7th International Conference on Modelling, Identification and Control (ICMIC 2015)
Sousse, Tunisia - December 18-20, 2015

Stock Selection Model Based on Advanced AdaBoost Algorithm

Sun Yutong[#1] Hanqing Zhao[#2]

[#]*International School, Beijing University of Posts and Telecommunications, Beijing, China*

[1]sunyutong@bupt.edu.cn [2]zhanqin@bupt.edu.cn

***Abstract*—Stock market is a complex non-linear dynamic system which is affected by many factors. Traditional analysis and forecasting methods are insufficient to accurately reveal the inherent pattern of the stock market, resulting in a big difference between expected and observed results. In recent years, machine learning analytical methods are applied to the stock selection model more often than before and have achieved good results so far.**

This paper introduces the application of machine learning in stock selection and conducts detailed research on AdaBoost algorithm. The aim is to establish a multi-factor stock selection model based on AdaBoost algorithm, by which we select stock through the analysis of various indicators of a stock. Furthermore, being aware of the flaws of the basic AdaBoost model, we optimize the stock selection model based on actual characteristics of stock selecting process. We also conduct an empirical analysis on Shanghai A-share Stock excluding the ones which were listed before 2010 and which are suspended. We compare the practicality and accuracy of the basic AdaBoost model and the advanced one. Based on experiment results, the advanced AdaBoost model outperforms the basic one by a substantial margin.

***Keywords*—Machine learning, AdaBoost, Stock selection factors**

Ⅰ.INTRODUCTION

The stock market is an evolutionary, complex, nonlinear dynamic system and there has been a lot of research done on this topic. Traditional methods including fundamental analysis and technical analysis are insufficient to enable better investment decisions because of huge amount of data and numerous impact factors in stock market. In recent years, artificial intelligence and machine learning are also gaining ground in the research of financial field.

AdaBoost, short for "Adaptive Boosting", is a mature machine learning meta-algorithm. It enjoys a high level of classification accuracy. The output of the learning algorithms ("weak learners") is combined into a weighted sum that represents the final output of the "strong learner" which can classify data. These weak learners are easy to calculate and to classify without the problem of over-fitting. Currently, there are not many attempts to apply AdaBoost to select high-quality stocks, especially when it comes to the A-share market.

Selecting good-quality stock is a vital part of stock investment. Good-quality stocks refer to these ones which have remarkable growth potential and high ROI (return on investment). Technically, stock selection is all about analyzing the influence that various factors exert on the future expected return of a stock. Expected return of a stock can be divided into several categories based on stock price growth rate over a certain period of time. Thus, our model is, in essence, a model of binary classification in multi-dimensional space.

Our model is a classification model composed of two elements. We divide stocks into two groups, good and bad, according to the rate of return over a certain period. The first half is marked as 1 while the second half is marked as -1. Thus, we can illustrate the problem by using the following formula,

$$f(X) = y, y \in \{1, -1\} \qquad (1)$$

That is, we have to find an expression, which enables a vector X that describes an impact factor of the stock to the right category Y. Y can only take two values, 1 or -1. The value 1 means good performance while -1 means the opposite.

AdaBoost Algorithm is an effective binary classification algorithm, whose input is the eigenvector of each sample and output is its degree of assurance. Higher degree of assurance means larger possibility that this sample belongs to that category. We predict the stock price growth rate in 2013 based on the data from 2010 to 2012 and then we compare the observed result with the predicted result. We put special focus on the stocks we failed to predict accurately and adjust the weight of impact factors accordingly in order to increase the accuracy of the prediction.

In chapter 2, we will introduce the theory of AdaBoost algorithm and try to improve it. In chapter 3, we utilize the data of A-shares in the past few years to test the effectiveness of our model.

