Macrothink Institute™

Business and Economic Research
ISSN 2162-4860
2018, Vol. 8, No. 4

Multi-factor Stock Selection Model Based on Adaboost

Ru Zhang

Finance Department of International Business School, Jinan University

Zhuhai, 519070, China

E-mail: zhangru@stu2016.jnu.edu.cn

Tong Cao (Corresponding author)

Finance Department of International Business School, Jinan University

Zhuhai, 519070, China

E-mail: monica@stu2016.jnu.edu.cn

Received: September 28, 2018 Accepted: October 17, 2018 Published: November 21, 2018

doi:10.5296/ber.v8i4.13942 URL: https://doi.org/10.5296/ber.v8i4.13942

Abstract

In this paper, we established multi-factor stock selection model based on Adaboost by using Adaboost to integrate the custom week classifier model, and Shanghai and Shenzhen 300 stocks are taken as the research object. During the stock retest, the first is make a comparative test between Adaboost multi-factor stock selection model and the traditional multi-factor model, among them, the factor large class isn't considered in the multi-factor stock selection model. And the results of two contrast experiment showed that the multi-factor stock selection model based on Adaboost has stronger profitability and less risk than the traditional multi-factor model.

Keywords: Quantitative investment, Multi-factor stock selection model, Adaboost

1. Introduction

Quantitative investment is an investment strategy. And on the basis of statistics, quantitative investment analyses and studies market data and information by mathematical model to excavate high value information for the later investment decision. Its advantages lie in objective rationality, accuracy, controllability, and efficiency and sensitivity (Liang Ou & Yongping Teng, 2018). As a classic model of stock investment, the combination of multi-factor stock selection model (Malkiel B. G. & Fama E. F., 1970. ASMESS C. S., 1997. Chen N, Zhang F, 1998. Mohanram P S, 2005) and machine learning algorithm has also attracted wide attention from academia and industry. Considered as a more effective algorithm of learning algorithm, Adaboost can be used to enhance the classification function of the weak classifier (Dhagat A & Hellerstein L, 1994), has more efficient performance in the problem of classification and regression (Yan Yan & Xiaoqing Ding,2008. Hongsheng Xie & Hong Zhang, 2009), and can be applied to the optimization of multi-factor stock selection model in view of the traditional regression method.

Therefore, we take Shanghai 300 stock as the research object and use the custom week classifier model integrated by Adaboost to established multi-factor stock selection model. In the stock retest, we first have a comparative test of Adaboost multi-factor stock selection model and the traditional multi-factor model without considering factor large class. And then, a robust model with high yield and low risk is built after analyzing the test results, which provides new ideas for the application of machine learning algorithm in quantitative investment.

2. Theoretical Model – Multi-factor Stock Selection Model Based on Adaboost

Adaboost is not a classifier itself, so, in order to use Adaboost to enhance classification algorithm, a weak classifier algorithm is required. The common classifier includes decision tree classifier, nearest neighbor classifier, neural network, support vector machine and so on, and all of there can be used as the weak classifier for Adaboost. In the paper, statistical classification is as the first method to be tried, which is the simplest and a classification method according to the statistical value of the exposure of each factor in T-1 period and the income in T period. Adaboost is an enhanced algorithm for statistical classification to observe whether the strong classification is effective under Adaboost. According to the definition of the weak classification algorithm, we have made the proper deformation to the standard Adaboost. The weak classifier in each layer is determined by one factor until the exit condition is reached. Then several weak classifiers can be got to form strong classifiers.

The content of Adaboost algorithm model is as follow:

(1) Division of the class standard

A stock training set $S=\{(x_1,y_1),...,(x_N,y_N)\}$ is given, and x_N is the data processed by the n factor of the N stock.

(2) Construct a weak classifier

The weak classifier h is established by a factor of x_i in the factor pool. The so-called weak classifier is a nonlinear function that maps factor values to the trust score space. For the k factor of stock i, the trust score is $f^k(x_i)$, and the weight of each stock is expressed as $w(x_i)$.

