Research on Quantitative Analysis Method of Financial Data Based on Machine Learning

Shulang Zhao

Faculty of Science, Beijing University of Technology, Beijing, China.


Received: March 15, 2023
Accepted: April 12, 2023
Published: May 6, 2023

Abstract

With the development of the Internet and the continuous innovation of artificial intelligence, the study of financial data based on quantitative analysis methods of machine learning has also emerged. Quantitative analysis of financial data is a method of using computers to analyse financial data to predict the direction of stock market fluctuations in order to obtain excess returns. Machine learning has become an important tool in quantitative analysis methods, and has shown better performance than traditional quantitative analysis methods. Investing in financial data through quantitative analysis methods based on machine learning has advantages that traditional investment methods do not have, such as objectivity and accuracy, and is thus used by a wide range of investors and financial institutions. This paper firstly explains the feasibility of forecasting financial markets through the efficient market hypothesis, followed by the characteristics of quantitative analysis methods and the application of machine learning in quantitative analysis methods.

Keywords

Quantitative analysis, financial data, SVM, Adaboost

Introduction

With the development of the social economy, the financial market has become an important part of the national economy and an important reflection of the country’s competitive strength. In recent years, China’s financial market has been developing rapidly from scratch. At the same time, the accumulation of massive data has put forward higher requirements for the effective organisation and management of financial data, and how to quickly extract effective information and effectively analyse and forecast it on the basis of data has now become an important issue for academic and industrial research. Quantitative analysis techniques have become the choice of many investors. While overseas research on quantitative analysis techniques has become very mature and has been widely used in the financial markets, domestic research on quantitative analysis methods for financial data is still in its infancy [1]. By comparing the domestic financial markets with those of other countries, we find that the differences between the domestic and international financial markets are significant, so we need to build a quantitative analysis model that is applicable to the Chinese financial market, which is of great importance to the Chinese economic market and investors.

1. Quantitative analysis of theoretical studies

1.1 The Efficient Market Hypothesis

The efficient market hypothesis [2] was proposed by the famous American economist Eugene in 1970. In this hypothesis, all investors involved in investment are rational and have sufficient rationality is its effective premise, under the
premise that all investors involved in investment can obtain the information displayed in the market in time and make corresponding responses, then in a stock market with sound laws, normal economic functions, open transparency and reasonable competition, all information carried in the stock has been fully exposed in time in the stock price. This information is valid and of high value. If there is no market manipulation in the market, then it is unlikely that an investor will be able to make a profit far in excess of the market benchmark by analysing the past price of a stock.

It is clear from the above analysis that the theory is based on an idealised financial market, which does not exist in the real world, so this hypothesis is not applicable to the huge financial markets of today. Rational investors. However, according to the degree of idealisation of financial markets in this hypothesis, financial markets can be divided into three main categories [3]: (1) Strongly efficient markets, in which stock prices fully respond to all information about stocks, which are ideal financial markets in which markets are unpredictable and therefore cannot be analysed quantitatively. (2) Semi-strongly efficient markets, in which stock prices respond to all publicly available information, in which quantitative analysis is possible as it still has uncertainty. (3) Weakly efficient markets, where stock prices respond to incomplete information about stocks, in which quantitative analysis is also possible.

1.2 Quantitative analysis features

Quantitative analysis of financial data [4] is a trading strategy constructed using computer science and technology as well as mathematical models. It is a method that uses artificial intelligence methods and statistics to make decisions instead of humans, and operates through quantitative analysis models that are constructed to generate returns. The quantitative analysis method uses computers to process data that cannot be processed by the human brain and to mine and summarise the underlying patterns of the market in order to build a repeatable investment strategy based on the data that guides the entire investment process of this paper. The subjective approach and the quantitative approach are essentially the same in that they are both based on the assumption that equity markets are weakly efficient or inefficient markets, but each has its own characteristics in terms of its analytical approach. The subjective investment approach is usually characterised by the following.

(1) Investments are usually made based on the investment manager's subjective judgement of the market.
(2) The manager's research into the macro environment, sectors and companies to predict future movements.
(3) More in-depth research, usually on a very small number of stocks.
(4) Concentrated holdings and poor investment stability.
(5) Trading perception relies on subjective judgement and is not replicable.

