ATLANTIS PRESS

Advances in Economics, Business and Management Research, volume 75

8th International Conference on Education and Management (ICEM 2018)

Stock Investment Selection Management Based on Bayesian Method

Zhixuan Gao
School of Economics
Shanghai University
Shanghai, China
gaozhixuan_1995@163.com

Abstract—**This paper aims to provide a stock selection management method based on bayesian in order to improve the investment management for investors. Firstly, the financial indicators of Shanghai A-shares were extracted, and those that had a significant impact on the stock increase were selected as the characteristic information of the stock by bayesian model average method. Secondly, the stock was classified into high yield stocks and other stocks by the stock characteristic information using naive Bayesian classification method. Finally, compare the increase of classified high yield stocks with the counterpart of benchmark. The results show that the classified high-yield stock by naive bayesian classification rose higher, indicates that the method provides the investors opportunity for higher returns on the stock investment, which is a meaningful method to improve their investment management.**

Keywords— Stock Investment Selection Management, Bayesian Model Average Method, Bayesian Naive Classification

I. Introduction

It has been 30 years since the Shanghai and Shenzhen stock exchanges have been officially established in 1990, and the 8 listed stocks in 1990 have grown to more than 3400 listed stocks now. By the end of 2017, the total market value of China's A share has reached 56.62 trillion yuan, and there are more than 100 million investors in the Shanghai and Shenzhen stock markets. It can be seen that with the development of the Chinese stock market, stocks will attract more and more investors, and thus become an important part of the daily life of more people.

Many people consider stock trading as a kind of financial management tool like treasury bonds, funds, and Yu Ebao, which have a much higher rate of return than saving accounts, to earn profits and to achieve the purpose of maintaining or even increasing personal wealth. Since China's Shanghai and Shenzhen stock markets have developed thousands of A-shares so far, among which it is even harder to select stocks that can help people preserve or increase their wealth. Judging from the market data in recent years, only half of the stocks‟ price are rising even in the bull market, and only a tiny amount of the stocks can exceed the market trend. Though some investors may catch the right trend, they can still be wrong about stock selection. As a result, investors get no profits and even lose their original investment. Above all, stock selection management is crucial for investors.

Therefore, this paper made some improvement based on the Bayesian classification by the predecessors, and construct a new naive Bayesian classification method to classify stocks.

II. Literature References

The research on stock investment selection management has been analyzed and studied by a large number of experts and scholars. Some scholars use statistical methods, for example: Qiangtai Lan (2017) used Principal Component Analysis to evaluate the financial status of listed companies‟ stocks recommended by the Institute, and made valuation predictions by combining BP neural network models [1]. Some scholars also used correlation analysis to examine the correlation between stock price and net profit or the broader market. The study drew a conclusion that investors should choose listed companies whose stock prices are highly related to net profits (Weihui Hu, 2018) [2]. Stocks selected by Lu Li (2017) using principal component analysis, cluster analysis and other multi-statistical analysis theory system method based on data mining had an overall trend that almost outperform the market [3]. Many scholars also use support vector machine(SVM)and multi-factor quantization methods for stock selection research: Chengxi Zhu (2017) and Xiang Li (2017) aimed at the multi-factor stock selection strategy [4,5]. In recent years, Bayesian has been used more and more frequently in the financial field. Daping Zhao et al. (2015) carried out regression analysis on long and short-term data in different markets based on Bayesian theory, which proves that Bayesian theory has positive effects on optimizing stock portfolios [6]. Kandel et al. (1995) and Winkler et al. (1975) used Bayesian inference to analyze the validity of stock portfolios [7,8]. Hui Zuo and Xinyuan Lou (2008) extracted the general characteristics of stocks recommended by securities analysts, and used Bayesian classifiers to classify stocks into two modes: one that meets the characteristics of recommended stocks, and one that does not [9]. Similarly, Hua Luo and Ximei Zhang (2015) clustered the energy stocks of the Shanghai and Shenzhen stock markets according to the characteristics and performance of US stocks, Tesla, and simplified the financial indicators, then used Bayesian classification to construct twomode classifier, and found that the cumulative return of the equal-weight portfolio of stocks selected by the classification has higher return rate than the benchmark portfolio [10].

III. THEORIES AND METHODS

A. Bayesian Model Averaging Method

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, Letter file.

