Advances in Economics, Business and Management Research, volume 155
Proceedings of the 2020 2nd International Conference on Economic Management and Cultural Industry (ICEMCl 2020)

Research on Quantitative Investment Based on Machine Learning

Kaiwen Zhang[1,*]

[1]*Capital University of Economics and Business, 100070, Beijing, China*
[*]*Corresponding author. Email: 13161510636@163.com*

ABSTRACT
The stock market is a complex nonlinear system with low signal-to-noise ratio. Machine learning is used to model fuzzy nonlinear data and has proved to be a powerful tool in many fields. Machine learning has been continuously improved, and the successful application of the algorithms in the fields of computer vision, expert systems, etc. makes it obvious advantages to use machine learning methods to construct quantitative investment strategies. Stock selection is essentially a sorting problem. Investors all want to pick relatively better performing stocks. Therefore, this article discusses how to choose a more appropriate investment strategy in the investment process. This paper analyzes the basic situation of the application of machine learning methods in the field of quantitative investment in conjunction with the relevant technical background, and studies and constructs a rate of return prediction model based on the analysis. As a product of the current fusion research of quantitative investment and machine learning methods, the subject is a research hotspot in the industry and has strong practical guidance and reference value.
Keywords: *machine learning, deep learning, quantitative investment, quantitative stock picking*

1. INTRODUCTION

The economic downturn has aggravated the problems of inflation and currency depreciation. Coupled with the advent of the era of negative interest rates in banks, a single way of depositing disposable income into banks to accumulate wealth has not achieved an increase in personal wealth. With the rapid development of financial management, simple bank deposits have long been unable to meet the needs of investors. Investors need products with higher returns and higher returns. Therefore, high-yield stocks have become the most popular product among investors.

The factors that generate risks are complex and changeable. In order to avoid investment risks as much as possible and obtain high returns, most investors and institutions need to analyze the risk factors involved in investment. Since the data in the information age is growing like a blowout, it is impossible to rely on traditional manual analysis methods to accurately and timely select enough high-quality stocks from such a large amount of information, so as to configure the optimal investment portfolio and ensure that the risk remains within an acceptable range and the expected return can be maximized. With the rapid development of computer science and technology, a large number of cumbersome data analysis and processing tasks have gradually changed from manual execution to computer automatic operation. This change is also happening quietly in the financial world in pursuit of precision and efficiency. Subjective securities investment, once considered an art, has been gradually replaced by quantitative investment strategies attached to computers.

Quantitative investment is a type of investment strategy that uses mathematical methods to analyze and model financial markets. Machine learning requires computer programs to improve their performance when processing specific tasks through learning on data sets. Both need to extract information from the data, so with the success of machine learning in recent years, the industry and academia have developed a strong research interest in the combination of quantitative investment and machine learning methods. The successful application of machine learning methods to quantify investment makes our financial market more rational. Machine learning strategy design and computing resource thresholds are far beyond the scope of ordinary investors. When machine learning methods show obvious advantages in terms of returns and risks, market transaction volume will be concentrated on institutions using machine learning methods.

2. MACHINE LEARNING RELATED TECHNOLOGIES

Machine learning is a computer that simulates human learning behaviors, continuously acquires new knowledge and new skills, and reorganizes its own existing knowledge structure to continuously improve its performance. It emphasizes learning itself, first determining the features through manual analysis, and then using the relevant data analysis algorithm to find the feature data, identify the data pattern and predict, the sample data volume increases, the

features are strengthened, the program automatically corrects its own errors and learns, thereby improving the recognition ability.

Machine learning algorithms can be divided into: supervised and unsupervised learning, semi-supervised and reinforcement learning. Unsupervised learning does not preset sample features, without training, directly analyzes the sample data, and learns to classify data features, such as cluster analysis. Semi-supervised learning is somewhere in between, or it can be considered a combination of the two. It determines the execution action by observing the variables in the current environment, and then enters another observation environment. In this way, a certain law is repeatedly obtained, and the purpose is to take a certain action according to the law to obtain a certain maximize returns. There are many algorithms for supervised learning, including linear discriminant analysis, Bayesian classifier, perceptron, backpropagation algorithm, CART decision tree, random forest, distance metric learning, etc. The decision tree appeared in the 1990s and is a type of supervised learning. The algorithm is simple but the application effect is very good and the interpretability is strong. Common algorithms for machine learning are shown in Figure 1.

