Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 2, No. 2, 2022

Pattern Classification of Stock Price Moving

Chenyu Wang

Department of TikTok e-commerce, Bytedance, Shanghai, China
Email: 121220091@smail.nju.edu.cn

Abstract: The stock is one of the most important instruments of finance. However, the tendency of stock always has a high level of irregularity. In stock market, the stock price moving is considered as a time series problem. Clustering method on stock data is one of the machine learning methods and it is one of the most important analysis methods of technical analysis. The aim of this project is to find an efficient unsupervised learning way to analysis the stock market data to make classification of the patterns on different stock price moving data and get useful information for investment decisions by implementing different clustering algorithms. For this aim, the research objective of this project is to compare several of clustering methods like K-means algorithm, EM algorithm, Canopy algorithm, specify the best number of clusters for each clustering method by several evaluation indexes, show the result of each clustering method and make evaluation on the results of these clustering methods on stock market data of standard S&P 500 stock marketing data. In addition, Weka 3 and Matlab are used to implement the clustering methods and evaluation program. Data visualization shows clearly that those public companies in the same cluster have similar stock price moving pattern. The experiment shows the result that K-means algorithm and EM algorithm perform effectively in stock price moving and Canopy algorithm can be used before K-means algorithm to improve the efficiency.

Keywords: Pattern Classification; Clustering; Evaluation index.

1. Introduction

Due to the high negotiability of stock, the stock is one of the most important instruments of finance. When the investors are making decision on stocks, it is fatal for them to have a better understanding on the stock price moving analysis. Because of the theory of demand and supply, the stock price data is easily influenced by the demand of market, and the demand of market is influenced by complex factors too. So that the tendency of stock always has a high level of irregularity.

Stock is easily influenced by theory of demand and supply and stock price is sensitive to market demand. Thus, it is a valuable subject to do research on stock price moving. It is worth considering that the stock market only shows the public data of stocks and the investors cannot directly observe and judge such a great scale of high dimension stock data. On the other hand, the different stocks which have similar tendency often have similar pattern.

Due to this situation, it is worth using different analysis methods such as clustering methods to find the potential information from the public stock market data so that it can help making investment plan. Effective investment portfolio is usually composed of the stocks which have high quality. And the first and important step is to dig out those stocks which have high quality, by an efficient method.

The aim of this project is to investigate and find an efficient unsupervised learning way to analysis the stock market data for making classification of different stocks and get useful information for investment decisions by implementing different clustering algorithms.

In this project, all the stock price data are collected from the standard S & P 500 public companies and several appropriate clustering methods are implemented to process on the data set. 4 evaluation indexes are used to evaluate the performance of different clustering methods.

2. Methodology and analysis

2.1. Research overview

The pattern classification of stock price moving is generally a time series clustering problem. In 2010, Hwang H and Oh J have described the stock price moving problem as a time series problem and used their fuzzy model to make analysis on the stock market [1]. Normally, the stock of public company can be classified into different kinds such as blue chip or trash stock. And in 2016, Chen T and Chen F have described their theory on pattern recognition on stock market for the investors to decide their best investment portfolio and reduce the risk [2].

As a time, series problem, some of different stocks often have the same pattern that they have similar tendency of stock price moving. In 2017, Nair B B, Kumar P K S and Sakthivel N R have briefly presented that the investors are more willing to buy those stocks with good performance that if they have known one of the stocks which have been proved to have a good performance, there is a high probability that another stock has the similar tendency will perform well too [3]. Choosing the stock which has good performance and build appropriate investment portfolio is the important guarantee of achieving good profit in the future.

However, in the stock market, there are quite a big number of stocks and each stock in the market changes dynamically on both trading volume and stock price. Thus, it is a valuable research problem to choose those stocks with good performance and build a reasonable investment portfolio. Ghadhab I have come up a conclusion in his research in 2016 that the trading volume, or say volume of transaction can only review the notability of each stock but not the tendency, and for this reason, the project uses stock price moving data to cluster different stocks [4].

From the great number of researches in the past, it is a feasible scheme to classify the stock and choose those stocks with good financial health by making use of the public

financial information of listed companies. Xu M, Lan Y and Jiang D have discussed the advantage of public financial information of listed companies in their academic paper in 2015 [5]. And then, they have also put forward that there are several problems for this kind of scheme which both considers public financial information and stock price moving. On the one hand, some researcher such as Wang J Y and Zhu Z X have discussed how to use Principal Component Analysis to measure public financial information due to the great number of indexes in 2017 [6]. And in another research dissertation in 2016, Gao T, Li X and Chai Y have implemented a two-dimensional deep learning way which includes principal component analysis on their stock price moving prediction system [7]. However, as the researchers put forward the opinion that the public financial information is a kind of nonlinear data and sometimes because of different purpose of each company, the financial information is incomplete and sometimes even not completely right. Therefore, most of the researches consider stock price moving as the main research data of clustering.

There are two main analysis orientations for the stock price moving analysis, and they are financial data analysis and technical analysis of which the advantages and disadvantages have been discussed by Cabañas R, Martínez A M, Masegosa A R in the year 2016 [8]. Financial data analysis is aimed to make evaluation on the potential value of stocks and in the year 2016, Papavassiliou V G put forward his improved algorithm on the individual stock price moving data by using financial data analysis and he has achieved a good result [9]. On the other hand, technical analysis is aimed to make analysis on the statistical data of marketing activity and it does not make assessment on the potential value of stocks. In the year 2017, Nazário R T F, e Silva J L and Sobreiro V A have briefly discussed the advantage and weakness of technical analysis which is implemented to make analysis on the stock price moving data [10]. In technical analysis, the researchers usually try to find patterns of tendency by using figures and tables. The experiment in this project is one of the technical analysis methods.