$\widehat{\beta}_m$ represents the parameter estimation in model m, M represents all possible model spaces. If there are p covariates, then each covariate has two possibilities for included and not included in a model, that is, there exists $M = 2^p$ possible models. Estimated value of $\widehat{\beta}_m$ is calculated by Bayesian model average as follow:

$$\widehat{\beta}_m = \sum_{m \in M} \widehat{\beta}_m \times P(m|\, y) \tag{1}$$

$P(m|\, y)$ is the posterior probability of model m given the data set y. $P\ (m|\, y)$ can be given according to Bayes' theorem as follow,

$$P(m|\, y) = \frac{P(m) \times P(y|m)}{\sum_{m \in M} P(m) \times P(y|m)} \tag{2}$$

$P(m)$ is the prior probability of model m, $P(y|m)$ is the marginal likelihood equation given model m.

$$P(y|m) = \int f(y|\beta_m, m) \times f(\beta_m|m)\, d\beta_m \tag{3}$$

$\beta_m = (\beta^1, \beta^2, \cdots)$ is a parameter vector, $f(y|\beta_m, m)$ is the likelihood function of the sample data given model m and parameter vector β_m, $f(\beta_m|m)$ is prior distribution of β_m.

B. Naive Bayesian Classification

The Bayesian classification is based on Bayesian theory and can be used to predict the probability that a certain sample with certain characteristics belongs to a certain category. The Bayesian theorem can be expressed as follows:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{4}$$

$P(C_i|X)$ is the probability of C_i given X, which is also posterior probability of C_i given X. $P(C_i)$ is the prior probability of C_i; $P(X)$ is the prior probability of X; $P(X|C_i)$ is posterior probability of X given C_i.

The word „naive‟ in the naive Bayesian classification means „simple‟ which comes from the fact that the classification assumes that the effect of a certain attribute on a certain classification is independent from other characteristic attributes. This assumption greatly simplifies the calculation, so this classification is considered as „simple‟ aka „naive‟. The „naive‟ attribute can be expressed as:

$$P(x_1, x_2, \cdots, x_n|C) = \prod_{i=1}^{n} P(x_i|C) \tag{5}$$

The analysis method of Bayesian classification is as follows:

1) Each sample is composed of an n-dimensional feature vector $X = \{x_1, x_2, \cdots, x_n\}$ that represents the observed value of the n attributes. Assuming that the sample can be divided into m different categories, then the prior probability of a certain category can be estimated by $P(C_i) = \frac{S_i}{S}$, where S is the total number of samples, S_i is the number of samples in category C_i.

2) Given categories $C_1, C_2, \cdots C_m$ and a sample X, the Bayesian classification will classify X to the category with the highest a posterior probability given X, that is, the Bayesian classification assigns the sample X of the unknown category to category C_i if and only if

$$P(C_i|X) > P(C_j|X) \qquad 1 \le j \le m, j \ne i \tag{6}$$

When $P(C_i|X)$ is the largest among all categories, C_i is the maximum posterior assumption.

3) For

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{7}$$

And $P(X)$ is constant, we only need to maximize $P(X|C_i)P(C_i)$. If the prior probabilities of each category is unknown, it is usually assumed equal, that is $P(C_1) = P(C_2) = \cdots = P(C_m)$. Then we only need to maximize $P(X|C_i)$. That is, sample X is classified into category C_i, if and only if:

$$P(X|C_i) > P(X|C_j) \qquad 1 \le j \le m, j \ne i \tag{8}$$

Theoretically, the Bayesian classification has the lowest error rate, but in practice, the inaccuracy of assumption leads to an increase in the error rate of the Bayesian classification due to our agreement on „naive"' or „simple"'. Compared with decision trees or neural network classification algorithms, however, that Bayesian classification is not always at disadvantage as various empirical studies have shown.

IV. Models and Data

A. Model Building

Firstly, getting data ready, using the financial indicators of the stock as the covariates, and stock rate of return as the Explained Variable. Select covariates that are robustly correlated with stock returns based on Bayesian model averaging linear regression as stock feature attributes values in the Bayesian classification model. Secondly, Classify the stocks into two types, high-yield stocks and common stocks according to the stock's rate of return performance. N-dimensional vector $X_i = \{x_1, x_2, \cdots, x_n\}$ is the financial indicator of the stock i. y_i is a binary dummy variable for return rate of stock, $y_i = 1$ if stock i is a high-yield stock, if not, $y_i = 0$. The Bayesian classifier above can be viewed as projecting a series of attributes values into a classification dummy variable $\{0,1\}$.The last step is tp test the classification effect of the Bayesian stock classifier.