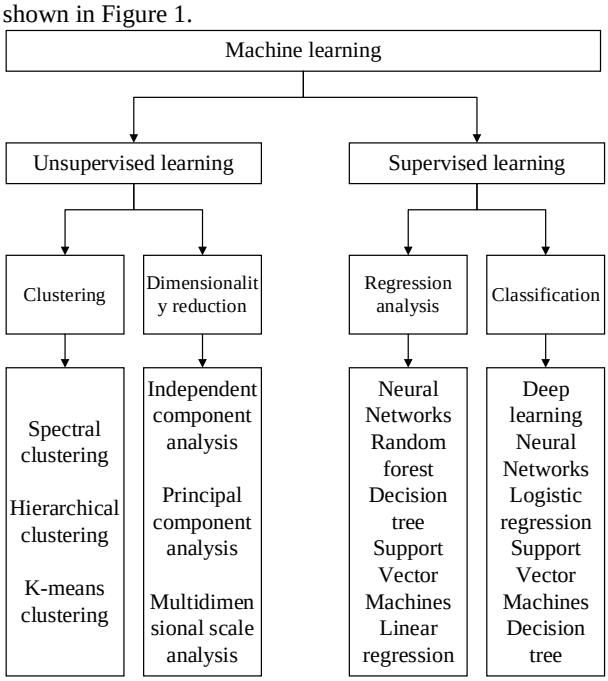

Figure 1 Common algorithms for machine learning

2.1. Machine Learning Ranking Method

This method applies the machine learning method to the relevance ranking of documents for information retrieval. It only needs to provide training data to the machine, and the machine can automatically learn to obtain the best ranking function, so as to comprehensively evaluate the relevance of the documents. Machine learning sorting is mainly divided into four steps: manually labeling training data, extracting document features, learning sorting functions, and applying machine learning models to reality.

Label training data. The actual sequence of the document collection under a given query can be obtained as training data through manual labeling or collection from query logs. When manually labeling the relevance of a document, the relevance can be expressed as a numerical sequence, and the user clicks the record to simulate the manual scoring mechanism.

Extract feature vectors. Before each document enters the machine learning system, it needs to be converted into a feature vector. For a query document pair (q, d), extract corresponding features, which can be expressed as feature vector $X = [X^Q, X^D, X^{QD}]$. Where X^Q is a feature vector that depends only on the query q; X^D is a feature vector that only depends on the document d; X^{QD} is a feature vector that depends on the relationship between the query q and the document d.

After converting the documents into force feature vectors X and combining the manually marked correlation score Y, each document is transformed into a feature vector and its corresponding correlation score (X, Y) form to form a specific training instance. The final training result ranking function is as follows.

$$y = f(X) \quad (1)$$

Where y is the relevance score of the document, f is the ranking function, and X is the feature vector. When the user searches, the ranking function f can be used to score the documents and form search results.

2.2. Decision Tree

With the layer-by-layer division, the sample categories included in the branches of the decision tree will gradually become consistent.

The decision tree algorithm can also be considered as a function that is infinitely close to the discrete target value. This algorithm is often used in predictive models. It can find classification rules from big data and mine valuable and potential information, and can learn the results. In the form of a tree. Its advantages are fast analysis speed, high accuracy, and simple generation mode. It recurs from the root node from top to bottom, classifies the sample data according to factor values, and obtains different branches and nodes. The final leaf node is the final conclusion of the decision tree. The generation process of the decision tree is shown in Figure 2.

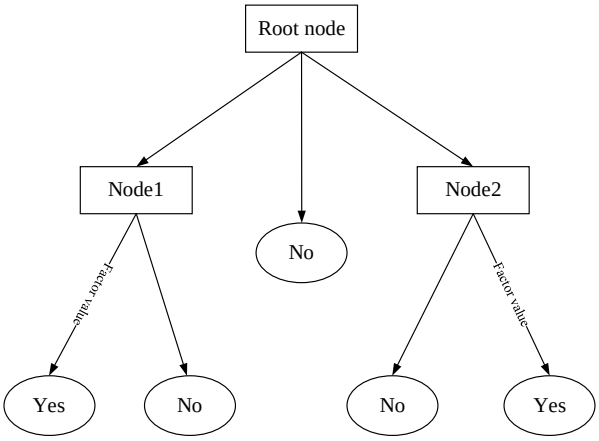

Figure 2 Decision tree generation process

The effectiveness of the decision tree can use a tree to test its ability to classify the sample set. The decision tree cannot grow indefinitely. In the extreme case, the growth of the decision tree is that there is only one leaf node left in the decision tree. However, the decision tree generated in this way is often too complicated and the prediction accuracy is not high, so the conditions for the decision tree to stop growing are usually set.