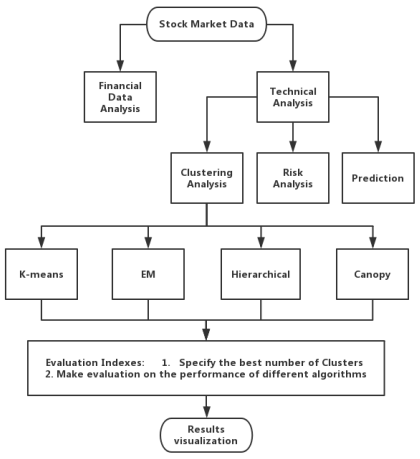


Figure 1. Brief flow diagram of literature review

The figure 1 below is used to show different kinds of analytical methods which are used to make analysis on the stock price moving data. And the next part of literature review would focus on the clustering analysis which belongs to technical analysis and have 4 main clustering algorithms: K – means algorithm, EM algorithm, Hierarchical algorithm, and Canopy algorithm. And then, several different evaluation indexes are discussed in order to specify the best number of clusters and make evaluation on the performance of different algorithms. After all, the results of different clustering methods are shown visually.

2.2. Technical analysis

In technical analysis, there are several different ways to do research on stock price moving such as prediction, risk analysis and clustering analysis and they have been discussed by Hilkevics S and Zablockis A in 2016 [11]. They have shown that prediction and risk analysis are two useful method on portfolio optimization. The prediction and risk analysis methods are briefly reviewed below, and the project mainly implements clustering analysis method.

Some of the researchers have discussed that implementing data mining method on stock price prediction. This prediction method relies on the development of data mining technology and analysis mechanism. In 2014, Wei L Y, Cheng C H and Wu H H have briefly established a prediction model and use the MATI predicting method to make prediction on the future stock price moving of TAIEX [12]. They have proved it to be an efficient prediction model. On the other hand, they put forward the opinion that for the investors, predicting the stock price moving accurately is one of the most useful way to help them making better investment portfolio. Most of the investors may utilize the historical stock price moving data to make prediction on the future tendency, but it usually does not work well. What is worth mentioning that, the investors can make deep analysis on historical data and establish efficient prediction model with the help of data mining technology. Due to the variety and complexity of stock market data, the data mining technology can be used to predict the future tendency by implementing decision tree classification algorithm or neural network algorithms. In 2016, Billah M, Waheed S, Hanifa A has used an improved neural network algorithm to make prediction on the stock price moving data [13]. Al Nasseri A, Tucker A, de Cesare S found an effective application on financial market by using decision tree algorithm in 2015 [14]. Al Nasseri A, Tucker A, de Cesare S have found that the decision tree algorithms have some disadvantages that it causes fragment problem easily, not stable enough and easy to get into local optimal.

There are also many researches on risk analysis of stock price moving data because the investors must undertake the risk of error in investment. In some way, stock margin trading has a characteristic of ‘throw out a minnow to catch a whale’. In 2004, Huang Y C and Lin B J have made analysis on Taiwan stock market by conditional risk analysis. They also explored and compare different risk analysis methods and found that the value at risk models performs better [15]. Considering an investment portfolio without stock margin trading has a profit of ‘r ‘. After the stock margin trading, considering the initial margin radio to be ‘a ‘and $0 < a < 1$. Then the real profit is ‘r / a ‘. Because ‘a ‘is from 0 to 1, the stock margin trading has a leverage effect which is discussed by Su J B in 2014 [16], and the scale of this leverage effect is influenced by the initial margin radio. Su J B has also discussed long memory and distribution effect of stock price moving data when he implements risk analysis on the stock market. What is more, Smith G P has discussed his opinion on leverage effect on individual stock price moving data and found that the leverage effect is one of the most essential mechanism in stock market in 2015 [17]. When the number of initial margin radio is smaller, this leverage effect is bigger and the risk level of investors are higher. In the risk analysis,

the investors are trying to choose those stocks which have a good development and smaller initial margin radio to reduce the risk of investment. In 2017, Mensi W, Hammoudeh S, Kang S H have shown their researches on the investment portfolio risk analysis on the stock price moving data and provided their method for make evaluation on investment portfolios [18].

In addition, some other algorithms, such as association rules and artificial neural networks are also common in use in some researches before. Association rules is a category of data mining technique and the Apriori algorithm is one of them which mines knowledge from historical data as knowledge patterns. The basic idea of association rules is discussed by Park J S, Chen M S and Yu P S in 1995 that, for example, there is a relationship between item A and B if a man who buys A and B at the same time and this rule is also appropriate to stock price moving too [19]. So association rule research is sometimes based on user's behavior in some way. What is more, Asadifar S and Kahani M have put forward their improved research achievement that they implement semantic data mining for prediction of stock price moving data by using association rules in 2017 [20]. In 2017, Selvanambi R and Natarajan J have applied their improved Apriori algorithm on performance evaluation which is used for association rule mining on the stock price moving data [21]. SOM algorithm is one of the artificial neural networks that imitates the function of the human brains that categorize items by groups and it is also well known as a self-organization algorithm. Isa D, Kallimani V P and Lee L H have discussed that SOM algorithm is measured to be useful in classification of text documents in 2009 [22]. And in 2007, the SOM algorithm is used in predicting stock price moving by Afolabi M O and Olude O [23]. It remains to be confirmed whether the SOM algorithm is suitable for classification of stock price moving data. On the other side, Lertyingyod W and Benjamas N have put forward their evaluation models on the stock price moving by implementing artificial neural networks in 2016 [24]. They are aimed to show the investors much more information but they do not make the decision for the investors. Their work implies that historical stock price moving data can be used to make classification on different stocks for the investors to come up their own decisions and make their best investment portfolio. In addition, there are some other existing methods that have been consider such as chaotic map synchronization by Basalto N, Bellotti R, De Carlo F in 2005 and enhanced index tracking portfolio by Dose C and Cincotti S in 2005 [25] [26]. However, it is not easy to assess the performance between these methods because different evaluation index benefits different method.

2.3. Clustering algorithms

Clustering method is one of the technical analysis methods. To do research on clustering methods is a hot issue for study in the data mining area and machine learning area. Thus, a great number of researches have been done aiming at giving a method of high performance and the academic researchers often use data mining or machine learning methods from historical data. Some of the researchers are aiming at improve existing clustering methods or put forward another new clustering method. For the traditional clustering method, such as K-means algorithm or CURE algorithm, the number of clusters are need to be specified by the researcher before implementing the clustering process. However, in the real stock price moving data, the number of clusters are unknown so that an appropriate number of experiments are useful to achieve the best number of clusters. The traditional clustering methods are generally suitable under certain condition, and there is not any clustering method which can be applied to all the cases. Now there are many clustering methods which can deal with a small amount of data and low dimension data. However, with the arrival of big data and massive amounts of information, it is a large and careful work to deal with the big data and some of the traditional methods are low efficiency in this case. So that it is important to consider the time cost of each program. It is worth mentioning that some of the clustering methods are theoretical and always under some kind of hypothesis such as no extreme value or the clusters can be separated easily. But the data in reality is usually complex enough and has a high level of noise. Thus, it is another important research problem to deal with how to eliminate the influence of noise. Clustering method on stock data is one of the machine learning methods and it is one of the most important analysis methods of technical analysis.

