ISSN 2616-5775 Vol. 4, Issue 8: 42-47, DOI: 10.25236/AJCIS.2021.040808

Stock Price Trend Analysis Method Based on Change Cycle Clustering

Yang Wang[1,a*], Linjie Huang[2,b], Yidan Zheng[3,c], Siuwa Lee[4,d], Yuanming Fu[5,e]

[1]*China University of Petroleum, Beijing, China*
[2]*Nanjing Agricultural University, Nanjing, Jiangsu, China*
[3]*University of Waterloo, ON, Canada*
[4]*Jinan University, Guangzhou, Guangdong, China*
[5]*Tianjin No.1 High School, Tianjin, China*
[a]*stellaworking@163.com,* [b]*272368225@qq.com,* [c]*y355zhen@uwaterloo.ca,* [d]*1007136294@qq.com,* [e]*fuyuanming2004@163.com*
**Corresponding author: stellaworking@163.com*
These authors contributed equally to this work

Abstract: *With the continuous development of financial informatization and the continuous improvement of China's economic system, analyzing and mining financial data has become an important means to study financial problems. Compared with other industries in the financial field, stock data is easier to collect and store, and its application is more convenient. Analyzing and predicting the changing trend of the future stock market through the historical data of stocks is helpful to reduce the risk of investors and increase income. It has become a research hotspot in the financial field. Using the historical price data of domestic stocks, a method based on a clustering algorithm is proposed to analyze the periodic characteristics of stock price changes. This method can provide a basis for the prediction of stock price and the detection of trading behaviour in the stock market.*

Keywords: *Moving average; K-means clustering; Stocks price prediction*

1. Introduction

As a derivative instrument of financial market, stock has the characteristics of profitability, risk, liquidity and permanence. In addition, due to the uncertainty and complexity of the market, the change of stock value is affected by many factors.

As the basis of stock measurement standards, the analysis and prediction of stock price can more effectively avoid future risks for decision makers. For regulators, it can strengthen the control of the stock market, timely regulate and guide the stock market while providing firm confidence and strong guarantee for the sustainable development of the economy. [1] Therefore, how to analyze and predict the volatility, trend and price of the stock market is a problem highly concerned by many scholars in the financial field.

The research on stock analysis and prediction can be divided into the traditional statistical methods in early research and the machine learning algorithm to analyze it with the development of digitization. In the initial research on stock trend and price prediction by traditional statistical methods, statistical knowledge are always used to establish the model, testing and analyzing the model by actual data. Liu, W. et al. in [5], Feng, P. et al. in [7], Zhang, B. in [8] and Wu, Y. et al. in [9] used the idea of autoregression to establish different models to analyze the data. It could be found that ARFIMA model failed in the research process, while ARIMA model and ARMA model had small errors in short-term regression and wear feasible. According to the analysis and prediction results of other regression models by Guo, G. et al in [6]and Zhang, S. in [3], it could be found that other regression analysis methods show more positive results. It could be seen from Shiba, T. et al. in [2], Fan, D et al. in [4] and Xia, L. in [10] that the combination model could be analyzed and predicted well, although pessimistic results happened in some data.

It can be found that there are few studies using traditional statistical methods in recent years, but more in the past. We can know that although the traditional statistical methods can get some optimistic results for the analysis and prediction of stocks, there are still some defects in accuracy due to the subjectivity of traditional statistical methods and the nonlinearity and complexity of stocks.

ISSN 2616-5775 Vol. 4, Issue 8: 42-47, DOI: 10.25236/AJCIS.2021.040808

In recent years, with the development of big data and information digitization, many scholars consider using machine learning algorithms when analyzing and predicting stocks. Xia, L. et al. in [12] and Hu, X. et al. in [11] used Markov model to analyze the prediction results, which found that there were also some parts that needed to be improved. Seo, J. et al. in [13] and Wang, Y. et al. in [22] found that decision tree could solve several problems at the same time and the effect was good. In [17], Xu, X used BP neural network to predict which found that due to the complexity of the stock market, other algorithms could be integrated on this basis to improve the accuracy of the stock market. Yang, X. et al. in [15], LV, Q. in [16], and Xie, G. in [19] used SVM and SVR to analyze and predict and came to the conclusion that compared with neural network and CAR model, support vector machine model had higher accuracy and was suitable for stock price prediction. Han, S. et al. in [21] and Prachyachuwong, K. et al. in [24] used deep learning to predict which came to the conclusion that such methods could accelerate learning and improve prediction performance. Zhang, Q. in [20] and Hu, Y. in [23] used a variety of methods or combined other theories with machine learning algorithms to analyze the stock price, and found that the combined algorithm had higher accuracy. Bai, M. et al. in [14] and Deng, N. et al. in [18] analyzed and studied the clustering algorithm. They found the effectiveness of clustering for big data processing and concluded that the clustering algorithm might become the research trend.

