Application of Cluster Analysis in Stock Selection in United States Stock Market

Shuo Wu
Dongbei University of Finance and Economics
Room 102, Unit 2, building 100, south district, airport courtyard, Urumqi, Xin Jiang, China
+8613201252709
wushuo0714@163.com

ABSTRACT

Quantitative trading plays a significant role in stock selection due to its great flexibility and operability. A stock selection strategy was introduced based on the K-means clustering model in machine learning. Some technical indicators, such as MA, KDJ, and MACD with short and long periods, were taken into consideration in our strategy. The United States market stocks were divided into several clusters. And stocks close to the center of the best cluster were chosen to construct a portfolio. Experimental results showed that the investment strategy has a higher excess return rate during the bull market and decreases synchronously with the market trend during the bear market. This strategy, however, is superior to the performance of the S&P500 index at any time. This paper proposes a feasible strategy, which could get a considerable rate of return, to solve the problem of US stock selection.

CCS Concepts

• Theory of computation → Theory and algorithms for application domains → Machine learning theory → Unsupervised learning and clustering.

Keywords

Stock selection; United States stock market; technical indicators; cluster analysis.

1. INTRODUCTION

Investors use the quantitative system to establish mathematical models and find profits by studying a great deal of historical data based on statistical methods. In developed country's capital markets, especially the United States (US) stock market, quantitative trade has been applied for a long time. More than 80% of their investment institutions have adopted a quantitative strategy to invest in the securities market.

Fundamental analysis and technical analysis are two principal methods in stock selection. Comparing with fundamental analysis, technical analysis has several conspicuous merits. Technical analysis could be used in any speculative field because of its extraordinary flexibility. By contrast, since it is difficult for analysis to collect information, fundamental analysis merely focuses on a particular field in general. Technical analysis could also be applied in the stock market for any time period. But fundamental analysis is often employed in the long-term analysis. Besides, technical analysis is operable because the factors that affect market can be reflected in the tendency clearly.

Cunha et al. used the cluster analysis to select stocks of major companies from North and South America. They put forward a technique to group stocks in spot markets according to a risk–return criterion [1]. Wang et al. compared the accuracy between a MACD histogram and a MACD-HVIX histogram. They found that the accuracy of MACD-HVIX histogram is approximate 55% higher than that of the MACD histogram when the buy-and-sell strategy is adopted [2]. Chong et al. showed that the RSI and MACD could generate returns higher than most strategies [3]. A new method was used to select stocks from the Thai stock market. This strategy could identify a group of stocks that has the best trend and momentum characteristics. Finally, it could outperform the market during a short time period [4]. Anghel considered the MACD as a good technical analysis investment method in the stock market. And it could help companies obtain abnormal cost and risk adjusted returns [5]. Suganthi et al. demonstrated the strength and accuracy of each algorithm for clustering in terms of performance and efficiency [6]. Steve et al. studied the selection and active trading of stocks by time series outlier analysis [7]. Based on analyzing the applicability and limitation of fundamental analysis, technical analysis, and portfolio investment analysis, Zhou et al. pointed out it is effective in the stock selection [8].

The clustering method is used to select the stocks in the US stock market in this paper. Section II shows the stock information in the US stock market. Section III reviews related works and methods on six main technical indicators, algorithmic trading, and model evaluation. The experimental results and conclusions are discussed in Section IV and V, respectively.

2. DATA RESEARCH

In August 2019, there are over 5000 stocks in the US stock market with a total market capitalization of 42,717 billion dollars. The S&P500 index, the main US market index, is an equity-weighted index of 500 stocks from the New York Stock Exchange, the US Stock Exchange and OTC. 400 industrial stocks, 40 utility stocks, 40 financial stocks, and 20 transportation stocks are in the S&P500 index [9]. This index accounts for almost 80% of the total stocks of the New York stock exchange. Various important factors of stocks should be considered to construct the S&P500 index, such as market value, liquidity, and industrial representativeness. Thus, it is an important reference index for the US stock market.

IC4E 2020, January 10–12, 2020, Osaka, Japan
© 2020 Association for Computing Machinery..
ACM ISBN 978-1-4503-7294-7/20/01…$15.00
DOI: https://doi.org/10.1145/3377571.3377628

In this analysis, we studied 5175 stocks monthly in the US stock market from August 1, 2009, to August 1, 2016. The stocks that suspended in the next three trading months are excluded. the stocks that have been listed for less than 3 months are also removed.