Ⅱ. ADVANCED ADABOOST

AdaBoost algorithm is composed of two parts, i.e. training different weak learners based on training sample and

combining these weak learners into a stronger classifier. In the first step, by changing the probability distribution of samples in the dataset, AdaBoost algorithm is used to increase the weight of these samples which are misclassified in the last iteration so that these particular data become prominent. However, on condition that the cost when data are misclassified is distinct of the two categories, the weight has to be different. For instance, when we select stocks, the cost when we misclassify good stocks as bad ones is different from that when we mistakenly attribute bad stocks to good ones. Usually, people only buy in the stocks with high return, thus they can tolerate taking good stocks as bad ones but they don't want to be cheated by fake good ones, because misclassifying a bad stock will reduce the overall yield rate. Thus, these two kinds of misclassified samples should be treated differently.

Based on understanding of the essence of AdaBoost algorithm, we introduce losses of incorrect classification and adjust the weight of samples according to misclassification coefficient to create Advanced AdaBoost algorithm.

A. Model Construction of Advanced Adaboost algorithm

Here are specific steps of Advanced AdaBoost algorithm:

Suppose that the input is $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where $x_i$represents the xith sample and y_i represents the value of the category of sample, $y_i \in \{-1,1\}$.

1) Initialize weight of the sample w_i: suppose that training dataset has a uniform distribution of weight, that is $w_i = 1/n$。
2) For every weak learner $y_m(\mathrm{m} = 1, \dots, \mathrm{M})$，we use the training dataset with current distribution of weight to train weak learners and its function of minimum weight error function is ：

$$\varepsilon_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) \neq t_n) \qquad (2)$$

Where I (·) is an indicator function. t_n represents the category of sample, when $y_m(x_n) \neq t_n$, $I(y_m(x_n) \neq t_n) = 1$, otherwise $I(y_m(x_n) \neq t_n) = 0$.

Calculate the weight of the weak learner α， which demonstrates how important this weak learner is in the boost classifier.

$$\alpha_m = \log\frac{1-\varepsilon_m}{\varepsilon_m} \qquad (3)$$

After generating a series of weak learners $y_m(x)$， m=1, 2, …, M with the value of 1 or -1, we combine them into a weighted sum $\mathrm{Y_M(x)} = \mathrm{sign(f(x))}$, $f(x) = \sum_{m=1}^{M} \alpha_m y_m(x)$ and α_m is calculated from (3)(4).

After rounds of iterative calculation, the new weight distribution of the training dataset needs to be updated:

Where $Z_m = \sum_{i=1}^{N} w_{m,i} \exp(-\alpha_m t_i y_m(x_i))$.

From the above formula, we can see that since $\exp(-\alpha_m t_i y_m(x_i))$ is an exponential function， when $t_i y_m(x_i) < 0$，which means the weak learner misclassifies the sample， $\exp(-\alpha_m t_i y_m(x_i)) > 1$, thus misclassified sample will have a larger weight and vice versa.

When it comes to the sensitive issue about misclassification cost, we have to highlight the weight of a misclassified sample. More specifically in selecting stocks, we should enhance the weight of the sample which is misclassified to -1 and, in the meantime, abate the weight of the sample which is misclassified to 1.

Because the updated distribution of the weight is a monotonically increasing exponential function, we can introduce a misclassification cost coefficient C_i to calculate the weight in different conditions, such as when cost is low or high, when misclassification happens in last iteration, or when misclassification does not happen.

Here, we calculate classification error rates, ${\varepsilon_m}^{(1)}$ and ${\varepsilon_m}^{(2)}$ respectively, of weak learner y_m when its output is 1 or -1：

$${\varepsilon_m}^{(1)} = \sum_{i:y(i)=1} w_i^{(m)} I(y_m(x_i) \neq 1) \qquad (4)$$

$${\varepsilon_m}^{(2)} = \sum_{i:y(i)=-1} w_i^{(m)} I(y_m(x_i) \neq -1) \qquad (5)$$

When updating the weight distribution，take LOC (or C_i) into consideration：

$$w_{m+1,i} = \frac{w_{m,i}}{Z_m} \exp\left(-C_i \alpha_m t_i y_m(x_i)\right), \ \mathrm{i} = 1,2, \dots, \mathrm{N} \qquad (6)$$