One of the research issues is to classify the different patterns of stock price moving. In other word, there are several different patterns such as rising or falling and each stock price moving has a pattern. There are two main sub problems. The first problem is that for each clustering method, what the best number of clusters is. And the second problem is that, for each stock, how to assign each stock into different clusters. Most researchers consider stock price moving analysis as a time series problem and there are many related works, in which clustering is considered (by Nanda, S. , Mahanty, B. and Tiwari, M. 2010) as a good general method for analysis of stock price moving that clustering methods classify a data matrix into several different classifications, in which data points belonging to one classification are similar to each other and not similar to those data points belonging to other classifications [27]. Clustering, as defined by Mirkin B in 1996, is a mathematical technique which is designed for the purpose of finding a scientific way to reveal the potential classification structure for the data which is collected or created from the real-world phenomena [28]. There is quite a number of researches showing using clustering to deal with stock market is a good choice and there are several different well-known clustering algorithms, such as K-means, EM, Hierarchical, Canopy, Cobweb, Farthest First, Filtered, Make Density Based clustering methods.

2.3.1. K – means clustering algorithm

The main research problems of K-means algorithm is shown in the figure 2 below.

Clustering method such as K-means can be used to set k different patterns from n stock price movings and classify each stock into a pattern which belongs to the k different patterns.

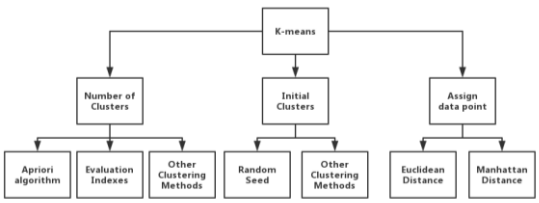


Figure 2. Main research problems of K-means algorithm

K-means algorithm is considered as one of the best basic clustering methods and in addition, it has been discussed in several studies. The number of clusters can be specified by Apriori algorithm or other clustering methods or in another

way, specified by evaluation indexes. In 2016, Sarma H K D, Mishra S have put forward their improvement on the research that to use Apriori algorithm to specify the best number of clusters on the time series data [29]. On the one hand, Cardoso M G M S, de Carvalho A P L have made their evaluation on the quality indexes which are used to make comparison on the performance of different clustering algorithms in 2009 and they draw the conclusion that K-means algorithm is a general and efficient algorithm in many cases and some of the evaluation indexes can be used to specify the best number of clusters on the stock price moving data [30]. On the other hand, in 2015, Amorim M J and Cardoso M G M S made their research on different adjusted paired indexes and used them to compare the performance on different clustering results [31] and after comparison, they have also concluded that K-means is efficient in many cases and some of the evaluation indexes can be used to specify the best number of clusters on the stock price moving data. For assigning each data point into different clusters, K-means algorithm can be used to implement both of the Euclidean distance and the Manhattan distance in order to calculate the distance between each data point and the center of each cluster. In 2014, Li L, He J and Sui X have discussed their conclusion that the Euclidean distance is useful in establishing analysis model on stock price moving data [32]. And in 2016, Kapil S and Chawla M have discussed several different distance metrics which are used to make performance evaluation on the clustering methods such as the most famous clustering algorithm, like K-means algorithm in detail [33]. The basic K-means algorithm uses seed to calculate the initial clusters and the number of initial clusters as it has been discussed by Arthur D, Vassilvitskii S in 2007 [34]. Zhu Z has implemented improved K-means algorithm which is based on the combination of independent component analysis in analyzing stock market data in 2016 [35]. Some of the improved K-means algorithm use other clustering method to calculate the initial clusters. In 2017, Kumar K M and Reddy A R M have improved the K-means algorithm by implementing density based clustering to calculate the initial cluster before the process of K-means [36]. What is more, in 2016, Xiong C, Hua Z and Lv K have improved K-means algorithm by optimizing initial cluster of the K-means algorithm too [37].

2.3.2. EM clustering algorithm

EM algorithm is based on the rule that, for each instance, it assigns a probability distribution value which can indicate the probability of it belonging to each of the clusters. In 1987, Hoenig J M and Heisey D M have described the method that using EM algorithm to make estimate on the stock price moving data [38]. In 2008, Shinozaki T and Ostendorf M put forward their maximum Log likelihood index evaluation algorithm to improve the EM clustering algorithm which has a better performance than the original EM clustering algorithm [39]. In 1991, Coakley K J described his cross-validation procedure which can be used to specify the best number of clusters in the EM clustering algorithm [40]. By implementing cross validation and using the Log likelihood index, the EM clustering can decide the best number of clusters to create itself which is implemented by Dempster A P, Laird N M, Rubin D B in 1977 [41]. In another way, Apriori algorithm can also be used to specify how many clusters to generate. In 2016, Sarma H K D, Mishra S have put forward their improvement on the research that to use Apriori algorithm to do data mining on the time series data [29]. And in 2003, EM algorithm is implemented to be used in market research by Karlis D and it is proved to have a better performance than traditional Hierarchical clustering algorithm on the stock market [42]. EM algorithm is a useful algorithm to calculate the probability of a data point belonging to a cluster from the data set which contains incomplete data or missing data.

