Adaboost-SVM Multi-Factor Stock Selection Model Based on Adaboost Enhancement

Ru Zhang[1], Zi-ang Lin[2], Shaozhen Chen[1], Min Zhao[2], Mingjie Yuan[2]

[1] Finance Department of International Bussiness School, Jinan University, Zhuhai, Guangdong Province, China

[2] Financial Management Department of International Bussiness School, Jinan University, Zhuhai, Guangdong Province, China

Correspondence: Shaozhen Chen, Finance Department of International Business School, Jinan University, Qianshan Road 206#, Zhuhai City, Guangdong Province, Post No. 519070, China. E-mail: 1813012994@qq.com

Received: May 28, 2018 Accepted: June 25, 2018 Online Published: June 26, 2018

doi:10.5539/ijsp.v7n5p9 URL: https://doi.org/10.5539/ijsp.v7n5p9

Abstract

In recent years, the applications of machine learning techniques to perfect traditional financial investment models has gained a widespread attention from the academic circle and the financial industry. This paper takes CSI300 stocks as the object of the research, uses Adaboost to enhance the classification ability of original linear support vector machine, and combines all major factors to build Adaboost-SVM multi-factor stock selection model based on Adaboost enhancement. In the backtesting analysis, the stock selection strategy of original linear support vector machine was compared with the Adaboost-SVM multi-factor stock selection strategy based on Adaboost enhancement. The result shows that the Adaboost-SVM multi-factor stock selection strategy based on Adaboost enhancement possesses stronger profitability and smaller income fluctuation than the original algorithm model.

Keywords: quantitative Investment, multi-factor stock selection model, Adaboost-support vector machine

1. Introduction

Quantitative investment has been accepted and applied by the majority of investors due to its stable performance and rational investment. As one of the most widely used machine learning models in the financial field, support vector machines have been deeply studied by many scholars. Zhang (2016) proposed a new SVM-GARCH forecasting model and proved that the new model has better denoising and forecasting ability than the traditional ARMA-GARCH model for time series data through experiment. Huang (2017) used support vector machines to improve the traditional Fama-French three-factor model and build a new stock selection strategy. Through empirical analysis of A shares, he proved that the new strategy model has stronger profitability and forecasting capabilities. However, in practical experience, linear support vector machines often have the defects of weak classification. Dong et al. (2018) proposed Adaboost-SVM algorithm, which uses SVM as a basic classifier and *Adaboost* as an integrated algorithm to further enhance the basic classifier and proved that the Adaboost-SVM algorithm has better classification performance and robustness through experiment.

In conclusion, this paper takes the CSI300 stocks as the research object, and constructs Adaboost-SVM multi-factor stock selection model based on Adaboost. In the back test, the comparison between the original SVM stock selection model and the Adaboost-SVM multi-factor stock selection model based on Adaboost further proves that the Adaboost-SVM multi-factor stock selection model based on Adaboost has stronger profitability and revenue robustness.

2. Theoretical Model

2.1 SVM Model

Support Vector Machine (SVM), the main content of which was formed in the 1990s, has achieved breakthrough development in theoretical research and algorithm implementation in recent years. The learning algorithm of SVM, due to its effectiveness in overcoming traditional difficulties such as "Curse of Dimensionality" and "over fitting", has attracted much attention in the statistical field. The most important feature of SVM is that it has changed the traditional principle of minimizing empirical risk, that is, it is proposed for the principle of minimizing structural risk, so it has good generalization ability. At the same time, when the support vector machine deals with nonlinear problems, through transforming the nonlinear problem into a linear problem in the high-dimensional space and replacing the inner product operation in the high-dimensional space with a kernel function, it skillfully solves the complex computational problems, and effectively overcomes curse of dimensionality and local minima problems.

1.1.1 SVM Classification Algorithm

Generally, the classification problem can be expressed as: considering the classification problem in the n-dimensional space, it contains n indices (i.e., $x \in R^n$) and l sample points. The aggregation of these l sample points is:

$$T = \{(x_1, y_1), ..., (x_l, y_l)\} \in (X \times Y)^l$$

Where $x_i \in X = R^n$ is the input indicator vector, or input, its components become features, or input indicators; $y_i \in Y = \{-1, 1\}$ is the output indicator, or output, $i = 1, \ldots, l$ .The set of these l sample points is called the training set. The question now is, for any given new pattern x and the underlying training set, whether the corresponding output y is 1 or -1.

