Stock Trading Based on Principal Component Analysis and Clustering Analysis

Yanru Guo
Computer Science and Technology Department, Chongqing University, Chongqing, China

Abstract. Predicting stock price becomes a hot problem in the realm of investment for investors, brokers and dealers. Stock selection in active project investment is a challenging and significant task. By using technical analysis, analysts make efforts to select stocks and set transaction rules. In this paper, 9 technical indicators were used to select appropriate stocks. Principal component analysis and cluster analysis were applied to estimate the stock returns in the day of the training time by selecting the earnings in the top 30 stocks. Finally through calculating annual yield, maximum retracement, Sharpe ratio, ratio of information, quantitative evaluation of a strategy was analyzed. The result showed the validation of technical indicators in the analysis of stock trading. This paper provides theoretical support for investors in technical analysis. Investors could make a reasonable judgment on the price trend of the stock market.

1. Introduction
Research of algorithmic trading system in stock market has been widely concerned. Algorithmic trading is an investment method that finds investment opportunities from large-scale data through mathematical and computer models and automatically completes securities trading[1]. Compared with common analysis of technical and fundamental analysis, algorithmic trading has the advantages of objectivity, discipline and accuracy. Algorithmic trading systems frequently and periodically decide when to buy or sell profitable stocks without human judgment[2]. The analysis methods of the stock market mainly include the fundamental analysis method and the technical analysis method. The fundamental analysis is mainly used to the selection of investment objects, while the technical analysis is applied to determine when to operate specific investments, as a beneficial supplement to raise the reliability and effectiveness of securities investment analysis[3].

The index class based on the historical data of price and quantity is one of the methods of technical analysis. It establishes a simple calculation formula and obtains an index value that reflects the intrinsic essence of a certain aspect of the securities market. The second method of technical analysis is model analysis. Both methods are widely used by investors. Hart et al. demonstrated the performance of value, momentum and return correction strategies can be improved by selecting stocks based on multiple characteristics and combining country selection to produce significant excess returns in 2003 [4]. Borodin et al. found out all the predictable statistical relations between stocks in the market and then proposed a new active trading algorithm for stocks[5]. Yan et al. were put forward a ranking method named PR in 2007 to improve stock returns [6]. Lee et al. in 2009 and Ni in 2011 demonstrated hybrid feature selection method for predicting the trend of stock market based on support vector machine (SVM)[7][8]. Huang et al. combined the method of hybrid support vector regression (SVR) and genetic algorithms (GÅ) to predict stock price[9]. Later, Huang et al. developed
an efficient stock selection method, using the fuzzy model and genetic algorithm (GA) [10]. Liu et al. were put forward a novel ANFIS time series model for stock prediction based on integrated nonlinear feature selection method in 2016 [11].

This paper proposed a method for stock selection and price prediction using principal component analysis and cluster analysis on technical indicators including trend indicator and momentum indicator. An experiment was conducted on 10-year historical stock price data of Shanghai and Shenzhen Stock Exchanges. Section II reviews latest works and related methods on HS300 index, technical analysis, principal component analysis, cluster analysis. Section III describes the experiment process in detail. Section IV shows experimental results and Section IV comes to a conclusion.

2. Data

2.1 Introduction of Hs300 Index

As of April 2015, Shanghai and Shenzhen Stock Exchanges have jointly released the HS300 index. HS300 index is a financial indicator, which not only provides basic conditions for index investment and derivative innovation, but also serves as an evaluation standard for investment performance, comprehensively reflecting the operating conditions and overall performance of listed companies in China's a-share market. Constituents of HS300 index is adjusted every six months according to the principle of sample stability and dynamic tracking, and the adjustment proportion is generally no more than 10%

This experiment studied more than 4,000 stocks that were included in the HS300 index from January 2007 to January 2017.

2.2 Roc

Momentum (MTM) is a technical analysis indicator which shows the difference between the closing price of today and n days ago. Momentum is the difference of the former:

\[ \text{momentum} = C_i - C_{i-1}, \]

where \( C_i \) is today's closing price while \( C_{i-1} \) is the closing price n days ago. Momentum can be regarded as a trend-following indicator. So when it peaked and began to fall, investors can consider to sell stocks. Momentum signals have been used in buy and sell recommendations of financial analysts [12].

For momentum, ROC is a simple technical analysis indicator that scales by the closing price n days ago, showing the increase by:

\[ \text{rate of change} = \frac{(C_i - C_{i-1})}{C_i} \]

Increasing ROC signals accelerates price growth. Negative and declining ROC signals accelerates price declines. In general, ROC measures the strength of a trend.