Where,

$$C_i = \begin{cases} {\varepsilon_m}^{(1)} \log\dfrac{1-{\varepsilon_m}^{(2)}}{{\varepsilon_m}^{(1)}}, \text{if } t_i = -1, \ y_m(x_i) = 1 \\ \dfrac{1}{{\varepsilon_m}^{(2)}} \log\dfrac{1-{\varepsilon_m}^{(1)}}{{\varepsilon_m}^{(2)}}, \text{if } t_i = 1, \ y_m(x_i) = -1 \\ 1 \quad , \qquad otherwise \end{cases} \qquad (7)$$

When $\log\frac{1-{\varepsilon_m}^{(2)}}{{\varepsilon_m}^{(1)}}$ is larger，the degree of assurance that -1 is misclassified to 1 is larger and the weight is larger too, so it will be more prominent in the next iteration. When classification error rates ${\varepsilon_m}^{(1)}$ is larger, the weight of the part with value of -1 which is misclassified is larger, so that it is more essential in the next iteration; when classification error rates ${\varepsilon_m}^{(2)}$ is larger, the weight of the sample with value 1 which is misclassified is larger, so that it is less essential in the next iteration.

B. Proof of model's appropriateness

From formula (4) (5) (6) (7),

$$\varepsilon_m{}^{(1)}\log\frac{1-\varepsilon_m{}^{(2)}}{\varepsilon_m{}^{(1)}}+\frac{1}{\varepsilon_m{}^{(2)}}\log\frac{1-\varepsilon_m{}^{(1)}}{\varepsilon_m{}^{(2)}}>0 \quad (8)$$

Hence,

$$\log\frac{1-\varepsilon_m{}^{(2)}}{\varepsilon_m{}^{(1)}}+\log\frac{1-\varepsilon_m{}^{(1)}}{\varepsilon_m{}^{(2)}}>0 \quad (9)$$

$$\frac{1-\varepsilon_m{}^{(2)}}{\varepsilon_m{}^{(1)}}+\frac{1-\varepsilon_m{}^{(1)}}{\varepsilon_m{}^{(2)}}>0 \quad (10)$$

Hence，from the definition of priori probability, the number of samples that are misclassified to 1 reduces and the number of samples that are misclassified to -1 rises, thus we have Pr[y=-1]>Pr[y=1].

From formula (8) (9) (10),

$$\frac{Pr[y=-1]}{Pr[y=1]}+\frac{Pr[Y_m(x)=-1|y=-1]}{Pr[Y_m(x)=-1|y=1]}+\frac{Pr[Y_m(x)=1|y=1]}{Pr[Y_m(x)=1|y=-1]}>0 \quad (11)$$

Suppose that all weak learners are mutually independent, then

$$Pr(y=1)\prod_{m=1}^{M}Pr[Y_m(x)|y=1]>Pr(y=-1)\prod_{m=1}^{M}Pr[Y_m(x)|y=-1] \quad (12)$$

$$Pr[Y_1(x),\cdots,Y_M(x)|y=1]\cdot Pr(y=1)>Pr[Y_1(x),\cdots,Y_M(x)|y=-1]\cdot Pr(y=-1) \quad (13)$$

Let both sides be divided by,

$$\frac{Pr[Y_1(x),\cdots,Y_M(x)|y=1]\cdot Pr(y=1)}{Pr[Y_1(x),\cdots,Y_M(x)]}>\frac{Pr[Y_1(x),\cdots,Y_M(x)|y=-1]\cdot Pr(y=-1)}{Pr[Y_1(x),\cdots,Y_M(x)]} \quad (14)$$

Formula (14) can be transformed into:

$$Pr[y=1|Y_1(x),\cdots,Y_M(x)]>Pr[y=-1|Y_1(x),\cdots,Y_M(x)] \quad (15)$$

It can be indicated from Bayesian optimal decision rate that the predicted value inferred from the advanced AdaBoost algorithm conforms to the maximum posteriori probability. That means that the final result is in accordance with our original intention, which is to reduce the probability that a sample is misclassified to 1.