Combining the factor characteristics of stock, it is to take each factor as a dimension and find a rule that divides the points on R^n into two parts. Specifically, the above classification is divided into two types of problems. Similarly, there are classification problems that are divided into many categories. The difference is the number of output results. This paper mainly studies the classification problems of dividing stocks into two categories by using SVM. There are generally three types of classification problems, and different classifiers may be used for different types of problems. As shown in Figure 1:

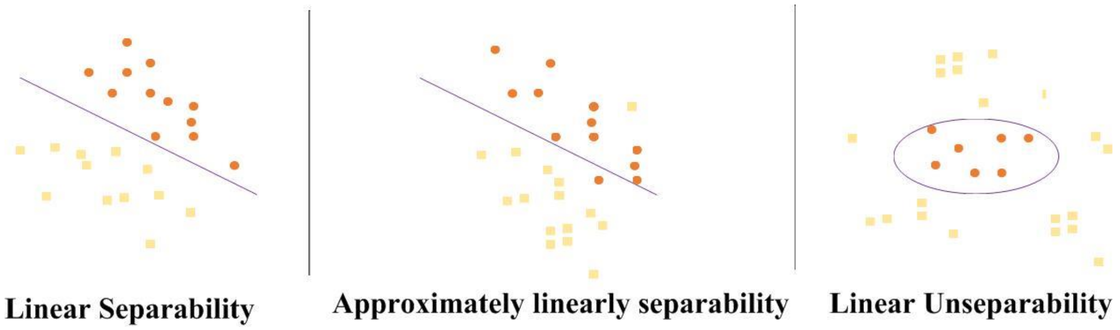

Figure 1. Classification problem types diagram

1.1.2 SVM Solves Linear Classification Problems

SVM algorithm is the process of finding the optimal hyperplane according to the training samples. In the case of two-dimensional coordinate points, the SVM algorithm is to find a straight line to separate the two types of coordinate points. However, there are countless lines, but in these lines, if they are too close to the coordinate points, the disturbance of noise will have a great impact on the classification results. So we can define that the SVM algorithm is to find the line that is farthest from the training sample, also called the optimal straight line.

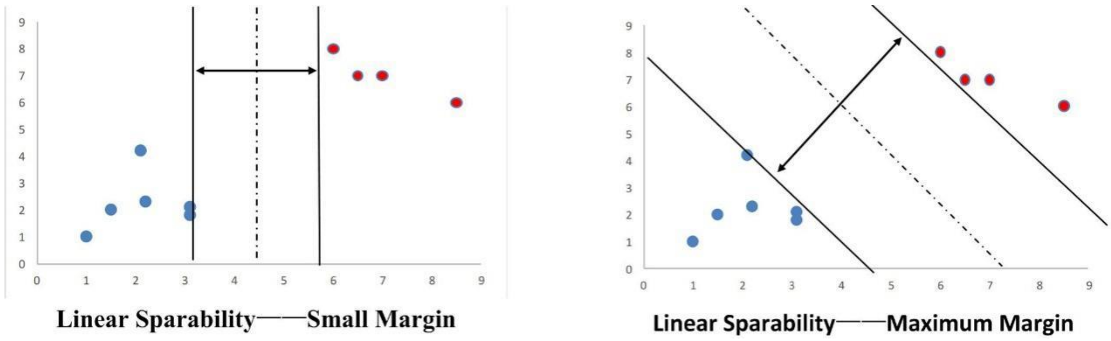
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9
Linear Sparability——Small Margin
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9
Linear Sparability——Maximum Margin

Figure 2. Classification algorithm principle

Starting from the definition of classification, assuming that the training sample set $T=\{(x_1,y_1),...,(x_l,y_l)\}\in(\mathrm{X}\times\mathrm{Y})^l$ $y_i\in Y=\{-1,1\}$ can be separated by hyperplane $g(x)=\langle w_0\bullet x\rangle+b_0=0$, and that the distance between the vector near the classification plane and the classification plane reaches the maximum, then G becomes the optimal classification hyperplane.

Eliminating the derivation of the optimization process, the resulting optimization problem is:

$$\max\sum_{j=1}^{l}a_j-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_ja_ia_j(x_i\bullet x_j)$$

$$s.t.\sum_{i=1}^{l}y_ia_i=0,c\geq a_i\geq 0,i=1,...,l$$

Among them, $x_i\in X=R^n,y_i\in Y=\{-1,1\}$; x_i is the input, and y_i is the corresponding output value; l is the number of samples, a is Lagrange coefficient, and c is regularization parameter. It is to achieve maximum margin and classification error reconciliation.