2.3 Rsi

The relative strength index(RSI) is a technical momentum indicator. With measuring the extent to which its recent closing price has changed, RSI can be used to assess whether a stock is overbought or oversold [13]. RSI makes use of the exponential moving average (EMA) so as to show an exponential weighting for the subsets. Among them, the latest data points of time series are more weighted [14].

The EMA is calculated by:

\[ \text{EMA} = \alpha \sum_{i=k}^{T} (1 - \alpha^{(i-k)}) C_i, \]

where \( \alpha \) is a weighting in the range of 0 to 1, \( k \) means the difference of \( T \) and \( n \) while \( n \) is the number of historical data points. A higher value for \( \alpha \) discounts the previous data points faster.

Closing prices in every trading period calculate an upward change \( U \) or downward change \( D \). The feature of up period is that the closing price today is higher than that several days ago:

\[ U = C_i - C_{i-1} \]
\[ D = 0, \]

where \( C_i \) refers to today’s closing price, \( C_{i-1} \) means the closing price \( n \) days ago.

Specifically, the feature of decline period is that \( C_i \) is lower than \( C_{i-1} \) in the past.

\[ D = 0 \]

\[ U = C_{i-1} - C_i \]

The up and down periods are given in the next step with using EMA, and relative strength (RS) is shown by the ratio of these averages:

\[ RS = \frac{EMA(U, n)}{EMA(D, n)} \]

where \( n \) refers to previous data points which are used in the EMA. Then the relative strength index (RSI) can be calculated by:

\[ RSI = 100 \times \frac{100}{1 + RS} \]

A relative strength index whose range is 0 to 100 then converts to a relative strength factor [15]. When RSI is of 70 or above, the stock may be overbought and price will fall. A RSI below 30 can indicate the stock is possible to be oversold and whether the market is bullish or bearish determines whether this is a good indicator [16].

2.4 Macd

Moving average convergence/divergence (MACD), is a trading technical indicator which is designed to judge the strength of the stock price changes, direction, momentum, and trend of energy and cycle, in order to grasp the timing of the stocks to buy and sell [17].

The MACD indicator is a collection of three time series based on price data in the previous period which mainly refers to the closing price. Two of the three series is the signal or average series and the MACD series proper while the last is divergence series which means the difference between the former. The MACD series is calculated by the difference between a slow-moving and a fast-moving EMA of the price series. EMA of the MACD series itself means the average series.

Take the parameter of EMA1 as the 12th and the parameter of EMA2 as the 26th and the parameter of DIF as the 9th as examples to see the calculation process of MACD. The first step is to calculate the moving average:

\[ EMA1 = EMA_{1\text{yesterday}} \times \frac{11}{13} + C_i \times \frac{2}{13} \]

\[ EMA2 = EMA_{2\text{yesterday}} \times \frac{25}{27} + C_i \times \frac{2}{27} \]

where \( EMA_{1\text{yesterday}} \) is the previous day’s EMA. The second step is to derive the differential item function (DIF):

\[ DIF = EMA_{1p} - EMA_{2p} \]

According to the deviation value through the calculation of 9 EMA, namely the average deviation, is the desires of the MACD values. In order not to confused with index formerly, this value is also known as DEA or DEM:

\[ DEA_p = DEA_{p-1} \times \frac{8}{10} + DIF_p \times \frac{2}{10} \]

where \( DEA_p \) is the number of today’s DEA, \( DEA_{p-1} \) is the number of yesterday’s DEA and \( DIF_p \) is today’s DIF. The calculated values of DIF and DEA are both positive or negative. MACD histogram is obtained by:

\[ MACD = (DIF - DEA) \times 2 \]

Therefore, MACD indicator consists of a set of curves and a histogram. The fast-moving line is DIF, while the slow-moving line is DEA. The histogram is MACD.
3. Model

