Analysis of Stock Data Based on Clustering Method

Tianhao Fu[1,*] , Jixin She[2,a], Yuhao Guo[3,b]

[1]School of Economy, Fudan University, Shanghai, China
[2]Academy of Reading, Nanjing University of Information Science and Technology, Nanjing, China
[3]Huaer Zizhu Academy, Shanghai, China
[a]2605039653@qq.com, [b]2194552791@qq.com
**Corresponding author:2512861153@qq.com*
These authors contributed equally to this work

Abstract: *Number of researches on stock data are based on machine learning method. However, former researches mainly applied all kinds of supervised learning method as well as some kind of basial clustering method, such as neuro net-work and k-means algorithm. We use an emerging clustering algorithm called AP algorithm to cluster the A-share stock data before 2017. And finally, according to the market performance, the selected 69 A-share stocks are divided into 9 classification clusters based on their market performances. Such classification has a strong guiding role for investment decision-making.*

Keywords: *Affinity Propagation Cluster, Investment, Stock Data*

1. Introduction

Since the setting up of Shanghai Stock market in 1990, China's stock market has been running out for about 32 years. As an important mechanism of enterprise financing, stock plays an important role in enterprise development and overall national macroeconomics. Since the birth of stock, people began to study the principle of its pricing behaviors. The earliest beta model, then Fama's three factors model, and then to the multi-factor model, the prediction of stock price is becoming more and more accurate, but it is impossible to get rid of the methodology of statistical regression. Nowadays, with the prosperity of machine learning methods, it has become a more powerful tool to predict stock price and analyses other complex decision systems compared with regression model.

The previous models of stock data analysis can be divided into two kinds of different models, regression models and machine learning models. For regression models, scholars mainly use various linear and nonlinear regression methods, such as OLS, penalty regression, PCR & PLS method, GLS method, etc. For ML models, scholars mainly use Neural Network algorithm and Decision Trees algorithm, etc. For example, Cai et al. (2019)[1] uses decision tree model and support vector machine regression to analyze the extra return rate of stock. Meng et al. (2019)[2] uses classification tree method to analyze the stock index. Sun et al. (2019)[3] analyzes stock suspension based on random forest model. Such branches of studies have proved that machine learning models is significantly better than the traditional regression models in analysis of stock data.

As a kind of unsupervised learning algorithm, clustering model is widely used in fields other than Finance. For example, Zhang et al. (2013)[4] used clustering algorithm model to analyze residents' electricity consumption behavior. Zhang et al. (2003)[5] used clustering algorithm to optimize logistics distribution path. Sun et al. (2011)[6] detects hot news of Weibo based on clustering algorithm. Because the clustering algorithm mainly carries out pattern recognition and prediction based on the training samples of discrete location categories, it is very suitable for dealing with the complex data set and solve diverse factors in stock data, since the actual mechanism is obscure. Therefore, this paper hopes to use clustering algorithm to analyze stock data, so as to fill the clustering blank for application of machine learning in analysis of stock data among academic fields.

The principle of Clustering is to divide the unknown sample set according to the internal similarity. The similarity within categories is large, and the similarity between categories is small. The well-known clustering algorithms include K-means and DBSCAN. K-means first artificially sets K center point, then calculate the distance from the sample point to K center points, then reassign the categories until the evaluation function converges. The optimized DBSCAN is a density-based algorithm. It uses two algorithm parameters (neighborhood radius and the minimum number of points) to determine the density, so as to divide different classification clusters according to different densities. Therefore, this method is less affected by extreme values.

The two clustering algorithms mentioned above have been used to analyze stock data in academic circles. For example, Jia et al (2008)[7] used the ICA method combined with the K-means algorithm to analyze stock time-series data, and distinguished 8 classification clusters according to different stock trends; Hanna (2011)[8] also discussed the application possibility of the K-means algorithm in time series analysis; Song et al (2018)[9] used fuzzy c-means clustering algorithm to analyze stock data; Hai(2015)[10] uses the k-means clustering algorithm to guide the financial decision-making of different companies. Therefore, the author uses the a-means clustering algorithm to measure the financial distance of different stock markets and finds that the a-means clustering algorithm is the best.