1.1.3 SVM Solves Nonlinear Classification Problems

The goal of SVM classification is to develop an effective approach to computing, so as to learn the "good" classification hyperplane in high-dimensional feature space. The research of SVM was originally proposed for the two types of linearly separable problems in pattern recognition. Because the data of the stock market is nonlinear, the classification ability of the hyperplane is limited, so the SVM performs nonlinear mapping on the data, through mapping: $\varphi:x\rightarrow f$,and maps the data into a higher-dimensional feature space so that the data is linearly separable and then the optimal hyperplane is constructed. Since both the optimization function and the classification function involve the inner product operation $\langle x_i\bullet x_j\rangle$ in the sample space, the inner product operation $\langle\varphi(x_i)\bullet\varphi(x_j)\rangle$ needs to be performed in the transformed high-dimensional feature space E .According to satisfying theory, corresponding the inner product in the linear transformation space, and adopting the appropriate kernel function, it can replace the nonlinear mapping in high-dimensional space, and realize the linear classification after nonlinear transformation .According to the optimality theory, the optimization problem is:`

$$s.t.\sum_{i=1}^{l}y_ia_i=0,c\geq a_i\geq 0,i=1,...,l$$

Here $f(x)=a_i^*k(x_i\bullet x_j)+b^*$,select b^* to make $y_if(x)=1$ work, and in terms of arbitrary i ,the decision rule $c\geq a_i^*\geq 0$ is set up which is given by $\mathrm{sgn}(f(x))$,and it's equal to the hyperplane within the characteristic space of kernel function $k(x\bullet z)$'s implicit definition which is designed to solve optimization problem, and the definition of slack variable is relevant to set interval, then:

$$\gamma=(\sum_{i,j\in w}y_iy_ja_i^*a_j^*Ker(x_i\bullet x_j))^{-\frac{1}{2}}$$

The corresponding decision function will be :

$$g(x) = \operatorname{sign}(\sum_{i=1}^{l} y_i a_i Ker(x_i \bullet x_j)) + b$$

What we can conclude from optimization function is that the training complexity index is unrelated to the method of SVM.

2.2 Original Stock Picking Strategies Based On SVM

According to economic definitions of index structure and references of BARRA model's factorial classifications, this paper divides 68 indexes into 11 factors including, Earnings variability, Growth, Leverage, Liquidity, Momentum, Size, Value, Volatility, Dividend yield, Financial Quality, etc.

The sample data standardization of SVM algorithm adopt ranking method, which means that ranking each stock by its corresponding factor's size, then dividing index by total stock numbers and standardized factor score will be in the domain $(0,1]$.

Later on , rank the next-term yield rate from the biggest to smallest, take the top 30% of stocks as strong stocks whilst the last as weak stocks ,and the former label as +1 whereas the latter label as -1; get the 40% of stocks in the middle out of training set because the 40% of stock yield in the middle is neither strong nor weak ,which serve as noisy data.

In order to optimize data to find stable and effective factors relatively as well as insure algorithm stability, this paper takes the past 12-month factor data as input samples.

What we can obtain from the SVM theoretical derivation is the samples was compartmentalized into two types of $\{-1,+1\}$ after getting the solutions to optimal plane, and the distance between samples and hyper plane represents the extent of how accurate the samples are classified. It can be represented by formula as:

$$r = \frac{w \cdot x + b}{\|w\|}$$

In this function, x is the new sample point, and w, b are the outcome of computing hyper plane.

On the basis of distance results, using the same way to classify stock portfolio into 10 types, then select the top type and the last type as the strong portfolio and the weak portfolio, the final step will be observing back-test results.

2.3 Adaboost Model

Adaboost is a kind of iterative algorithm, whose core idea is training different weak classifiers at the aim of a same training set, especially doing repeated trains to the data that are hard to be classified accurately, then gather all the weak classifiers to form a stronger strong-classifiers. The Adaboost algorithm itself is achieved by changing the layout of data, it adjusts each sample's weight by whether the classification of sample is correct in training set as well as by the accuracy rate of total classification in the last time, which make the hard-to-classify data trained. Send the new data set with amended weight to be trained by Bottom, classifiers, lastly aggregate the classifiers obtained from each training as the final Decision analyzer. There are Steps of Algorithm:

1) Input training set $\{(x_1, y_1), ..., (x_m, y_m)\}$, among there including $y_i \in Y = \{1, ..., k\}$, T represents the number of iterations, and $WeakLearn$.

Initialize: $\forall i : D_1(1) = \dfrac{1}{m}$

2) Traverse all the time points: $\forall t \in T : WeakLearn \to D_t$

Back to hypothesis: $h_t : X \to Y$

Calculate errors of h_t that will be $\varepsilon_t = \sum_{i:h_t(x_i)\neq y_i} D_t(i)$. If $\varepsilon_t > \dfrac{1}{2}$,then $T = t-1$,jump out of current loop .

Set $\beta_t = \dfrac{\varepsilon_t}{1-\varepsilon_t}$

Update $D_t : D_{t+1}(i) = \begin{cases} \frac{D_t(i)}{Z_t} \times \beta_t, if \ \ h_t(x_i) = y_i \\ \frac{D_t(i)}{Z_t} \times 1 \ , otherwise \end{cases}$, in here Z_t is a constant used to standardize, which makes D_{t+1} be a distribution.