3. MODEL

3.1. Technical Indicators

3.1.1. MA

MA is a technical indicator that averages the prices of securities over a period of time. It connects the mean value of different times to observe the trend of securities prices. The formula is as follow:

$$MA = \frac{\sum_{t=1}^{n} C_n}{n}$$

where C_n is the closing price. The short period and long period of MA we choose are 4 and 8 months.

3.1.2. KDJ

KDJ is an innovative and practical technical analysis indicator, which is widely used in the short-term and medium-term trend analysis of the stock market and futures market [10]. According to statistics principle, the highest price, lowest price and closing price within a specific period are used in KDJ. The K value, D value, and J value could depict the stock trend. KDJ is mainly a technical indicator that uses the real volatility of price fluctuation to reflect the price trend and the phenomenon of overbuying and overselling. The buy and sell signals are shown before the price changes. The main formula of KDJ is as follows [11]:

$$RSV_n = \frac{C_n - L_n}{H_n - L_n} \times 100$$

$$K_n = \frac{2}{3} \times K_{n-1} + \frac{1}{3} \times RSV_n ,$$

$$D_n = \frac{2}{3} \times D_{n-1} + \frac{1}{3} \times K_n ,$$

$$J_n = 3 \times K_n - 2 \times D_n ,$$

where C_n is the closing price on the n day, Ln is the lowest price during n days, and H_n is the highest price during n days. KDJ is so sensitive to market signals that it is suitable for short-term trading. On the other hand, the KDJ indicator has high accuracy, especially when KDJ is used with other indicators. The short period and long period of MA are 4 and 8 months.

3.1.3. MACD

Based on the principle of moving average, MACD is a technical indicator that utilizes the convergence and separation between the short-term and long-term moving averages of closing prices to determine the timing of buying and selling. The formula of MACD is as follows [12]:

$$EMA_t^m(S_t) = (1 - \frac{2}{m+1}) \times EMA_{t-1}^m + \frac{2}{m+1} \times S_t ,$$

$$DIF_t = EMA_t^m(S_t) - EMA_t^n(S_t) ,$$

$$DEA_t = EMA_t^p(DIF_t) ,$$

$$MACD_t = 2 \times (DIF_t - DEA_t) ,$$

where m=5, n=10, p=4. MACD is a widely used indicator with many advantages. MACD can be used to study the medium-term and long-term trends. Since MACD could determine whether it is a long market or a short market, it is often used to avoid contrarian operations [13]. And it could adopt corresponding trading strategies to avoid frequent operations.

3.2. K-means Clustering

K-means clustering is one of the unsupervised learning methods and it is an iterative clustering analysis algorithm [14]. K-means clustering randomly selects K objects as the initial clustering center, then calculates the distance between each object and each clustering center. After that, it assigns each object to its nearest clustering center. Each time a sample is allocated, the cluster center of the cluster will be recalculated until reach the maximum iteration number.

K-means clustering has many advantages. It is efficient, simple and easy to implement. It is highly consistent with our intuitive classification cognition. Also, when we deal with a mass of data, K-means could still maintain scalability and efficiency. The basic K-means algorithm for clustering into K groups is shown in the previous paper [15].

3.3. Evaluation

The Annualized Rate of Return is calculated by converting the current rate of return, such as the daily rate of return, the weekly rate of return, and the monthly rate of return, into an annual rate [15]:

$$Annualized\ Rate\ of\ Return = \frac{I}{P \times N} \times 365 \times 100\% ,$$

where I is investment income, P is principal, N is investment days.

Maximum Drawdown is the maximum value of the rate of retreat of the yield when the product's net worth reaches the lowest point, at any historical point in the selected period:

$$Drawdown = \max \frac{D_i - D_j}{D_i} ,$$

where D_i is the net value of a certain day and D_j is the net value of the day after Di.

Sharpe Ratio is the excess return on the portfolio for each unit of total risk [16]:

$$Sharpe\ Ratio = \frac{E(p) - R_f}{\sigma_p} ,$$

Information Ratio is the excess return from the unit's active risk [16]:

$$IR = \frac{\alpha}{\omega} ,$$

where α is the combined excess return, ω is the active risk.

4. RESULTS

In this paper, short-term MA, long-term MA, short-term KDJ, long-term KDJ, and MACD are 5 important factors to construct a portfolio based on K-mean clustering. The result is shown in Figure 1 from August 1, 2009, to August 1, 2016. The cluster we chose (ncluster) is 10; the number of stock (nstock) is 20; the short period of the characteristic factor is 4, and the long period the characteristic factor is 8. Figure 1 shows the return of the proposed strategy (blue line) and S&P500 (yellow line).

