Prediction of Stock Index Based on Ensemble Learning

Mengfei Xia[1], Ruihan Qiu[1], Chenyu Wang[2] and Yilin Zhang[1]

***Abstract*— Stock index prediction has long been paid close interest to since the growth of modern stock market. Machine learning algorithms play a significant role in stock analysis and quantitative trading. There are several papers on relevant topic, each correlated with several classification, regression or clustering models. Besides, deep learning methods are also essential to this topic. Based on the implications of former papers and certain knowledge about stock market and stock index, a stock index prediction method is developed in this article. Two main steps are involved in this research. As the first step, clustering is utilized to separate different types of stocks in order to improve the accuracy of the following step, classification. Furthermore, voting method is used to ensemble various classification algorithms. At last, we get a fairly good accuracy score on the test set compared with average stock prediction accuracy.**

I. INTRODUCTION

The research on investment analysis using machine learning originated in the United States at the end of the 19th century. With the deepening of research, machine learning has been recognized in the financial field and has been widely used in the field of stock market trading. It is mainly used to forecast stock trends. With the help of proper algorithms, investment institutions manage to choose the right portfolio which can lead to a higher return and a relevantly lower risk by cross hedging.

Domestic institutions did not use machine learning methods for investment analysis until the 21st century. By now, various hedge fund and quantitative trading department are becoming more and more common. According to statistical results, up to 70% of trading in the US is made by quantitative method, with the proportion being only 5% in China but in a rapidly growth period nowadays. The machine learning algorithms utilized in stock market forecast range from Support Vector Machine (SVM)[7], Deep Neural Network (DNN)[2] to Random Forest (RF). Other methods like time series analysis[1] and Principal Components Analysis (PCA) are also effective ways in China given that Chinese stock market is just under the weak-form market efficiency.

This project is an implementation of the paper[5], and we will mainly use clustering algorithms and classifier models to predict the types of stock index changes on the $T+1$ trading day. The clustering evaluation is based on both the output score and the real-life situation. Besides, in order to get a better result, a voting process of various algorithms will be utilized.

[1]Mengfei Xia, Ruihan Qiu and Yilin Zhang are undergraduate students in Department of Mathematical Sciences, Tsinghua University

[2]Chenyu Wang is an undergraduate student in School of Economy and Management, Tsinghua University

II. RELATED WORKS

Domestic research on the use of machine learning for investment analysis did not begin to emerge until the 21st century, so there are fewer related studies than abroad. Most of the research methods are based on existing models and then tested in the Chinese stock market.

In [4], the experiment combines the wavelet theory with the SVM method and combines the advantages of both to propose a new machine learning method called Wavelet Support Vector Machine (WSVM). This method introduces a wavelet basis function to construct the kernel function of SVM, so that in addition to the advantages of SVM, the model can also eliminate high-frequency interference of data and has good noise immunity. This article applies this method to the prediction of stock price index, and obtains better prediction accuracy.

In [3], in order to deeply dig out the information of technical analysis indicators, the article designed a set of machine learning and technical analysis (ML-TEA) based on machine learning and technical indicators. This model uses machine learning algorithms to mine a variety of Common technical indicators to predict the ups and downs of the stock price in a few days (up or down), and then to build an investment portfolio based on the predicted direction. The strategy not only outperforms the market index in annualized returns but also outperforms the broad market index in nearly all risk indicators. It is a robust strategy with high returns and low risks.

III. PROBLEM DEFINITION

In our experiments, we will predict the rise and fall of the stock index, but redefine them: when the market's rise and fall reach a certain level, the individual stocks in the market will have a clear trend, and the individual stocks will have a relatively large probability follows the market trend. Therefore, when the stock index rises or falls sharply, the guiding role of the broader market index begins to appear. If the trend of the broad market index can be predicted correctly, it will have a certain guiding role for investors to avoid risks and invest in profits.

IV. ALGORITHM

The main idea of the algorithm is simplification, that is, by converting the complex change of stock index to a two-label problem "rise" and "fall". In order to do so, we first cluster the data with respect to two features: (1) change of Shanghai and Shenzhen 300 index, (2) the ratio between the numbers of rising stocks and falling stocks. We also compared the

cluster results between `KMeans` and `SpectralClustering` and list them below.

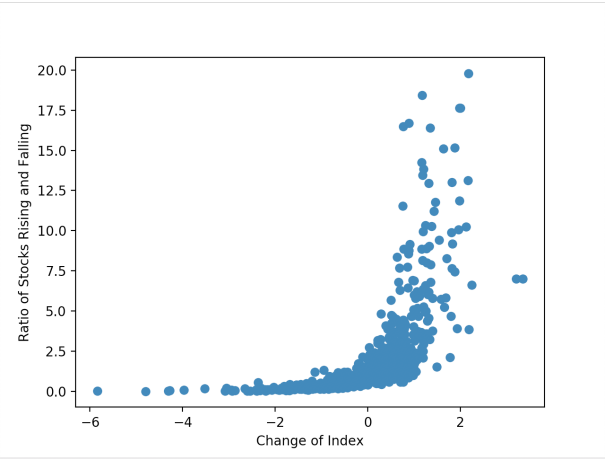

Fig. 1: Original Data

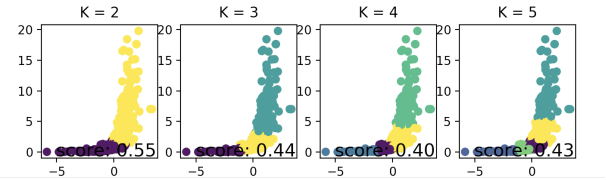

Fig. 2: Cluster Results with Distinct Hyperparameters

In view of both the `silhouette_score` and the cluster results, we set the hyperparameter `n_clusters` to be 4.

