ツイートデータとインデックスデータを併用した株価騰落予測

黒崎 地大(19X4026)　山田 壮雄(19X4134)　指導教員 劉 慶豊

1. はじめに

近年，寿命の増加や NISA，iDeCo の拡充などの影響により自己資産形成の必要性が増している. それに伴い様々な投資家や機関によって株価変動を予測する研究が進められている. 機械学習の分野ではニュートラルネットワークやブースティングなどの方法が株価予測の研究に活用されている. 機械学習により株価や為替レート, 経済指標などの大量のデータを学習することでより高度な予測が可能になると期待されている. 本研究では，機械学習を用いて Twitter の大量のテキストデータから, 投資家の心理状況を反映するシグナルを抽出した上で, インデックスデータと併用して株価騰落予測を行い,ツイートデータ併用の効果を検証した.

2. 利用モデルおよびアルゴリズム

2.1 BERT

BERT(Bidirectional Encoder Representations from Transformers)[1]は注目されている最新な自然言語処理モデルの一つである. Transformer による双方向のエンコーディングを行うことで，Biderectional の文字通り文脈を加味した文章の解釈が可能になっている. BERT モデルの作成は，事前学習とファインチューニングの 2 ステップで行われる．ステップ 1 の事前学習では 2 つのタスクがある．1 つ目は Masked Language Model で，入力文の 15%を[Mask]に置き換え元の単語を予測させる．2 つ目は Next Sentence Prediction であり，2 つの入力文に対して「その 2 文が隣り合っているか」を当てるよう学習させる. 事前学習が完了すれば， ステップ 2 で目的タスクに合わせた教師あり学習を行う（ファインチューニングを行う）ことでモデルが完成する. 本研究では，BERT モデルによってツイートデータのセンチメントを数値化したものを LightGBM に組み込んで株価の予測に使用する.

2.2 LightGBM

LightGBM[2]とは複数の決定木を用いた勾配ブースティング（Gradient Boosting）の機械学習フレームワークである. 勾配ブースティングとは複数の弱学習器,（LightGBM の場合は決定木）を 1 つにまとめるアンサンブル学習の一種である. アンサンブル学習をすることで 1 つの学習モデルで予測するより予測精度を向上させることが可能である.

2.3 TPE

TPE はベイズ最適化を用いて機械学習モデルのハイパーパラメーターチューニングを行う方法である. ハイパーパラメーター最適化の方法は他に手動でパラメータを調節し, 実験的に最適なパラメータを見つける方法とグリッドサーチと呼ばれる総当たりのアルゴリズムによりパラメータを探索する方法がある. しかし, 前者の方法ではモデルやパラメータに対する深い理解と経験による勘が必要であり, 後者の方法では組み合わせが増えると計算量が増加し時間がかかるという欠点がある．そこで, 本研究では経験や勘の必要がなく, かつグリッドサーチよりも高速な TPE アルゴリズムを用いる.

3. 使用データの概要

分析に使用するテクニカルデータとして CFD ブローカーである Capital.com が配信するデータを用いる. 予測対象はナスダック 100 指数の騰落である. 対象期間は 2021 年 4 月 1 日から 2021 年 10 月 1 日までとし，各 1 時間足の終値を分析に使用する. 特徴量(説明変数)としてナスダックの価格データ, 出来高データ, ダウ平均, ドルインデックスおよび金の価格データのラグを分析に利用する. 分析対象のツイートデータは, Twitter 社が提供するツイートの参照や検索を行うサービスである Twitter API を用いて収集した. 2021 年 4 月 1 日から 2021 年 10 月 1 日までに投稿された,「Nasdaq」の文言を含む英文ツイート 479,293 件を用いる. 取得したツイートについて, BERT モデルを用いてセンチメントスコアを抽出したものを特徴量として用いる. 加えて, 各ツイートの Like 数も投資家の銘柄への注目度を表す 1 つの特徴量として利用した.

4. 実験概要

4.1 分析の概要

LightGBM を用い，現在から 5 期後の騰落予測を行う. 最初にテクニカルデータのみを使用して予測を行ったベースモデルを作成したのち, ツイートデータを特徴量として加えた併用モデルを作成しベースモデルとの比較を行い併用による効果を検証する．評価関数には Logloss を用いる．交差検証を行うために，期間の異なるデータを 5 つの Fold にして，ハイパーパラメーターのチューニングとモデルの学習を行う[図 1]．テストデータによる評価では 5 モデルの予測値の平均をとったアンサンブル結果を最終の予測とする．

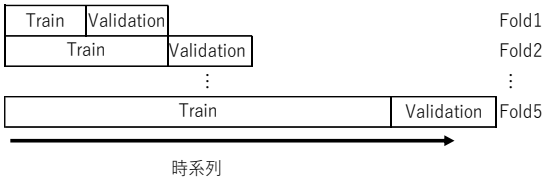

図 1 5-fold

5. 実験結果

テクニカルデータのみのベースモデルとセンチメント，ツイートの Like 数から作成した特徴量を加えた併用モデルに関して，各 Fold での Validation データの損失関数 Logloss の値，および損失値の平均を表 2 にまとめた．

表 2 分析結果

	併用モデル	ベースモデル
Fold1	0.68476	0.68737
Fold2	0.68037	0.68224
Fold3	0.68180	0.68193
Fold4	0.68091	0.68234
Fold5	0.68076	0.68409
損失値の平均	0.68172	0.68360

ベースモデルと比較すると, ツイートデータを併用したモデルでは各 Fold で損失値が小さくなり, 損失値の平均も小さくなる結果となった. ツイートデータ併用による効果が見られたと言える. 次に，テストデータにおいて評価を行った結果を示す[図 2]. ベースモデル(図 2 左側)では，上昇予測の正解率は 54.61%，下落予測の正解率は 53.40%であった．これと比較して，併用モデル（図 2 右側）では上昇予測の正解率が 58.61%，下落予測の正解率は 56.83%と上下どちらのクラスも予測精度が向上する結果となった．

ベースモデル

実際＼予測	UP	DOWN
UP	148	151
DOWN	123	173

ツイートデータ併用モデル

実際＼予測	UP	DOWN
UP	160	139
DOWN	113	183

図 2 混同行列による評価結果

6. 分析結果の考察

モデルの特徴量重要度を比較すると, ツイートデータを用いて作成した特徴量のうち Like 数の特徴量はモデルの予測に大きく寄与している結果となった. それと比較し, 感情スコアの特徴量は重要度寄与度が小さくあまり重要な特徴量ではなかったことがわかった. これについて, 今回作成したセンチメントの数値化指標では投資家の心理情報を完全には表現できていない点が考えられる. ツイートデータの処理にはまだまだ改善が必要であるといえる.

7. 終わりに

本研究ではナスダック 100 指数に対し，テクニカルデータとツイートデータを併用し, 数量データに加えてテキストデータを活用した機械学習による騰落予測モデルを作成した. 結果として, 併用モデルではより高い予測精度を示し, ツイートデータを取り入れたことの有効性を示した. 本研究ではツイートデータのセンチメントと Like 数が株価に影響しているという仮定の下で分析を行ったが，この 2 つの特徴量ではテキストデータのもつ情報をすべて表しているとは言えない. 今後は，より多くの特徴を抽出できるようなテキストデータの処理を行うことでさらなる精度向上が期待できると考えられる.

参考文献

[1] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[2] LightGBM 徹底入門 – LightGBM の使い方や仕組み，XGBoost との違いについて https://www.codexa.net/lightgbm-beginner/ アクセス日　2022 11/23