A weak classifier h is defined as the following piecewise function:

$$h(x)=\frac{1}{2}\ln(\frac{W_{+}^{j}+\varepsilon}{W_{-}^{j}+\varepsilon}) \tag{1}$$

Among which, $\varepsilon=\frac{1}{N}$, $j=1,2,...Q$ is the number of quantile segment. In this paper, $Q=5$, and according to the value of the k factor and their quantiles, the training samples are divided into Q group. W_{+}^{j} represents the sum of the weights in the classification j.

$$W_{+}^{j}=\sum_{y_i=\mathrm{m},\, f(x)\in quantilej} w(x_i) \tag{2}$$

The index Z is used to measure the quality of the weak classifier:

$$Z=\sum_{j=1}^{Q}\sqrt{W_{+}^{j}W_{-}^{j}} \tag{3}$$

Intuitively, if the weight of a strong stock is greater in a category, the value $h(x)$ is greater. If the factor of a stock falls in this classification, then there is reason to believe that the stock will perform better, because the sum of all weights is 1. If the different between W_{+}^{j} and W_{-}^{j} is large, the value of index Z is small, so the smallest Z corresponding classifier can be selected as weak classifier.

(3) Update the weight

After each round, the corresponding weight of each stock should be updated:

$$w_{l+1}(x_i)=w_l(x_i)e^{-y_i h_l(x_i)} \tag{4}$$

Among which, l represents the weak classifier in the l layer. If the current weak classifier is correctly classified, then the next round of classification will reduce the weight of the stock.

(4) Combined strong classifier

The final classifier is obtained by summing up all weak classifiers:

$$H(x)=\sum h_l(x) \tag{5}$$

3. Empirical analysis

3.1 Data Preprocessing

Because most of the machine learning algorithms are sensitive to the input data, for example, to the normalization of data to, if there is a lot of noise data, it is easy to cause over fitting. Because the aim of this paper is to seek the relative advantage between strong and weak potential shares and do not pay much attention to the absolute size of the data, we calculate the ranking of each stock by a factor, and then, we divide it by the total stock number to achieve the normalization of factor values.

Then, the next period of return is ranked from small to large, and the first 30% is taken as a strong stock, the second 30% is taken as the weak share. The strong stocks classified as +1, the disadvantaged stocks classified as -1. Besides, the middle 40% of the stock are removed from the training concentration, because the share of the stock is not strong and not weak and can be treated as noise data. At the same time, in order to make full use of the data, we find the relatively stable and effective factors to ensure the robustness of the algorithm. At last, we use the panel data of up to 12 months to build the training set.

3.2 Retest Analysis

The parameters of the machine stock selection model are, and Shanghai and Shenzhen 300 index are as stock pool. Based on the trust score, the stock pool is divided into ten stalls. The top one is taken as a strong combination, and the lowest one is a weak combination. The data for the test is selected from 2011-01-31 to 2016-03-31

3.2.1 Adaboost Stock Selection Model without Considering the Factor Categories

With the previous statement of the Adaboost algorithm, the net value of the combination of all commonly used factors into the algorithm, regardless of the factor categories, is as follows:

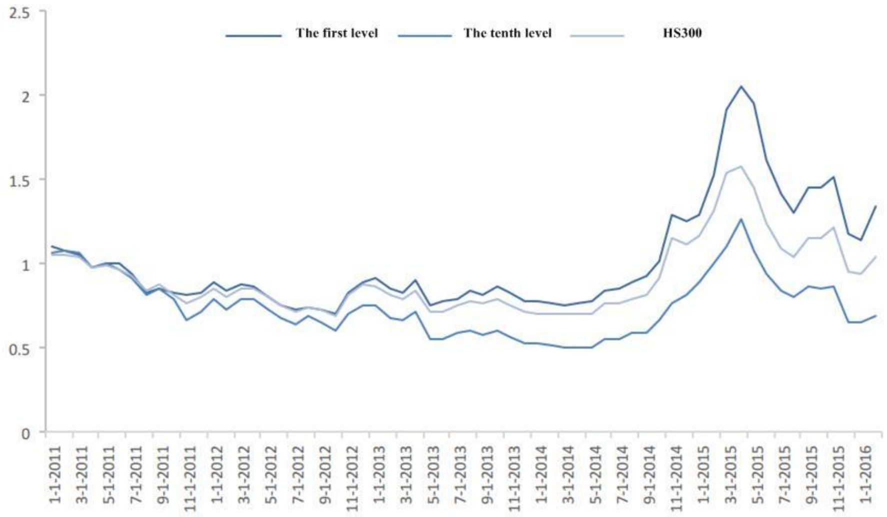

Figure 1. Net value of Adaboost multi-factor stock selection

From Fig.1, it can be seen that the strong combination obtained by the Adaboost algorithm can win the market, and the net value difference is obvious between strong combination, market index and disadvantaged combination. The combination of the algorithms has clear regional diversity. Besides, we find that there is a certain interval between all ten file combinations, which shows that the algorithm is effective.