Above all, most of the accuracy of machine learning algorithm is high. What's more, it is feasible to study the stock price by using a combination of various methods which has further improved accuracy. While sorting out the literature, it can be found that there are more studies on the use of neural network algorithms, and less literatures on the use of clustering to study the stock price. Therefore, this paper will use k-means clustering algorithm to analyze and study the stock price, so as to provide reference for subsequent research in this regard.

The rest of this paper is organized as follows: the second section introduces the core idea of this paper and all the methods involved in the research. In third section the methods introduced in this paper are run in Python and the data results are shown and analyzed. The fourth section summarizes the content of this paper.

2. The Proposed Methods

In order to predict the trend of a specific stock price, a large amount of data will be collected, smoothed and extracted the features for clustering analysis eventually. Thus, the methods of this research will be divided into four parts separately: data collection, data preprocessing, extraction of data features and data clustering especially K-means clustering.

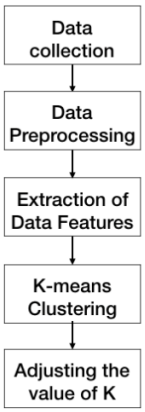

Figure 1 Flow chart representation of the methods

2.1 Collection of data

In this research, yfinance of Yahoo is used to obtain the stock historical data. Besides, four stocks are chosen to be our research objects: ZhongGuoShiHua (600028.SS), BaoGangGuFen (600010.SS), QingSongJianHua (600425.SS) and LanGuangFaZhan (600466.SS) among which LanGuangFaZhan will be treated as the main object. The data of these four stocks will be collected from January 1st, 2009 to June 30th, 2021 to avoid the disturbance of abnormal stock characteristics caused by the financial crisis in 2008.

ISSN 2616-5775 Vol. 4, Issue 8: 42-47, DOI: 10.25236/AJCIS.2021.040808

2.2 Preprocessing of data

Stock data are updated on a daily basis and the stock price fluctuates everyday which shows its volatility and uncertainty. In order to decrease the noise that the stock data brought in a more efficient way, there are two steps will be performed to preprocess the data. Firstly, remove the null values. Secondly, smooth the data.

Null values are removed from the data because nulls are meaningless. Deleting null values at the early stage of data analysis can simplify the workload of data processing in the future.

Moving averages are often used to smooth stock prices and highlight underlying trends. Meanwhile, moving average is one of the most popular tools for predicting the trend of each stock. The method of moving average could be explained as

$$X'_n = (X_{n-m+1} + X_{n-m+2} + \ldots + X_n)/m. \quad (1)$$

Here X_n is the closing price of that day, m is the size of the window and X'_n is the closing price of that day after moving average. Therefore, the moving average method is suitable for decreasing noises of the data, and the moving average method in the data smoothing process will be used.

2.3 Extraction of data features

The extraction of data features includes acquiring oscillation cycles of the data, the determination of extreme value and finding specific data features (the initial value of each oscillation cycle, the duration of each oscillation cycle and the change amplitude of each oscillation cycle).

Periodic pattern analysis not only helps to understand the behavior of data, but also helps to predict future trends of data. [25] Therefore, the acquisition of data oscillation cycle is an essential task which highlights the time dependence in the data. Oscillating period is defined as a time duration for a stock to complete a maximum (minimum) to the minimum (maximum) value to the maximum (minimum). Among them, the maximum and minimum values refer to the local maximum and minimum values in a certain period of time. In this study, local extremums were identified by the highest or lowest values in the 30 days before and after.