Canopy clustering algorithm and Hierarchical clustering algorithm

Hierarchical clustering is a traditional clustering method and there are lots of researches related. It is one of the clustering methods that calculate the similarity between different classifications of data points and create a clustering tree which has many levels. It always combines the most similar two data points together by calculating the distance such as Euclidean distance. There are mainly 3 different linkage types to calculate the distance between two data point sets and they are single linkage, complete linkage and average linkage. In 2017, Zhu D, Guralnik D P and Wang X have shown the advantages and disadvantages of single linkage in hierarchical clustering algorithm in unsupervised data analysis [43]. And Großwendt A, Röglin H have briefly described their improved complete linkage method in the hierarchical clustering algorithm in 2017 [44]. In 2016, Gagolewski M, Bartoszuk M and Cena A improved the average linkage method in Hierarchical clustering algorithm [45]. In 1981, Srivastava R K, Leone R P and Shocker A D discussed how to use Hierarchical clustering to make analysis on market products [46]. In 2016, Lahmiri S has discussed how to examine the co-movement of stock data by implementing hierarchical clustering but hierarchical algorithm is considered to be not very useful in clustering of stock price moving data [47]. It is considered that clustering is a kind of unsupervised learning algorithms that there is no definitely correct or incorrect classification on the stock price moving data. Because of this situation, there are quite a number of indexes being created to make evaluation on the performance of different clustering methods.

The principle of Canopy clustering algorithm is to firstly classify the data set into several groups call canopy by approximate distance algorithm. And then, it uses strict distance algorithm to calculate the data points in the same canopy and assign them in the most appropriate cluster. In 2014, Sharma S and Tiwari R discussed their improved Canopy algorithm [48]. In 2012, He H, Guo L and Geng Y have proved that using Canopy algorithm before K-means algorithm have a better performance than the traditional K-means algorithm [49]. In this project, the canopy clustering algorithm is used to specify the best number of clusters of K-means clustering algorithm and calculate the initial cluster centers of K-means clustering algorithm too.

2.4. Evaluation indexes

There are several evaluation indexes to provide validity measures for each classification in order to specify the best number of clusters and also evaluate the performance of clustering methods. The evaluation indexes can also provide a clear line chart that shows the performance of each clustering method. One of the simplest way is to use the list index which is defined by Wang Y F, Chuang Y L and Hsu M H in 2004 [50]. There are also various slightly more complex evaluation indexes and functions. The Calinski criterion defined by Shu G, Zeng B, Chen Y P in 2003 calculates the similarity of data points which are within the same cluster and the dissimilarity of those who are not in the same cluster in

order to evaluate the clustering method [51]. The Davies Bouldin index defined by Kasturi J, Acharya R, Ramanathan M in 2003 and the Silhouetee index defined by Chen G, Jaradat S A, Banerjee N in 2002 are also two useful indexes [52][53]. There are also Dunn's index defined by Bezdek J C, Pal N R in 1995, Krzanowski Lai index, Alternative Dunn index, Xie and Beni's index, Partition index, Separation index (Raghuvanshi A S, Tiwari S, Tripathi R, 2009) [54][55]. Because some of the indexes have similar functions, the project selects several typical indexes among them.

2.5. Critical analysis

As there are many methods to be implemented for analysis of stock price moving, but there is seldom research of using appropriate evaluation indexes to specify the best number and make evaluation on the clusters of K-means algorithm, EM algorithm, Canopy algorithm + K-means algorithm on pattern classification of stock price moving. So the project will focus on implementing K-means algorithm, EM algorithm, and Canopy algorithm to make analysis on the stock price moving data. In this project, the Calinski criterion index, Davies Bouldin index, Silhouetee index and Gap index are implemented in the project to specify the best number of clusters of K-means algorithm. The Log likelihood index is used to specify the best number of clusters of EM clustering.

2.6. Research procedure

The first stage is Data processing and it has 4 steps. The first step is clarifying the problem and achieving the data and it means that achieving the relational data for concrete analysis based on the specific problem. The second step is tools select. In this project, the Matlab, Weka 3 are chosen as main softwares. The third step is data process method selection. As it is shown in the research procedure, the K-means, EM and Canopy clustering algorithm are chosen as the main analysis methods and it is discussed in the later part that the Calinsky criterion, Davies Bouldin index, Gap index, Silhouette index and log likelihood index are chosen to be the evaluation indexes. The fourth step is Data normalization and it means that both doing normalization on the original data and eliminating incomparability which is derived from the difference between indexes. After the first stage, the original data has been transformed into a data set (matrix) which is easily to be used to implement clustering method in programs.

The second stage is implementing clustering algorithms and it has 3 parts running at the same time that each part is independent from each other. The first part is K-means analysis and it means that using K-means to do clustering on the data matrix after pre-processing and record the results. The second part is EM analysis and it means that implementing EM clustering method on the data matrix. The third part is Canopy + K-means analysis and it means using the Canopy algorithm to calculate the best number of clusters and initial clusters which can be used in the K-means algorithm.

The third stage is experiment results analysis and evaluation, that the K-means, EM clustering methods and Canopy + K-means algorithm are three main parts of evaluation. Because clustering for stock market price moving is an unsupervised learning problem, it is difficult to demonstrably prove that for example, which one of K-means and EM clustering method is more accurate than the other one. The indexes for evaluation on the performance of K-means algorithm, EM algorithm, and Canopy plus K-means method are Calinsky criterion, Davies Bouldin index, Gap index, Silhouette index. The same indexes are used to specify the best number of clusters of K-means algorithm and Canopy plus K-means algorithm too. Different from them, the index to specify the best number of clusters on EM method is Log Likelihood index. The time cost is another reasonably important index for these methods. Because this project involves an unsupervised learning problem, the clustering results would have some visualization graphs which are presented in the experiment result analysis and evaluation part to show the performance of different clustering methods.

3. Results and evaluation

3.1. K – means analysis

The first clustering method is K-means algorithm to generate disjoint and homogeneous clusters from a large set of data. The main process of K-means algorithm is in the following order.

In the first step, there are number K of initial clusters and each of them has a data point which is randomly selected from all the number N of original data points. In this step, the algorithm used in randomly choosing data point can be replaced by some other algorithms. The value of K is specified by Apriori algorithm and the initial clusters is generated by using randomly seed to call the table of data points.

In the second step, there are totally number N - K data points remaining and for each of these data points, using one of the distance functions to calculate the distance between one data point and all the center of the clusters. The purpose of this step is to calculate the distance that from each data point to each cluster, and then assigning each data point to its nearest cluster. The distance function which is used to calculate the distance can be replaced too. The most popular distance function is Euclidean Distance which is used in the project. There are also some other distance functions such as Chebyshev Distance, Filtered Distance, Manhattan Distance and Minkowski Distance. These distance functions are not implemented in the project.