3.1 Principal Component Analysis
Principal component analysis (PCA) transforms multiple indexes into a few comprehensive indexes and USES them to explain variance-covariance structure of multivariate variables [18]. The few principal components obtained should have no connections with others and should retain a large amount of content of the original variables. PCA is a common mathematical transformation method, which transforms a given set of variables into a set of unrelated variables through linear transformation. In this transformation, the total variance of the variable is kept the same, and the first principal component F1 is the maximum variance[19]. Once the information of the original p indicators is not shown enough in the first principal component F1, the second linear combination F2 is then selected which is called the second principal component. The information F1 includes has no need to appear again in F2 for effectively reflecting the original information. In this way, the third, the fourth and even the first P a principal component:

\[ F_p = a_{i1} \times Z_{i1} + a_{i2} \times Z_{i2} + \ldots + a_{ip} \times Z_{ip}, \]

where \( a_{i1}, a_{i2}, \ldots \), the \( a_{pi} \) (\( i = 1, \ldots, m \)) to the covariance matrix of \( X \) eigenvectors corresponding to eigenvalue, \( Z_{i1}, Z_{i2}, \ldots, Z_{ip} \) is the value of the original variable that has been standardized. In practice, there are always different dimensions of indicators, so the influence of dimensions should be eliminated before calculation and the original data should be standardized:

\[ A = (a_{i}) \quad p \times m = (a_{i1}, a_{i2}, \ldots, a_{im}) \]

\[ R_{ai} = \lambda a_{i}, \]

where \( \lambda_{i} \) is the corresponding characteristic value, \( a_{i} \) is the unit feature vector while \( R \) is the correlation coefficient matrix. If there are \( p \) variables in total, it is usually enough to find \( m \) principal components instead of \( p \) principal components in practical application, as long as these \( m \) principal components can reflect the variance of most of the original variables.

The regression variables were screened by principal component analysis. The selection of regression variables has an important practical significance. In order to make the model itself easy to conduct structural analysis, control and prediction, the optimal variable is selected from the sub-set of original variables so as to make the optimal variable set. Through principal component analysis the variables are screened and less computational efforts can be used to select variables and the best variable quantum set can be selected [20].

K-means algorithm has been used to choose stocks in active portfolio management. For example, analyze similar technical charts through the k-means method to predict stock trends[21].

3.2 Clustering
Cluster analysis is a kind of sample that is clustered into a class according to the distance between the samples or the similarity, and the smaller and smaller samples are grouped together, finally a plurality of clusters are formed in order that the samples are similar to each other within the same cluster, and the difference among the clusters is high.

There are many clustering analysis algorithms. We choose the classical k-means clustering analysis algorithm. K-means algorithm is a method to divide data based on the centroid. Given a data set \( D \) and the number of clusters \( k \) to be divided, we can divide the data set into \( k \) clusters by this algorithm. In general, each data item can belong to only one of these clusters. The specific method can be described as follows:

Assuming that the data set is in a mm-dimensional Euclidean space, at the beginning, we can randomly select \( k \) data items as the centroid \( C_i \) of the \( k \) clusters, I in \( \{1,2 \ldots K \} \), each cluster center represents a cluster, that is, a set of data items.

Then, for all \( n \) data items, the distance between these data items and \( C_i \) is calculated (in general, in Euclidean space, the distance between data items is represented by Euclidean distance). For example, for data item \( D_j, j \in \{1 \ldots N \} \), which is closest to one of its cluster center \( C_i \), then \( D_j \) is classified as
cluster $C_i$. With this step, we have initially divided $D$ into $k$ classes. Now recalculate the centroids of these $k$ classes. The method is to calculate the mean of each dimension of all data items in the class. In this way, a new centroid is formed and the centroid of the class is updated. This is done once for each class to update the centroid.

4. Result

The period values are 2, 4 and 6 months for ROC, EMA and MACD. Based on python programming, the principal component analysis was used. After calculation, the contribution rates of the three principal components are 35.1%, 15.6% and 8.8% respectively, and the cumulative contribution rates exceed 75%. The 5 principal components can be selected to comprehensively represent 9 participating situations. Then make a observation on the price factor of the first principal component, the trend factor of the second principal component, and the profit factor of the third principal component.

According to the k-means stock selection strategy constructed by us, when the cluster number is 8, the number of share votes is 6, the period length is 5, the characteristic factor short period is 2, the middle period 4, and the long period 6, the total return rate of the portfolio and the return rate of the stock market (HS300) are shown in the Figure 1

By comparing the above data, we can conclude that the return of our constructed portfolio is significantly better than the market index. In most cases, the return rate of the portfolio is more than twice the return rate of the market index. After the second half of 2015, the return rate of the portfolio has reached 700%, much higher than the market. This shows that our investment strategy is basically successful. The annualized return rate of the optimal investment portfolio reached 15.2%, the Sharpe ratio reached 1.12, the information ratio reached 1.21, and the maximum withdrawal rate was 0.53.