Output: final hypothesis: $h_{fin}(x) = \arg \ \max\limits_{y \in Y} \sum\limits_{t:h_t(x)=\mathrm{y}} \log \frac{1}{\beta_t}$

2.4 The Enhanced Adaboost-SVM Stock-Picking Model Based on Adaboost

Firstly pick up 12-month stock data ,set initiative weight $D_1(1) = \frac{1}{N}$, N as the number of sample , consider selecting sample t monthly , then train SVM classifier h_t in the basis of sample t and classify all samples by h_t . Set the errors in classification as:

$$\varepsilon_i = \frac{1}{N}(\sum_j w_j \delta(\mathrm{C}_i(x_j) \neq \mathrm{y}_j))$$

Then renew the weight of all samples:

$$D_{t+1}(i) = \begin{cases} \frac{D_t(i)}{Z_t} \times e^{-a_t}, if \ \ h_t(x_i) = y_i \\ \frac{D_t(i)}{Z_t} \times e^{a_t} \ \ , otherwise \end{cases}$$

Here, $a_i = \frac{1}{2} \ln \frac{1-\varepsilon_i}{\varepsilon_i}$

Put $T = 12$ into the number of training layers, next the final $Adaboost - SVM$ sorting algorithms can be represented as:

$$H(x) = \sum_{t=1}^{12} a_t h_t(x)$$

Utilize the newer sample data into $H(x)$ and obtain new scores of each stock, then divide them into ten classes and observe the top and last classes.

3. Empirical Analysis

3.1 Test Based on SVM Original Stock Picking Strategies Model

In the case of taking no account of non-linear analysis, the consequences of rolling backtest of sample data in 12 months reveal better classified effects. apparently strong portfolio outstrips weak portfolio.