3. STOCK SELECTION MODEL AND DATA PROCESSING BASED ON CART

3.1. Stock Selection Model Framework

All machine learning methods want to have as much data as possible, and deep learning models are no exception. Generally, when constructing a portfolio of stocks, a viable stock pool needs to be defined first. The usual practice is to at least remove new and sub-new stocks, and select stock pools in index constituent stocks or industry classifications. In short, considering the constraints of the real world and the characteristics of financial markets, the data available for deep learning is quite limited. Unlike general machine learning tasks, the evaluation criteria for applying deep learning models in the financial field are quite vague. The basic framework of machine learning for stock selection is shown in Figure 3.

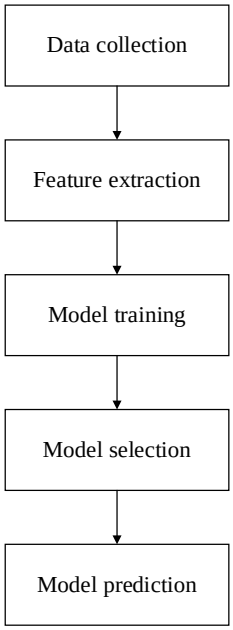

Figure 3 The basic framework of machine learning for stock selection

3.2. Build CART Stock Selection Model

The CART method is used to build a decision tree classification model, mainly for creating root nodes, splitting, reaching the end condition to stop splitting, pruning, finding stocks with predicted excess returns, building a portfolio, setting transaction frequency and transaction costs, and finally returning measurement.

The CART model is a tree-like structure constructed from known examples, and each path and node have clear text or numbers. Decision trees can handle high-latitude variables very well, and can quickly screen out important variables.

3.3. Data Processing

The data of the China Securities 500 Index is selected as the research object, and it reflects the listed companies with medium market value. Through the CART algorithm for variable selection, the importance of variables is preliminarily calculated, the variables of low importance are eliminated, and the remaining variables are sorted into the training model for gradual call and verification, and finally a prediction set is formed.

Outliers in real financial market data are not uncommon. Given that data sets are at the foundation of machine learning methods, outlier processing methods have an important impact on model training and prediction. It is necessary to standardize and introduce outlier processing methods. Since the models are allowed to accept input for a period of time, it is necessary to consider the effect of missing values in the data on input continuity.

3.4. Determination of Forecast Indicators

We define the forecast indicator Y as a binary 0-1 variable. If the excess return rate of the next month exceeds 3%, that is, the absolute return rate of the stock of the next month is 3% more than the increase or decrease of the CSI 500 index in that month, then the forecast The indicator Y is 1; conversely, if the excess return of next month is less than 3%, the forecast indicator Y is 0.

$$Y_t^i = \begin{cases} 1, \ R_{t+1}^i > (R_{t+1}^x + 3\%) \\ 0, \ R_{t+1}^i \leq (R_{t+1}^x + 3\%) \end{cases} \quad (2)$$

Among them, Y_t^i is the i-th stock, the forecast indicator of the t-th month, R_t^i is the i-th stock, the t-th month's return, and R_t^x is the x-month return of the CSI 500 Index.

3.5. CART Stock Selection Decision Tree Model Algorithm Criterion

In the process of generating a decision tree is to recursively construct a binary tree, the square error minimization criterion is used for the regression tree. In the classification problem, the CART algorithm uses the Gini index for optimal feature selection. While selecting the optimal feature, the optimal bisection point is selected.

In the classification problem, assuming that there are K categories, and the probability that the sample point belongs to the kth category is P_k, the Gini index of the probability distribution is defined as follows.

$$Gini(P) = \sum_{k=1}^{K} P_k(1 - P_k) = 1 - \sum_{k=1}^{K} P_k^2 \quad (3)$$

For a binary classification problem, if the probability that the sample point belongs to the first category is P, the Gini index is as follows.

$$Gini(P) = 2P(1 - P) \quad (4)$$

For a set U, its Gini index is as follows.

$$Gini(P) = 1 - \sum_{k=1}^{K} (\frac{C_K}{U})^2 \quad (5)$$

Among them, C_K is the sample subset of U belonging to the kth class, and K is the number of classes.

