Feature Selection for Stock Market Analysis

Yuqinq He, Kamaladdin Fataliyev, and Lipo Wang

School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore

Abstract. The analysis of the financial market always draws a lot of attention from investors and researchers. The trend of stock market is very complex and is influenced by various factors. Therefore to find out the most significant factors to the stock market is very important. Feature Selection is such an algorithm that can remove the redundant and irrelevant factors, and figure out the most significant subset of factors to build the analysis model. This paper analyzes a series of technical indicators used in conventional studies of the stock market and uses various feature selection algorithms, such as principal component analysis, genetic algorithms, and sequential forward search, to find out the most important indicators.

Keywords: Stock Market Analysis, Principal Component Analysis, Genetic Algorithm, Feature Selection.

1 Introduction

Stock market analysis has always been a hot area for researchers and investors. People have come up with a lot of theoretical foundation in mathematics, and developed a variety of methods to analyze the stock market with the help of modern computer technology. Among them, the Feature Selection method is a very important research field. It evaluates a lot of factors that are considered important to the stock market, and selects out the most significant ones for people to depict the market trend. The function of Feature Selection method is to discard the dross and select the essence. The computational time could be dramatically reduced, since the significant factors are pointed out for the investors. In order to figure out the prominent features in the stock market, researches on effective Feature Selection methods are extremely needed.

This paper begins with a literature review of the stock price, analysis of the stock market and feature selection algorithm. Subsequently in chapter 3, technical indicators, principal component analysis, genetic algorithm and sequential forward feature selection are given. Then, in chapter 4 the results are discussed and analyzed. Lastly, in chapter 5, we draw our conclusion and talk about future works.

M. Lee et al. (Eds.): ICONIP 2013, Part II, LNCS 8227, pp. 737–744, 2013.

2 Related Work

2.1 Stock Price

A stock itself has no values, but it can be set a certain price as a commodity for selling and buying, and this price is called stock price [1]. Stock price is also known as the quotation of a stock market, refers to the price of the stock traded in the securities market. The stock price is divided into two categories, theoretical-price and market-price. Theoretical-price of a stock not only provides a significant basis to predict changes in the trend of the stock market-price, but also is a basic factor for the formation of market-price in the stock market. Market-price of a stock refers to the actual price of the stock traded on the stock market. Since the stock market can be categorized into issuing market and circulation market, the market-price also has two types: the issue price and the circulation price [1]. The issue price is the price determined by the issuing company and the securities underwriter, when the stock enters the market at the first time. Therefore, when people talk about the market-price, they normally mean the circulation price of the stock. The market-price contains a lot of details, such as the opening price, closing price, highest price and many other records. The closing price is the most important one among them. It is the basic data that is used in analysis of the stock market [1].

2.2 Analysis about the Stock Market

Investors have to learn to understand a series of theories and scientific methods, in order to analyze all kinds of the information of the stock market. The most famous theory is the Efficient Market Hypothesis Theory (EMH), and the analysis method is generally divided into two types, technical analysis and fundamental analysis [2].

Efficient Market Hypothesis Theory, proposed by E.F. Fama in 1965, is a perfectly competitive market model with an entirely rational basis. It is the cornerstone of the traditional mainstream financial theory [3]. The kernel of this theory is that the stock price always tells all the relevant information accurately, adequately, and in time in an effective market.

The purpose of fundamental analysis is to determine whether the current stock price is reasonable and depict the long-term development space, whereas the technical analysis to predict the short term ups and downs of the trend of the stock price. With fundamental analysis, people can be aware of what stocks they should buy, while technical analysis helps them to detect the timing of specific purchase.

2.3 Feature Selection Algorithm

Feature Selection, also known as the feature subset selection (FSS), or attribute selection (Attribute Selection), is a method to select a feature subset from all the input features to make the constructed model better. In the practical application of machine learning, the quantity of features is normally very large, in which there may exist irrelevant features, or the features may have dependence on each other

Feature Selection can remove irrelevant or redundant features, and thus decrease the number of features to improve the accuracy of the model. The purpose of reducing the running time can also be achieved. On the other hand, selecting the really relevant features can simplify the model, and make the data generation process easy-to-understand for the researchers.

