has been improved by 12.2%, from 55.1% to 67.3%.

1. はじめに

金融市場の動向を予測する研究は古くから行われている。これまで、金融市場の動きを予測する研究として、過去の取引実績の時系列パターンから株価を予測するテクニカル分析や、財務諸表などから株価を予測するファンダメンタル分析が主流であった。これらの手法はいずれも定量的データを用いた分析である。しかし、定量的データを用いた分析だけでは予測できない市場の動きがある。例えば、ある企業の不正会計が発覚したとき、その企業の株価は急落する。このような株価の動きは定量的データの分析だけで予測することはできない。そこで、近年、テキスト分析を用いて金融市場の動きを予測する研究が盛んに行われている [和泉 11, 藏本 13]。

テキスト分析による金融市場の動向予測において、株価や金利といった市場データである目的変数が時系列を成すように、金融市場の動きを説明するのに用いられるテキストも時系列をなしている。そのため、新聞などの定期的に発行されるテキスト・データは、一種の時系列データであると考えることができる。これを利用して、前日の新聞記事と当日の新聞記事を比較し、その差分に基づいて特徴語の出現パターンを抽出し、SVM を用いて東証株価指数 (TOPIX) が上昇するか下落するかを予測する手法が提案されている [松井 11]。

また、近年、パターン認識の分野において、近年深層学習が注目を集めている。深層学習は、画像認識や音声認識において高い認識精度を発揮しており、テキスト分析においても従来手法と比べ高い性能であることが確認されている [Socher 13]。そこで我々は、SVM の替わりに深層学習を用いたところ、時系列テキスト分析を用いた市場動向予測においても、深層学習を用いた予測は、2008 年から 2014 年までの予測正解率が 91.3%であった [松本 17]。しかし、2015 年、2016 年を対象とした実験では、2015 年は 59%、2016 年は 55.1%と予測正解率がそれ以前の予測正解率から大幅に低下してしまった（図 1）。

連絡先: 松本一樹，中部大学 大学院 工学研究科 情報工学専攻，
tp17028-2903@sti.chubu.ac.jp

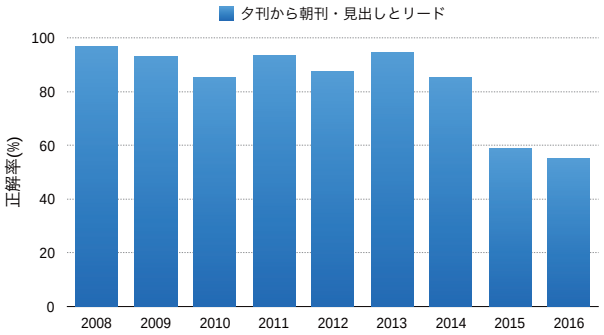

図 1: 予測正解率 [松本 17]

そこで本論文では、新聞記事を時系列データと捉えてテキストの差分に着目した分析を行い、深層学習を用いて株価が上昇するか下落するかを予測する手法において、2015 年以降において予測正解率が低下した原因について調査し、予測精度向上を図る。

2. 深層学習を用いた時系列テキスト分析による市場動向予測

まずはじめに、新聞記事を対象とした、深層学習を用いた時系列テキスト分析によって市場動向を予測する手法について説明する。

対象とする新聞記事は、日本経済新聞の見出しとリードとし、市場動向に影響を与えるテキストであると考えられる予測対象日の前営業日の夕刊から当日の朝刊を結合したテキストを用いる。

テキスト分析に用いる深層学習は、ニューラルネットワークを多層化したものであり、画像認識や音声認識で高い認識精度を発揮している。テキスト分析においても従来手法と比べて高