B. Data

This paper used 10 financial indicators of 1418 SSE A-share stocks on 2017 and the first quarter of 2018 including amount of increase and earnings per share(EPS), net assets per share, capital reserve per share, Rate of Return on Common Stockholders"' Equity (ROE), Operating Profit Ratio, return on assets(ROA), Debt Asset ratio, total asset turnover, Net Assets Growth Rate and return on invested capital(ROIC) for Bayesian model averaging. Financial indicators that are robustly related to stocks return rate of the first quarter of 2017 and amount of increase were selected as training set, and those of the last three quarters of 2017 and the first quarter of 2018 were a test set for the classification experiment. All data analysis processes in this paper were completed by R 3.5.0.

V. EMPIRICAL ANALYSIS

A. Bayesian Model Averaging

TABLE I is the average result of the Bayesian model of stock price increase, and each set of estimations is the average of the $2^{10} = 1024$ models, including the posterior inclusion probability (PIP) of each covariate, the mean estimated coefficient (Post.Mean)], the coefficient posterior standard deviation (Post.SD), and the posterior inclusion probability of positive coefficient under the condition of the variable inclusion (Cond.Pos.Sign). The papers of predecessors on the Bayesian model average usually believe that if a posterior inclusion probability of a variable is greater than 0.9, then the variable is a robust predictor of the outcome, that is, this variable is robustly related to the explanatory variable; otherwise when less than 0.1, it is recommended to excludethe variable. Post.Mean reflects, as the name suggests, the average coefficient level of a certain covariate to the Explained Variable. Post.SD shows the degree of fluctuation of the coefficient, which can also reflect the sign of coefficients to some extent. Cond.Pos.Sign is the probability that a coefficient is positive when it is included in the model.

TABLE I BAYESIAN MODEL AVERAGE RESULTS OF STOCK INCREASE (FIRST QUARTER IN 2017)

	PIP	**Post.Mean**	**Post.SD**	**Cond.Pos.Sign**
EPS	0.3313	-0.0535	0.0866	0.0044
net assets per share	0.9879	0.1716	0.0879	1.0000
capital reserve per share	0.3677	-0.0389	0.0580	0.0011
ROE	0.2184	-0.0153	0.0330	0.0000
Operating Profit Ratio	0.0339	0.0005	0.0062	1.0000
ROA	0.9951	0.1948	0.0519	1.0000
Debt Asset ratio	1.0000	0.2181	0.0320	1.0000
total asset turnover,	0.0791	-0.0035	0.0148	0.0000
Net Assets Growth Rate	0.0332	-0.0005	0.0060	0.0000
ROIC	0.0377	-0.0001	0.0157	0.2649

For example, the average posterior coefficient of net assets per share for the first quarter of 2017 is 0.1716, and its posterior standard deviation is 0.00899, then its coefficient is greater than zero anyway; likewise, its Cond.Pos.Sign Equal to 1 also proves that fact.

The table contains the results of the model average from the first quarter of 2017 (the results of last 3 quarters of 2017, first quarter of 2018 and the combined five quarters are not listed). Since the main purpose of our Bayesian model average is to find covariates that are robustly related to the increase in stocks, we mainly observe the value of PIP.

The results for the first quarter of 2017 show that only the PIP of net assets per share, ROA, and Debt Asset ratio is greater than 0.9, which means that these three variables are robust predictors of stock increase and the coefficient is positive; the data recommended that four variables including Operating Profit Ratio, total asset turnover, Net Assets Growth Rate ROIC should be excluded. Similarly, it is recommended to exclude six variables in the second quarter of 2017, and EPS, ROA and Debt Asset ratio are robust predictors; data in the third quarter keeps net assets value per share, ROA and Debt Asset ratio as robust predictors, and recommend to exclude five variables; in the fourth quarter, only one of the predictors is recommended to be robust and exclude five variables. In the first quarter of 2018 only keep one variable and recommend to exclude a large number of variables. Data of the five quarters in all recommend to retain 4 variables and exclude five variables.

