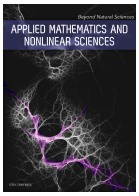
sciendo Applied Mathematics and Nonlinear Sciences 8(1) (2023) 2413–2424



Applied Mathematics and Nonlinear Sciences

https://www.sciendo.com

Application of machine learning in stock selection

Pengfei Li[1†], Jungang Xu[1], Mohammad AI-Hamami[2]

[1]School of Computer Science & Technology, University of Chinese Academy of Sciences, Beijing, China

[2]Management Information Systems, College of Administrative Sciences, Applied Science University, Bahrain

Submission Info

Communicated by Juan Luis García Guirao
Received February 22nd 2022
Accepted May 15th 2022
Available online September 20th 2022

Abstract

With the development of artificial intelligence technology, machine learning has achieved very good results in the field of stock selection. This paper mainly studies the application of linear model, clustering, support vector machine, random forest, neural network and deep learning methods in the field of stock selection. The main contribution of this paper is to provide a new idea for traditional quantitative investors, so that they can build a more efficient stock selection model in practical application. The experimental results show that the stock selection model constructed by these six machine learning methods can obtain higher return and stability.

Keywords: Stock selection, machine learning, clustering, SVM, random forest, neural network, deep learning

1 Introduction

Quantitative investment is a process that allows the investment idea to be transformed into a transaction model with its profitability is analysed by data, with which to guide the transaction. Since the 1990s, some quantitative investment researchers and financial practitioners have got mathematical logic formulas extracted from the complicated transaction logic, and analysed the market behaviour through mathematical methods such as statistics and probability, in order to find the law and form a stable and reusable transaction strategy that can be used in actual transactions. Especially in recent years, with the rise of artificial intelligence, the researches of the new-generation artificial intelligence methods such as machine learning and deep learning in voice, text and image appear increasingly healthy in the market, more and more people try to apply machine learning algorithms to the quantitative investment field. It is expected to use the complex structure of machine learning algorithm to explain some complex market trading phenomena, so as to construct a more stable and effective quantitative strategy.

†Corresponding author.
Email address: lipengfei@ucas.ac.cn

sciendo ISSN 2444-8656 doi:10.2478/amns.2022.1.00025

As a strategic form of quantitative investment, stock selection has been studied and applied for a long time. From the traditional statistical method to the current machine learning method, the stock selection method has also changed greatly with the development of computer technology. In this paper, we will focus on the specific application of machine learning method in stock selection, without considering the countries, markets and groups to which the method applies. We hope that through the literature research on linear model, clustering, support vector machine, random forest, neural network and deep learning methods in the field of stock selection, and through the comparison of the results of various methods, people can more intuitively compare and refer to various methods, in order to better understand how to develop and optimize their own stock selection strategies according to specific needs.

The rest of the paper is structured as follows: in Section 2, we briefly introduce the basic concept of stock selection. In Section 3, we will focus on the six most widely used machine learning methods in the field of stock selection, and analyse the specific research and application of various methods through the literature. In Section 4, we compare the six models with SSE 500 from three aspects: excess return rate, information ratio and max drawdown. The experimental results verify the effectiveness and stability of the six models. Finally, we summarize the application and development of stock selection method, and put forward the prospect of future work.

2 Stock selection

Stock selection is one of the quantitative investment strategies. The key to stock selection is to mine the driving factors behind the stock price, and then analyse the internal relationship between these factors, so as to find the portfolio relationship between stocks and build the corresponding portfolio. Hua [1] defines stock selection as "the process of establishing a model by using data-based methods, screening listed stocks, and obtaining income by selecting stocks with better performance in the future." In short, stock selection is to use quantitative method to select stock portfolio, hoping that the stock portfolio can obtain investment behaviour that exceeds the benchmark rate of return.

There are various types of stock selection data, which we call stock selection factors. We divide different stock selection strategies into two categories: Fundamental Stock Selection and Technical Stock Selection according to different stock selection factors. Among them, fundamental stock selection is a stock selection method that uses the existing public information, obtains the dynamic P/E ratio and growth results of the stock after comprehensive analysis, and obtains whether the stock valuation is reasonable with reference to the current stock price, so as to determine the purchase and sale of the stock. The data involved in fundamental stock selection mainly include macroeconomic data, microeconomic data and financial data. Among them, macroeconomic data include basic national conditions, market and industrial policies; Microeconomics includes basic industry demand and price; Financial data analysis mainly includes enterprise assets and liabilities, cash and income. Fundamental stock selection mainly includes multi factor model, style rotation model and industry rotation model. Technical stock selection is also called stock selection based on trading data or stock selection based on technical analysis. It mainly analyses all price related data in the stock market. This kind of data is generally in the trading instruction book of the stock market. Compared with the fundamental data, the price data required by technical stock selection has faster update frequency, larger data volume, and obvious time series characteristics. Due to the strong time series characteristics of stock data, the data noise and uncertainty are large, and the repeatability is random, which makes the stock selection analysis more difficult. Technical stock selection mainly includes capital flow model, momentum reversal model, consistent expectation model, trend tracking model and chip stock selection model.