After dividing the data into 4 different sets, we combine them into 6 groups, i.e., there are 6 methods in all to sample 2 sets out of 4. Then we focus on each group and train a binary classifier step by step. Instead of choosing a simple model, we here design a classifier by voting to further improve the accuracy, consisting of three individual classifiers (1) `KNeighborsClassifier`, (2) `GradientBoostingClassifer` and (3) `RandomForest`. We have tried all the voting algorithms and finally what we use is the weighted voting one, i.e., the final prediction is a linear combination of each prediction of the models together with a weight value proportion to the accuracy and a hyperparameter. As for the whole classification task, what one needs to do is to get the final result from another voting of the results pf the 6 distinct vote-classifiers. We show the pseudo code below.

V. Experiments

Different from the common testing process, the prediction of stock index most focuses on the accuracy of prediction in the future. Rather than splitting the training and testing data randomly, we simply choose the former and latter parts of the data to be used in training and testing respectively. The baseline is set to the accuracy of the three models used in voting in 1. Besides the total accuracy, we pay higher attention to the accuracy of the predictions of the target

Algorithm 1 The hyperparameter `n_clusters` in cluster step is set to 4 for better performance, hyperparameters of weights for each base classifier is set to be 1, 10, 1 corresponding to `KNeighborsClassifier`, `GradientBoostingClassifer` and `RandomForest`.

```
Initialize the cluster results and combine into 6 distinct groups
for Each group do
    Training the three base binary classifiers
    Calculate accuracy and weight of each base classifier
    Get the prediction by weighted voting
end for
Get the final prediction by major voting
```

labels, i.e., the label of "rise" and "fall", so we list below this two metrics in the tabular.

Models	Accuracy	Target Accuracy
KNN	0.474	0.094
GradiantBoosting	0.590	0.562
RandomForest	0.549	0.375
Voting	0.636	0.594

One can see that our model raises the total accuracy by 5% and the target accuracy by 3%, which is of great practical significance in stock investment. Compare to our former results of 0.711 total accuracy, we focus on all the 29 stock features rather than only some of them, which makes the prediction far more difficult. However, our voting model still can fit this project and shows a great extensive applicability.

One can also notice that, there is a big gap of total accuracy and target accuracy, which is probably leading from the lack of samples with target labels and the sparsity. In fact, there are only less than 20% of the data with the target labels and the data with target labels are far more sparse than the one with the other two labels. When checking the final 6 predictions of each group, predictions of the other two labels account for a great proportion, which means that the base classifiers "believe" the data near the boundaries between the target labels and the others tends not to be with the target labels. This also implies the difficulty of stock predictions.

VI. Conclusions

The innovation of this paper lies in: First, we dig out the internal relationship between the change of index and the ratio between the numbers of rising stocks and falling stocks through `KMeans` clustering algorithm. Then we select the categories *sharp rise* and *sharp fall* that have a greater impact on the market as the focus of prediction. We define the rise and fall as a multi-category problem. Since we focus on the sharp change rather than the small change of index, the prediction accuracy is improved and our result has more practical significance.

Secondly, we apply the ensemble learning to the stock index prediction. The experiment result shows that

`RandomForest` is better than the other classifiers and the weighted voting algorithm performs better than other classifiers. Our result, the accuracy of 0.636 is significantly higher than the prediction result by SVM[6].

Thirdly, we select 29 technical indicators (of the T-1 day) as the *features* in the classification. In previous studies[7][6], The correlation between index rise-and-fall and technical indicators was higher than 0.95, which means they can reflect the change of index well. We select these 29 features to avoid over-fitting, under-fitting and linearity between features as much as possible and then normalize them.

In general, considering Chinese stock market itself a weak efficient market, influenced by policy and unexpected events, our model has performed very well.

VII. Outlook

We have not studied stock technical indicators thoroughly, and therefore our prediction model still has some limitations. The next step is to improve the model by trying more indicators. Furthermore, given that stock technical indicators are calculated by the daily *high, low, and closing prices* in a period of time, we consider predict the next day's stock price with time series. There is co-integration between high, low, and closing prices therefore we can use the VAR model to predict. Then we can calculate the technical indicators of the day and use them to classify instead of the previous day's indicators. This may further improve accuracy.

VIII. Source Code

Source code and all files for the whole project is available at `https://github.com/XIA943188/FML_Project`.

References

[1] Yilan Bao. Research on financial time series analysis and prediction algorithm based on support vector machine. *Dalian Maritime University*, 2013.

[2] Debao Dai, Yusen Lan, Tijun Fan, and Min Zhao. Research on stock index prediction and decision-making based on text mining and machine learning. *China Soft Science*, 04:160–164, 2019.

[3] Bin Li, Yan Lin, and Wenxuan Tang. Ml-tea: a set of quantitative investment algorithms based on machine learning and technical analysis. *Systems engineering – theory & practice*, 37(5):1089–1100, 2017.

[4] Yuancheng Li. Research on wavelet support vector machine in stock market prediction. *Computer Science*, 30(10):215–217, 2003.

[5] Ye Meng, Zhongqing Yu, and Qiang Zhou. Prediction of stock index based on ensemble learning. *Modern Electronics Technique*, 42(19):115–118, 2019.

[6] Donghai Ren. Predicting direction of stock price index movement based on support vector machines. *Shandong University*, 2016.

[7] Dengming Zhang. Optimization of technical index investment strategy and its application in quantitative trading. *Huazhong University of Science and Technology*, 2010.

IX. Acknowledgements

This research was encouraged by Prof. Mingsheng Long and all TAs in the course *Fundamental Machine Learning*.