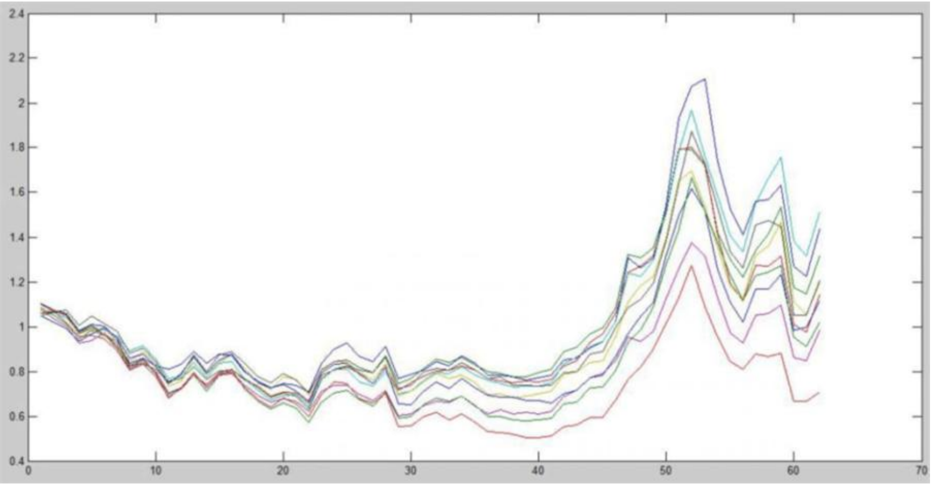

Figure 2. The next value comparison result in the Adaboost ten file combinations

From the above empirical results, we can see that the multi-factors stock selection model based on Adaboost is quite effective. Classifier is an algorithm in supervise learning for machine learning, and another main algorithm type in supervised learning is regression, which is in accordance with common multi-factor stock selection method. To further test the effectiveness of Adaboost, the next is to compare the multi factor stock selection model of Adaboost with traditional regression method.

In the regression method, we return the original factor data and the next rate of return within the same time range, and according to the factor regression coefficient value as the basis for selection. To compare with Adaboost algorithm, the original data used are consistent, but each method of standardization is different which is based on their requirement. And the regression method also selects 12 factors. After selecting factors by the regression method, we predict the lower income of each stock, and also take the highest gear of the ten gears as the strong combination and the lowest grade as the weak combination. The result is as follow:

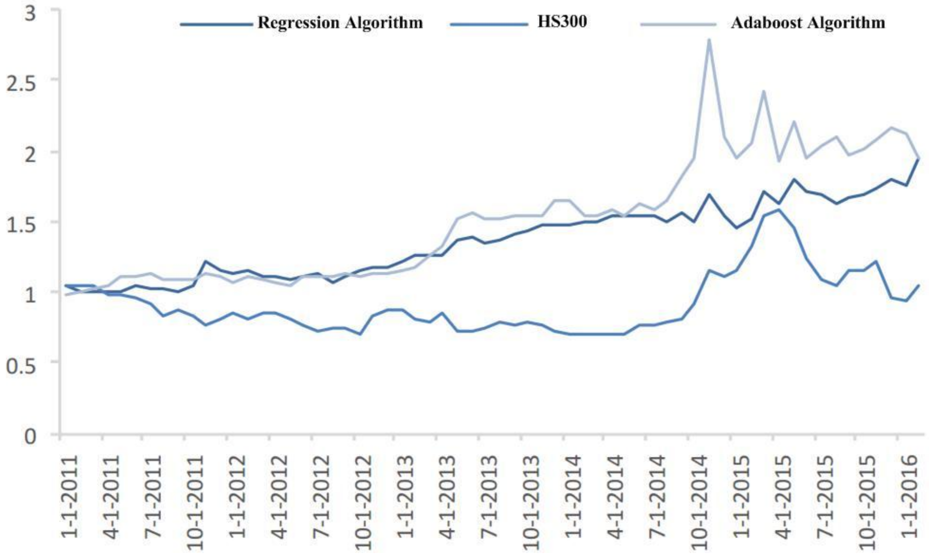

Figure 4. The next value comparison result in two strategies

Compared the result of using Adaboost, the net value in the combination of traditional regression are better than Adaboost in both strong and weak group. Based on the Figure 4, although the two methods have the same net value in the return period, the highest net value produced by the Adaboost method is higher.

3.2.2 Stock Selection Model Considering Factor Categories

Adaboost algorithm can add all weak classifiers and change them into final strong classifier. The strong classifier will give a corresponding trust score for each factor value, so as to measure its strong shares. Because can be regarded as a composite factor, we construct Adaboost factor by Adaboost algorithm.

Therefore, we construct a strong classifier for the index of every major factor in the conventional 68 factors based on Adaboost algorithm, and then get the corresponding Adaboost factor. If the index of this class factor is more than 4, the number of layers of takes half of the index number. At the same time, the index of the weak classifier is recorded, and the mean value of the class factor is calculated. To rank, Take the corresponding to the first 4 minimum values as the final Adaboost factor. Then we predict the stocks in the current data and build a strong and weak combination.

The traditional factors are constructed as the control group. In each month, the next period returns to each index, and obtains the regression coefficient value, thus getting the value sequence for several months. The average value of the absolute value sequence of value is greater than 2, and the ratio of absolute value sequence greater than 2 is greater than 20%, and an effective risk index is found. Based on the large class attribution of Table 1, we construct a large class factor. Then we can predict the stocks in the current data and build a strong and weak combination.

Table 1. Broad class factors involved in the stock-picking model

Type of factor	**Type of factor**
Valuation	Share price
Growth	Beta
Financial quality	Turnover rate
Gearing	Mood
Market value	Shareholder
Momentum of the inversion	Technology
Volatility	

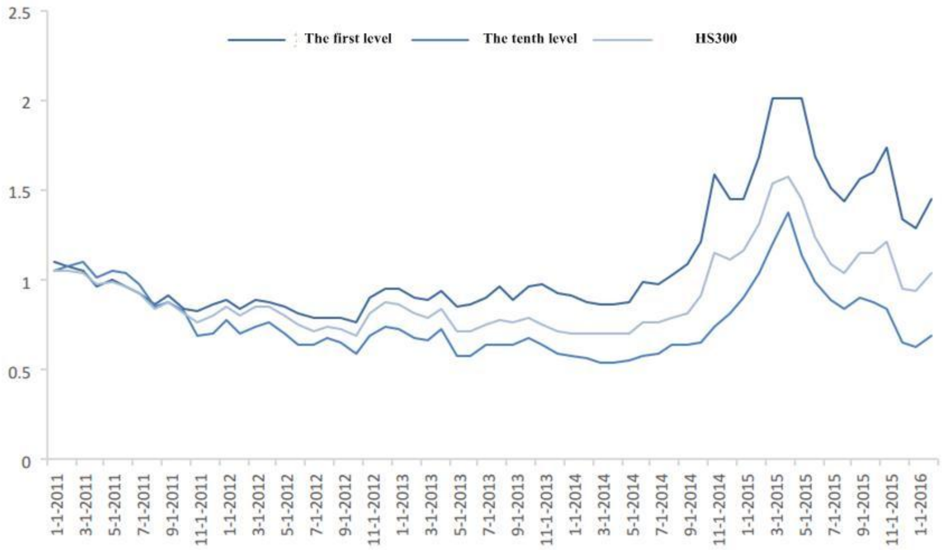

Figure 5. Net value of Adaboost multi-factor stock selection

From the net value curve, we can find that, after considering the large class of factors, the algorithm has a higher degree of discrimination in the stock selection combination, and the net value difference between the strong combination and the weak combination is obvious.