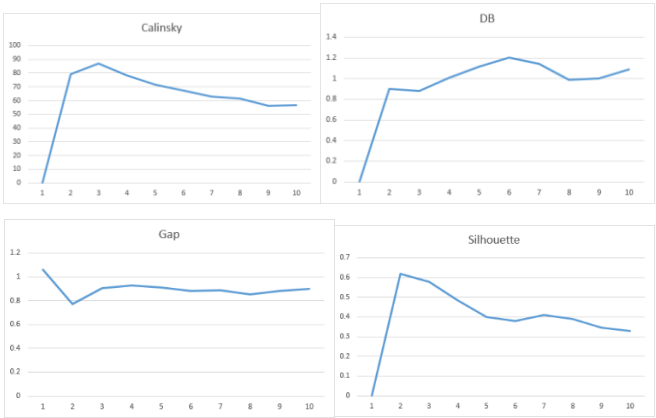


Figure 3. K-means

In the third step, the center of each cluster is recomputed and like what the program does in the step two, the distance between each data point and each center of cluster is calculated by the same distance function. After that, by judging whether each data point is in the nearest cluster or not, all the data points are reassigned to its nearest cluster again. The step two and step three repeat again and again until all the data points have been assigned to the best cluster. This

project implements the K-means algorithm on the data set, and calculates the evaluation indexes by MATLAB.

The best number of clusters and 4 evaluation indexes are shown in the figure 3 below.

To sum up, the best number of clusters of K-means algorithm is specified to be 3. Further detailed analysis data and the visualization results of K-means and the comparison of different clustering methods are shown in the section 3.4 later.

3.2. EM analysis

EM is the second clustering method used in this project. For each data point, the EM algorithm assigns a probability distribution and by doing so, it indicates the probability that one data point is belonging to one cluster. As it is mentioned in the part of K-means algorithm above, there is one same problem in the EM algorithm too. The problem is that, for the purpose of achieving the best performance of EM clustering, what the best number of clusters is.

As it is the same in the K-means algorithm, the number of clusters can also be specified by some algorithms such as Apriori algorithm. The researcher can also use control variable method and draw curve graph to show the relation between number of clusters and the evaluation index. Not like the case that in K-means algorithm, EM algorithm can specify the number of clusters itself by cross validation process. And the cross validation which is used to specify the number of clusters is processed in the following steps. Firstly, the initial number of clusters is set to 1. Secondly, the whole training set is randomly split into n segments and the number n can be specified by the researcher. This number is normally 10 which is the default setting value in Weka 3. Thirdly, implementing EM algorithm on the n segments. And next, an evaluation index, which is called Log likelihood is used to evaluate the performance of each cluster and the total Log likelihood is averaged over all the n results. And then, if the total Log likelihood index has increased, the program is going to the next circulation which is from step 2 to step 5 and increase the number of cluster by 1.

The EM algorithm has two steps. In the E step, it calculates the posterior probability of latent variables from initial parameter or the last turn's value and in the M step, it maximums the likelihood function to gain the new parameters.

For EM, this project implements Log likelihood index and cross validation to specify the best number of clusters. After processing, the best number of clusters by EM algorithm is 5 and the value of Log likelihood index in this case is 306.21262. The time cost is longer than the other two clustering methods.

3.3. Canopy + K-means analysis

Canopy plus K-means algorithm is the third clustering method used in this project. The Canopy clustering algorithm can be implemented by only one pass over the stock price moving data. And it can also be implemented to use in both clustering by itself or in the pre-processing part of other clustering algorithms. In this project, the Canopy algorithm is used as the pre-processing part of the K-means clustering algorithm to specify the best number of clusters and calculate the initial centers of clusters too.

The principle of Canopy clustering algorithm is to firstly classify the data set into several groups that called canopy by approximate distance algorithm which is a low cost method. It can be some overlaps among these canopies. And then, it uses strict distance algorithm to calculate the data points in the same canopy and assign them in the most appropriate cluster. The detailed process of creating a canopy is shown below:

Step1. If there is a data point set called ‘ S ‘. And there are two initial distance parameter called ‘ T1 ‘ and ‘ T2 ‘, T1 is larger than T2.

Step2. Then, the Canopy clustering algorithm chooses a data point called ‘ P ‘ and calculate the distance between it and each of the other data points belonging to ‘ S ‘ by a low cost distance algorithm.

Step3. Next, the Canopy clustering algorithm puts those data points belonging to ‘ S ‘ whose distance from the ‘ P ‘ is smaller than ‘ T1 ‘ into a canopy. At the same time, the Canopy clustering algorithm removes those data points belonging to ‘ S ‘ whose distance from the ‘ P ‘ is larger than ‘ T2 ‘ from the ‘ S ‘.

Step4. The process is continuing until there is no data point in the data set ‘ S ‘.

It can be considered that the Canopy clustering algorithm is simply using circles to classify the data set and the center of each canopy is not too close because the distance between them is larger than ‘ T2 ‘. In this project, the T1 is set to be 1.25 and T2 is set to be 1.0 as the default setting.

The best number of clusters of Canopy + K-means clustering algorithm can be specified by different indexes which are the same as what are used in the K-means clustering algorithm.

The best number of clusters and 4 indexes are shown in the figure 4 below.

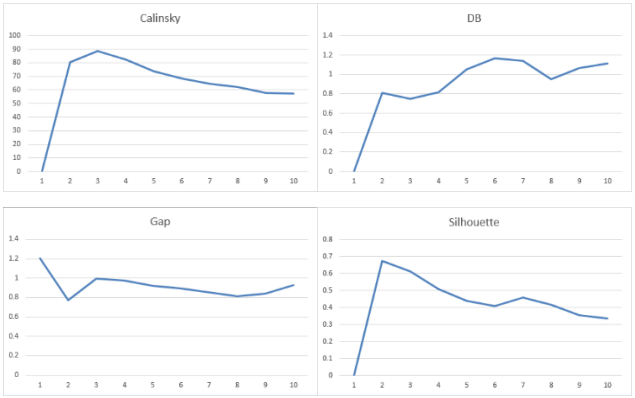


Figure 4. Canopy + K-means

To sum up, the case of 3 clusters have the best comprehensive performance due to the evaluation of these 4 indexes shown above. So the best number of clusters of Canopy + K-means algorithm is specified to be 3. And then, the visualization results of K-means and the comparison of different clustering methods are described in the section 3.4 later.