Under the condition of feature A, the Gini index of set U is expressed as follows.

$$Gini(U, A) = \frac{U_1}{U} Gini(U_1) + \frac{U_2}{U} Gini(U_2) \quad (6)$$

Among them, $U_1 = (x, y) \in U | A(x) = a, U = U_1 + U_2$, $Gini(U)$ represents the Gini index of the set U, and $Gini(U, A)$ refers to after $A = a$, Gini index of collection U.

4. STOCK SELECTION EFFECT OF CART BASED ON STATIC TREE

In the static model, the data set is divided into two groups of samples: 1) observations from 2011 to 2015, and 2) observations from 2016 to 2019. The first group is the test "in-sample" data used to train the model, and used to classify cases that have never been seen before. Out-of-sample testing can effectively test the predictive ability of the model. Since there is only one tree model during the entire period, this method is called a static tree method. Figure 4 is part of a static tree, showing only the results of the first four splits.

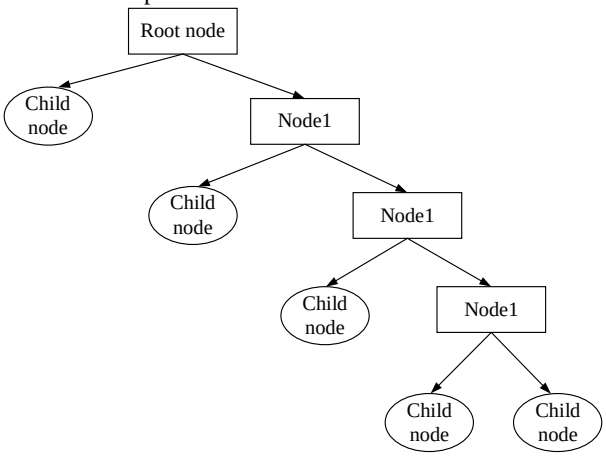

Figure 4 Part of the static tree structure

5. CONCLUSION

This paper discusses the research work of machine learning in terms of quantification, explores the decision tree, and selects the stocks with high excess returns for the 500 constituent stocks of CSI. The advantage of CART algorithm is that the generated result is intuitive and easy to understand, it can process discrete data and continuous data at the same time, and it has stronger fault tolerance for outliers. However, this article does not empirically explore the application effect of the machine learning method, nor does it further improve the decision tree algorithm. In the follow-up research, we will further verify the specific implementation effect of the method through experimental analysis, and compare with other popular deep learning algorithms.

REFERENCES

[1] B. Li, X.Y. Shao, Y.Y. Li, Research on Fundamental Quantitative Investment Driven by Machine Learning, China Industrial Economics, Issue 08, 2019, pp. 61-79.

[2] B. Li, Y. Lin, W.X. Tang, ML-TEA: A set of quantitative investment algorithms based on machine learning and technical analysis, Systems Engineering-Theory & Practice, 2017 05, pp. 1089-1100.

[3] R.D. Chen, H.H. Yu, support vector machine stock selection model based on heuristic algorithm, Systems Engineering, 2014, 02, pp. 40-48.

[4] H.F. Xiong, Thinking of Quantitative Analysis in the Course System of "Investment", Research of Finance and Education, 2016, 02, pp. 85-88.

[5] Y. Lu, Credit Risk Assessment Model Based on Decision Tree Algorithm, Science & Technology Information, Issue 36, 2018, pp. 18-19.

[6] Y.N. Jiao, J. Ma, an improved MEP decision tree pruning algorithm, Journal of Hebei University of Technology, 2019, 06, pp.24-29.

[7] Z.X. Xu, Y. Qian, Y. Su, Research on Decision Tree Recommendation Algorithms Combining Time Series, Modern Computer, 2019, Issue 34, pp.20-23+27.

[8] Y. Pan, Unbalanced Data Set Classification Method Based on Improved Decision Tree Algorithm, Journal of Changchun Institute of Technology (Natural Sciences Edition), 2019, 04, pp. 95-98+102.

[9] D.H. Yang, B.G. Wu, X.C. Sun, Research on the Prediction Model of the Usefulness of Network Comment Information Based on Machine Learning, Information Science, Issue 12, 2019, pp. 34-39+77.