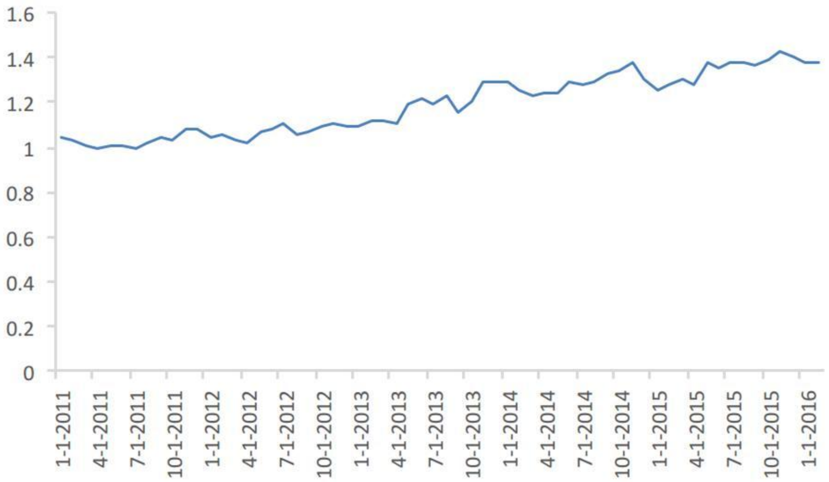

Figure 6. Relative index excess return curve

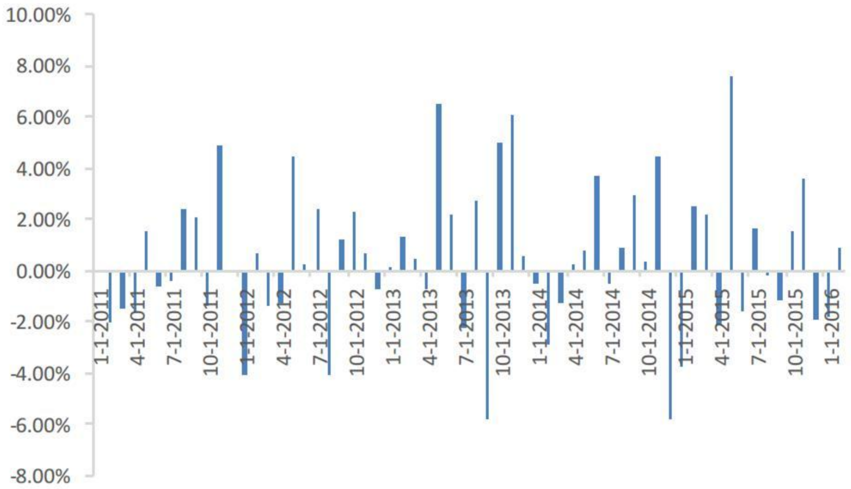

Figure 7. Relative index monthly excess returns

The maximum return net value curve of the combined relative HS300 index is relatively stable, the maximum retracting occurred in December 2014, and the maximum retracement was 9%. Before this, the maximum return of excess return was about 5%. From the monthly excess returns, it is obvious that the combined monthly victory rate is over 50%, and reaches 56%, which indicates that the overall effect is good. Compared to the excess returns of the disadvantaged combination, the net value of the multi space strategy is more volatile. The main retracement also occurs at the end of 2014, but the winning rate of the strategy remains above 50%, reaching 58%.