1.3 The application of machine learning in quantitative analysis methods

Machine learning [5] belongs to the core of artificial intelligence, he is the product of a multidisciplinary intersection field, its main idea is to achieve the human learning ability through the computer, which is what makes it different from the traditional algorithm. In the implementation of traditional algorithms, the goal is to obtain the results needed for this paper through the computational process of that algorithm, which is understood for traditional algorithms. In the case of machine learning, however, there is no need to know how the computation is carried out to obtain the results, because machine learning is the process of exploring the data and finding patterns in it and learning from it. Compared to traditional algorithms, machine learning is more autonomous and the implementation of the algorithm code is relatively simple, which increases the computing power of the code. It is also more sensitive to changes in data than traditional algorithms, and can be improved through continuous learning [6].

2. Machine learning quantitative analysis models

2.1 Decision tree models

Decision Tree (DT) [7] is a basic classification and regression algorithm that can be divided into classification trees and regression trees according to the method applied. In classification problems, decision trees are essentially a binary classification process, whereby all data sets can be divided into two completely different parts. The goal of a classification tree is to generate a classifier that generalises well to all samples by generating a classifier that has good classification power. The basic idea of a decision tree is to build a top-down tree with the best possible metrics through information gain or other measures. The process of generating a decision tree progresses from the root node to the leaf nodes, layer by layer, using the final result of the metric classification. Overall, the learning process of a decision tree consists of the selection of metrics, the production of the decision tree and the pruning of the decision tree. Classification trees can be measured by a variety of metrics, such as the information gain in the ID3 algorithm, the information gain ratio and retrograde evaluation in the C4.5 algorithm, and the Gini coefficient in the CART decision tree. A decision tree is generated based on the input training set \( D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{n-1}, y_{n-1}), (x_n, y_n)\} \) and the feature set \( A = \)
\{a_1, a_2, a_3, \ldots, a_{m-1}, a_m\}. The following is the process of generating a decision tree using information gain as a discriminant.

1. Generate the root node according to the training set D and calculate its information entropy \(H(D)\).
2. Calculate the conditional entropy according to the features in the feature set A, and derive the optimal division feature \(a_i\).
3. Each class \(a_i^j\) in the partition feature \(a_i\) is treated as a branch, so that \(D_{ij}\) denotes the subset of training set D that takes the value \(a_i^j\) on \(a_i\), and the node where \(D_{ij}\) is located is a branch node.
4. In each branch node, use \(D_{ij}\) as the training set and the feature set \(A-a_i\) with \(a_i\) removed as the feature set to repeat the above operation until the feature set is empty, or the training set is empty, or the training set does not change its position after the training set is classified according to the feature set, and the branch node generated by it is used as the leaf node.

In step 2, if the current node contains samples that all belong to the same category, then no division is needed. In step 4, if the sample set of the current node is empty, it means that it can no longer be divided, so the current branch node is transformed into a leaf node; if the feature set of the current node is empty, it means that all the features have been divided into branches and there are no features that can be further divided, so the division is ended and the current node is transformed into a leaf node. The higher the correlation between attributes and labels, the more critical it is for data partitioning. However, in the process of finding the optimal attributes, the model is prone to overfitting, which leads to a reduction in the generalisation performance of the model, and the decision tree can be pruned to improve the generalisation performance [8].

2.2 Ada Boost Algorithm

The Ada Boost algorithm [9-11] is one of the Boosting integrated learning models. It is a serial combination of multiple weak classifiers that can achieve a good balance between variance and bias of the model, and has been developed in recent years. The core of the Adaboost algorithm is the construction of weak classifiers, which are combined in a serial way to form a strong classifier. The process of classification by the algorithm can be described as a process of enhancement of the weak classification algorithm. It is adaptive in the sense that samples that were incorrectly classified by the previous weak classifier are strengthened and the weighted whole sample is used again to train the next weak classifier. At the same time, a new weak classifier is added in each round until some predetermined sufficiently small error rate is reached or a pre-specified maximum number of iterations is reached.

Overall, the iterative process of the Adaboost algorithm has the following three steps.