The selection of stock selection factors is multifaceted, which is also a problem worthy of study. This paper mainly studies the application of machine learning in the field of stock selection, but mostly analyses the stock selection factors. In our experiment, we only use the typical stock selection factor as our experimental data.

3 Machine learning in stock selection

The traditional qualitative investment and quantitative investment are essentially based on the market efficiency hypothesis or non-effective theoretical basis. By analysing the potential factors affecting the stock price, we can establish a portfolio that can produce excess returns. Quantitative investment relies more on computer analysis than traditional human observation to screen out strategies that can bring excess returns from massive data. Some of these strategies are based on the experience summary of historical data, and some are based on reasonable market experience, which can avoid emotional interference to the greatest extent and repeat past wrong decisions in the face of extreme market instability.

Over the past 30 years, stock selection has been a hot spot in stock market investment research. With the increasing amount of stock related data and the increasing number of data dimensions (factor types), the traditional stock selection model is greatly limited in terms of performance and efficiency when dealing with these data, and the machine learning method has inherent advantages in dealing with this multi-dimensional and large amount of data, which can better meet the new requirements in the field of stock selection.

Machine learning is not a single method, but a general term of a class of methods. Machine learning includes many different methods. According to different model training methods, machine learning methods can be divided into supervised learning and unsupervised learning. The classification of machine learning methods is shown in Figure 1.

As shown in Figure 1, there are many machine learning methods. By consulting the relevant literature on stock selection, we have selected the six most widely used methods: linear model, clustering, support vector machine, random forest, neural network and deep learning for specific introduction. There are many other machine learning methods that we have not mentioned, and you can consult them by yourself, it will not be repeated here.

3.1 Linear model

Linear model is the simplest and most commonly used model in machine learning. There are many kinds of linear models, such as regression model (Linear Regression, Ridge Regression and Lasso Regression), classification model (Logistic Regression), Principal Component Analysis, Linear Discriminant Analysis, etc. The linear model is characterized by simple structure, easy analysis and understanding, and difficult over fitting. The disadvantage is that the assumption is strong, and the complex financial signal may not meet the linear assumption.

3.1.1 Regression model

Regression model includes Linear Regression, Lasso Regression and Ridge Regression, etc. Linear regression is the most common machine learning algorithm in traditional multi factor models. Simple univariate linear regression can be used in the single factor model. According to the known “characteristic factors” and “labels”, the model reflecting the relationship between the two can be trained. The model can be used to predict the future trend and select the stocks (portfolios) with good trend.

However, single factor is difficult to effectively predict the trend, and multi factor multiple linear regression is needed. The model is constructed by constructing the linear relationship between the rise and fall range and

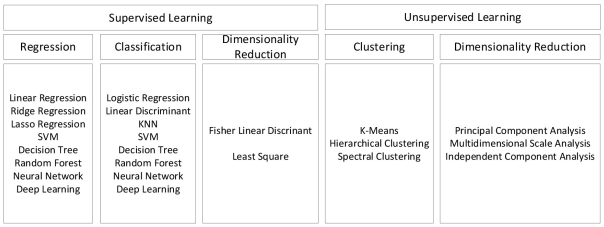


Fig. 1 Machine learning methods.

"characteristic factors":

$$y = \omega_0 + w_1 x_1 + w_2 x_2 + \cdots + w_P x_P \tag{1}$$

where, y means the rise and fall range, P is the number of "characteristic factors", the estimator of the coefficient vector $\omega = (\omega_0, \omega_1, \omega_2, \ldots, \omega_p)$ is obtained by the least square method, and the loss function is defined as the sum of squares of fitting residuals of all n samples:

$$J(\omega) = \Sigma_{i=1}^{N} (y_i - w_0 - \Sigma_{j=1}^{p} w_j x_{ij})^2 \tag{2}$$

where, x_{ij} represents the jth characteristic factor of the ith sample, and y_i represents the label of the ith sample. Estimation of model coefficients $\hat{\omega}$ is the minimum loss function ω value of:

$$\hat{\omega} = minJ(w) \tag{3}$$

When the sample size is small, it can be calculated directly $\hat{\omega}$. Otherwise, the gradient descent algorithm is usually repeatedly used to obtain $\hat{\omega}$.