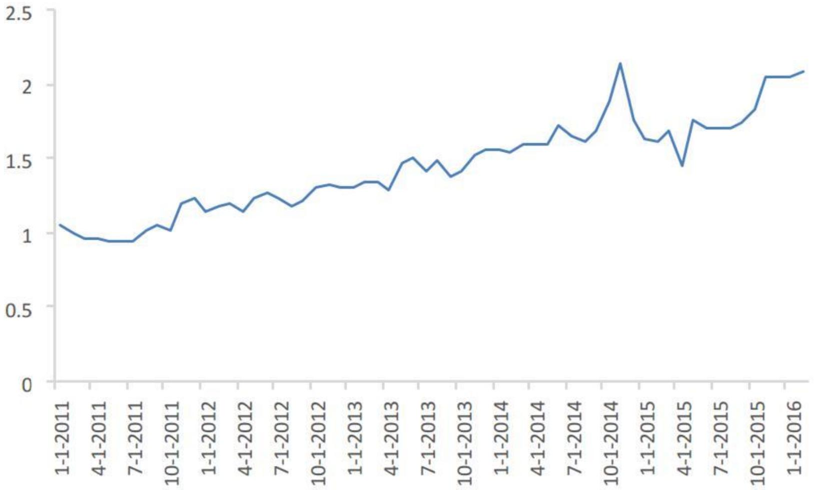

Figure 8. Excess return curve of BBI strategy

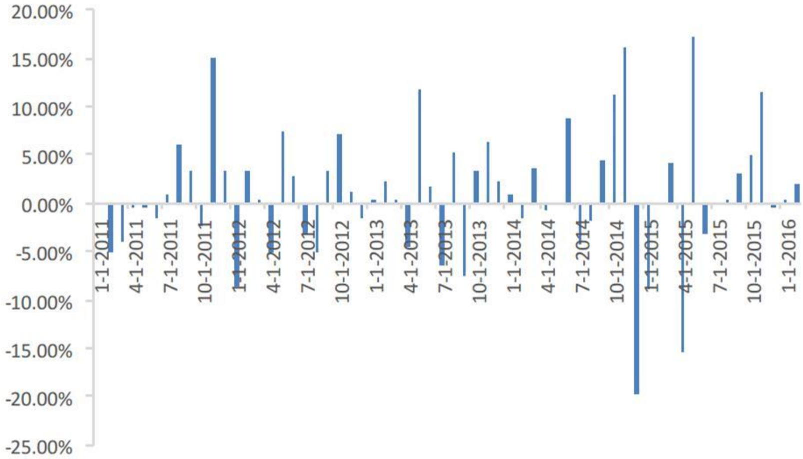

Figure 9. Relative disadvantaged combination of excess returns

As a comparison of regression algorithm, the stock selection effect is relatively poor.

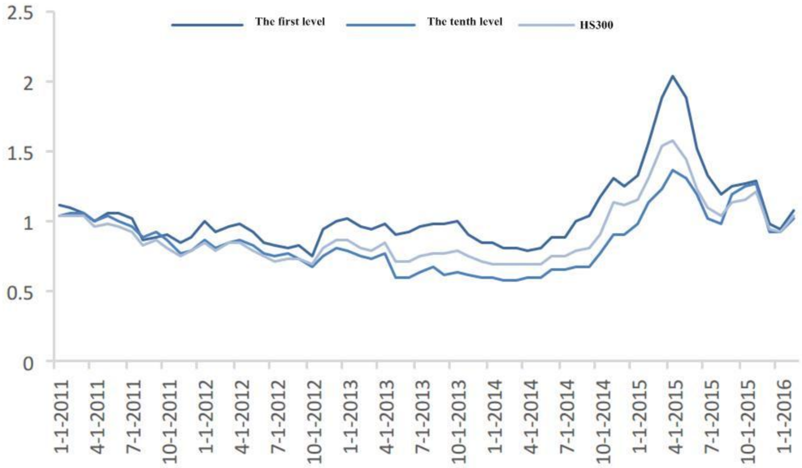

Figure 10. Net value of multiple factor selection combination with the regression method

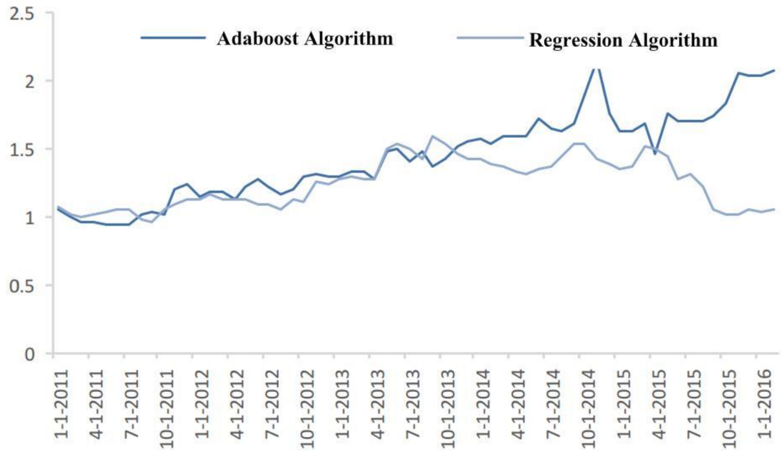

Figure 11. Comparison of excess income of relatively weak combination in two methods