3 Theory and Methodology

3.1 Technical Indicators

Technical indicator is a proper noun in finance [4]. It refers to a collection of stock data calculated by mathematical formulas. This kind of indicators needs to take all the aspects of the market's behavior into consideration, and build a mathematical model, giving out the calculation formula, and then get a number reflecting the intrinsic essence of a certain aspect of the stock market.

There are 12 indicators in total used as the input factors for the stock market. These indicators include SMA (simple moving average), EMA (exponential moving average), ALF (Alexander's filter, which is used to estimate the percentage changes in the prices of financial varieties within a specific period), Relative Strength (it is used to compare the stock price with the whole market in a certain period), RSI (relative strength index), MFI (money flow index, which evaluates the selling and buying pressure with the help of trading price and volume), %B Indicator, Volatility, Volatility Band, CHO (Chaikin Oscillator, which measures the change of the average range of prices in a certain period), MACD (Moving Average Convergence-Divergence), %K Indicator (it focuses on the relationship between the day's high, day's low and the closing prices in the calculation process), Accumulation and distribution (AD) oscillator and Williams %R indicator (analyzes the short-term trend of the market by forecasting the high and low points in a cycle period and picking up the effective signals).

All the 12 indicators are calculated as a line vector. To form the original input feature set, all the 12 vectors are put together into a matrix called *feature*, and after the selecting process of the Feature Selection algorithms, some of them will be chosen to be the optimized feature subset. The elements of the optimized subset are exactly the most significant factors to the stock market.

3.2 Principle Component Analysis

Principle Component Analysis (PCA) is a statistical analysis method to extract the principle contradiction of things, proposed by K. Pearson in 1901 [5]. The essence of this method is to reveal the nature of things by resolving the main factors and simplifying complex problems. PCA is mainly used for dimensionality reduction of data.

The calculation purpose of PCA is to make a projection from the main components of high-dimensional data, onto a lower dimensional space. A multidimensional vector is composed of a series of examples of the characteristics, and some of the elements

have no distinction themselves. The goal is to find elements with huge changes, i.e. the dimensions with high variances, and removing those with little changes, in order to reduce the computation time.

The detailed steps of doing PCA are:

- Calculate the covariance matrix S of the sample matrix X.
- Obtain the eigenvalues (1, 2, …, N) and eigenvectors (e_1, e_2,…, e_N) of the covariance matrix X. Sort the eigenvalues in descending order.
- Project the sample data onto the space formed by the eigenvectors, and get the new sample matrix.

3.3 Genetic Algorithm

Genetic Algorithm (GA) is a kind of random search methods, which is inspired by the laws of evolution in the biosphere (the survival of fittest) [6]. This algorithm has internal implicit-parallelism and better global optimization capability. GA can automatically access and guide the optimized search space, with the probabilistic optimization methods, and there is no need of certain search rules, GA can adjust the search direction self-adaptively. These properties of genetic algorithm have been widely used in combinatorial optimization, machine learning, signal processing, adaptive control and artificial life. It is one of the key technologies in the area of modern intelligent computation. The process of the genetic algorithm includes Initialization of the Population, Individual Evaluation, Selection, Crossover and Mutation steps.

3.4 Sequential Forward Feature Selection

Sequential Forward Selection (SFS) algorithm is one kind of Heuristic searching methods. This method starts with an empty feature subset X, and adds a feature x to make the Criterion function J (X) optimal at each decision step [7]. Simply speaking, it selects a feature that could give the optimal value of the evaluation function each time, and in fact, it is a simple greedy algorithm.

However, there is a disadvantage that it could result in nesting problem, since once a selected feature is added to the subnet X, it would not be discarded any more. For example, if one feature A is completely dependent on the other two features B and C, then it is obvious that A is superfluous since B and C are the dominant factors. Suppose that the Sequential Forward Selection algorithm select feature A first, and then adds B and C, therefore the selected subset will be redundant since it could not remove feature A.

Another sequential Feature Selection algorithm called Sequential Backward Feature Selection (SBS) method has the opposite mechanism. It starts with the full set X = Y, and each time removes one feature so that the evaluation function achieves optimal. Similarly, this algorithm also has the drawback of nesting problem. It cannot add back the removed feature, and needs more computation than SFS. Therefore, the Sequential Forward Selection is used to do the selection in this part.