In the ordinary least squares method, we do not make any prior assumptions on model coefficients ω in fact, ω is impossible to take a maximum positive number or a minimum negative number. Moreover, when there are many features, it is likely that only a few features have predictive effect. Therefore, we introduce the important idea of regularization and add a penalty term after the loss function of the least squares method. When the penalty term is a coefficient ω, the regression method is called Ridge Regression (also known as L2 regularization), and the loss function is:

$$J(\omega) = \Sigma_{i=1}^{N} (y_i - w_0 - \Sigma_{j=1}^{p} w_j x_{ij})^2 + \lambda \Sigma_{j=1}^{p} w_j^2 \tag{4}$$

When the penalty term is a coefficient ω. This regression method is called Lasso Regression (also known as L1 regularization), and the loss function is:

$$J(\omega) = \Sigma_{i=1}^{N} (y_i - w_0 - \Sigma_{j=1}^{p} w_j x_{ij})^2 + \lambda \Sigma_{j=1}^{p} |w_j| \tag{5}$$

where, free parameter λ is the regularization coefficient. When setting a larger λ, even for smaller ω, it will also be punished, the fitting results ω closer to 0.

In the framework of the multi factor model constructed by the above method, ω represents the factor yield. After the model parameters are determined by the least square method, given the factors $x_0, x_1, x_2, \ldots, x_p$ of any stock at the end of a month, we can predict the return y of the stock in the next month, so as to select the stock (portfolio) with large return y.

3.1.2 Classification model

Many times, we don't need to predict the rise of stocks next month, but we want to predict whether stocks will rise or fall next month. In this way, it is no longer a regression problem, but a classification problem. Logistic regression is one of the simplest machine learning classification methods.

Logistic regression can be composed of two parts, one is the same as linear regression, and the other is Sigmoid Function. Logistic regression does not strictly divide the result into 0 or 1, but gives the probability that the result is between 0 and 1. The logistic regression model is adopted:

$$p_{(x_1)} = \frac{e^{\omega_0 + w_1 x_1 + \omega_2 x_2 + \cdots + w_p x_p}}{1 + e^{\omega_0 + w_1 x_1 + \omega_2 x_2 + \cdots + w_p x_p}} \tag{6}$$

Another common equivalent form of the model is:

$$log\left(\frac{P(x)}{1 - P(x)}\right) = \omega_0 + w_1 x_1 + \omega_2 x_2 + \cdots + w_p x_p \tag{7}$$

where, parameter ω is obtained by maximum likelihood estimation method. Similar to linear regression, regularization method is also used to avoid over fitting.

Logistic regression is mainly applied to the construction of multi factor stock selection model, which selects stocks (portfolio) by judging the rise and fall of stocks. However, in practical application, stock trend includes rise, fall, shock and other forms. In this way, stock selection is no longer a simple two classification problem, but a multi classification problem. Multiple classification logistic regression method includes many classification and OvR (one vs rest) strategies, which have relevant applications in stock selection.

Hiemstra [2] compares linear regression and back propagation network to predict the quarterly excess return of the stock market. The experimental results verify the effectiveness of the two models. Better Stocks (combinations) can be selected through the excess return. Huang [3] makes a comparative study on the traditional regression model of stock scoring and the linear model based on ml. in the ML based model, genetic algorithm (GA) is a famous search algorithm in the ML field, which is used to optimize the model parameters and select the input variables of the stock scoring model. The experimental results show that the proposed genetic method is obviously better than the traditional regression method and benchmark. Huang [4] proposes a multiple linear regression model based on fundamental indicators and technical indicators, and constructs a multi factor stock selection model through the relative rate of return. The model can distinguish the relative rate of return of stocks according to factors, so as to select stocks (combinations) with high rate of return. Chaudhary [5] developed a regression system to predict the stock value of the company. Through linear regression and logistic regression, stock prediction and stock analysis are carried out to select stocks with higher returns. Chen [6] proposed a logistic stock selection model based on financial driving factors, and verified that the model performs best in the threshold of 5% return through the test of different test sets.

The above literature studies different stock markets with different linear models. The research results confirm that the linear model has obvious advantages over different comparative models and can be improved to varying degrees.

3.2 Clustering

Clustering is a kind of unsupervised machine learning method, which classifies similar (related) objects into the same cluster. The biggest difference between clustering and classification is that the we know the target of classification in advance, but we do not know the target of clustering. Clustering can be applied to almost all objects. The more similar (related) the objects in the cluster, the better it effects.

K-means clustering is the most classical clustering algorithm. It groups the samples according to the similarity between samples and divides them into K clusters. This is a very effective and practical idea in stock selection.

Clustering is simple and intuitive. It can provide multiple possible solutions for exploratory research. Selecting the final explanation requires subjective judgment and subsequent analysis. Stock selection is to select a high-yield stock portfolio from multiple stocks, which is completely consistent with the clustering characteristics.