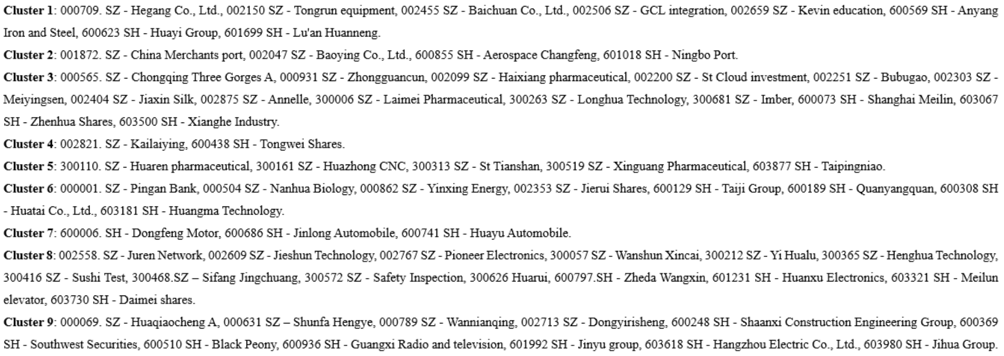
Based on the above summary, it is not difficult to see that at present, the application of clustering algorithm in stock time series analysis is mainly k-means algorithm and its derivative pair mean calculation and distance calculation correction algorithms. Compared with the traditional clustering algorithm, the AP clustering algorithm is more suitable for the processing of high-latitude data, such as stock time series. Therefore, this paper considers applying this new algorithm to stock data analysis and analyzes the effect of this algorithm.

2. Algorithm and result analysis in this paper

2.1. Algorithm introduction and data source

The analysis of this paper is based on the affinity propagation clustering method, hereinafter referred to as the AP clustering algorithm. Compared with the traditional clustering algorithm mentioned above, the AP clustering algorithm is suitable for high-dimensional and multi-data fast clustering without specifying the number of final clustering families. And this method is not sensitive to the initial value of the data, and there is no requirement for the symmetry of the initial similarity matrix data. Therefore, based on the above two points, the AP algorithm is more suitable for data analysis in which the classification characteristics of stock data are fuzzy and it is difficult to directly determine the number of clustering clusters. In addition, compared with the K-centers clustering method, the square difference error of its results is smaller (Frey, 2007)[11].

The data of this study is from the website https://tushare.pro and the method of pro.stock_basic is used to obtain the basic information (stock code, name, listing date, delisting date, etc). There are 69 stock samples in total. At the same time, due to the huge amount of data, the author only intercepts the relevant data of stocks listed before December 31, 2017, and then intercepts some stocks for analysis in step 20 to obtain the selected stock data.



Cluster 1: 000709. SZ - Hegang Co., Ltd., 002150 SZ - Tongrun equipment, 002455 SZ - Baichuan Co., Ltd., 002506 SZ - GCL integration, 002659 SZ - Kevin education, 600569 SH - Anyang Iron and Steel, 600623 SH - Huayi Group, 601699 SH - Lu'an Huanneng.

Cluster 2: 001872. SZ - China Merchants port, 002047 SZ - Baoying Co., Ltd., 600855 SH - Aerospace Changfeng, 601018 SH - Ningbo Port.

Cluster 3: 000565. SZ - Chongqing Three Gorges A, 000931 SZ - Zhongguancun, 002099 SZ - Haixiang pharmaceutical, 002200 SZ - St Cloud investment, 002251 SZ - Bubugao, 002303 SZ - Meiyingsen, 002404 SZ - Jiaxin Silk, 002875 SZ - Annelle, 300006 SZ - Laimei Pharmaceutical, 300263 SZ - Longhua Technology, 300681 SZ - Imber, 600073 SH - Shanghai Meilin, 603067 SH - Zhenhua Shares, 603500 SH - Xianghe Industry.

Cluster 4: 002821. SZ - Kailaiying, 600438 SH - Tongwei Shares.

Cluster 5: 300110. SZ - Huaren pharmaceutical, 300161 SZ - Huazhong CNC, 300313 SZ - St Tianshan, 300519 SZ - Xinguang Pharmaceutical, 603877 SH - Taipingniao.

Cluster 6: 000001. SZ - Pingan Bank, 000504 SZ - Nanhua Biology, 000862 SZ - Yinxing Energy, 002353 SZ - Jierui Shares, 600129 SH - Taiji Group, 600189 SH - Quanyangquan, 600308 SH - Huatai Co., Ltd., 603181 SH - Huangma Technology.

Cluster 7: 600006. SH - Dongfeng Motor, 600686 SH - Jinlong Automobile, 600741 SH - Huayu Automobile.