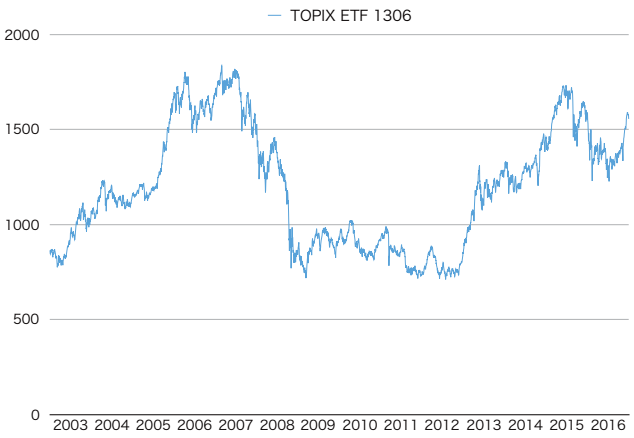

図 2: TOPIX ETF 1306 の株価推移

い性能であることが確認されている。深層学習への入力は、先行研究 [松井 11] で作成された特徴量ベクトルのデータを CSV 形式に変換したものを用いる。深層学習の出力は、予測対象日の日中の利益率が正または 0 のとき $(0,1)$、そうでないときは $(1,0)$ とする。新たに予測を行うときは、テキストを特徴量ベクトルに変換してニューラルネットワークへ入力し、その出力 (o_1, o_2) を求め、$o_1 < o_2$ なら利益率は正になると予測し、そうでないなら負になると予測する。

3. 提案手法

本論文では、従来手法において訓練データの期間を直近過去 5 年間であったのを、直近過去 3 年間として、特徴量ベクトルの作成と深層学習の学習を行う手法を提案する。

従来手法の問題点として、2015 年、2016 年を対象とした実験において、予測正解率が 2015 年以前と比べて大幅に低下したことがあげられる。2015 年、2016 年を対象とした実験では、2015 年は 59%、2016 年は 55.1%であった [松本 17]。2016 年を予測する際、訓練データの期間を直近過去 5 年間とすると、訓練データに 2013 年以前のデータが含まれる。しかし、2012 年にアベノミクスが始まり、2013 年頃からアベノミクスの効果が現れ始めたため、2013 年以前と以降では市場の動きが異なると考えられる（図 2)。

そこで、訓練データの期間を過去 5 年間より短くすることで、2016 年の予測における訓練データに 2013 年以前のデータが含まれないようにする。このように訓練データの期間を短くすることによって、2015 年以降における予測精度の向上が期待できる。

4. 実験結果と考察

提案手法の有効性を確認する為に検証実験を行った。日本経済新聞を対象として、予測対象日の前営業日の夕刊から予測対象日の営業日の朝刊までを 1 つのテキストとし、その見出しとリードを用いた。予測対象は 2016 年の東証株価指数 (TOPIX) 連動型上場投資信託 (ETF) とし、予測対象日の寄りから引けにかけて TOPIX ETF の取引価格が上昇するか下落するかを予測した。訓練データの期間を予測対象日の直近過去 3 年間とした。

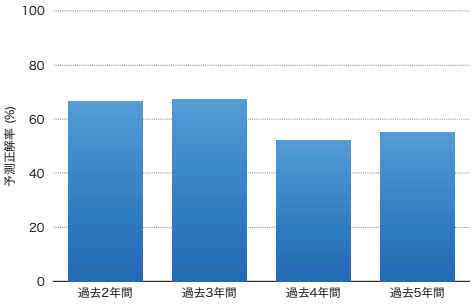

図 3: 予測正解率 (2016 年)

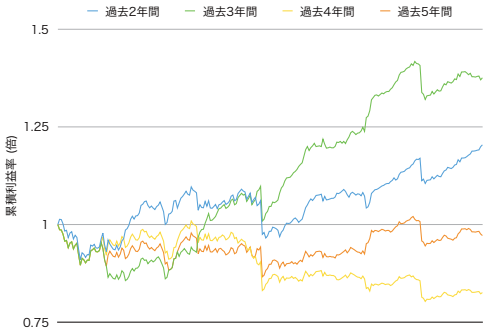

図 4: 累積利益率 (2016 年)

特徴語はテキストの差分に基づいて新出、続出、消滅の 3 つの出現パターンに分類する。分類した特徴語を出現パターンごとに出現回数をカウントし、30 回以上出現したパターンのみを対象とし、パターンが出現した日に上昇した割合が 0.55 以上または 0.45 以下のパターンを特徴として用いた。

深層学習のツールは TensorFlow を使用した。深層学習のユニット数は、入力層を特徴語数分のユニット、出力層を 2 ユニット、隠れ層は 2 層とし、1 層のユニット数を 64 ユニットとした。また、活性化関数には ReLU を用いた。