Fig. 1 is the cumulative model inclusion probability of the optimal 500 models (1024 models in total). The dark(blue) parts of the figure indicates positively related, the light(orange) parts indicates negatively related, and the white color indicates irrelevance (the coefficient is zero, that is, the model does not contain the variable). For example, the optimal model for the first quarter of 2017 contains three covariates, the posterior probability of the model is 33%, and the three covariates are positively correlated with the stock increase. Similarly, the second quarter includes three variables at the probability of 67% and in the third quarter was 49%; the posterior probability of the fourth quarter‟s optimal model including two variables is 31%, and that of the first quarter in 2018 and combined sample is 36% (1 variable) and 68% (4 variables).

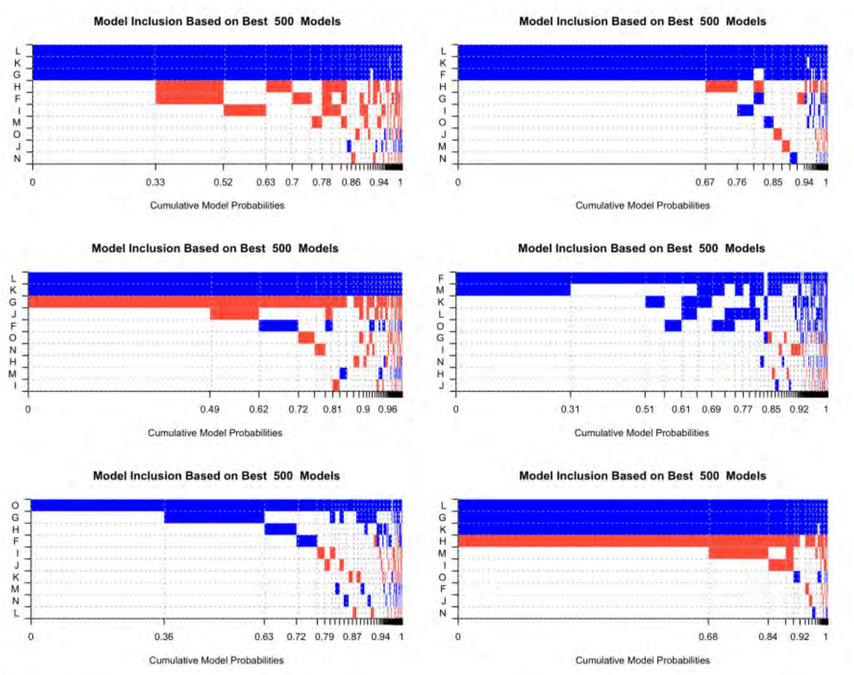

Fig. 1 Cumulative model contains probability

Fig. 2 shows the differences in size between posterior distribution and prior distribution, in which a solid line represents posterior distribution, the dotted line represents a prior distribution. The prior distribution of model averaging in our paper is set to be a uniform distributed, so the model here shows a symmetric distribution with a symmetry axis at 10/2=5 (10 variables in total). With the increases of sample data, the posterior probability updated constantly. At the same time, Fig. 2 also shows the average size of the model. The five quarters in total and the first quarter of 2017 have the largest average model size, and the average size in the first quarter of 2018 is the smallest. Although the average model size of the combined samples for five quarters and the first quarter of 2017 reached 4.0843 and 4.3219, respectively, their proposed model sizes were 3 and 4, for the fat tails of the posterior distribution of the sample size which shifted the mean value to the right; similarly, the average model size in the first quarter of 2018 is 1.801, but the proposed model size is 2, for the posterior distribution of sample size has a thick tail on the left side which results in a left mean value.