Ⅲ. Empirical Analysis

A. Background

In our empirical research, we need to build a model to predict the performance of a stock next year based on financial data this year. For example, we employ 2013 Annual Report of a listed company to predict their stock price in 2014. In our paper, we choose the financial data of 2010 to 2012 and stocks' performance from 2011 to 2013 as a training dataset while financial data of 2013 and stocks' performance of 2014 are taken as testing dataset.

B. Data preparation

Before we compare basic AdaBoost and advanced AdaBoost algorithm empirically, we have to pre-process these data to improve the accuracy and efficiency of the model.

The data source of this paper is Wind Financial Terminal. The stocks' data is daily from January 1st, 2011 to December 31st, 2014. Financial data is from the annual reports from 2010 to 2013. Considering public listing and the integrity of data, we exclude 2461 stocks which were listed before 2010 and were suspended between 2010 and 2013.

There are 6985 stocks in the training dataset and 2461 stocks in the testing dataset in total.

In our model, the property (impact factor) of a sample comes from the financial indicators that appear in the financial statements of the firm and the criteria of measuring a stock's performance is based on the annualized rate of return. This is how we calculate yield rate:

Yield rate = (closing price in the end of the semester– closing price at the beginning of the semester)/ closing price at the beginning of the semester

We divide stocks into two categories according to the yield rate, high yield or low yield. The specific step is: after filtering the stocks that were suspended and delisted, we mark the stocks whose current yield sits in the top 1/2 of the whole market as high-yield stocks and mark the other half as low-yield stocks.

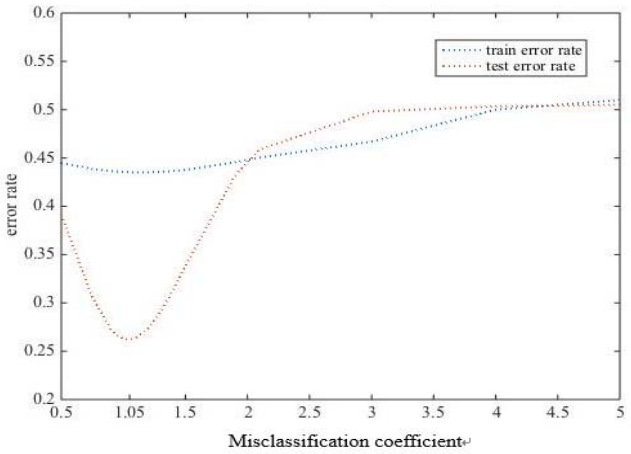

Figure 1Traning set error rate under different misclassification coefficients

According to Bayes' theorem and experiment results, when misclassification coefficient is 1.05, degree of assurance reaches its maximum value. As shown in the picture above,

when weight coefficient is 1.05, error rate of training and testing reach their minimum value. When weight coefficient is larger than 1.05, error rate of training and testing increase monotonically. Hence, we select 1.05 as the most appropriate misclassification coefficient and all the rest data is generated when the misclassification coefficient equals 1.05

C. Factor selection

The selection of factors has a great influence on classification result. Different investment strategies will result in a distinct group of selected factors. Traditional value investment reckons that the price of stock is decided by the intrinsic value, thus analyst should try to find stocks that are underestimated. However, growth investment focuses more on firms whose investment revenue and margin are above the average and have great growth potential. Since our model uses annualized yield rate as the metric for classification, we have to consider both the value and the growth attribute of the stock. To get a comprehensive assessment of individual stock, we choose ten themes (debt-paying ability, growth capacity, Share Index (ASX), earnings quality, cash flow, profit cost, profitability, operating capacity, capital structure and others), 79 factors in total for modeling. They are listed in the following appendix:

D. Data regularization

Because most of the factors oscillate a lot and are greatly influenced by the market, we have to normalize the data. Here is the normalization function:

$$y = \frac{x - x_v}{x_{max} - x_{min}} \qquad (16)$$

Where x is the value before normalization，y is the value after normalization，x_v is the average value of this factor in all samples，x_{max} is the maximum value of this factor in all samples，x_{min} is the minimum value of this factor in all samples.

After normalization, all the data will be mapped to the interval [-1, 1].

In the meantime, since annual reports from listed company do not include all the factors we choose, we will give the average value to the missing factor.