Cluster 8: 002558. SZ - Juren Network, 002609 SZ - Jieshun Technology, 002767 SZ - Pioneer Electronics, 300057 SZ - Wanshun Xincai, 300212 SZ - Yi Hualu, 300365 SZ - Henghua Technology, 300416 SZ - Sushi Test, 300468.SZ – Sifang Jingchuang, 300572 SZ - Safety Inspection, 300626 Huarui, 600797.SH - Zheda Wangxin, 601231 SH - Huanxu Electronics, 603321 SH - Meilun elevator, 603730 SH - Daimei shares.

Cluster 9: 000069. SZ - Huaqiaocheng A, 000631 SZ – Shunfa Hengye, 000789 SZ - Wannianqing, 002713 SZ - Dongyirisheng, 600248 SH - Shaanxi Construction Engineering Group, 600369 SH - Southwest Securities, 600510 SH - Black Peony, 600936 SH - Guangxi Radio and television, 601992 SH - Jinyu group, 603618 SH - Hangzhou Electric Co., Ltd., 603980 SH - Jihua Group.

Figure 1: Stock of 9 data clustered categories (Obtained by AP clustering algorithm)

2.2 The Structure of Algorithm

The specific algorithm is developed in the following order:

1) Getting the interface of the web address called http://tushare.pro.

2) Obtaining data, setting to display Chinese font.

3) Obtaining stock code, names and the basic information (stock code, name, listing date, delisting date etc.).

4) Intercepting stock data and printing the length of selected stock data.

5) Setting the start time, integrating the market data (analysis indicators) of non-reinstated, pre reinstated and post reinstated stocks, indexes, digital currencies, financial funds, futures and options.

6) Data filtering (filter out items with empty data).

7) Eliminating the filtered empty data items.

8) Calculating correlation coefficient and the basic attributes of sample, using AP clustering algorithm to process the data.

9) Generating the result and obtaining the cluster obtained according to the AP clustering algorithm (Figure 1).

10) Visualizing the result. Figure 1 is the contour coefficient diagram. The closer contour coefficient diagram is to figure 1, the better the stock quality is; Figure 2 shows the general clustering view.

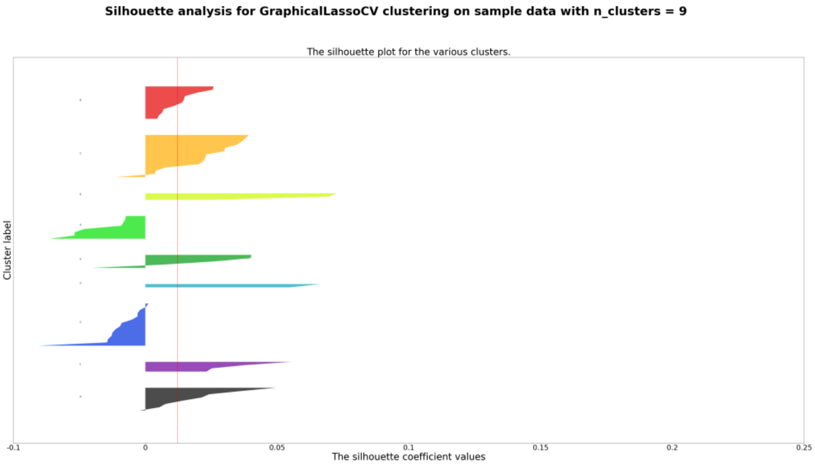


Figure 2: Evaluation of different stock clusters

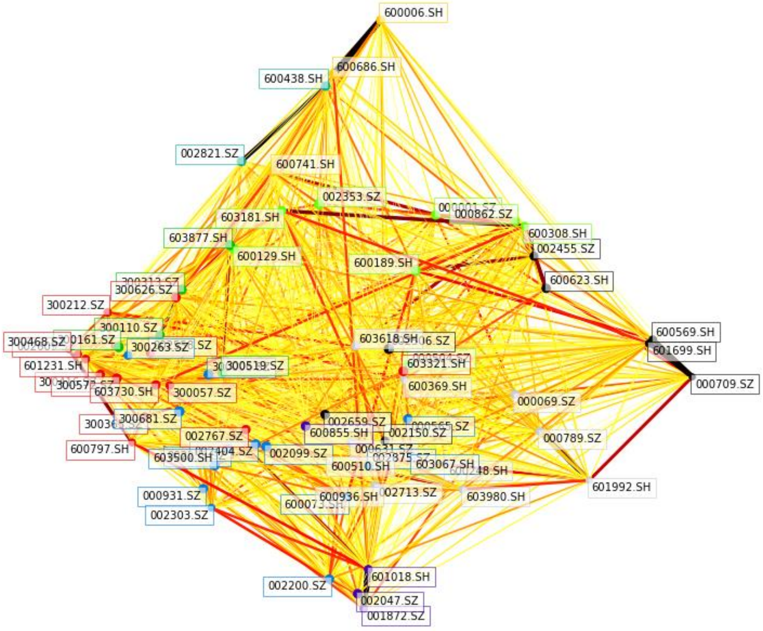


Figure 3: Visualization results of classification clusters

2.3 Results of Analysis

1) Observing the clustering clusters obtained by AP clustering algorithm (Figure 1), it can be seen