The selecting process of SFS is similar to a searching process. The algorithm starts with an empty set, and each time chooses one feature to make the evaluation function optimal until all the features are added into the subset or the evaluation function could not be better any more. The detailed procedures of doing the selection are illustrated below:

- Define an empty feature subset S.
- Evaluate each feature in the input feature set and choose one feature that could make the evaluation function have the optimal value.
- Add this feature into the subset S, and choose the second feature out of all the other left features with the same evaluation criterion as the previous step.
- Repeatedly evaluate the left features and add them into S until there is no improvement when adding a new feature or the criterion is met.

4 Discussion and Analysis

All the results obtained from the three algorithms are shown below together with the result got in [8]. This proposed data was evaluated by Sui et al. using Genetic Algorithm with the criterion of measuring classification complexity. From Table 1, it could be concluded that the results of the four programs turned to be good.

The result got from PCA should be relatively "correct", since the mathematics is designed to find out the most efficient dimensions to describe the target. The order of the output has specific meaning, since it is ordered according to each factor's contribution rate. From Table 1, it can be seen that the most significant feature is Volatility, and the least ones are Chaikin Osillator and Williams %R. This result is the same as the one that proposed in [8], it selects 10 features out of 12, excluding Chaikin Osillator and Williams %R as well.

In the meanwhile, the Genetic Algorithm has some randomness, which will influence the result if the iteration times are not enough. The number of features could not be controlled in the first version since there might be crossover and mutation happening at any time. Besides, the result might be different due to the randomness of Genetic Algorithm, however, if the number of generation and the size of population are large enough, the result should go into a steady situation and have optimal solution. Although this program only selects five features to form the output subset, the chosen ones are the first five features in the output of PCA. The accumulative contribution of the first five features is 94% (in Part 3.2.5), which is much higher than the normal criterion 70%. Therefore, the result of GA_1 can be concluded trustworthy and concise.

According to the explanation before, the output of GA_2 only denotes which features are selected by the subset, and there is no difference in the importance. The number of features to be selected to form the feature subset can be controlled with a parameter feaN, and the output with feaN = 10 has the same ten features as that of [8], meaning that this program also works well.

Table 1. Comparison of the results

	Technical Indicator	Proposed Result	PCA	GA_1	GA_2	SFS
1	ALF	(1)ALF	(6)V	(1)ALF	(7)VB	(4)MFI
2	RS	(4)MFI	(5)B	(3)RSI	(3)RSI	(6)V
3	RSI	(10)%K	(1)ALF	(5)%B	(11)ADO	(9)MACD
4	MFI	(5)%B	(3)RSI	(6)V	(5)%B	(7)VB
5	%B	(7)VB	(9)MACD	(9)MACD	(2)RS	(1)ALF
6	V	(6)V	(10)%K		(9)MACD	(8)CHO
7	VB	(3)RSI	(7)VB		(4)MFI	(3)RSI
8	CHO	(9)MACD	(2)RS		(10)%K	(5)%B
9	MACD	(2)RS	(4)MFI		(6)V	(2)RS
10	%K	(11)ADO	(11)ADO		(1)ALF	(10)%K
11	ADO		(8)CHO			(12)%R
12	%R		(12)%R			(11)ADO

The output of SFS has the discarded feature "Chaikin Osillator" on the sixth position, which makes the subset seem not good as those of the others. This is due to the drawback "nesting problem" of SFS which is explained in Part 3.4.1. If the feature has been selected, it could not be removed any more. Therefore, SFS might not be very efficient during actual practice.

Overall, PCA is the most suitable method in this paper, since it is reliable and accurate. However, the computation time might be very long if the input data has too many factors. In such a situation Genetic Algorithm will have a better performance since it takes the advantage of randomness.

5 Conclusion

In this paper, we did researches on the principles and theories in the field of financial market, and basic technical analysis methodologies about the stock market was studied and practiced with the help of Feature Selection algorithms. We used the data of Shanghai Stock Exchange Composite Index (SSECI) from 24/03/1997 to 23/08/2006 to measure twelve technical indicators for later research. The twelve chosen technical indicators were calculated, and the results were taken as the input of the Feature Selection algorithms. The three kinds of Feature Selection algorithms, Principle Component Analysis (PCA), Genetic Algorithm (GA) and Sequential Forward Selection (SFS) were studied. According to the results and analysis, PCA is the most reliable, but might be time-consuming if the input has very large dimensions. Genetic Algorithm will have a better performance since it takes the advantage of randomness in such a situation. SFS could generate the local optimal solution, but with a risk of "nesting problem".