3.4. Clustering methods comparison

As the best clusters of each clustering method has been specified. The method, number of clusters and evaluation index table including time cost are shown in the table 1 below.

As it is shown in the table above, the performance of EM algorithm and Canopy plus K-means algorithm is better than the K-means algorithm because all the evaluation index are shown that K-means algorithm is not well enough. On the other hand, the Canopy + K-means clustering algorithm is to some degree having little gap with the EM clustering algorithm. The Calinsky and Davies Bouldin index show that

EM algorithm is better than the Canopy + K-means algorithm however the Gap index and Silhouette index shown that Canopy + K-means algorithm performs better than the EM clustering algorithm. However, there is another important index to make evaluation on the performance of each clustering method, and it is the time cost. As it is shown in the table above, the K-means algorithm has the lowest time cost among them and the Canopy + K-means algorithm has nearly twice time cost as the K-means clustering algorithm. The EM algorithm has a really high time cost and that is because it spends a lot of time on implementing cross validation step.

As an unsurprised learning problem, the evaluation index can only be used to measure the performance of each clustering method in one aspect. So that the result of each clustering method is shown visually below. Because it is too complex and crowded to show all the data of each cluster in one graph. The visualization graph randomly selects several different data points of each cluster.

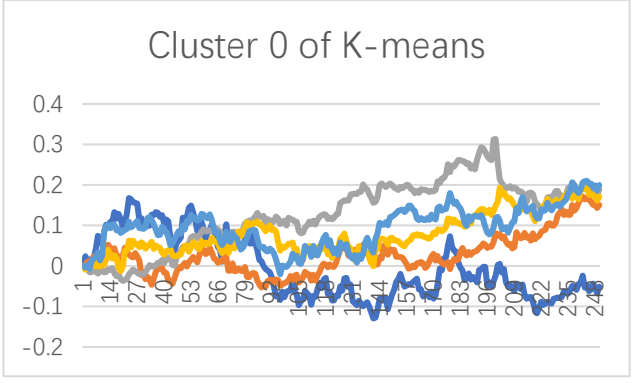
Table 1. Result

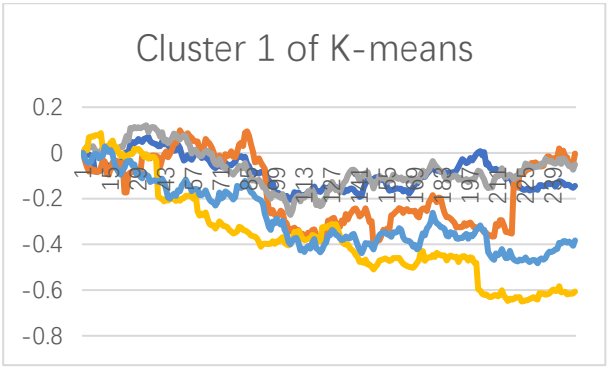
Method	Num Clusters	Index				Time cost
		Cal	DB	Gap	Sil	
K-means	1	null	null	1.0634	null	0.01
	2	79.421	0.9034	0.7744	0.6205	0.01
	3	**86.8927**	**0.881**	**0.9045**	**0.5789**	**0.01**
	4	78.3009	1.006	0.9268	0.484	0.01
	5	71.4057	1.1133	0.91	0.3985	0.01
EM	1	null	null	0.8798	null	0.01
	2	72.1659	0.9245	0.8978	0.5548	0.05
	3	83.2513	0.9413	0.9344	0.5412	0.11
	4	87.4578	0.8634	0.9726	0.5937	0.14
	5	**92.7942**	**0.7343**	**0.9832**	**0.6043**	**0.17**
Canopy + K-means	1	null	null	1.2056	null	0.02
	2	80.6157	0.8094	0.7692	0.6752	0.02
	3	**88.6259**	**0.7462**	**0.9967**	**0.6141**	**0.01**
	4	82.6124	0.8172	0.9723	0.5086	0.02
	5	73.5691	1.0528	0.9211	0.4396	0.02

The K-means algorithm creates 3 clusters and the sample are shown in the figure 5 below.

As it is shown in the figure 5, the result of K-means is not very well that some of the data points in the cluster 2 are obviously more likely to be in cluster 1. K-means clustering algorithm has the worst performance among the three clustering methods.

The EM clustering algorithm creates 5 clusters and the sample are shown in the figure 6 below.