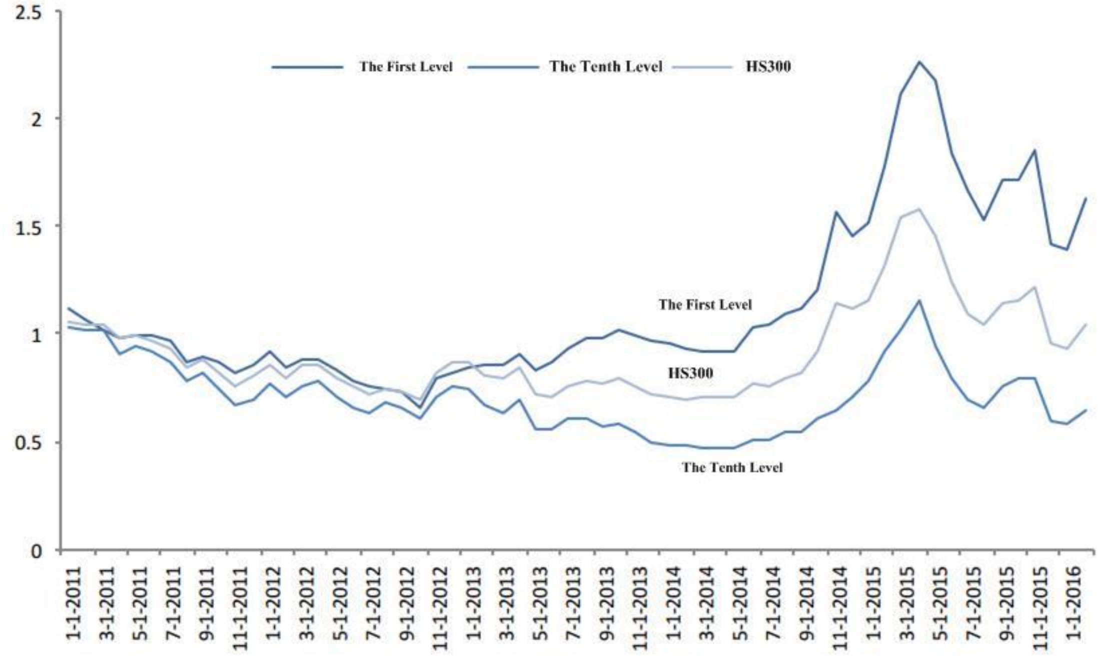

Figure 3. Net value of SVM algorithm stock picking

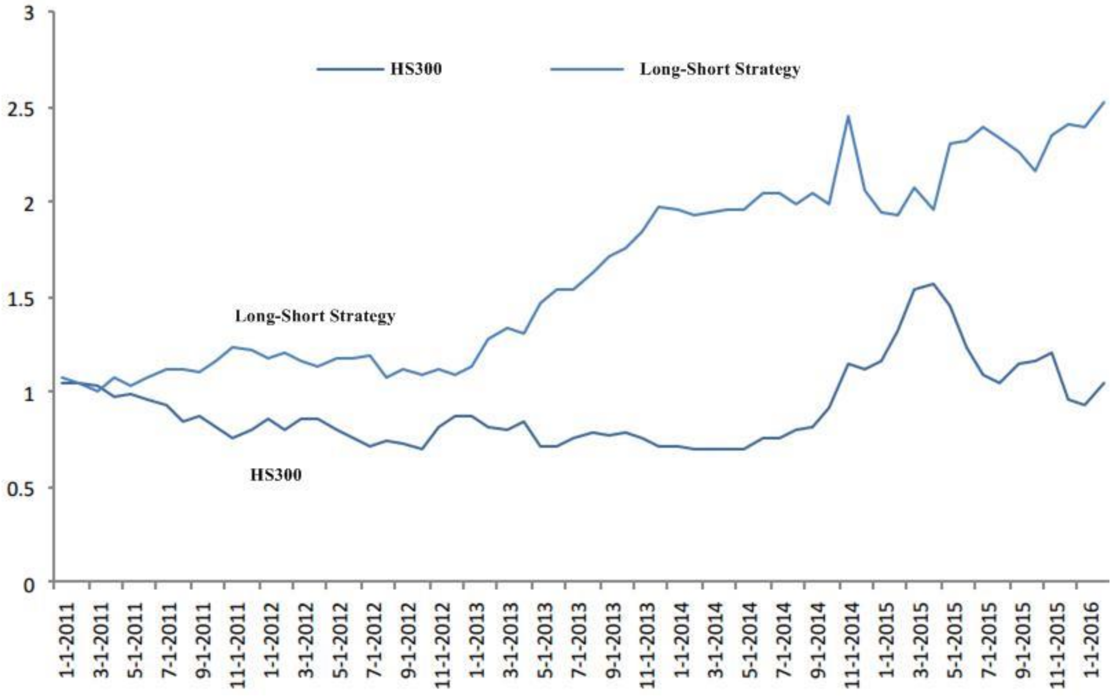

Figure 4. Net value of long-short strategy

Make contrasts in net value of long-short portfolio from Adaboost algorithm, traditional regression method, SVM algorithm, as we can see, at last net value of SVM algorithm gets highest while its fluctuation degree is lower than traditional regression method.