1. The training dataset is initialised and each training set is given the same initial weights at the very beginning.
2. The weak classifier is trained, and the weights of each training dataset are updated during the training process. If a sample point has been accurately classified during the training of the dataset, the weight held by that sample point is reduced during the construction of the next training set. If, in the process, a sample point is not well classified, the weight of that sample point is raised in the next training set. In this way, the updated set of weights is used as the new set of samples for the next weak classifier.
3. The weak classifiers completed by the above iterations are combined into strong classifiers. After the training of each weak classifier is completed, the weaker classifiers with accurate classification are given more weights so that they can play a dominant role in the final strong classifier. For weak classifiers with poor classification performance, the weights are reduced so that they have less role in the classification, thus reducing the classification error.

The essence of the Adaboost algorithm is that the newly constructed weak classifier needs to focus more on the samples that the previous classifier failed to classify correctly. Based on the performance of the previous classifier, the sample weights are updated to increase the number of samples that were not correctly classified, and the new classifier is trained based on the updated set of data samples. The training process of classifier can be shown in Figure 1. Each weak classifier in the final strong classifier will determine its weight ratio according to its performance.

2.3 Support vector machine models

Support vector machine (SVM) is a machine learning method based on statistical learning theory. It was first proposed by Vapnik in 1992, and since then SVMs have been widely used in classification and regression problems where the machine has shown excellent performance [12]. The SVM algorithm can be seen as dividing the sample into two parts through a plane. For high-dimensional data, there is also a hyperplane that cannot be drawn but can divide the sample data out. The support vector machine divides the sample data by finding that plane. For a given data set \(D = \{(x_i, y_i) | i = 1, 2, 3, \ldots, n\}, y_i \in \{-1, +1\}\) the hyperplane obtained by partitioning using SVM is shown in Figure 2. Where \(H_1\) and \(H_2\) are the planes where the two sample points are nearest to the hyperplane \(H\).
In the sample space, the square of the plane towards which the samples are divided can be described by the linear equation \( w^T x + b = 0 \) where \( w = (w_1, w_2, \ldots, w_n) \) is the normal vector, which determines the direction of the hyperplane, and \( b \) is the displacement of the hyperplane, which determines the distance between the hyperplane and the origin of the coordinates. Also for the plane \( H_1 \) and \( H_2 \), this can be expressed in terms of \( w^T x + b = 1 \) and \( w^T x + b = -1 \). Clearly the division of the hyperplane is determined by the normal vector \( w \) and the displacement \( b \). The distance to the hyperplane \( H \) for any sample point can then be expressed by Equation (1).

\[
r = \frac{|w^T x + b|}{\|w\|}
\]  

(1)

For the hyperplane \( H \), if all training sample points can be correctly classified, then for sample point \((x_i, y_i)\), if \( y_i = +1 \), then \( w^T x + b > 0 \), whereas \( w^T x + b < 0 \). Equation 2 can be obtained here:

\[
\begin{align*}
& w^T x + b \geq 1, y = +1 \\
& w^T x + b \leq 1, y = -1
\end{align*}
\]  

(2)

The idea of SVM is to find the two sample points closest to the hyperplane by dividing the hyperplane so that the sum of the distances from the hyperplane to the plane where the two dissimilar sample points are located is maximized, and these two dissimilar sample points are called support vectors [13]. The sum of the distances of the two dissimilar sample points to the hyperplane becomes the interval, which is expressed by Equation 3:

\[
r = \frac{2}{\|w\|}
\]  

(3)

Then for the SVM, the task becomes finding the hyperplane with the maximum interval, that is, finding the appropriate parameter normal vector \( w \) and displacement \( b \) such that \( r \) is maximized, as in Equation 4.

\[
\max \frac{2}{\|w\|} \quad \text{s.t.} \quad y_i (w^T x_i + b) \geq 1
\]  

(4)
This equation indicates that in order to maximize the interval is equivalent to maximizing $\|w\|^{-1}$, that is, minimizing $\|w\|$. 

$$
\max \|w\| \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1
$$

This is the basic type of support vector machine.

3. Conclusion

This paper provides an introduction to theories related to financial markets, arguing for the feasibility of forecasting financial markets through the efficient market hypothesis and the theory of random walk. Next, the characteristics of quantitative analysis methods applied in finance are introduced to demonstrate the advantages of conducting quantitative analysis methods. Finally the paper provides a detailed study of commonly used quantitative analysis methods based on machine learning, and it can be seen that the classification algorithms are currently the most used in the study of quantitative analysis methods.

References