from the table that the AP clustering algorithm used in this paper can effectively classify the stock data with complex and miscellaneous indicators. Because the AP algorithm is suitable for high-dimensional and multi data fast clustering and does not need to specify the number of final clustering families, the 69 stocks used in this paper can be quickly and effectively clustered into 9 classification clusters.

2) By observing the contour coefficient diagram in Figure 2, it can be seen that there are obvious differences in the contour coefficient of each cluster. Among these coefficient diagram, the stocks with a coefficient closer to 1 have better quality and are more suitable for stable investment. Therefore, this classification cluster can be used as a reference for investment portfolio.

3) Observing the result visualization in Figure 3. Labels of different colors represent different clusters, from which the density relationship of different classification clusters are clearly seen.

To summarize, the AP clustering algorithm selected in this paper can efficiently identify and analyze the stock data, and the stock quality can be effectively screened according to the analysis results.

3. Conclusion

From the upward concerns about our analysis of stock data based on clustering method, we can argue several conclusions on our research. We have filled in the gap of analysis of stock data based on cluster models, and we have tested the behavior of such model.

Firstly, based on the results that all 69 pieces of stock are classified into 9 disparate clusters showed in figure 1 and figure 3, we can conclude that the AP-clustering model we used are effective in classification of A-share stocks.

Secondly, based on the evaluation difference of different clusters showed in Figure 2, we can conclude that this clustering is based on market behaviors of different stock and such clustering can reveals the instinct similarity of different stocks' quality.

Finally, based on the first and second conclusion, we can conclude that such clustering is useful in investment. Inverter can use stock in different clusters to generate a hedging portfolio, or can use stock in the same clusters to generate portfolios with different behaviors.

References

[1] Cai Qingquan, Ma yunyun, and Li Jinmei (2019) Prediction Model of Stock Excess Return Based on Machine Learning. Information System Engineering, 6-17.
[2] Meng Ye (2019) Research and Application of Machine Learning Stock Selection Algorithm Integrating Behavioral Financ. Qingdao University, DOI:10.27262/d.cnki. gqdau. 2019.000890.
[3] Sun Fuxiong, Liu Guangming, Zeng Zixuan, and Peng Mengqi (2020) Research on Stock Suspension Prediction Based on Portfolio Model. Computer Engineering and Application, 18, 272-278.
[4] Zhang Suxiang, Liu Jianming, Zhao Bingzhen, and Cao Jinping (2013) Research on analysis Model of Residential Power Consumption Behavior Based on Cloud Computing. Power Grid Technology, 06: 1542-1546 DOI:10.13335/j.1000-3673. pst. 2013.06.010.
[5] Zhang Qian, Gao Liqun, Hu Xiangpei, and Wu Wei (2003) Clustering Improved Genetic Algorithm for Multi-objective Optimization of Logistics Distribution Path. Control and Decision Making, 04: 418-422 DOI:10.13195/j.cd. 2003.04.34. zhangq. Seve
[6] Sun Shengping (2011) Research on Hot Topic Detection and Tracking Technology of Chinese Weibo. Beijing Jiaotong University.
[7] Guo Chonghui, Jia Hongfeng, and Zhang Na (2008) Time Series Clustering Method Based on ICA and Its Application in Stock Data Analysis. Operations Research and Management, 05, 120-124.
[8] Hannah (2011) Research and Application of Clustering Algorithm in Time Series. Guangdong University of technology.
[9] Song Zongxiang (2017) Application of Fuzzy C-means Clustering in Stock Investment. Northeast Petroleum University.
[10] Hai Mo, Niu Yihan, and Zhang Yuejin (2015) Application of Parallel Clustering Algorithm For Big Data in Stock Sector Division. Big data, 04, 9-17.
[11] Frey, BJ, Dueck, D (2007) Clustering by Passing Messages between Data Points. Science, pp.972-976.