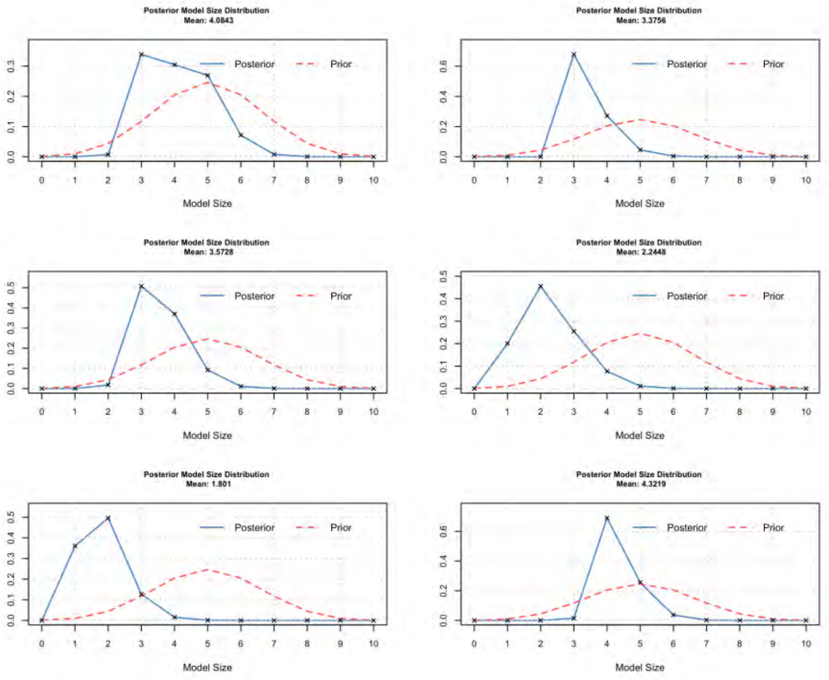

Fig. 2 Model scale posterior distribution

According to the results above, we choose 5 covariates as the stock feature attributes of the following Bayesian classification. The five covariates are: EPS, net assets per share, ROA, Debt Asset ratio and ROIC.

B. Bayesian Classification

We use stock sample data of the first quarter in 2017 as training set. Calculate the increase for the entire quarter by the closing price of the first and the last trading day in the first quarter, and we define the top 50% of the stocks listed from high to low as high-yielding stocks, marked as $y_i = 1$; the rest of the stock are defined as common stocks, marked as $y_i = 0$. Each stock characteristic attribute consists of five covariates robustly related to the stock increase from the result of Bayesian model averaging above. The features of stock i can be listed as $X_i = \{x_1, x_2, x_3, x_4, x_5\}$, x_1 to x_5 represents EPS, net assets per share, ROA, debt Asset ratio and ROIC.

Train the naive Bayes classifier with the training set, then put the data of the rest 5 quarters into the trained classifiers, the classification results are shown in TABLE II. y_i is the original category, and $\widehat{y_i}$ is category classified by trained naive Bayesian classifier. It can be seen from the example that the predictions of these six stocks are quite well, and only one stock ranked in the top 50% according to the increase of stock, which belongs to the high-yield stock category, is classified as one of the common stocks.

TABLE II EXAMPLE OF FORECASTING AND CLASSIFICATION RESULTS

Stock i	EPS	net assets per share	ROA	Debt Asset ratio	ROIC	y_i	$\widehat{y_i}$
1	0.017	2.63	0.6573	10.7835	0.6094	0	0
2	0.43	6.5496	6.8147	13.9442	6.8436	1	1
3	0.065	4.4087	1.2518	25.3495	1.3074	0	0
4	0.36	4.0635	8.5057	16.8734	9.2715	1	1
5	0.011	2.5782	0.4319	18.8409	0.3549	1	0
6	0.24	9.7759	2.3982	16.8153	2.4665	1	1

TABLE III is the aggregated result of Bayesian classification. The total number of stock samples used in each quarter is different, and there is generally an upward trend. There are great differences between classified high-yielding stocks and common stocks. The accuracy of classifier is calculated by comparing actual results based on stock increase with classification results by Bayesian classifiers. The total accuracy is the mean of high-yield stock accuracy and common stock accuracy. We can see from the table that the classification accuracy of high-yield stocks is between 50-65 percent, while the common stock classification accuracy is about 50-60%. The total accuracy is between 50% and 63% approximately.

TABLE III NAIVE BAYES CLASSIFICATION TEST RESULTS

	Q1,2017	Q2,2017	Q3,2017	Q4,2017	Q1,2018
Total stocks	1067	1169	1167	1292	1269
Classified as high yield stock	169	486	680	958	253
Classified as common stock	898	683	487	334	1016
Accuracy of high yield stock [%]	59.7633%	65.4321%	49.7059%	53.4447%	56.9170%
Accuracy of common stock [%]	51.8931%	61.0542%	49.6920%	59.8802%	51.9685%
Total accuracy [%]	55.8282%	63.2431%	49.6989%	56.6625%	54.4428%
increase in high yield stocks [%]	3.4486%	-5.0918%	5.2978%	-7.0430%	-2.2316%
increase in benchmark stocks [%]	0.4418%	-9.4486%	4.7807%	-8.4485%	-4.8577%

The last two lines of TABLE III shows the increase of high-yield stocks and the benchmark stocks. The price of high-yield stocks selected by the classification are always higher than the benchmark – when the benchmark stocks rise, and they fall less when the benchmark stocks fall. This conclusion can also be intuitively derived from Fig. 3.