2.4 Data normalization

Data normalization, also known as data standardization, is a very important step in K-means. The function of data normalization is to facilitate subsequent data analysis. Considering that the data object of this study is the change of stock prices over a long period of time, it can be seen that the stock price fluctuates greatly during the data preview, with the maximum value reaching 13.87857056. The minimum value is only 1.065714002. Dhillon, I. et al. [26] explained: "If the magnitude of each dimension in the data is different, it will lead to inconsistent contribution of each dimension to the Euclidean distance." It means that different dimensions have different weights, which leads to errors in the clustering results.

For example, the starting price and range of fluctuation are usually within 10, while the duration of fluctuation is sometimes as long as several months. In this way, when calculating the distance, the size of the distance between two points mainly depends on the duration of the fluctuation. The impact of the clustering results is minimal. Therefore, the data needs to be normalized before cluster analysis. Yu, L. [27] pointed out that the commonly used positive index standardization methods are divided into linear standardization and non-linear standardization. Principle of equal maximum value. Considering the characteristics of the data in this research, this paper uses linear standardization to normalize the stock data.

3 Experiment process

3.1 Collection result of data

After importing the yfinance into python, the closing prices of four given stocks among the given time range (2009.1.1 to 2021.6.31) was outputted. Here, stock codes of ZhongGuoShiHua, BaoGangGuFen, QingSongJianHua and LanGuangFaZhang are 60028.SS, 600010.SS, 600425.SS, 600466.SS separately.

ISSN 2616-5775 Vol. 4, Issue 8: 42-47, DOI: 10.25236/AJCIS.2021.040808

According to the figure that Python imported for the raw prices of these four stocks varying with respect to time, the unprocessed prices of stocks are fluctuating sharply and noisy which indicates that they require preprocessing for further analysis.

3.2 Preprocessing of data

3.2.1 Deleting nulls and extracting valid values

All the nulls of those closing prices we collected will be deleted by defining DeleteNan in Python. Here, Nan represents the missing value. After processing, the result shows there are not any nulls among the cases we studied.

Results shows that the highest closing price of LanGuangFaZhan is 13.879; Lowest close: 1.066; Average close: 5.192

3.2.2 Smoothing data: moving average

In this study, let m=6. This method deals with special data. For example, if January 5, 2009 is the first value collected, and the data before this date is not collected, then the closing price on January 5, 2009 remains unchanged after modification. Before January 6, 2009, only January 5, 2009 was collected, so the modified value on January 6, 2009 was the two-day average of January 5 and 6, and so on. The modified value on January 7, 2009 was the three-day average of January 5 to 7.

3.3 Extraction of data features

In this experiment, when the closing price of a day is higher (or lower) than the closing prices of 30 days before and after, we define the closing price of that day as a local maximum (or minimum). Furthermore, we define an oscillation period as a process from local minimum to local maximum and then to local minimum.

Taking LanGuangFaZhan as the main object for further study, the local maximum value (or local minimum value) and the starting and closing date of each oscillation period and the total number of oscillation periods.

Due to the large amount of data being collected, the first five local minima are selected to analyze. According to these five local minima, four oscillation periods are contained one from 2009-01-05 to 2009-07-28, one from 2009-07-28 to 2009-11-23 and the rest periods can be deduced in the same way.

The starting and closing date of each oscillation period are shown distinctively in Figure 3-6 as well. 1.065714 in 2009-01-05 is the start value of the first oscillation period, and 6.651428 in 2009-07-28 is the start value of the second oscillation period (and the end value of the first oscillation period).

The operation results help to judge the duration of the oscillation period. For example, the days contained in 2009-01-05 to 2009-07-28 are the duration of the first oscillation period, and the days contained in 2009-07-28 to 2009-11-23 are the duration of the second oscillation period.

3.4 Experiment analysis

3.4.1 Effective clustering method

Xiong, Z. [28] pointed out that K-means is known for its wide range of data adaptability. No matter what type of dimension data, K-means can adapt well. Next, this article will use K-means as the basic algorithm for stock data analysis. Using the K-means algorithm that comes with python, set K to 2, and perform cluster analysis on the normalized starting price, range of fluctuation, and duration of fluctuation to obtain a better classification method, so as to Provide basis for stock forecasts and market dealer testing. Since the end price of volatility is not clearly distinguishable, in order to make the results more impressive, two characteristic dimensions are used to visually analyze the data respectively, and the volatility change range and volatility duration, volatility start price and volatility change range, and volatility duration are obtained. The results of the three clustering methods with the starting price of volatility, the two-dimensional visualization results are as follows.