In the case of the same setting conditions of the two algorithms, the strong combination constructed by the Adaboost factor can obviously win the market, the strong combination, the market index, the disadvantaged combination have obvious distinction, and the combination effect of the traditional factor construction is a little worse. From the two methods of strong weak combination net value comparison chart, the effect of the Adaboost factor is better and more stable. The strong combination can maintain the advantage over the weak combination, especially in the rapid recovery after the big rise and fall of the abnormal market.

4. Conclusion

The Adaboost algorithm discussed in this paper is not a classifier algorithm, but a synthetic algorithm related closely to the classifier. Because the degree of coincidence of factor and classified thought is high, we hold the view that, in addition to the traditional multiple factor regression, the improvement of classifier algorithm and classification can provide new ideas for factor selection.

From the result, Adaboost stock selection is effective. Starting with the simplest probability and statistics, 68 factors information are incorporated into the Adaboost algorithm, and the result shows that the combined income of BBI can be distinguished clearly and excess return can be gained than market index. For the refinement of the model, the factors can be selected according to the large class, which can further improve the division ratio of the multi space combination. In all kinds of retest results, we compare Adaboost with the traditional regression method, and the result shows that the method of Adaboost and probability statistics classification can keep high validity, so it is feasible to consider the multi factor selection from the point of view of classifier.

There are two directions worth considering improving model:

First, the stock selection model in paper does not take into account the general optimization such as the general neutral problems in the multifactor stock selection. So, the market value neutral, the industry neutral, and the portfolio risk optimization are all the directions that can be further strengthened.

Second, the classification method used in paper is a probability statistic, and the factor problem is highly compatible with the classifier in fact. Therefore, a better classifier can expect to get better back test results. However, because Adaboost is not a classifier algorithm, the enhancement of Adaboost for most classifiers may further improve the effectiveness of multi-factor stock selection.

References

Asmess, C. S. (1997). The Interaction of Value and Momentum Strategies. *Finance Analysis Journal*, 29-36. https://doi.org/10.2469/faj.v53.n2.2069

Chen, N., & Zhang, F. (1998). Risk and Return of Value Stocks. *Journal of Business, 71*, 501-535. https://doi.org/10.1086/209755

Dhagat, A., & Hellerstein, L. (1994). PAC learning with irrelevant attributes. Symposium on

Foundations of Computer Science. *IEEE Computer Society*, 64-74. https://doi.org/10.1109/SFCS.1994.365704

Lawrence, S. et al. (2001). Persistence of Web References in Scientific Research. *Computer*, *34*, 26-31. http://dx.doi.org/10.1109/2.901164

Liang, O., & Teng, Y. P. (2018). Application of quantified investment in futures market. *Inner Mongolia coal economy*, 81-82. http://dx.doi.org/10.13487/j.cnki.imce.011412

Malkiel, B. G., & Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, *25*(2), 383-417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x

Mohanram, P. S. (2005). Separating Winners from Losers among LowBook-to-Market Stocks using Financial Statement Analysis. *Review of Accounting Studies*, *10*, 133-170. https://doi.org/10.1007/s11142-005-1526-4

Smith, J. (1999). One of Volvo's core values. [Online] Available: http://www.volvo.com/environment/index.htm (July 7, 2007)

Strunk, W., Jr., & White, E. B. (1979). *The elements of style* (3rd ed.). New York: Macmillan, (Chapter 4).

Van der Geer, J., Hanraads, J. A. J., & Lupton, R. A. (2000). The art of writing a scientific article. *Journal of Scientific Communications*, *163*, 51-59

Xie, H. S., & Zhang, H. (2009). Integration of SVM and Boosting methods in content-based image retrieval applications. *Computer Applications*, *29*, 979-981. https://doi.org/10.3724/SP.J.1087.2009.00979

Yan, Y., & Ding, X. Q. (2008). Improved AdaBoost algorithm based on multi-step correction. *Journal of Tsinghua University: Natural Science Edition*, 1613-1616.

Copyright Disclaimer