5. Conclusion

In this paper, 9 indicators of monthly frequency of 4000 a-shares were adopted to obtain the principal component factors through comprehensive principal component analysis. The daily returns of stocks during the training period are calculated according to the k-means stock selection strategy. The top 30 gainers of the day will be selected to buy and sell the remaining stocks.

Improvements should be made in the following areas. In addition to clustering used in this study, better artificial intelligence algorithms such as CatBoost and XGBoost are mainly applied to the prediction of stock prices. Some technical indexes, such as RSI, are bought if they exceed a certain
value and set to 1, while others are set to 0. Based on experience, it may be better to try to discretize some values.

References
[1] Hendershott T, Jones C M, Menkveld A J. Does algorithmic trading improve liquidity? [J]. The Journal of Finance, 2011, 66(1): 1-33.
[2] Borkovec M, Heidle H, Sinclair R. Algorithmic trading system and method: U.S. Patent 8,140,416[P]. 2012-3-20.
[3] J. Murphy, Technical analysis of the financial markets, New York Institute of Finance, p 23, 1999.
[4] Van der Hart J, Slagter E, Van Dijk D. Stock selection strategies in emerging markets[J]. Journal of Empirical Finance, 2003, 10(1-2): 105-132.
[5] Borodin A, El-Yaniv R, Gogan V. Can we learn to beat the best stock[C]//Advances in Neural Information Processing Systems. 2004: 345-352.
[6] Yan RJ, Ling C X. Machine learning for stock selection[C]//Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2007: 1038-1042.
[7] Lee M C. Using support vector machine with a hybrid feature selection method to the stock trend prediction[J]. Expert Systems with Applications, 2009, 36(8): 10896-10904.
[8] Ni L P, Ni Z W, Gao Y Z. Stock trend prediction based on fractal feature selection and support vector machine[J]. Expert Systems with Applications, 2011, 38(5): 5569-5576.
[9] Huang C F. A hybrid stock selection model using genetic algorithms and support vector regression[J]. Applied Soft Computing, 2012, 12(2): 807-818.
[10] Huang C F, Chang B R, Cheng D W, et al. Feature Selection and Parameter Optimization of a Fuzzy-based Stock Selection Model Using Genetic Algorithms[J]. International Journal of Fuzzy Systems, 2012, 14(1).
[11] Su C H, Cheng C H. A hybrid fuzzy time series model based on ANFIS and integrated nonlinear feature selection method for forecasting stock[J]. Neurocomputing, 2016, 205: 264-273.
[12] Low R K Y, Tan E. The role of analyst forecasts in the momentum effect[J]. International Review of Financial Analysis, 2016, 48: 67-84.
[13] Jordy Bolton, Sven T.von Boetticher. Momentum Trading on the Johannesburg Stock Exchange after the Global Financial Crisis[J].Procedia Economics and Finance, 2015, 24: 83-92
[14] Bolton J, von Boetticher S T. Momentum Trading on the Johannesburg Stock Exchange after the Global Financial Crisis[J]. Procedia Economics and Finance, 2015, 24: 83-92.
[15] Rodríguez-González A, García-Crespo Á, Colomo-Palacios R, et al. CAST: Using neural networks to improve trading systems based on technical analysis by means of the RSI financial indicator[J]. Expert systems with Applications, 2011, 38(9): 11489-11500.
[16] Kannan K S, Sekar P S, Sathik M M, et al. Financial stock market forecast using data mining techniques[C]//Proceedings of the International Multiconference of Engineers and computer scientists. 2010, 1: 4.
[17] Appel G. Become your own technical analyst: how to identify significant market turning points using the Moving Average Convergence-Divergence indicator or MACD[J]. The Journal of Wealth Management, 2003, 6(1): 27-36.
[18] Jolliffe I. Principal component analysis[M]. Springer Berlin Heidelberg, 2011.
[19] Abdi H, Williams L J. Principal component analysis[J]. Wiley interdisciplinary reviews: computational statistics, 2010, 2(4): 433-459.
[20] Lu H, Plataniotis K N, Venetsanopoulos A N. MPCA: Multilinear principal component analysis of tensor objects[J]. IEEE transactions on Neural Networks, 2008, 19(1): 18-39.
[21] K. Wu, Y. Wu, H. Lee, “Stock trend prediction by using K-means and AprioriAll algorithm for sequential chart pattern mining,”, Journal of Information Science and Engineering, 30 (2014), pp. 653667.