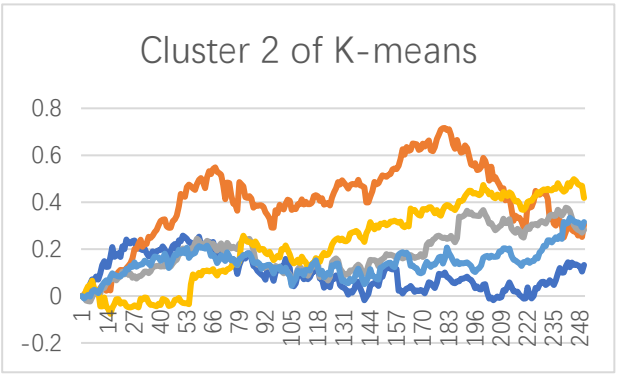
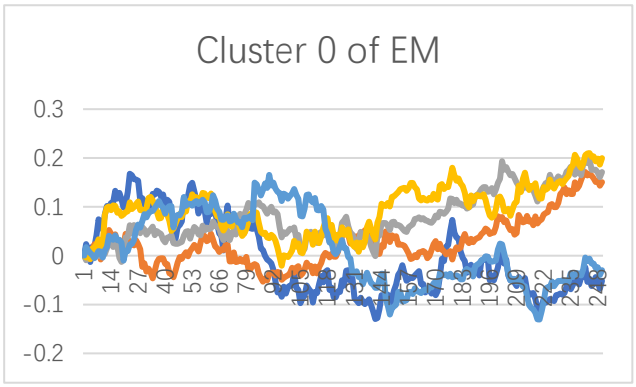
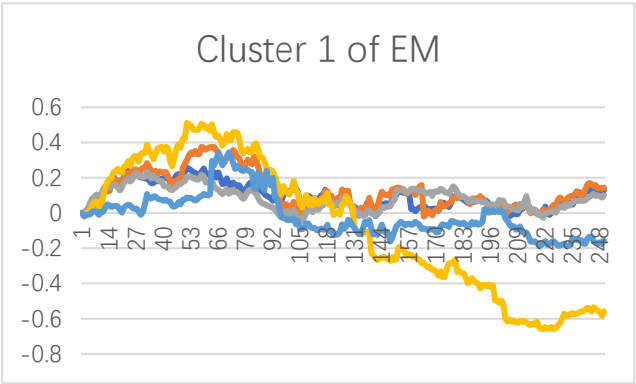
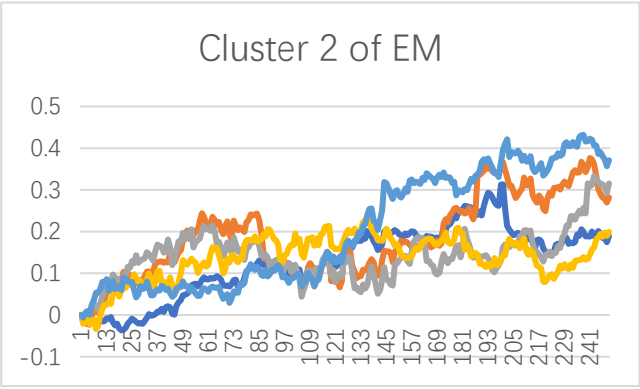
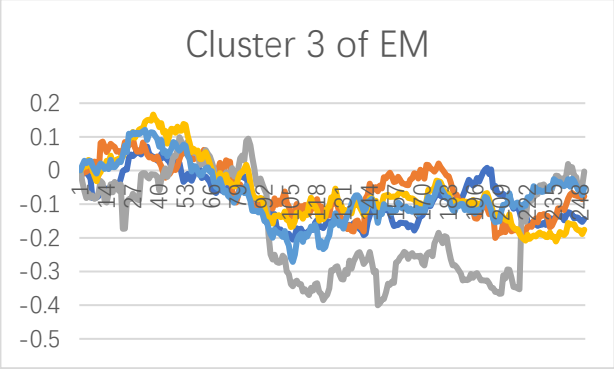


Figure 5. Visualization of K-means









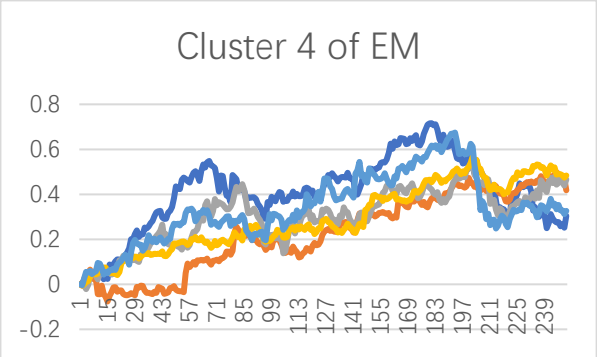
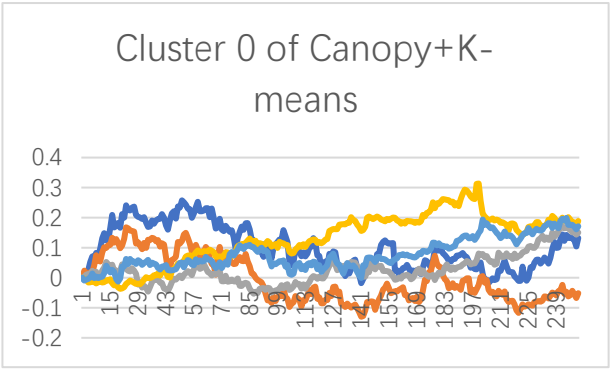


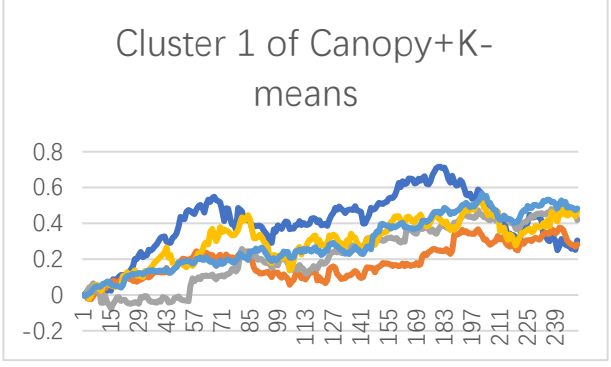
Figure 6. Visualization of EM

As it is shown in the figure 6 above, the performance of EM clustering algorithm is obviously better than the K-means clustering algorithm. What is worth noticing is that the cluster 0 of EM clustering algorithm represents those stock price having a fluctuation.

The Canopy + K-means clustering algorithm creates 3 clusters and the sample are shown in the figure 7 below.

As it is shown in the figure 7 above, the performance of Canopy + K-means clustering algorithm is obviously better than the K-means clustering algorithm. On the other hand, it is slightly inferior to the EM clustering algorithm above. However, the time cost of Canopy + K-means clustering algorithm is much lower than the EM clustering algorithm.





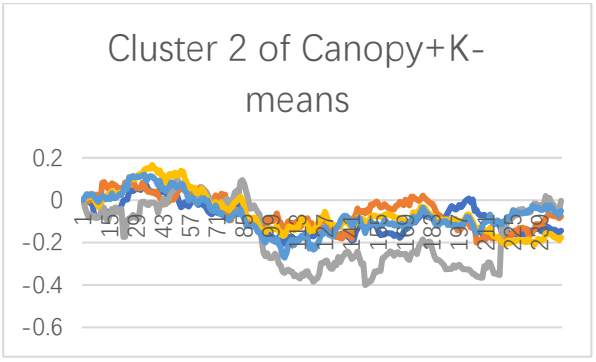


Figure 7. Visualization of Canopy + K-means

4. Conclusion

The aim of this project is to compare the performance of different clustering methods on the pattern classification of stock price moving data. To sum up, the three different clustering methods and several different indexes which are implemented in this project shows that the clustering performance of EM algorithm is better than the Canopy + K-means clustering algorithm on the stock price moving pattern classification and the K-means clustering algorithm has the worst performance. On the other hand, the Canopy + K-means clustering algorithm have a much lower time cost than the EM clustering algorithm. In addition, the project provides an efficient method to cluster the stock price moving data. It is better to choose different clustering algorithms in different circumstances. If time cost is the most important parameter, the Canopy + K-means algorithm is a useful method in pattern classification, otherwise, the EM algorithm has a better performance.