比較として、訓練データの期間を過去 5 年間、4 年間、2 年間にしたものを用いた。この時、それぞれの特徴量ベクトルは、特徴として用いる出現パターンの出現回数を、過去 2 年間の場合 20 回以上、過去 4 年間の場合 40 回以上、過去 5 年間の場合 50 回以上のパターンを対象として特徴量ベクトルを作成した。

実験結果を図 3、図 4 に示す。 実験の結果、従来手法である訓練データの期間が直近過去 5 年間では予測正解率が 55.1%であったのに対して、訓練データの期間を直近過去 3 年間にした場合では 67.3%となり、従来手法と比べて予測精度が 12.2%上昇した。また、過去 4 年間では 52.2%、直近過去 2 年間では 66.5%となった。

訓練データの期間を直近過去 4 年間にした場合と直近過去 2 年間にした場合において、直近過去 3 年間にした場合と比べて予測正解率が低下した原因として、直近過去 4 年間の場合では、アベノミクスの影響を受ける前である 2012 年の新聞記事が訓練データに含まれているためであると考えられる。

また、直近過去 2 年間の場合では、2013 年以前の新聞記事が訓練データに含まれていないものの、同じく 2013 年以前の

新聞記事が訓練データに含まれていない直近過去 3 年間の場合に比べて、訓練データが減っているために、直近過去 3 年間の場合に比べて予測正解率が低くなっていると考えられる。

本手法による予測に基づいて 1 年間の取引を行った場合の累積利益率を図 4 に示す。TOPIX ETF の価格が上昇すると予測された場合には、寄り付きに始値で買い、引けに終値で売る、下落すると予測された場合には、寄り付きで始値で信用売り、引けに終値で買い戻すことを毎日続けた場合のシミュレーションである。ただし、取引手数料は考慮していない。

1 年後の累積利益率は、直近過去 5 年間を訓練データの期間とした場合は約 0.97 倍、直近過去 3 年間を訓練データの期間とした場合では約 1.38 倍となった。このため、提案手法は有効であると考えられる。

5. まとめ

本論文では、深層学習を用いた新聞記事分析による市場動向予測と、2015 年以降において予測正解率が低下した原因について調査し、訓練データの期間を短くすることによって予測精度の向上を図った。日本経済新聞を対象として、夕刊から朝刊までを 1 つのテキストとし、その見出しとリードに対して時系列テキスト分析を行い、TOPIX ETF の取引価格が上昇するか下落するかを予測した。実験の結果、提案手法である訓練データの期間を直近過去 3 年間にした場合では予測正解率が 67.3%となり、訓練データの期間が直近過去 5 年間だった従来手法と比べて予測精度が 12.2%上昇した。よって、提案手法は有用であると考えられる。

今後は、2016 年以外の予測期間においても訓練データを短くすることが有効であるか確認したい。

参考文献

[和泉 11] 和泉 潔, 後藤 卓, 松井 藤五郎: 経済テキスト情報を用いた長期的な市場動向推定, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3309–3315 (2011)

[藏本 13] 藏本 貴久, 和泉 潔, 吉村 忍, 石田 智也, 中嶋 啓浩, 松井 藤五郎, 吉田 稔, 中川裕志: 新聞記事のテキストマイニングによる長期市場動向の分析, 人工知能学会論文誌, Vol. 28, No. 3, pp. 291–296 (2013)

[松井 11] 松井 藤五郎, 石田 智也, 中嶋 啓浩, 和泉 潔, 吉田 稔, 中川 祐志: 新聞記事を対象とした時系列テキスト分析による市場予測, 第 7 回人工知能学会ファイナンスにおける人工知能応用研究会 (SIG-FIN), pp. 44–47 (2011).

[Socher 13] Socher,R., et al : Recursive deep models for semantic compositionality over a sentiment treebank. *EMNLP 2013*, pp. 1631–1642 (2013).

[松本 17] 松本 一樹, 松井 藤五郎: 深層学習を用いた新聞記事分析による市場動向予測. 第 79 回情報処理学会全国大会, 6N-08 (2017).