Development of Quantitative Investment Model Based on Selecting Stocks with AI

Hanyue Liu[1], Aoqi Xie[2]

[1]Beijing Normal University, Beijing, 100875, China

[2]Jilin University, Changchun, 130000, China

Keywords: Quantitative investment model, Selecting stocks, AI

Abstract: The objectivity and rapidity of quantitative investment are more and more loved by investors. The stock market is a complex nonlinear system with low signal-to-noise ratio, and artificial intelligence has been proved to be a powerful tool for modeling fuzzy nonlinear data. This paper analyzes the common algorithms of artificial intelligence, gives the idea and process of the development of quantitative investment model based on artificial intelligence, which provides references for the relevant researchers.

1. Introduction

With the development of artificial intelligence technology, the financial circle that pursues precision and efficiency is also quietly changing [1]. The subjective securities investment, once regarded as an art, has been gradually replaced by the quantitative investment strategy attached to the computer. In essence, both quantitative investment and traditional qualitative investment are based on the theory of market inefficiency to establish a portfolio that can defeat the market and generate excess return. However, unlike the traditional qualitative investment method, quantitative investment does not rely on people's feelings to manage assets, but on the basis of people's investment ideas and investment experience to build a mathematical model, and use computers to process a large number of historical data, in a short time to verify the effectiveness of the model. Only when the performance of the model in the historical data meets the requirements, can it be further applied to the real deal. A large number of research results show that the stock market is a complex non-linear system, and the stock price involves many uncertain factors and the relationship between each factor is complex. However, a large number of facts show that there is a certain law in the stock price fluctuation, and the historical data and other information of the stock price contain the information that can predict the future stock price. Machine learning has been proved to be a powerful tool for modeling nonlinear data in many fields, such as search, personalized recommendation, speech recognition and natural language processing. Therefore, using machine learning to build quantitative investment strategy has certain natural advantages. A large number of previous works has also made a different degree of exploration in this direction, and has proved that machine learning is indeed useful in the field of quantitative investment. In this paper, the more mature sorting algorithm based on machine learning in the field of information retrieval is applied to quantitative stock selection, and a new quantitative stock selection strategy based on machine learning is constructed, which provides a new perspective and thinking for the construction of quantitative stock selection strategy.

2. Common Algorithms of Artificial Intelligence

2.1. Support Vector Machine.

As a common machine learning method, support vector machine has been widely used in many scientific fields [2]. The core idea of SVM is to build a hyperplane in the feature space, so that the distance between different types of sample points and the hyperplane is the largest. SVM uses the method of mapping the feature space to a higher dimension to solve the problem of linear separability in the original feature space. Different from the traditional statistical methods, SVM minimizes the structural risk and reduces the confidence interval as much as possible while controlling the empirical risk. Since our purpose is to predict the price of the underlying asset, we consider the binary classification of support vector machine. In machine learning, support vector machine is a supervised learning model and related learning algorithm for analyzing data in classification and regression analysis. Given a set of training instances, each training instance is marked as belonging to one or the other of two categories. SVM training algorithm creates a model that assigns a new instance to one of two categories, making it a non-probability binary linear classifier. In SVM model, instances are represented as points in space, so the mapping makes the instances of individual categories separated by as wide and obvious intervals as possible. Then, the new instances are mapped to the same space, and their categories are predicted based on which side of the interval they fall on. Suppose the sample set is $(x_i,\ y_i),\ x_i \in R^m,\ y_i \in 0,1, i \in 1,2,\dots,n$. Considering the setting of error bandwidth and penalty function, the maximum interval is transformed into optimization problem.

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

$$s.t. y_i(w^T \phi(x_i) + b) + \xi_i - 1 \geq 0 \ and \ \xi_i \geq 0$$

By solving the above optimization problem, the required hyperplane can be obtained, and the classification of the data can be predicted by inputting the features of the sample data. In this paper, we use the technical index value as the input eigenvalue, and try to use linear SVM to predict the stock price rise and fall in the next five days.

2.2. Neural Network.

Convolutional neural network is a multilayer perceptron designed for the recognition of two-dimensional shape [3]. It can also be said to be a special feedforward neural network model, which has the characteristics of local connection and weight sharing. In addition, the convolution neural network has high invariance in the case of translation, and invariance in the case of scaling and tilting. These invariance characteristics make a large number of neurons in the convolution neural network connected in a certain way in an organized way, so as to respond to the overlapping area of the image. Since the development of deep learning, convolutional neural network has made rapid progress in image processing. Convolutional neural network is similar to biological vision neural network, which has the characteristics of hierarchical and local perceptual region to extract features. In addition to input and output, a standard convolutional neural network mainly consists of convolution layer, subsampling layer and full connection layer. These three layers constitute the main body of the convolutional neural network and are the core part of the convolutional neural network. Therefore, the determination of the algorithm of each layer is particularly important. A convolution layer may contain multiple convolution surfaces, each of which is associated with a convolution kernel. Each convolution layer calculation will produce several weight parameters related to it, and the number of these weight parameters is related to the number of convolution layers, that is, they have a direct relationship with the functions used in the convolution layer. In this paper, sigmoid function is chosen as the transfer function of hidden layer. Take the calculation of the hidden layer of the first layer as an example, and the calculation formula is as follows:

$$a_0^{(2)} = \mathrm{g}(\theta_{10}^1 x_0 + \theta_{11}^1 x_1 + \theta_{12}^1 x_2 + \cdots + \theta_{1n}^1 x_n)$$

$$a_1^{(2)} = \mathrm{g}(\theta_{20}^1 x_0 + \theta_{21}^1 x_1 + \theta_{22}^1 x_2 + \cdots + \theta_{2n}^1 x_n)$$

$$\cdots\cdots$$

$$a_m^{(2)} = \mathrm{g}(\theta_{m0}^1 x_0 + \theta_{m1}^1 x_1 + \theta_{m2}^1 x_2 + \cdots + \theta_{mn}^1 x_n)$$

In these equations, a is the excitation value of each neural node. Function g represents Sigmoid function.

$$\mathrm{g(x)} = \frac{1}{1 + e^{-x}}$$

Finally, the output value of the output layer is calculated. The output layer is determined by the neurons of the hidden layer connected to it.

2.3. Adaboost Ensemble Algorithm.

Promotion learning is a kind of machine learning technology, which can be used in regression and classification problems [3]. It generates weak prediction model in each step and accumulates it into the total model. If the generation of weak prediction model in each step is based on the gradient of loss function, it is called gradient lifting. In the application of machine learning, researchers usually use ensemble algorithm to enhance the prediction effect of a single weak classifier, which is a representative of ensemble algorithm. By increasing the weight of the error samples and reducing the weight of the correct samples in the iteration, the weighted average of the final iterative classifier is the prediction of the whole system. The existing research found that for the complex nonlinear time series prediction, the neural network after using AdaBoost integration has a significant improvement in the classification effect. The working mechanism is to train a weak learner with the initial weight from the training set. The weight of training samples is updated according to the performance of learning error rate of weak learning, so that the weight of training samples with high learning error rate of weak learning becomes higher. We make these points with high error rate get more attention in the following weak learners. Then we train the weak learners based on the training set after adjusting the weight. This is repeated until the number of weak learners reaches the specified number in advance. Finally, we integrate multiple weak learners through the set strategy to get the final strong learner. Specifically, for each integrated weak classifier, the initial weight is set, and the weight is updated according to the classification results. The final integrated classifier function is:

$$Y_m(x) = sgn(\sum_{m=1}^{M} \alpha_m y_m(x))$$

The model in this paper integrates the neural network with AdaBoost algorithm and adjusts its weight according to the error of the neural network in the iteration.

3. Development of Quantitative Investment Model

3.1. Basic Ideas.

Stock selection model is the key link in the process of stock value investment. We have extracted a relatively simple and comprehensive set of stock value features that contain stock value information. According to the feature set of stock value, an effective stock selection model is designed and trained through appropriate samples. We should choose the appropriate pattern classifier and what kind of method to train and test the pattern classifier, so as to establish the stock selection model based on value investment. Specifically, the stock selection problem under the new value investment framework can be understood as the segmentation problem of the midpoint in the geometric space. Suppose there is an indicator in the characteristic space of the stock sample, the number of stocks with investment value is, to form a set, and the number of stocks without investment value is, to form a set. Obviously, the indexes in the characteristic space of stock samples can form a dimensional Euclidean space. In this space, the indexes are distributed in the dimensional space in a scattered state with this index as the coordinate. They can be reduced to two sets, one is the set of stocks with investment value, the other is the set of stocks without investment value. There may be intersection between two sets. The purpose of stock selection model is to separate the two sets with the minimum error rate and the maximum efficiency.

3.2. Structure Design.

The process of establishing stock selection model includes two processes: training and testing. After training and testing, the model has good recognition ability. At this time, it can be applied to the specific process of stock selection in investment practice to further verify the stock selection ability of the model. The first step of model building is to establish training samples and test samples, then input training samples to train pattern classifier, after training, input test samples to verify the training results. Only after the test can the pattern classifier be used to select stocks. If the actual stock recognized by the pattern classifier can get an ideal return on investment, it can prove that the design of the stock selection model is successful. The eigenvector of stock selection model can be composed of the vector of this index. The purpose of stock investment is to make profits, so this paper ranks the average weekly return of stocks in, and divides the samples into two categories with equal numbers Multi factor stock selection model, in fact, assumes that the financial indicators and some market indicators of listed companies will have an impact on the future return of the stock. At the same time, the multi factor stock selection model believes that history can be repeated. The model hopes to find effective influence factors from historical data, and then find those companies with investment value through these effective influence factors. That is to say, the multi factor stock selection model is based on certain selection criteria, and selects a portfolio that can overcome the market and obtain stable excess return. The traditional multi factor model needs to test the effectiveness of the factors and remove the redundant factors. Because of the advantages of its own algorithm, the algorithm model used in this paper does not need to use all the characteristic factors of each tree, and at the same time, it uses the original stock factors to process to get new comprehensive indicators, so it can directly build the model.