Based on this paper, there are still many improvements that can be made to get better results. Here are some recommendations we summarized for a better work.

- This paper only studies on three kinds of Feature Selection algorithms, so other algorithms [13]-[19] can be researched and compared with these three to provide more convenient and reliable method for building models. Moreover, the study on a hybrid algorithm will also have great potential to come out a better solution.
- Pattern recognition techniques like Artificial Neural Network and Support Vector Machine could be used to do further analysis. By training the learning machine with the selected features, the resultant model will be more reliable and practical.

References

[1] Wise, B.O.: Fundamentals of the stock market. Machinery Industry Press (2012)
[2] Roberts, H.V.: Stock-Market "Patterns" And Financial Analysis: Methodological Suggestions. The Journal of Finance (1959)
[3] Fama, E.F.: Random walks in stock market prices. Financial Analysis Journal 21, 55–59 (1965)
[4] Blume, L., Easley, D., O'hara, M.: Market Statistics and Technical Analysis: The Role of Volume. The Journal of Finance (1994)
[5] Jolliffe, I.T.: Principle Component Analysis. Springer, New York (1986)
[6] Huang, C.L., Wang, C.J.: A GA-based feature selection and parameters optimiza-tion for support vector machines. Expert Systems with Applications 31, 231–240 (2006)
[7] Liu, B., Wang, Y., Lu, L., Li, Y., Cai, Y.: Prediction the protein SUMO modification sites based on Properties Sequential Forward Selection. Biochemical and Biophysical Research Communications (2007)
[8] Sui, X.S., Qi, Z.Y., Yu, D.R., Hu, Q.H., Zhao, H.: A novel feature selection approach using classification complexity for SVM market trend prediction
[9] Wang, L.P.: Wavelet neural networks for stock trading and prediction. In: Invited Oral presentation, SPIE Defense, Security, and Sensing, Bal-timore, USA, 29 April - 3 May (2013)
[10] Gupta, S., Wang, L.P.: Stock Forecasting with Feedforward Neural Networks and Gradual Data Sub-Sampling. Australian Journal of Intelligent Information Processing Systems 11, 14–17 (2010)
[11] Zhu, M., Wang, L.P.: Intelligent trading using support vector regression and multilayer perceptrons optimized with genetic algorithms. In: 2010 International Joint Conference on Neural Networks, IJCNN 2010 (2010)
[12] Gupta, S., Wang, L.P.: Neural networks and wavelet de-noising for stock trading and prediction. In: Pedrycz, W., Chen, S.M. (eds.) Time Series Analysis (2012)
[13] Wang, L.P., Zhou, N., Chu, F.: A general wrapper approach to selection of class-dependent features. IEEE Transactions on Neural Networks 19(7), 1267–1278 (2008)
[14] Zhou, N., Wang, L.P.: Effective selection of informative SNPs and classi-fication on the HapMap genotype data. BMC Bioinformatics 8, 484 (2007)
[15] Zhou, N., Wang, L.P.: A modified T-test feature selection method and its application on the Hapmap genotype. Genomics, Proteomics & Bioinformatics 5, 242–249 (2007)

[16] Wang, L.P., Chu, F., Xie, W.: Accurate cancer classification using expres-sions of very few genes. IEEE-ACM Transactions on Computational Biology and Bioinformatics 4(1), 40–53 (2007)
[17] Liu, B., Wan, C.R., Wang, L.P.: An efficient semi-unsupervised gene selection method via spectral biclustering. IEEE Transactions on Nano-Bioscience 5(2), 110–114 (2006)
[18] Chu, F., Wang, L.P.: Applications of support vector machines to cancer classification with microarray data. International Journal of Neural Systems 15(6), 475–484 (2005)
[19] Fu, X.J., Wang, L.P.: Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance. IEEE Trans. System, Man, Cybern, Part B-Cybernetics 33(3), 399–409 (2003)