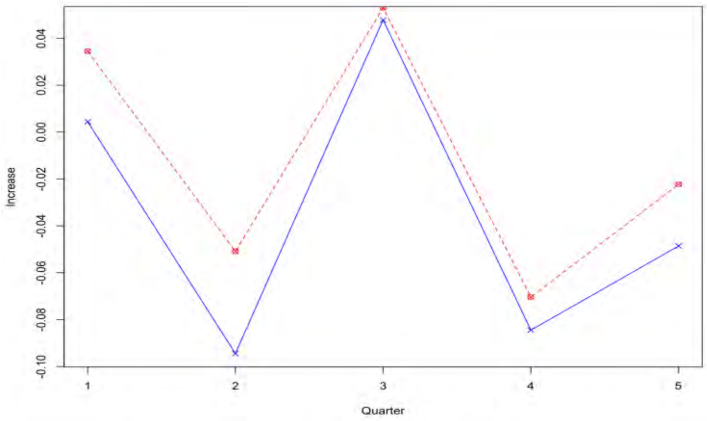

Fig. 3 High-yield stocks and benchmark stocks

In Fig. 3, the dotted line represents high-yield stocks, the solid line represents the benchmark stock gains, we can clearly see that high-yield stocks rose significantly higher than the benchmark stocks in the first and second quarter of 2017 and the first quarter of 2018, the difference in third or fourth quarters of 2017 is small, but it is still the selected high-yield stocks that is in advantage.

VI. Conclusions

In this paper, the Bayesian model averaging method is used to select the covariates that are significantly related to stock increase as the stock characteristic attribute for naive Bayesian classification. We can see it clearly that the naive Bayesian classification is effective for stock investment selection. Our results show that the SSE A shares, the cumulative increase in high-yield stocks of the five quarters is higher than the benchmark stocks by 11.9123%, which shows that the classification of stocks using the naive Bayesian classification method has certain practical significance for investors' stock investment selection management.

References

[1] Q.T. Lan, “An Empirical Study on Comprehensive Stock Selection Based on Principal Component Analysis and BP Neural Network,” Jinan University, China 2017.

[2] W.H. Hu, “Industry and individual stock selection methods for value investment: correlation analysis,” Chinese Township Enterprise Accounting, (2018), pp.26.

[3] L. Li, “Research on Quantitative Stock Selection Strategy Based on Data Mining,” Tianjin University of Commerce, China 2017.

[4] C.X. Zhu, “An Empirical Analysis of Multi-factor Quantitative Stock Selection Model in China‟s A-share Market,” Capital University of Economics and Business, Beijing, China 2017.

[5] X. Li, “Muti-factor Quantitative Stock Option Planning Based on XGBoost Algorithm,” Shanghai Normal University, China 2017.

[6] D.P. Zhao, C.L. Zhang and Y. Fang, “Portfolio Selection of Unequal Histories od Returns with Bayesian Tramework,” Chinese Journal of Management Science, Vol. 23(2015) No.11, pp.504-509.

[7] S. Kandel, R. McCulloch, R.F. Stambaugh, “Bayesian Inference and Portfolio Efficiency,” The Review of Financial Studied Spring, Vol.8(1995) No.1, pp.1-53.

[8] R.L. Winkler, C.B. Barry, “Bayesian Model for Portfolio Selection and Revision,” the Journal of Finance, (1975) No.1, pp.179-192.

[9] H. Zuo, X.Y. Lou, “Stock Selection Using Naive Bayesian Classifier,” Computer engineering application technology, Vol. 20 (2008) No.10, pp.173.

[10] H. Luo, X.M. Zhang, “Research on Stock Selection Model Based on Bayesian Classifier ,” Journal of Zhejiang University Sci-Tech University (Natural Science), Vol.33 (2015) No.3, p.418.

[11] Y. Luo, H.T. Pan, Statistical Institute, Vol.24 (2011) No.4, p.272.

[12] J. Traczynsk, “Firm Default Prediction: A Bayesian Model-Averaging Approach,” Journal of Financial and Quantitative Analysis, Vol. 52, No. 3, June 2017:1211-1245.