Steve [7] uses clustering algorithm and time series outlier analysis, uses PAM clustering algorithm to constrain the initial set of stocks, and uses outlier analysis to define two independent active trading strategies. These results are compared with the passive strategy of fully investing in the S&P 500 index, the stock portfolio constructed by the combination of clustering algorithm and time series outlier analysis is better than the pure passive index strategy. Newton [8] uses hierarchical clustering algorithm for cluster analysis, groups the stocks in the spot market according to the risk return criterion, and then uses Ward method to sort to form the final stock portfolio. Wang [9] conducted cluster analysis on the constituent stocks of SSE 180 index. By dividing the constituent stocks of SSE 180 index into 8 categories, using conditional stock selection and cluster analysis, the success rate and return rate were tested. It was proved that the return rate of stock portfolio selected by cluster analysis method was higher than that of stocks without cluster analysis. Ratchata [10] proposed a stock selection method combining the moving average of the index, trend and momentum technical indicators to identify the

stock selection method that is most likely to be better than the market index, and determined the best combination of these indicators through cluster analysis. Li [11] improves the semi supervised k-means algorithm for the multi factor stock selection model, adds the modified Gaussian kernel function on the basis of introducing the gravitational image of the marked data to the unmarked data, and constructs the improved Gaussian kernel function and the semi supervised K-means clustering method based on the gravitational image factor. Without increasing the complexity of the algorithm, it greatly improves the ability of K-means model to deal with high-dimensional and linear inseparable problems.

Clustering method has a lot of research and application in stock selection, which is mainly due to the fact that the characteristics of clustering are very similar to stock selection. Different clustering combinations have been studied in the literature, and the stock selection results have also been significantly improved compared with the comparison model.

3.3 Support vector machine

Support vector machine is one of the most widely used supervised machine learning methods. Support vector machine can be divided into linear support vector machine and kernel support vector machine. The former aims at linear classification problem, and the latter belongs to nonlinear classifier.

The idea of support vector machine is to construct "hyperplane classification", and calculate the hyperplane by introducing kernel. For the linear case, the equation of the hyperplane is:

$$f_{(x)} = \beta_0 + \Sigma_{i=1}^{n} \alpha_i \langle x_i, x \rangle \tag{8}$$

where n represents the number of training samples, x is the characteristic of the new sample, $\langle x_i, x \rangle$ is the inner product of the new sample and all training samples, and β_0 and α It is calculated by training samples. After adding the core, replace the inner product to obtain a new equation:

$$f_{(x)} = \beta_0 + \Sigma_{i=1}^{n} \alpha_i K(x_i, x) \tag{9}$$

There are usually three kernel functions of support vector machine:

1. Linear kernel:

$$K(x_i, x_i') = \Sigma_{j=1}^{p} x_{ij} x_{i'j} \tag{10}$$

2. Polynomial kernel:

$$K(x_i, x_i') = (1 + \Sigma_{j=1}^{p} x_{ij} x_{i'j})^d \tag{11}$$

where d is the order of the polynomial.

3. Gaussian kernel:

$$K(x_i, x_i') = \exp(-\gamma(\Sigma_{j=1}^{p}(x_{ij} - x_{i'j})^2)) \tag{12}$$

Support vector machine has many advantages. And it has many models in dealing with linear and nonlinear problems. However, the training time of support vector machine is generally long, especially when the number of training samples is too large. Moreover, the kernel method requires to store the kernel matrix, resulting in the high spatial complexity of the algorithm. Therefore, support vector machine is more suitable for the task of small batch samples.

Lee [12] proposed a hybrid feature selection method, which combines the advantages of filtering method and packaging method, selects the best features subset from the original feature set. And then, it finds the best

parameter value of SVM kernel function. Through this method, the stock selection prediction of SVM is realized, and its accuracy is better than the traditional BP network. Zikowski [13] uses volume weighted support vector machine to select short-term stocks and then construct stock portfolio. Ru [14] uses AdaBoost enhancement method to process parameters and uses SVM as classifier. The addition of AdaBoost makes the model more profitable and less income fluctuation than traditional SVM. Zhang [15] tested SVM models with linear kernel, polynomial kernel, Gaussian kernel and sigmoid kernel, and constructed different K-SVM stock selection models on this basis. The experiments proved that K-SVM has strong prediction ability of stock selection income. Gao [16] proposed an integrated model of fuzzy dynamic SVM for stock selection. Experiments show that the average rate of return of the model is higher than that of the industry. Meng [17] proposed a multi-level stock selection model of GBDT-SVM based on a large number of factors, which uses machine learning technology to optimize factor selection and dynamic adjustment of factor weight, so as to improve the acquisition ability of multi factor model to stock excess return.