ISSN 2616-5775 Vol. 4, Issue 8: 42-47, DOI: 10.25236/AJCIS.2021.040808

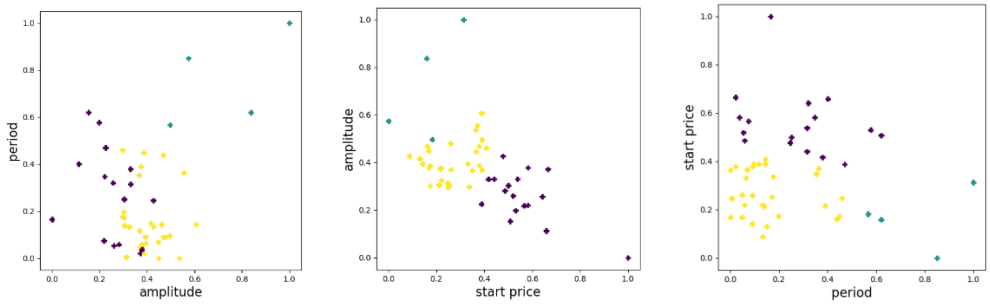

Figure 2 Two-dimensional clustering results

3.4.2 Optimal K value acquisition

However, since the choice of K value in the K-means algorithm is related to the quality of the clustering effect and requires pre-input, the indicators for evaluating the effect of the clustering result include: Sum of the Squared Errors (SSE) and Silhouette Coefficient (Silhouette Coefficient) And CH indicator (Calinski-Harabaz) and so on. In this paper, the contour coefficient calculation formula draws the contour coefficient curve, and then finds the optimal clustering result of the data. By traversing the K values from 2 to 9, it is found that when K=4, the contour coefficient is the largest and the clustering effect is the best.

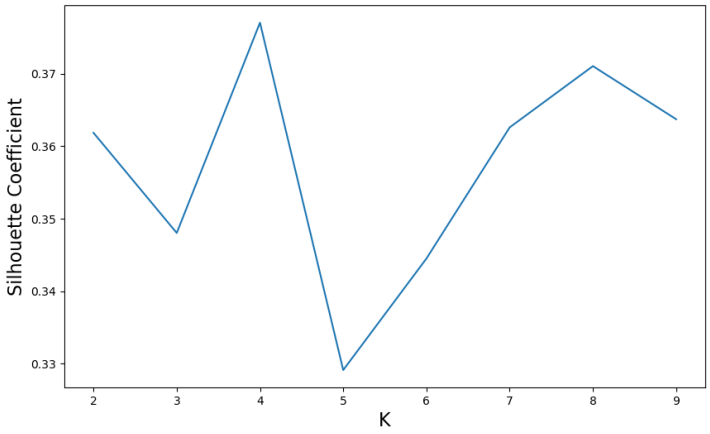

Figure 3 Contour factor

4. Conclusions

This paper describes the method of data smoothing and K-means clustering algorithm and takes LanGuang development stock as an example to show the statistical analysis method. The research work has brought about a discovery of the change cycle of some stocks that do have periodicity. We can find this periodicity by using the method in this paper, which proves that the method proposed in this paper is effective. The findings should make an important contribution to the field of statistical analysis. The novelty of it is that using the clustering algorithm of machine learning to analyze domestic stock price trends from the perspective of the change cycle. In addition, the research of this paper has practical significance for the general investors or makers. Because we can make statistical analysis of specific domestic stocks through the method in this paper. The visual results are very helpful for the investors who want to predict the stock price or the bankers who want to operate.

References

[1] Xu, C. (2020). Summary of Stock Price Forecasting Methods. China Market, 4(09), 42-43+68.
[2] Shiba, T., & Takeji, Y. (1994). Asset price prediction using seasonal decomposition. Financial Engineering and The Japanese Markets, 1(1), 37-53.
[3] Zhang, S. (1995). A Mathematical Model for Increment Ratio of Rational Function and Its