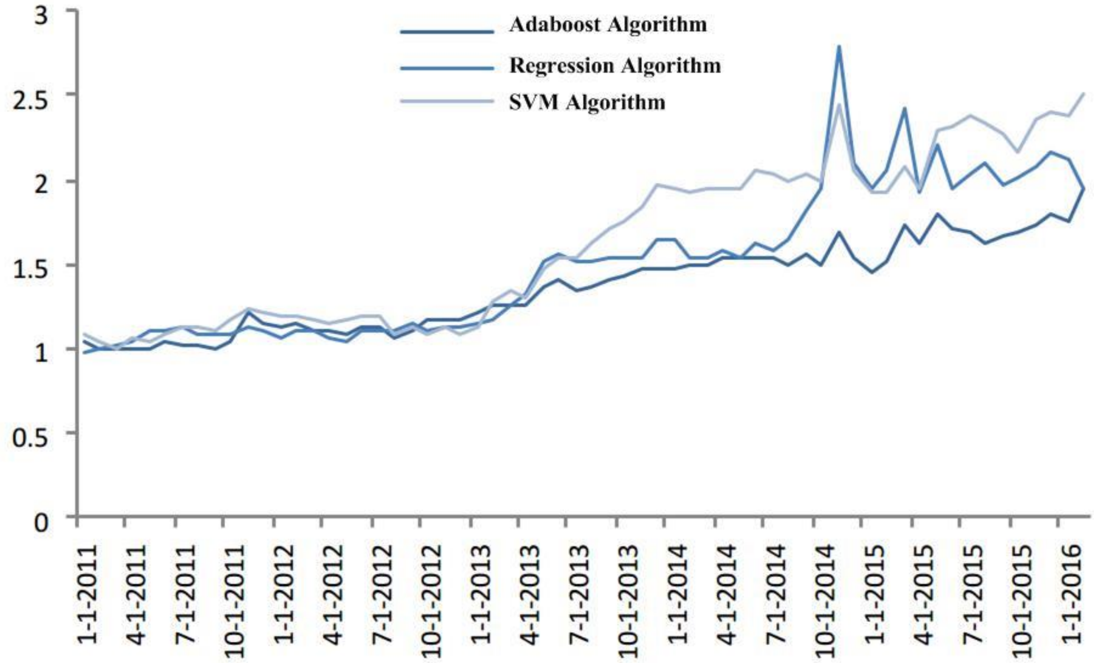

Figure 5. Contrasts of long-short strategy's net value

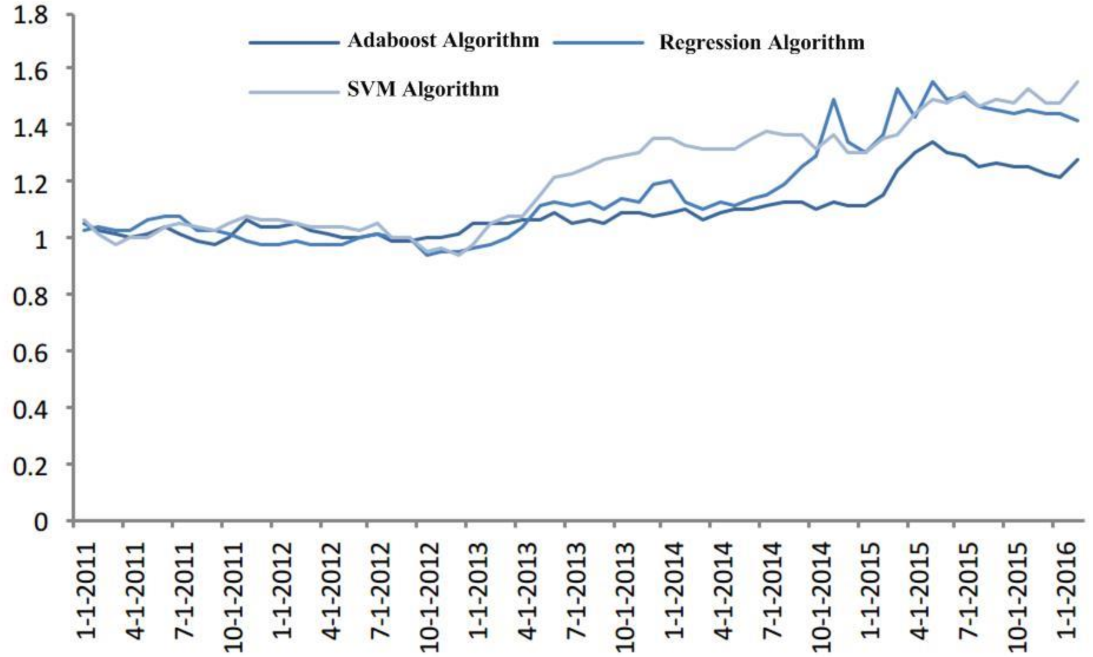

Figure 6. Contrasts of strategy/index's net value

3.2 Testing Of Adaboost-SVM Stock Picking Model Based On Enhanced Adaboost

From the classified outcomes of linear SVM, Adaboost portfolio with 12-layer data is more effective than monthly SVM, and the profits of long-short portfolio can be distinguished easily.

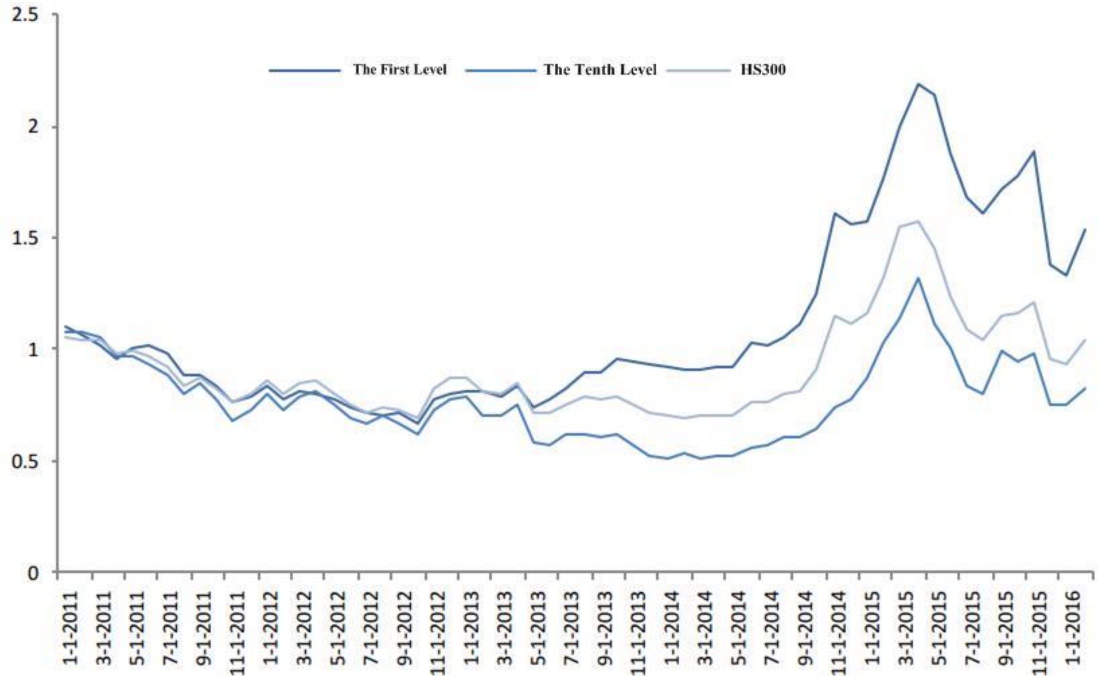

Figure 7. Net value of Adaboost-SVM algorithm stock picking

Besides input 12-month sample data simultaneously by a traditional method, using Aadaboost to enhance it layer by layer will make a big difference on the multi-factorial stock picking problem.