Due to the introduction of kernel method, support vector machine model can better solve the nonlinear classification problem than the previous machine learning model, which is very effective in complex stock selection data, solves the problem of linear indivisibility, and the obtained model is more effective and stable.

3.4 Random forest

Random forest is a classifier containing multiple decision trees. It trains a series of decision trees with Bagging method. It is a strong classifier that can be better applied to different scenarios of classification and regression.

Specifically, the random forest algorithm constructs each decision tree according to the following two-step method. The first step is called "row sampling", which is to sample back from all training samples to obtain a bootstrap data set. The second step is called "column sampling", m features (m less than m) are randomly selected from all m features, and M features of bootstrap data set are used as a new training set to train a decision tree. If it is classified prediction, the category or one of the categories with the most votes cast by N decision trees is the final category. In case of continuous numerical regression prediction, the value obtained by arithmetic averaging the regression or 0–1 classification results obtained from n decision trees is the final model output.

Random forest has good accuracy, can effectively run on large data sets, and can process input samples with high-dimensional features. Random forest can evaluate the importance of each feature. Therefore, dimensionality reduction is not required in the construction of random forest model. These characteristics make the training and prediction of random forest faster. The performance of random forest in stock selection is better than the traditional stock selection methods in many literatures, and it is widely used in the construction of stock selection model.

Random forest algorithm can solve two kinds of problems: classification and regression. Tan [18] and Li [19] uses random forest to select stocks by multiple factors through financial data, industry data, stock data and other data to build a stock portfolio, and its return rate is higher than the average return. Zheng [20] uses the minimum spanning tree method to investigate the relationship between stocks, explore the structure of stock network, use Kruskal algorithm, Prim algorithm and weight matrix method to solve the minimum spanning tree of stock network, model and solve the factors of stocks according to grouping technology, find out the similarities and differences between stocks, and select stocks. Jia [21] proposed a random forest support vector machine model. The model uses random forest to reduce the dimension of the original data, which improves the reliability and effectiveness of decision-making. By comparing the model with PCA-SVM, it is confirmed that the stock selection accuracy of the model is higher.

As an ensemble learning model, random forest can solve the instability of a single model in classification. Random forest can process higher dimensional data, ensure the diversity of stock selection factors, and make the stock selection results more comprehensive and effective.

3.5 Neural network

Artificial neural network is a model that imitates the information transmission between brain neurons. It can approximate any function with any accuracy. It can deal with all kinds of complex nonlinear relations. It is mostly used to deal with classification problems. Artificial neural network can be divided into multi-layer and single-layer. Each layer contains several neurons and they are connected by a directed arc with variable weight.

In the process of model training, we calculate the value a of the hidden layer according to the input feature X, and then calculate the predicted value y of the output layer according to the value of the hidden layer. This step is called forward propagation algorithm. We hope that the closer the predicted value y is to the real value T, the better, and the closer the output value is to the label depends on the connection weight in each layer.

Theoretically, there is a close relationship between the number of hidden layers and nodes and the fitting degree of the model to the data. More number leads to more accuracy. However, more number of layers and nodes will bring many problems, such as the number of weight parameters, which makes the solution space of the optimization problem too large and the algorithm is difficult to converge. Secondly, the back-propagation algorithm will also fail. The error gradient becomes minimal after several layers of transmission, and it is almost impossible to modify the connection weight of the previous layers. In addition, it also brings the problem of over fitting.

Asriel [22] used multilayer feedforward neural network to predict the stock excess return through technical and fundamental factors, and constructed a hedging portfolio composed of the same capitalized long and short positions, with a return higher than that of S&P 500 index. Ghosn [23] and Quah [24] and Stanley [25] and Yeh [26] and Kim [27] and Huang [28] use artificial neural network to select stocks in different stock markets and build stock portfolios, and their returns are higher than those of the general index. Tsai [29] and Lan [30] use back propagation neural network to conduct quantitative stock selection in Taiwan stock market and Shenzhen stock market. Principal component analysis is added to select indicators, and its accuracy and yield are higher than that of BPNN network [30]. Tong [31] compared multilayer perceptron, adaptive neuro fuzzy inference system and general growth radial basis function, and proposed how to systematically select stocks by using relative operating characteristic curve (ROC).

Neural network can build more complex stock selection model. Through the study of different literature, we find that the organization of different neurons has a great impact on the effect of stock selection model. And different stock selection factors also have a great impact on the results of the model, and the performance of the same model in different stock pools is also quite different.

3.6 Deep learning

As the most important branch of machine learning, deep learning has developed rapidly in recent years and attracted extensive attention at home and abroad. As one of the hottest trends in machine learning and artificial intelligence research, deep learning realizes the feature extraction of external input data from low-level to high-level by establishing and simulating the hierarchical structure of human brain, so as to achieve the purpose of interpreting external data. In recent years, as a new branch of machine learning, deep learning has achieved great success in many fields, including image processing, computer vision, speech recognition, machine translation, natural language processing and so on.