E. Evaluation

In the evaluation system of classification model, we always use accuracy rate and recall rate.

Accuracy rate refers to the percentage of all measured values that the measured values which meet the limited condition account for under certain experiment conditions. That is, accuracy rate = the number of measured values that meet the conditions/total number of measured values.

Recall rate refers to the percentage of all sample values that the measured values which meet the condition accounts for under certain experiment conditions. That is, recall rate = the number of measured values that meet the conditions divided by total number of sample values.

To a specific category, if we use TP to represent the number of samples of this category that are classified correctly, FP to represent the number of samples of the other category that are misclassified, FN to represent the number of samples of this category that are misclassified, TN to represent the number of samples of the other category that are classified correctly, then the formulas of accuracy rate and recall rate will be:

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (17)$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (18)$$

When selecting stocks, we care more about the accuracy rate and recall rate of positive samples which behave well and seldom care about negative samples. Thus we emphasize more on the accuracy rate and recall rate of positive samples to evaluate the model.

Besides, in real investment scenario, people care about whether their return on the investment is above average. For low risk appetite investors, it is vital for them to consider whether retracement can happen. Then, how to evaluate our model also includes the yield rate of groups of stocks that our model selects and the percentage that stocks whose yield rate is below 0 accounts for.

We can calculate evaluating indicator from training dataset as follow: (10)

Table 3-1 Evaluation of testing data

Yield rate of stocks in 2014	33.04%
The number of stocks	2461
The number of stocks whose yield rate is above 0	2003
The number of stocks whose yield rate is below 0	458
The percentage that stocks whose yield rate is above 0 accounts for	81.4%
The percentage that stocks whose yield rate is below 0 accounts for	18.6%

Ⅳ. Result analysis

A. The analysis of the result of AdaBoost algorithm

Using basic AdaBoost algorithm to predict on the testing set, the result is as follow:

Table 4-1 The result of basic AdaBoost algorithm

Label	Accuracy rate	Recall rate
1	54.5%	59.8%
-1	55.4%	50.1%

Label 1 shows that the yield rate of the stock is in the first half while label -1 means that the yield rate of the stock is in the second half. It is rather obvious that the basic AdaBoost algorithm is quite effective to select the right stock. The average yield of all the chosen stocks is 37.71%, 4.67% higher than the overall average yield. There are 1350 samples that are classified to category 1, among which 1131 of them has yield rate above 0 and accounts for 83.8% of all the chosen samples.

From previous descriptions on basic AdaBoost algorithm, we know that the absolute value of the model's output represents degree of assurance. The higher the absolute value, the more confidence AdaBoost has in its decision. We rank the absolute value of the model output and choose two intervals with degree of assurance in the top 1/4 and bottom 1/4.

Table5-2 the result of basic AdaBoost algorithm with different level of confidence

Degree of assurance	Label	Accuracy rate	Recall rate
Top 1/4	1	63.5%	63.0%
	-1	56.8%	56.1%
bottom 1/4	1	45.3%	52.4%
	-1	49.3%	42.3%

From the above table, when we choose the classification result with high degree of assurance, the accuracy rate will be advanced. When label equals to 1, the accuracy rate is 63.5% and average yield is 45.2%, which rises 12.16% over the overall average yield.

B. The analysis of advanced AdaBoost algorithm

Advanced AdaBoost algorithm uses the same training dataset and testing dataset as basic AdaBoost algorithm.

We test out model using testing dataset and following is our results:

Table 5-3 The result of Advanced AdaBoost algorithm

Label	Accuracy rate	Recall rate
1	62.1%	52.7%
-1	52.4%	60.1%

We can tell that the accuracy rate of label = 1 improves from 54.5% to 62.1%. The average yield is 44.3% and 6.6 % higher than that of basic AdaBoost algorithm. There are 1052 samples that are classified into category 1 among which the samples whose yield rate is above 0 accounts for 86.5%.
We use the same method to analyze the samples of basic AdaBoost model and the advanced one.