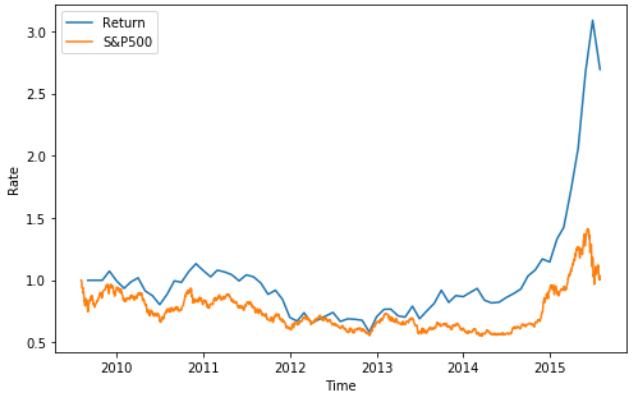

Figure 1. The return of the proposed strategy and S&P500.

As shown in Figure 1, the return of the proposed portfolio we constructed is significantly better than the return of the S&P500 index. In most cases, the return of the portfolio is twice and more than the S&P 500 index. After the second half of 2015, the return on the portfolio will reach 300%, much higher than the return of the S&P500 index. It shows that our investment strategy is basically successful. The Annualized Rate of Return for the optimal portfolio reached 16.4%, Sharpe Ratio reached 1.11, Information Ratio reached 1.21, and Maximum Drawdown was 0.42.

The selection of the number of clusters and stocks may affect the results of the analysis. For the proposed method, we calculated the Annualized Rate of Return, Maximum Drawdown, Sharpe Ratio, and Information Ratio separately when clusters are divided into 5, 10 and 20 respectively. As the results are shown in Table 1, when the number of clusters is 5, 10, and 20, the return of the strategy is higher, the Sharpe Ratio and Information Ratio are higher, and the Maximum Drawdown is smaller. When ncluster is 20, Sharpe Ratio and Information Ratio are highest in the strategy. When ncluster is 5, the Rate of Return and Maximum Drawdown is the smallest.

Table 1. Results of four indicators at 5, 10, 20 cluster

	Annual Rate of Return	Maximum Drawdown	Sharpe Ratio	Information Ratio
5	14.4%	0.36	1.03	1.15
10	16.4%	0.42	1.11	1.21
20	17.4%	0.46	1.12	1.27

According to Table 2, taking these 5 indicators, we found that when we selected 30 stocks, the results were optimal. When the number of stocks is 20, the strategy's return reaches 0.164. Sharpe Ratio is lower than Information Ratio. It can be seen that the information ratio and the Sharpe Ratio have a relatively linear relationship with the number of stocks and decrease as the number of stocks increases.

Table 2. Results of four indicators at 10, 20, and 30 stocks

	Annual Rate of Return	Maximum Drawdown	Sharpe Ratio	Information Ratio
10	13.1%	0.34	1.01	1.11
20	16.4%	0.42	1.11	1.21
30	20.0%	0.46	1.15	1.35

In general, under monthly frequency data, when the number of clusters is 20, and the number of stocks is 20, the strategy works best.

5. CONCLUSION

This paper introduced a strategy for selecting stocks based on the K-means clustering model in machine learning. We divided the US stock market into several clusters and selected stocks that are closest to the clustering center to construct the portfolio. The portfolio and the performance of the portfolio's earnings and sensitivity factors outperformed the market portfolio. Our proposed investment strategy based on K-means clustering has a higher excess return rate during the bull market, and it decreases synchronously with the S&P500 index during the bear market. However, our proposed strategy is better than the S&P500 index at any given time.

To improve, other indicators may become good feature factors, such as market capitalization factors, momentum reversal factors and so on. According to other scholars' investigation and research [17-20], the characteristics of these factors may lead to better results and higher returns. In addition, other machine learning models, taking random forests, gradient enhancements, support vector machines for example, may be needed to make more profits.

6. ACKNOWLEDGMENTS

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed.

7. REFERENCES

[1] Da Costa Jr, N., Cunha, J., and Da Silva, S. 2005. Stock selection based on cluster analysis. *Econ. Bull.*, 13, 1 (Oct. 2005), 1-9.

[2] Wang, J., and Kim, J. 2018. Predicting Stock Price Trend Using MACD Optimized by Historical Volatility. *Math. Probl. Eng.*, 2018, (Dec. 2018), 1-12. DOI= https://doi.org/10.1155/2018/9280590

[3] Chong, T. T. L., and Ng, W. K. 2008. Technical analysis and the London stock exchange: testing the MACD and RSI rules using the FT30. *Appl. Econ. Lett.*, 15, 14 (Nov. 2008), 1111-1114. DOI= https://doi.org/10.1080/13504850600993598

[4] Peachavanish, R. 2016. Stock selection and trading based on cluster analysis of trend and momentum indicators. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Hongkong, China, March 16 - 18, 2016). IMECS, Hong Kong, 317-321.