3.3. Assessment and Training.

Due to the fact that the data in the real world is very disorderly, the information cannot be obtained temporarily, the information is omitted, some properties of some objects are unavailable, some information is considered unimportant, the cost of obtaining these information is too high, the real-time performance of the system is high, and so on, the data will be missing. The processing of missing values should be analyzed in detail, because sometimes the missing attribute does not mean the missing data, and the missing itself contains information. Therefore, it is necessary to fill in the missing values according to the information that may be contained in different application scenarios, rather than delete them all. Data standardization is to scale the data to a small specific range. It is usually used to process some indexes that need to be compared and evaluated, so as to eliminate the limitation of unit data and convert it into dimensionless values, so as to compare and measure indexes of different units or orders of magnitude. And normalized data processing. Because all the numerical data in this paper contain a variety of information about the stock, so there are great differences in the numerical aspects, so it is necessary to make the data dimensionless and normalize the data when doing the model calculation. In this paper, we define a positive example as a rising stock and a negative example as a falling stock. If the price is constant, considering that there is a handling fee for each transaction, we also define a stock with constant price as a falling stock. The stock is converted into a binary problem. Put the data into the model for data training and parameter setting. The accuracy rate is increasing with the increase of training times.

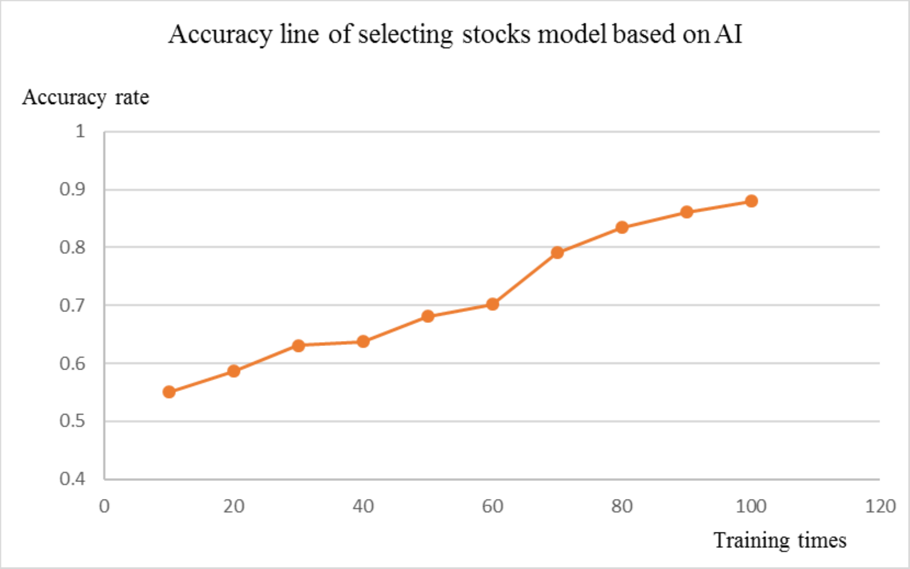

Figure 1. Accuracy line of selecting stocks model based on AI

4. Conclusion

In the stock data back test results also show that the artificial intelligence model in the stock market is able to obtain good returns, the model in the stock investment market is effective and feasible. The data used in this paper only uses the stock and financial data. In the follow-up research, we will add the characteristic factors of futures, bonds and so on. A more comprehensive factor is believed to improve the prediction ability of the model.

References

[1] Li Bin, Lin Yan, Tang Wenxuan, et al. ML-TEA:A set of quantitative investment algorithms based on machine learning and technical analysis[J]. Systems Engineering-Theory & Practice, 2017, 37(5): 1089-1100.

[2] Baum. Das Stabilisierungspotential staatlich-administrierter Preise[J]. Jahrbücher Für Nationalökonomie Und Statistik, 2017, 190(4):349-376.

[3] Mohammad M Ghassemi, Edilberto Amorim, Tuka Alhanai,et al. Quantitative Electroencephalogram Trends Predict Recovery in Hypoxic-Ischemic Encephalopathy[J]. Critical care medicine, 2019, 47(10): 1416-1423.

[4] Fu Hangcong, Zhang Wei . The Application of Machine Learning Algorithms in Stock Movements Forecasting[J]. Software Guide, 2017, 16(10): 31-34+46.