Table 5-4 The result of Advanced AdaBoost algorithm with different confidence level

Degree of assurance	label	Accuracy rate	Recall rate
TOP 1/4	1	67.4%	55.1%
	-1	54.8%	63.6%
Bottom 1/4	1	47.2%	53.0%
	-1	49.7%	48.9%

It is obvious that when choosing results with high degree of assurance, accuracy rate will be advanced. When label = 1, accuracy rate reaches 67.4% and average yield of selected stock with high degree of assurance is 47.2%.

At the same time, when the number of training dataset reaches a specific number, Advanced AdaBoost algorithm has higher degree of assurance than basic AdaBoost algorithm and the training effect is more stable.

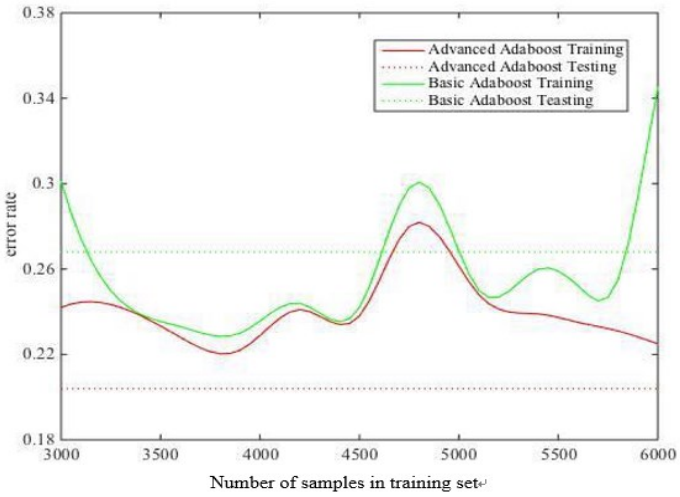

Figure 2 Error rates of Advanced Adaboost and Basic Adaboost under different number of training set

It is easy to see that error rate of testing of AdvancedAdaBoost algorithm is smaller than that of basic AdaBoost. In other words, Advanced AdaBoost algorithm is more convincing in training dataset and improves average yield more efficiently. Also, the curve decreases rapidly when the number of training dataset reaches 4700. It means that error rate of testing will be smaller with more training dataset. However, the curve of basic AdaBoost algorithm does not have the tendency to decrease and oscillate a lot. All this leads up to the conclusion that Advanced AdaBoost algorithm is much better in learning ability than basic AdaBoost algorithm.