[5] Anghel, G. D. I. 2015. Stock market efficiency and the MACD. Evidence from countries around the world. *Proc.*

Econ. Finance, 32, (Dec. 2015), 1414-1431. DOI= https://doi.org/10.1016/S2212-5671(15)01518-X

[6] Suganthi, R., and Kamalakannan, P. 2015. Analyzing stock market data using clustering algorithm. *Int. J. Future Comput. Commun.*, 4, 2, (Apr. 2015), 108. DOI= https://doi.org/10.7763/IJFCC.2015.V4.366

[7] Craighead, S., Klemesrud, B., Financial, N., and Plaza, O. N. 2002. Stock selection based on cluster and outlier analysis. In *Fifteenth International Symposium on Mathematical Theory of Networks and Systems University of Notre Dame* (South Bend, USA, August 2002). University of Notre Dame, South Bend.

[8] Zhuo-hua, Z. H. O. U., Wen-nan, C. H. E. N., and Zong-yi, Z. H. A. N. G. 2015. Application of cluster analysis in stock investment. *J. Chongqing Univ.*, 2002, (Jul. 2015), 122-126.

[9] Kawaller, I. G., Koch, P. D., and Koch, T. W. 1987. The temporal price relationship between SandP 500 futures and the SandP 500 index. *J. Finance*, 42, 5, (Dec. 1987), 1309-1329. DOI= https://doi.org/10.1111/j.1540-6261.1987.tb04368.x

[10] Chen, J. 2010, October. SVM application of financial time series forecasting using empirical technical indicators. In *2010 International Conference on Information, Networking and Automation (ICINA)* (Kunming, China, November 15, 2010). (Vol. 1, IEEE, Kunming, 1-77.

[11] Wu, M., and Diao, X. 2015, December. Technical analysis of three stock oscillators testing MACD, RSI and KDJ rules in SH and SZ stock markets. In *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)* (Harbin, China, 16 June, 2016). IEEE, Harbin, 320-323. DOI= https://doi.org/10.1109/ICCSNT.2015.7490760

[12] Asness, C. S., Moskowitz, T. J., and Pedersen, L. H. 2013. Value and momentum everywhere. *J. Finance*, 68, 3, (Jun. 2016), 929-985. DOI= https://doi.org/10.1111/jofi.12021

[13] Fernández-Blanco, P., Bodas-Sagi, D. J., Soltero, F. J., and Hidalgo, J. I. 2008, July. Technical market indicators optimization using evolutionary algorithms. In *Proceedings of the 10th annual conference companion on Genetic and evolutionary computation* (Atlanta, GA, USA, July 12 - 16, 2008). GECCO '08. ACM. New York, NY, 1851-1858. DOI= https://doi.org/10.1145/1388969.1388989

[14] Wagstaff, K., Cardie, C., Rogers, S., and Schrödl, S. 2001, June. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning* (Williamstown, MA, USA, June 28 - July 1, 2001). Williams College, Williamstown, 577–584.

[15] Hartigan, J. A., and Wong, M. A. 1979. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)*, 28, 1, (Jan 1979), 100-108. DOI= https://doi.org/10.2307/2346830

[16] Israelsen, C. L. 2005. A refinement to the Sharpe ratio and information ratio. *J. Asset Manage.*, 5,6, (Nov. 2004), 423-427. DOI= https://doi.org/10.1057/palgrave.jam.2240158

[17] Patel, J., Shah, S., Thakkar, P., and Kotecha, K. 2015. Predicting stock market index using fusion of machine learning techniques. *Expert Syst. Appl.*, 42, 4, (Mar. 2015), 2162-2172. DOI= https://doi.org/10.1016/j.eswa.2014.10.031

[18] Khaidem, L., Saha, S., and Dey, S. R. 2016. Predicting the direction of stock market prices using random forest. *arXiv preprint,* (Apr. 2016), arXiv:1605.00003.

[19] Baetje, F., and Menkhoff, L. 2016. Equity premium prediction: Are economic and technical indicators unstable?. *Int. J. Forecasting*, 32, 4, (Dec. 2016), 1193-1207. DOI= https://doi.org/10.1016/j.ijforecast.2016.02.006

[20] Cervelló-Royo, R., Guijarro, F., and Michniuk, K. 2015. Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data. *Expert Syst. Appl.*, 42, 14, (Aug. 2015), 5963-5975. DOI= https://doi.org/10.1016/j.eswa.2015.03.017