G-24

機械学習を用いた株価の予想

Stock price prediction using machine learning approaches

○山崎将吾[1]
*Shougo Yamazaki[1]

Abstract: Artificial intelligence is pervasive nowadays and has been exploited in stock markets for various purposes. In this study, some machine learning methods are applied to predict stock prices. It is then investigated which method provides a better predicting of stock price.

１．はじめに

近年, 機械学習の分野において深層学習法を用いた人工ニューラルネットワークモデルが目覚ましい発展を遂げ, 機械学習において現在では深層学習法が代表的な手法の一つであるが, その理由は, 深層学習法が適切な特徴量を獲得できる点が従来の機械学習の手法を上回ることに大きく貢献したからである[1]. 一方, 深層学習を含むニューラルネットワーク系の機械学習はブラックボックス型で, 実行結果について説明することができないといった問題がある[2].

本論文では, そのようなブラックボックス型の深層学習ではなく, 特徴量をあらかじめ設定するような従来の機械学習法である決定木分析による手法により株価の予想を行い, どのような特徴量が株価の予想において効果的であるかを検証し報告する.

２．機械学習とは

機械学習とは文字通り機械による学習を実現しようとする技術であり, 従来の手法ではルールのすべてをあらかじめ設定した上で用いられていたのとは異なり, 与えられたデータのみから機械に学習させるところに特徴がある[1]. これにより目視のみでは気づくことが容易ではないとされる規則性などを発見することが可能となり, より高度な予測を行うことが可能であると考えられる.

３．決定木分析

本研究において検証に用いたアルゴリズムは決定木分析である. 決定木分析とは, データを複数のクラスに分類する教師あり学習の１つで, 「樹々モデル」と呼ばれる木構造を利用した分類アルゴリズムである[1].

決定木分析のパラメータは数多くあるが, 今回用いたパラメータは決定木の弱点である過学習を防ぐため, 過学習に影響が大きいと思われるもののみを抽出し, またその抽出した全てのパラメータの組み合わせについてグリッドサーチによりパラメータの最適化を行った.

Table 1 は決定木分析で用いたパラメータである[3].

パラメータ	意味 (パラメータの範囲)
max_depth	木の最大深度(1~10)
min_samples_split	ノード(節)を構成するために必要なサンプルサイズ(1~10)
min_samples_leaf	葉を構成するために必要なサンプルサイズ(1~10)
criterion	"gini"か"entropy"の選択
splitter	"best"か"random"の選択

Table 1. 決定木分析を行う際に用いられるパラメータ

４．評価に用いる指標

本検証の評価を行うための指標として正答率と F 値を用いた. なお, 指標はこれら 2 つのほかに, 適合率と再現率があるが, これらは互いにトレードオフの関係にあり, 他のアルゴリズムとの比較する上で適合率と再現率は適さないと判断されたため, 本検証では用いらなかった.

まず, 予測に対して正解がどの程度あるのかを比較するために正答率を選択した. 次に, 適合率と再現率の調和平均で両者を総合的に比較できる F 値を用いた.

５．検証方法

本研究では予測する銘柄として NSD, DTS, NTT データの 3 社の株価を選択した. またその下準備として, それぞれの銘柄の始値, 高値, 安値, 終値, 出来高の 6 項目を 100 日分～500 日分を 100 日間隔で 5 個の csv ファイルを作成した[4]. なお csv ファイル作成に際し, [4]のサイトより直接作成した.