Deep learning can be regarded as a neural network structure with multiple hidden layers. It can form more abstract high-level representation attributes or features by combining low-level features, so as to find the distributed feature representation of data.

Deep learning can also be regarded as a general term of a series of methods. Typical deep learning models include deep feedforward neural network, deep belief network, convolutional neural network, generative countermeasure neural network, graph convolution neural network, etc. There are great differences in the structure of each model. According to our research on stock selection literature, we find that cyclic neural network is widely used in quantitative investment, and there are many applications of cyclic neural network in stock

selection.

Yang [32] proposed a multi-index feature selection method based on multi-channel convolutional neural network structure. The maximum information coefficient feature selection method is used to select candidate indicators to ensure the correlation with stock trend and reduce the redundancy between different indicators, so as to build a high-yield stock portfolio. Zhang [33] and Zhang [34] and Sun [35] is based on the LSTM model for stock selection. The deep stock ranker model proposed in [34] is based on the LSTM model, which can predict the future earnings ranking of stocks and make stock selection. Sun [35] selects stocks based on the sequence of yield prediction, and studies LSTM and GRU series prediction models. The prediction and stock selection accuracy of the two models is better than the traditional forward neural network model, and GRU model is slightly better than LSTM model. Zhou [36] constructs a multi-factor stock selection model based on adaptive recurrent neural network algorithm. Compared with RNN, the model has less calculation time, low cost and high accuracy. Compared with LSTM, the accuracy and convergence speed of the model are significantly improved.

The deep learning model can better process high-dimensional data and build more complex stock selection models. Moreover, for the stock trend with frequent changes, the deep learning model has better fault tolerance, can make better adjustments, and the stock selection results are more efficient and stable.

4 Results and discussion

By studying stock selection models constructed by machine learning algorithm in different literatures, we find that the performance of stock selection model based on machine learning is better than traditional stock selection model and comparison index. This is mainly because the characteristics of machine learning method for high-dimensional data processing are better than traditional methods, which can meet the needs of high-dimensional information extraction and analysis. Financial data has the characteristics of very low signal-to-noise ratio, less samples and unstructured data, high-dimensional information input and market dynamics. It has high requirements for the model. Machine learning method can deal with these characteristics of financial data well, which leads to the better effect of stock selection model constructed by machine learning method than traditional stock selection model.

We select six machine learning models in all a shares of China's stock market, and compare them with SSE 500 in excess return, information ratio and max drawdown. Among them, in the deep learning model, we choose RNN model as the stock selection method.

Excess rate of return is the difference between the actual rate of return of the stock and its normal rate of return, expressed as:

$$R_{it} = \alpha_i + \beta_i R_{im} + \varepsilon_{it} \tag{13}$$

where, R_{it} is the actual rate of return of stock i at time t, R_{im} is the rate of return of the market at time t, ε_{it} is a random perturbation term. The comparison of excess rate of returns is shown in Figure 2.

Information ratio can measure the heterogeneity of stocks and represent the excess return brought by unit active risk.

$$\mathrm{IR}_i = \frac{\overline{TD_i}}{TE_i} \tag{14}$$

where, IR_i is the information ratio of stock i. $\overline{TD_i}$ represents the sample mean of tracking deviation of stock i. TE_i is the tracking error of stock i. The greater the information ratio, the higher the excess return obtained by tracking error. Therefore, the performance of stocks with high information ratio is better than that of stocks with low information ratio. The comparison of information ratios is shown in Figure 3.

Max drawdown refers to the drawdown range of the return rate from the maximum net value to the lowest

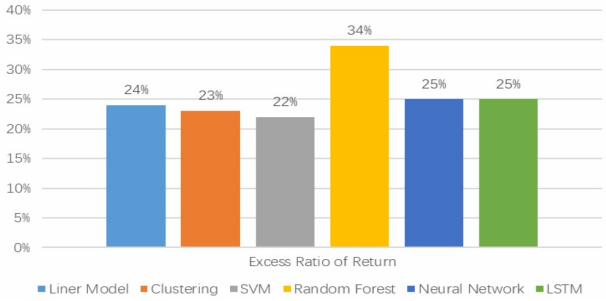


Fig. 2 The comparison of excess rate of returns.