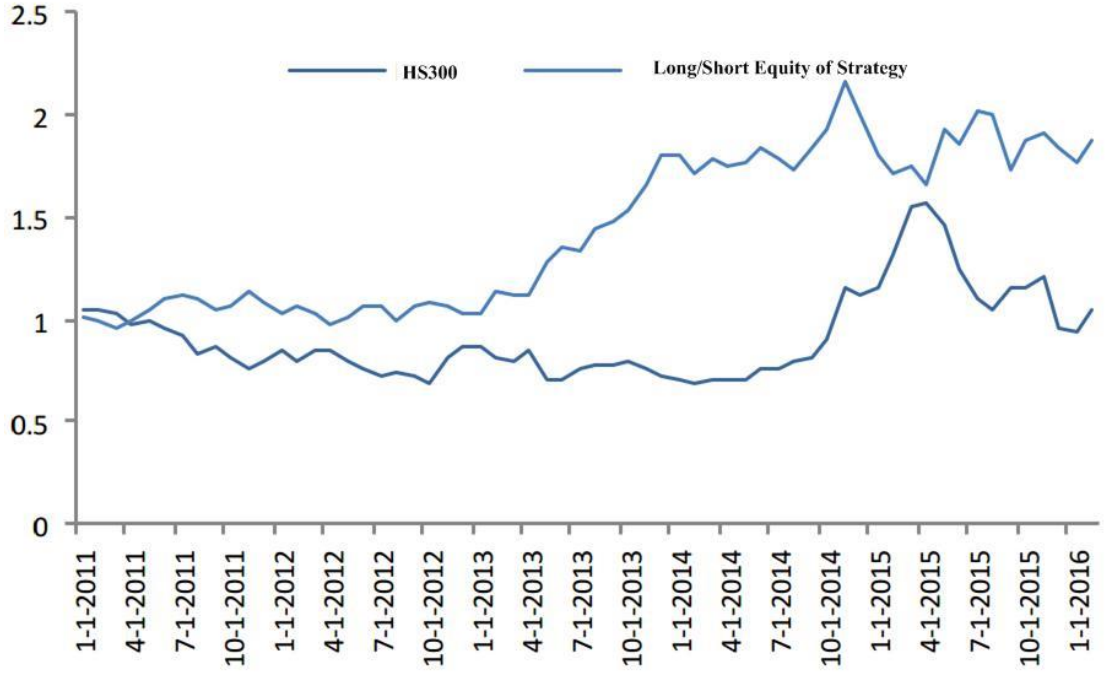

Figure 8. Contrasts of long-short equity of strategy in Adaboost-SVM algorithm

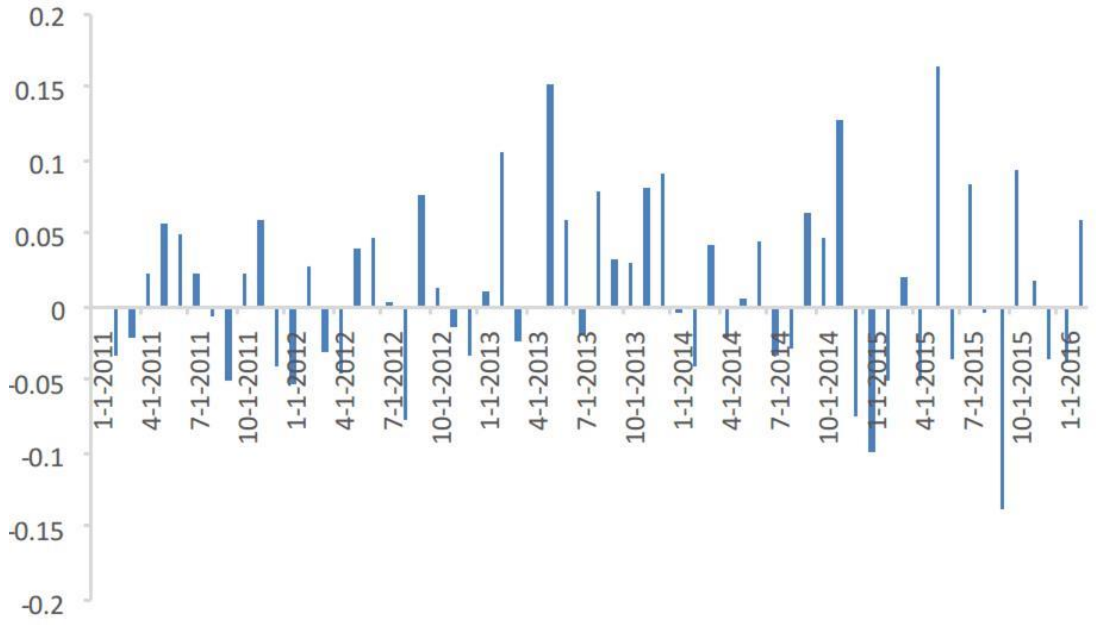

Figure 9. Monthly excess earnings

Compared to the traditional SVM algorithm discussed above, the profit net value of long-short strategy from the enhanced Adaboost-SVM algorithm classification model based on Adaboost increases significantly, which means that it is superior than traditional SVM model holistically.

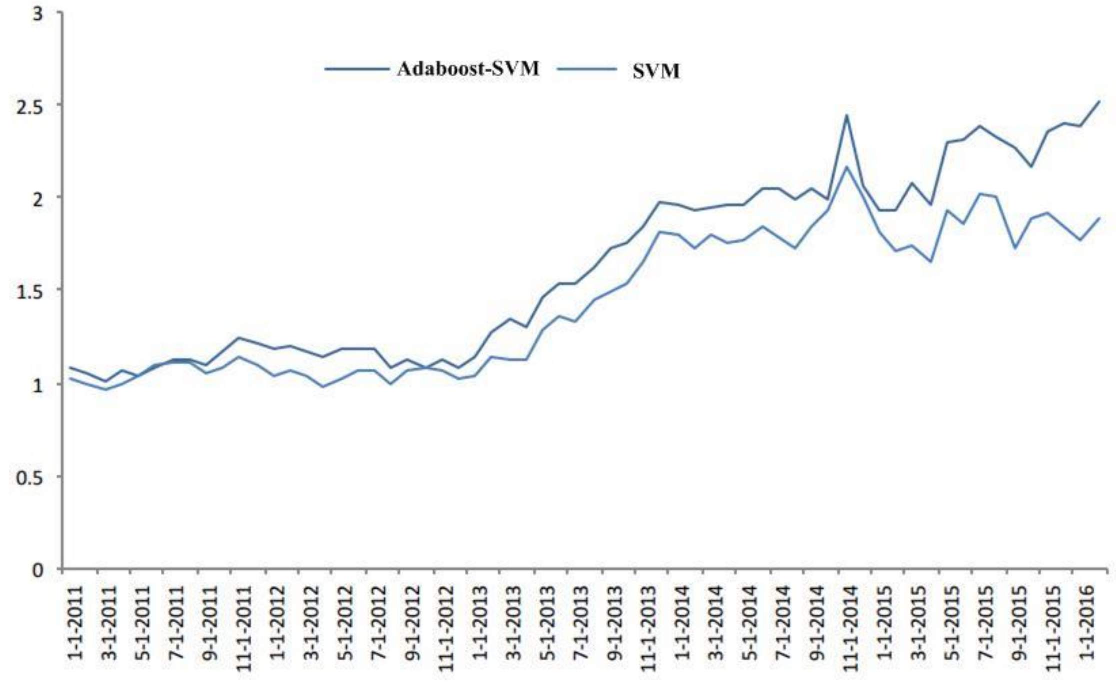

Figure 10. Comparison of net value between Adaboost-SVM algorithm and SVM algorithm

4. Conclusions

This paper makes a new attempt on multifactorial stock-picking strategy—Adaboost-SVM algorithm classification based on enhanced Adaboost, feature each factor in each layer, with the dimensions of 68 factors, Adaboost-SVM algorithm can make classifications and predictions on the tags of portfolio effectively. Compared to the initiative portfolio net value of linear support vector algorithm, Adaboost-SVM algorithm yield higher net value profits.

For the further direction of the model, pure SVM algorithm has better effects on binary classification problems, although from a perspective of practice, it doesn’t achieve better back-test outcomes for multiple or continuous classification variables, however, the idea of Adaboost algorithm is also applicable for multivariate or continuous variable problems. Hence, the portfolio of Adaboost-SVM algorithm deserves further research and test on multivariate problems.

Reference

Zhang Gui-Sheng, & Zhang Xin-Dong. (2016). A SVM-GARCH Model for Stock Price Forecasting Based on Neighborhood Mutual Information[J]. *Chinese Journal of Management Science, 24*(9),11-20.

Huang Qin. (2017). Research on the application of support vector machines in China a-share market quantitative strategy -- based on fama-fench three-factor model [J].*Times Finance*, (11),172-173.

Dong Zhi-qiang, Liu Yong-nian, & Wei Li-hua. (2018). Fault detection of automatic car equipment based on *AdaBoost* and SVM combination [J]. *Electronics World*, (2),17-19.

Copyrights