株価予測における TCN-LightGBM 複合モデルの適用

中居 幸太[redacted]　李 径直[redacted]　菓子尾 伊織[redacted]　指導教員 劉 慶豊

1. はじめに

これまで，株価予測の研究では AR モデルや ARMA モデルなどの自己回帰モデルを用いた分析が行われてきた．

近年では，RNN や LSTM などのディープラーニングを用いた研究が盛んに行われており，文献 [1] では，LSTM の予測値を LightGBM モデルの特徴量として用いることで，電力消費予測モデルの精度を向上させている．

本研究では，文献 [1] で用いられた LSTM の代わりに，畳み込みニューラルネットワーク (CNN) の一種である時間的畳み込みネットワーク (TCN) を用いることで，株価変動 (上昇・下落) の予測精度を向上させられるか検証を行う．

2. データについて

2.1 データの入手方法

本研究では，Investing.com[2] という金融データを提供するポータルサイトから入手した，2014 年〜2018 年の S&P500 株価指数のデータを用いる．

また，分析に用いる S&P500 に関するニュースのヘッドラインは，kaggle の S&P 500 with Financial News Headlines (2008–2024)[3] というコンペティションのデータを使用する．

2.2 特徴量

本研究では，目的変数を当日の終値が前日の終値より上昇したか下落したかの二値変数とする．

特徴量は，終値・S&P500 に関するニュースヘッドライン [3] のセンチメントスコアの 1 日から 5 日前までのラグデータ，前日の米国債利回り，ボリンジャーバンド・SMA・MACD・RSI といったテクニカル指標を使用する．

センチメントスコアについては，FinBERT という金融分野に特化した大規模言語モデルを使用して感情分析をおこない，-1〜1 の範囲で数値化する．

3. 使用モデル

3.1 TCN(時間的畳み込みネットワーク)

TCN(時間的畳み込みネットワーク) とは，株価のような時系列データに対して CNN(畳み込みニューラルネットワーク) を用いたアルゴリズムである [4].

TCN は，未来のデータを使わないことで因果性を守り，入力の幅を空けることで，広い範囲を効率的に捉えることができる．本研究では，筆者が公開している GitHub のリポジトリ [5] を参考に，プログラムを作成した．

3.2 LightGBM

LightGBM とは，Microsoft によって開発された，複数の決定木を使った勾配ブースティングアルゴリズムである．

計算速度が速いうえに予測精度も高く，kaggle や SIGNATE といったデータ分析コンペティションなどで幅広く使用されている．

同様のアルゴリズムを用いた XGBoost は，決定木の分岐を深さごとに増やしているが，LightGBM では葉単位で目的関数を減らせるように分岐を増やしている．

3.3 TCN-LightGBM 複合モデル

TCN-LightGBM 複合モデルは，本研究で提案するモデルである．

文献 [1] では，LSTM の予測値を用いていたが，本研究で使用する提案モデルは LightGBM の特徴量に TCN の予測値を追加したものである．

実験では，TCN の予測値を使わないモデル (以下，LightGBM 単体モデルと定義する) と，提案モデルの予測性能を比較する．

4. 実験概要及び結果

提案する TCN-LightGBM 複合モデルの有効性を検証するために，LightGBM 単体モデルとのアブレーション実験を行った．

4.1 2018 年の株価予測

はじめに，2014 年〜2017 年を訓練データ，2018 年をテストデータとして，株価予測を行った．結果を表 1 に示す．

LightGBM 単体モデルの上昇予測の正解率は 52.2%，下落予測の正解率は 50.9%，全体の正解率は 51.9%となった．

一方，提案モデルの上昇予測の正解率は 53.1%，下落予測の正解率は 55.3%，全体の正

表 1: 2018 年の株価予測結果

LightGBM 単体モデル

実際＼予測	上昇	下落
上昇	96	28
下落	88	29

TCN-LightGBM 複合モデル

実際＼予測	上昇	下落
上昇	103	21
下落	91	26

解率は 53.5%と，単体モデルの精度をやや上回った．また，TCN の予測値の特徴量重要度は特徴量 18 個のうち 15 番目であった．

予測結果から，正解率は当て推量から少し改善されただけという課題が残った．これに関しては，テストデータである 2018 年の株価の変動がこれまでの傾向と変わっていることが原因である可能性があるため，テストデータを 2017 年に設定して同様の実験を再度行う．

4.2 2017 年の株価予測

次に，2014 年〜2016 年を訓練データとし，株価の変動が似ている 2017 年をテストデータとして予測を行った．結果を表 2 に示す．

表 2: 2017 年の株価予測結果

LightGBM 単体モデル

実際＼予測	上昇	下落
上昇	119	24
下落	86	22

TCN-LightGBM 複合モデル

実際＼予測	上昇	下落
上昇	123	20
下落	89	19

LightGBM 単体モデルの上昇予測の正解率は 58.0%，下落予測の正解率は 47.8%，全体の正解率は 56.2%となった．

提案モデルの上昇予測の正解率は 58.0%，下落予測の正解率は 48.7%，全体の正解率は 56.6%と，こちらも単体モデルの精度をやや上回った．また，TCN の予測値の特徴量重要度は特徴量 18 個のうち 14 番目であった．

2018 年の株価変動を予測した際と比べると，下落予測の精度が下がったものの，全体の正解率は上昇した．

5. 考察

2018 年の株価と比べて，2017 年の方が予測精度が高かったのは，変動傾向が訓練データと似ていたからであろう．

レジームスイッチなどの影響により，株価の傾向は急に変わることがあるため，4.1 節の実験ではモデルドリフトに影響を受けた可能性がある．

本モデルを運用するのであれば，継続的に学習サイクルを繰り返す必要があろう．

また，特徴量に TCN の予測値を加えたが，特徴量重要度は 2017 年で 18 個中 14 番目，2018 年で 18 個中 15 番目とあまり高くなかった．文献 [1] の実験は回帰問題であったが，本研究は分類問題であったため寄与度が低かった可能性がある．回帰的に株価を予測するのであれば，TCN の予測値の寄与度が高くなるかもしれない．

6. おわりに

本研究では，提案した TCN-LightGBM 複合モデルを用いて，S&P500 株価指数の変動について予測を行った．

結果として，予測精度は向上したものの，株価の傾向が変わると精度は下がってしまった．モデルドリフトの影響を軽減するためにも，株価予測モデルを定期的に再訓練する必要があろう．

また，本研究では TCN の予測値を用いたことにより予測精度の向上が見られたが，特徴量重要度はあまり高くなかった．

重要度も高く，予測精度向上に大きく寄与する特徴量を調査することが今後の課題である．

参考文献

[1] You Zhou et al., Application of LSTM-LightGBM Nonlinear Combined Model to Power Load Forecasting, 2022

[2] Investing.com, https://jp.investing.com/indices/us-spx-500-historical-data+, (参照 2025 年 12 月 31 日).

[3] kaggle: S&P 500 with Financial News Headlines (2008–2024), https://www.kaggle.com/datasets/dyutidasmahaptra/s-and-p-500-with-financial-news-headlines-20082024, (参照 2025 年 12 月 31 日).

[4] Shaojie Bai et al., AnEmpirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, 2018

[5] TCN, https://github.com/locuslab/TCN, (参照 2025 年 12 月 31 日).