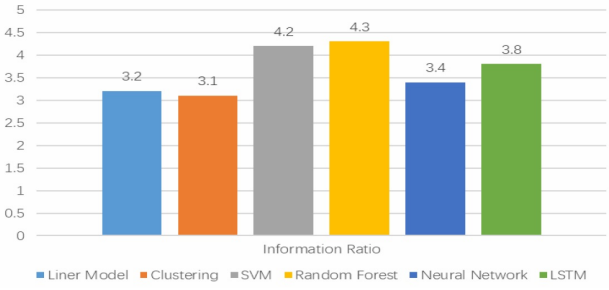


Fig. 3 The comparison of information ratio.

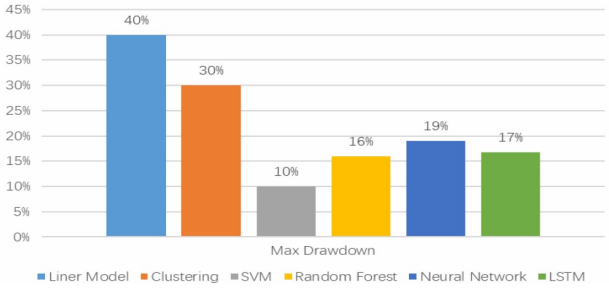


Fig. 4 The comparison of max drawdown.

net value in the cycle. It is an important risk index to evaluate the stability of the model. The comparison of max drawdown is shown in Figure 4.

Through the comparison of six machine learning models, we find that each model has different performance in different evaluation indexes. However, in general, random forest performs best in these evaluation indicators. Firstly, the excess ratio of return is significantly higher than other models, the information ratio is also higher than other models, and the max drawdown is only higher than support vector machine model. The performance of neural network model and RNN model is also better than linear model and clustering model, and effective experimental results are obtained.

5 Conclusion

Over the years, as an important part of quantitative investment strategy, stock selection methods have changed with the continuous development of computer technology. The behavior of the stock market has not changed in essence. The traditional stock selection method has been fully verified in theory and practice. No matter how the stock market changes, its basic principle has not changed fundamentally, as long as the data

and strategies during stock selection are consistent with the corresponding market mechanism, the traditional stock selection method can still be used in the current stock market. With the development of machine learning, quantitative investment methods have also been promoted. Through the use of new machine learning methods, stock selection methods also have new research directions and ideas. We briefly introduce the six most common machine learning methods in the field of stock selection, and introduce the application examples of different machine learning methods through the analysis of literature. Through the experimental research in China's stock market, it is compared with the SSE 500 in three aspects: excess return, information ratio and max drawdown. The experimental results show the effectiveness of the machine learning method. Through this paper, we hope that people can study more practical stock selection methods according to the actual situation of the stock market and various applicable machine learning methods, so as to obtain more efficient and stable returns in the stock market. In our future work, we will deeply study the construction of stock selection models of machine learning methods in different stock markets, especially the application of integrated learning methods represented by random forest in the field of stock selection.

References

[1] Yu Hua. Empirecal Analysis of Quantitative Investment Trend Strategy [J]. Financial Management, 2019, (04):189.
[2] Hiemstra Y. Linear Regression versus Back Propagation Networks to Predict Quarterly Stock Market Excess Returns [J]. Computational Economics, 1996, 9(1):67–76.
[3] Huang C F, Hsieh T N, Chang B R, et al. A Comparative Study of Stock Scoring Using Regression and Genetic-Based Linear Models [C]. IEEE International Conference on Granular Computing. IEEE, 2012.
[4] Huang HY, Wang M, Zhu J M. Multi factor stock selection model based on multiple regression analysis [J]. Journal of TongHua Normal University, 2016, (8):3.
[5] Chaudhary S, Arora V, Singh V. Regression based on Stock Selection Market Prediction [J]. IJARIIE, 2018, 4(3).
[6] Chen M X, Wu M, Wu H, et al. Logistic Forecasting Stock Selection Model based on Financial Driving Factors [J]. Economic and Trading Practice, 2018, (12X):2.
[7] Steve Craighead, Bruce Klemesrud. Stock Selection Based on Cluster and Outlier Analysis [C]. IMA Conference. 2002.
[8] Newton Da Costa, Jefferson Cunha, Sergio Da Silva. Stock Selection Based on Cluster Analysis [J]. Finance, 2005, 13(1):1–9.
[9] Ruizhong Wang. Stock Selection Based on Data Clustering Method [C]. 2011 Seventh International Conference on Computational Intelligence and Security, IEEE, 2011.
[10] Ratchata Peachavanish. Stock Selection and Trading Based on Cluster Analysis of Trend and Momentum Indicators [J]. Proceedings of the International MultiConference of Engineers and Computer Scientists, 2016.
[11] Li Wenxing, Li Junqi. Improvement of Semi-supervised Kernel Clustering Algorithm Based on Multi-Factor Stock Selection [J]. Statistics & Information Forum, 2018, 33(3):30–36.
[12] Lee M C. Using Support Vector Machine with a Hybrid Feature Selection Method to the Stock Trend Prediction [J]. Expert Systems with Applications, 2009, 36(8):10896–10904.
[13] Zikowski K. Using Volume Weighted Support Vector Machines with Walk Forward Testing and Feature Selection for the Purpose of Creating Stock Trading Strategy [J]. Expert Systems with Applications, 2015, 42(4):1797–1805.
[14] Ru Z, Zi-Ang L, Shaozhen C, et al. Adaboost-SVM Multi-Factor Stock Selection Model Based on Adaboost Enhancement [J]. International Journal of Statistics & Probability, 2018, 7(5):9–18.
[15] Zhang R, Lin Z A, Chen S, et al. Multi-factor Stock Selection Model Based on Kernel Support Vector Machine [J]. Journal of Mathematics Research, 2018, 10(5):119–129.
[16] Gao Anjing. Research on Stock Selection Based on The Method of Dynamic Fuzzy Integration Support Vector Machine [D]. Harbin Institute of Technology, 2017.
[17] Meng Qingyan. Optimizing Multi-Factor Stock Selection System Using GBDT-SVM Multi-Level Model [J]. Statistics and Application, 2019, 8(1):184–192.
[18] Tan Z, Yan Z, Zhu G. Stock Selection with Random Forest: An Exploitation of Excess Return in the Chinese Stock Market [J]. Heliyon, 2019, 5(8):e02310.
[19] Li Qi, Yang Junqi. Application on Random Forest Algorithm in Multi-Factor Stock Selection [J]. Manager Journal, 2017, (2):243.
[20] Xiangkun Zheng. Multi-Factor Model Based on the Minimal Spanning Tree [D]. Hebei University of Technology, 2016.
[21] Jia Xiujuan. Quantitative Stock Selection Based on Support Vector Machine of Random Forest [J]. Journal of Regional