ISSN 2616-5775 Vol. 4, Issue 8: 42-47, DOI: 10.25236/AJCIS.2021.040808

Application in Share Prediction. Journal Of Wuhan University of Technology(Transportation Science & Engineering), 19(1).
[4] Fan, D., Lau, K., & Leung, P. (1996). Combining ordinal forecasts with an application in a financial market. Journal Of Forecasting, 15(1), 37-48.
[5] Liu, W., Liu, B., & Zhang, W. (2002). Uselessness of ARFIMA Model in Forecasting Chinese Stock Market. Journal Of Systems & Management, (02), 94-100.
[6] Guo, G., Chen, L., Luan, C., & Lu, Z. (2003). Application of Regression Analysis to the Construction of New Stock' s Price Mode. Journal Of South China University Of Technology (Natural Science Edition), 31(3), 57-59.
[7] Feng, P., & Cao, X. (2011). An Empirical Study on the Stock Price Analysis and Prediction Based on ARMA Model. Mathematics In Practice and Theory, 41(22), 84-90.
[8] Zhang, B. (2012). Trend Analysis and Forecast of Shanghai A-share Index. Modern Business, 4(21), 45-47.
[9] Wu, Y., & Wen, X. (2016). Short Term Stock Price Forecast Based on ARIMA Model. Statistics & Decision, (23), 83-86.
[10] Li, X. (2019). Application of Multiple Linear Regression and Time Series Model in Stock Forecasting. Pioneering With Science & Technology Monthly, 32(02), 153-155.
[11] Hu, X., Han, D., & Zhu, W. (1997). Regression Markov Chain Analysis and Prediction of Stock Price. Forecasting, (05), 67-69+73.
[12] Xia, L., & Huang, Z. (2003). Application of Markov Chain in Stock Price Forecasting. Commercial Research, (10),62-65
[13] Seo,J&Jang,H.(2004) .A Development for Short-term Stock Forecasting on Learning Agent System using Decision Tree Algorithm[J]. Journal of the Korea Safety Management and Science,6(2).
[14] Bai, M., & Liu, W. (2009). Study on Stock Prediction Based on Clustering Technology. World Sci-Tech R & D, 31(03), 553-555.
[15] Yang, X., & Huang, X. (2010). Study about Application of Stock Price Forecasting Based on Support Vector Machine. Computer Simulation, 27(09), 302-305.
[16] Lv, Q. (2011). The Stock Market Forecast System Based on SVM. Journal Of Jilin Teachers Institute of Engineering and Technology, 27(07), 48-49.
[17] Xu, X., & Yan, G. (2011). Stock Price Trend Analysis Based on BP Neural Network. Zhejiang Finance, (11), 57-59+64.
[18] Deng, N., & Su, W. (2012). The Case Study of the Price Prediction in Stock Market Based on Data Mining. Technology And Enterprise, (18), 272-274.
[19] Xie, G. (2012). Short—term Forecasting of Stock Price Based on Support Vector Regression. Computer Simulation, 29(04), 379-382.
[20] Zhang, Q., & Zhu, H. (2013). Application of Grey Model and Neural Network in Stock Prediction. Computer Engineering and Applications, 49(12), 242-245.
[21] Han, S., & Tan, S. (2018). Design and Implementation of Deep Learning Model for Stock Forecasting Based on TensorFlow. Computer Applications and Software, 35(6), 267-271+291.
[22] Wang, Y., Chen, D., & Tang, Y. (2019). A Stock Prediction Model Based on Cart and Boosting Algorithm. Journal Of Harbin University of Science and Technology, 24(06), 98-103.
[23] Hu, Y. (2021). Stock Forecast Based on Optimized LSTM Model. Computer Science, 48(S1), 151-157.
[24] Prachyachuwong, K., & Vateekul, P. (2021). Stock Trend Prediction Using Deep Learning Approach on Technical Indicator and Industrial Specific Information. Information, 12(6), 250.
[25] Rasheed F, Alhajj R. Periodic pattern analysis of non-uniformly sampled stock market data. Intelligent data analysis. 2012;16(6):993-1011. doi:10.3233/IDA-2012-00563.
[26] Dhillon, I., Guan, Y., & Kulis, B. (2004). Kernel k-means. Proceedings Of the 2004 ACM SIGKDD International Conference on Knowledge Discovery And Data Mining - KDD '04.
[27] Yu, L., Pan, Y., & Wu, Y. Academic journal comprehensive evaluation data standardization method research [J]. Books intelligence work,2009,53(12):136-139.
[28] Xiong, Z. (2016). A Study on Cluster Analysis of Comprehensive Stock Index Data Based on K-means (degree of Master). Shanghai Jiao Tong University.