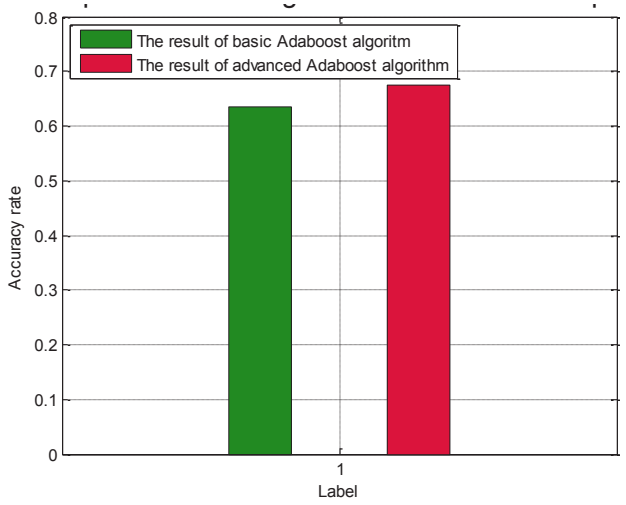

Figure 3 Degree of assurance of top1/4 stock

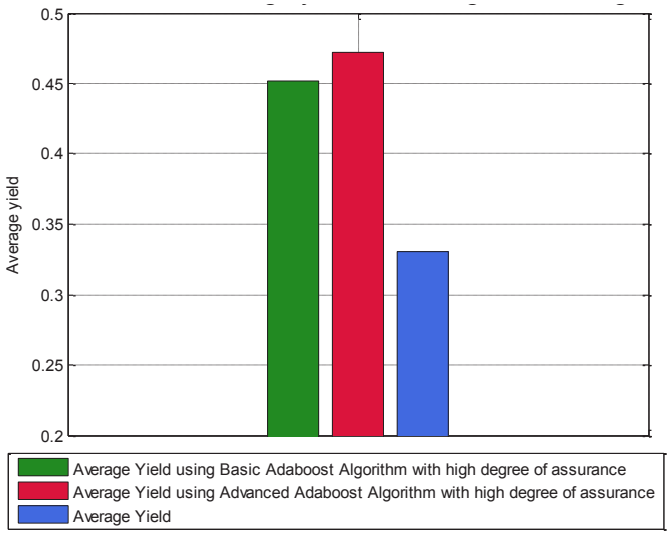

Figure 4 Degree of assurance of different algorithms

As shown in figure 3 and 4, using Advanced AdaBoost algorithm to select stocks whose yield rate is in the top 1/2 or larger than 0 is much more effective than using AdaBoost algorithm. It is also true that the advanced model is more suitable for selecting good stocks.

V. CONCLUSION

In this paper, we raise the idea of an advanced AdaBoost algorithm model and realize it in order to select stocks that outperform peer stocks. Through thorough analysis, the return of stocks selected by this model is higher than the average return in stock market. Such result shows that AdaBoost algorithm enjoys high efficiency and accuracy in selecting stocks. The result of output is more reliable and the return is higher than the average. This model avoids the problem of overriding and long training period. In fact, in most case, we only buy the stock which performs well, so the cost of different type of category's error is different. Hence, in this paper the weight of different error type is different to increase the accuracy and efficiency.

However, we found that the advanced AdaBoost algorithm is not good enough because the accuracy of the testing dataset is not so satisfying, at a rate of 54.5%. Moreover, we found that the influence of factors is notable and the simple weak learners perform badly towards a single factor, which lead to slow decrease of error rate. At the same time, factors calculated in this model are not independent which makes result unstable and boosted classifier perform badly.

Since machine learning is limited to some extent, the advanced model based on machine learning has many shortcomings. In subsequent work, we will utilize analyzing tools such as fundamental analysis and industry analysis to find high-quality stocks. In addition, targeting at the shortcomings of this stock selection model, we will distinguish and classify stocks into different industries, choose proper factor and conduct model training with regard to a specific industry in order to obtain better results and versatility.

Reference

[1] Markowitz, Harry. "Portfolio selection*." *The journal of finance* 7.1 (1952): 77-91.
[2] Shanmin Li, Pei Xu Markowitz portfolio management model application research Economic Science 1 (2000): 42-51
[3] Merton, Robert C. "An intertemporal capital asset pricing model."*Econometrica: Journal of the Econometric Society* (1973): 867-887.
[4] Huberman, Gur. *Arbitrage pricing theory*. No. 216. Staff Report, Federal Reserve Bank of New York, 2005.

[5] Makridakis, Spyros, and Michele Hibon. "ARMA models and the Box–Jenkins methodology." *Journal of Forecasting* 16.3 (1997): 147-163.
[6] Engle, Robert F. "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation." *Econometrica: Journal of the Econometric Society* (1982): 987-1007.
[7] Fama, Eugene F., and Kenneth R. French. "Common risk factors in the returns on stocks and bonds." *Journal of financial economics* 33.1 (1993): 3-56.
[8] Kimoto, Takashi, et al. "Stock market prediction system with modular neural networks." *Neural Networks, 1990. 1990 IJCNN International Joint Conference on*. IEEE, 1990.
[9] Fan, Alan, and Marimuthu Palaniswami. "Stock selection using support vector machines." *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*. Vol. 3. IEEE, 2001.
[10] Chang, Pei-Chann, et al. "An ensemble of neural networks for stock trading decision making." *Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence*. Springer Berlin Heidelberg, 2009. 1-10.
[11] Deyang Liang, Mingfei Niu, Stock index forecasting model based on BP AdaBoost (2013)
[12] Valiant, Leslie G. "A theory of the learnable." *Communications of the ACM*27.11 (1984): 1134-1142.
[13] Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." *ICML*. Vol. 96. 1996.
[14] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).

978-0-9567157-5-3 © 2014 IEEE