Financial Research, 2019, (1):27–30.
[22] Asriel E. Levin. Stock Selection via Nonlinear Multi-Factor Models [C]. Advances in Neural Information Processing Systems 8, Nips, Denver, Co, November. DBLP, 1996.
[23] Ghosn J, Bengio Y. Multi-Task Learning for Stock Selection [C]. Advances in Neural Information Processing Systems 9, Nips, Denver, Co, Usa, December. DBLP, 1997.
[24] Quah T S, Srinivasan B. Improving Returns on Stock Investment through Neural Network Selection [J]. Expert Systems with Applications, 2006, 17(4):295–301.
[25] Stanley, G, Eakins, et al. Can Value-Based Stock Selection Criteria Yield Superior Risk-Adjusted Returns: An Application of Neural Networks [J]. International Review of Financial Analysis, 2003.
[26] Yeh, I-Cheng, Liu, et al. Using Mixture Design and Neural Networks to Build Stock Selection Decision Support Systems [J]. Neural Computing & Applications, 2017.
[27] Kim G H, Kim S H. Variable Selection for Artificial Neural Networks with Applications for Stock Price Prediction [J]. Applied Artificial Intelligence, 2018:1–14.
[28] Yuxuan Huang, Luiz Femando Capretz, Danny Ho. Neural Network Models for Stock Selection Based on Fundamental Analysis [J]. 2019 IEEE Canadian Conference of Electrical and Computer Engineering, IEEE, 2019.
[29] Tsai C F, Hsiao Y C. Combining Multiple Feature Selection Methods for Stock Prediction: Union, Intersection, and Multi-Intersection Approaches [J]. Decision Support Systems, 2011, 50(1):258–269.
[30] Lan Taiqiang. An Empirical Study on Comprehensive Stock Selection Based on Principal Component Analysis and BP Neural Network [D]. Ji'nan University, 2017.
[31] Tong-Seng Quah. Using Neural Network for DJIA Stock Selection [J]. Engineering Letters, 2007.
[32] Hui Yang, Yingying Zhu, Qiang Huang. A Multi-Indicator Feature Selection for CNN-Driven Stock Index Prediction [J]. Springer Nature Switzerland AG 2018, 2018:35–46.
[33] Zhang X, Tan Y. Deep Stock Ranker: A LSTM Neural Network Model for Stock Selection [M]. Data Mining and Big Data. 2018.
[34] Zhang R, Huang C, Zhang W, et al. Multi Factor Stock Selection Model Based on LSTM [J]. International Journal of Economics & Finance, 2018, 10(8):36–42.
[35] Sun J. A Stock Selection Method Based on Earning Yield Forecast Using Sequence Prediction Models [J]. Papers, 2019.
[36] Zhou Zhiyuan. Research and Application of Multi-Factor Stock Selection Model Based on RNN-ACT Algorithm [D]. Kunming University of Technology, 2018.