まず, それぞれの項目に対し前日に対する上昇率を計算する. また, それぞれの項目を説明変数とし, 上昇率

１：日大理工・学部・数学

が正なら 1，負なら 0 を目的変数として格納し，それらをもとに学習し，予測を行い，すべてのパラメータの結果を excel ファイルにまとめる。

次に，3 社とも結果が良かった日数に順位をつけた後に，3 社の合計順位を出し，1 位の日数の周辺で再度 csv ファイルを数個作成する．

上記を繰り返し行い，何日分の株価データを用いる場合が最も良い結果が得られるか，また，その際に用いられたパラメータを最終的に excel プログラムの MAX 関数により求める．

6．random_state

決定木分析を行うにあたり，random_state という乱数シードを固定するための Table 1 にはないパラメータが存在する．このパラメータの引数を整数にすると，与えられる整数値により実行結果が異なる．デフォルトの場合は numpy.random を使った RandomState のインスタンスでメルセンヌツイスタにより乱数が生成される[5].

よって乱数シードを固定することや，乱数シードをデフォルトのままで検証を行った場合には同じパラメータでも毎回実行結果が異なるため，他のアルゴリズムとの比較ができないと考えられる．そこで，乱数の影響を極力なくすため，乱数シードを固定せず，同じパラメータで 100 回実行し，その平均を計算することにより， 用いたパラメータによる結果として評価し検証した．

7．検証結果

NSD の検証結果を Figure 1 に示す． Figure 1 は上述の検証方法の通りに検証に用いるデータの個数(日数)を狭めていった最終段階で，各データの個数(日数)ごとにグリッドサーチを行い，その結果の箱ひげ図である。Figure 1 で示されるように data70(データの日数が 70 日分)の時が最も良い結果が得られた．

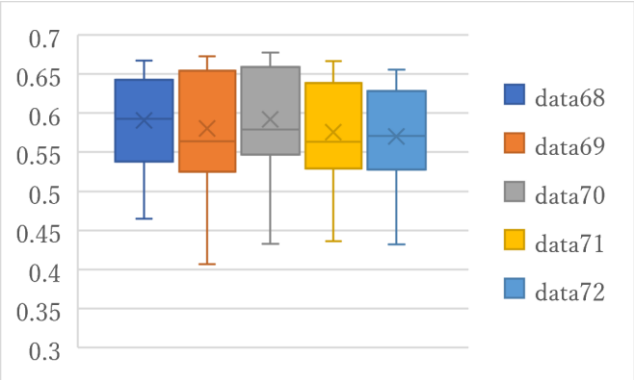

Figure 1. NSD の正答率

次に残りの 2 社も同様に検証し，データ数ごとの順位付けをすると data70 の時が最も良い結果が得られることが確認できた．

なお，data70 における各社の最適値は Table 2 の通りである．

銘柄	正答率の最大値
NSD	0.677289377
DTS	0.69543956
NTT データ	0.647619048

Table 2. 各銘柄の正答率の最大値

ここで，他の手法と比較する際の正答率は上記 3 銘柄の最大値の平均を取り，0.673449328 とした．

さらに，F 値についても正答率と同様に検証したところ，正答率と同様に data70 の時が最も良い結果が得られ，その場合 3 銘柄の平均は 0.673258852 となった．

以上の検証結果より，最も結果が良い時のパラメータは銘柄やデータの数によって全く異なるものと考えられる．

学術講演会当日では，本研究で用いた決定木分析の他，ランダムフォレストや SVM(サポートベクターマシン)による検証を行い，それらとの比較，および検討したことについて報告する予定である．

8．参考文献

[1] 株式会社システム計画研究所 ：「Python による機械学習入門」， オーム社 ， pp 31, 3, 56， 2016 年.

[2] 福島 俊一, 藤巻 遼平, 岡野原 大輔, 杉山 将 ：「ビッグデータ×機械学習の展望：最先端の技術的チャレンジと広がる応用」，情報管理, Vol.60, No.8, pp.543-554, 2017 年.

[3] http://data-analysis-stats.jp/2019/01/14/決定木分析のパラメータ解説/ (パラメータの選択).

[4] https://kabuoji3.com/stock/ (株価データの抽出).

[5] https://docs.scipy.org/doc/numpy-1.12.0/reference/routines.random.html#random-generator .