References

[1] Hwang H, Oh J. Fuzzy models for predicting time series stock price index [J]. International Journal of Control, Automation and Systems, 2010, 8(3): 702-706.

[2] Chen T, Chen F. An intelligent pattern recognition model for supporting investment decisions in stock market [J]. Information Sciences, 2016, 346: 261-274.

[3] Nair B B, Kumar P K S, Sakthivel N R, et al. Clustering stock price time series data to generate stock trading recommendations: an empirical study [J]. Expert Systems with Applications, 2017, 70: 20-36.

[4] Ghadhab I. The effect of additional foreign market presence on the trading volume of cross-listed/traded stocks [J]. Journal of Multinational Financial Management, 2016, 34: 18-27.

[5] Xu M, Lan Y, Jiang D. Unsupervised Learning Part-Based Representation for Stocks Market Prediction[C]//Computational Intelligence and Design (ISCID), 2015 8th International Symposium on. IEEE, 2015, 2: 63-66.

[6] Wang J Y, Zhu Z X. The relationship between firm characteristic variables and stock returns: An empirical study based on principal component analysis[C]//Service Systems and Service Management (ICSSSM), 2017 International Conference on. IEEE, 2017: 1-6.

[7] Gao T, Li X, Chai Y, et al. Deep learning with stock indicators and two-dimensional principal component analysis for closing price prediction system[C]//Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on. IEEE, 2016: 166-169.

[8] Cabañas R, Martínez A M, Masegosa A R, et al. Financial Data Analysis with PGMs Using AMIDST[C]//Data Mining

Workshops (ICDMW), 2016 IEEE 16th International Conference on. IEEE, 2016: 1284-1287.

[9] Papavassiliou V G. Allowing for Jump Measurements in Volatility: A High-Frequency Financial Data Analysis of Individual Stocks [J]. Bulletin of Economic Research, 2016, 68(2): 124-132.

[10] Nazário R T F, e Silva J L, Sobreiro V A, et al. A Literature Review Of Technical Analysis On Stock Markets [J]. The Quarterly Review of Economics and Finance, 2017.

[11] Hilkevics S, Zablockis A. THE COMBINATION OF FUNDAMENTAL AND TECHNICAL ANALYSIS IN PORTFOLIO OPTIMIZATION [J]. Regional Review/Regionalais Zinojums, 2016 (12).

[12] Wei L Y, Cheng C H, Wu H H. A hybrid ANFIS based on n-period moving average model to forecast TAIEX stock [J]. Applied Soft Computing, 2014, 19: 86-92.

[13] Billah M, Waheed S, Hanifa A. Stock market prediction using an improved training algorithm of neural network[C]//Electrical, Computer & Telecommunication Engineering (ICECTE), International Conference on. IEEE, 2016: 1-4.

[14] Al Nasseri A, Tucker A, de Cesare S. Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms [J]. Expert Systems With Applications, 2015, 42(23): 9192-9210.

[15] Huang Y C, Lin B J. Value-at-risk analysis for Taiwan stock index futures: fat tails and conditional asymmetries in return innovations [J]. Review of Quantitative Finance and Accounting, 2004, 22(2): 79-95.

[16] Su J B. Empirical analysis of long memory, leverage, and distribution effects for stock market risk estimates [J]. The North American Journal of Economics and Finance, 2014, 30: 1-39.

[17] Smith G P. New evidence on sources of leverage effects in individual stocks [J]. Financial Review, 2015, 50(3): 331-340.

[18] Mensi W, Hammoudeh S, Kang S H. Dynamic linkages between developed and BRICS stock markets: Portfolio risk analysis [J]. Finance Research Letters, 2017, 21: 26-33.

[19] Park J S, Chen M S, Yu P S. An effective hash-based algorithm for mining association rules [M]. ACM, 1995.

[20] Asadifar S, Kahani M. Semantic association rule mining: A new approach for stock market prediction[C]//Swarm Intelligence and Evolutionary Computation (CSIEC), 2017 2nd Conference on. IEEE, 2017: 106-111.

[21] Selvanambi R, Natarajan J. Performance Evaluation of Association Rule Mining with Enhanced Apriori Algorithm Incorporated with Artificial Bee Colony Optimization Algorithm [J], 2017.

[22] Isa D, Kallimani V P, Lee L H. Using the self organizing map for clustering of text documents [J]. Expert Systems with Applications, 2009, 36(5): 9584-9591.

[23] Afolabi M O, Olude O. Predicting stock prices using a hybrid Kohonen self organizing map (SOM) [C]//System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on. IEEE, 2007: 48-48.

[24] Lertyingyod W, Benjamas N. Stock price trend prediction using Artificial Neural Network techniques: Case study: Thailand stock exchange[C]//Computer Science and Engineering Conference (ICSEC), 2016 International. IEEE, 2016: 1-6.

[25] Basalto N, Bellotti R, De Carlo F, et al. Clustering stock market companies via chaotic map synchronization [J]. Physica A: Statistical Mechanics and its Applications, 2005, 345(1): 196-206.

[26] Dose C, Cincotti S. Clustering of financial time series with application to index and enhanced index tracking portfolio [J]. Physica A: Statistical Mechanics and its Applications, 2005, 355(1): 145-151.

[27] Nanda S R, Mahanty B, Tiwari M K. Clustering Indian stock market data for portfolio management [J]. Expert Systems with Applications, 2010, 37(12): 8793-8798.

[28] Mirkin B. Mathematical classification and clustering: From how to what and why [M]//Classification, data analysis, and data highways. Springer, Berlin, Heidelberg, 1998: 172-181.

[29] Sarma H K D, Mishra S. Mining Time Series Data with Apriori Tid Algorithm[C]//Information Technology (ICIT), 2016 International Conference on. IEEE, 2016: 160-164.

[30] Cardoso M G M S, de Carvalho A P L. Quality indices for (practical) clustering evaluation [J]. Intelligent Data Analysis, 2009, 13(5): 725-740.

[31] Amorim M J, Cardoso M G M S. Comparing clustering solutions: the use of adjusted paired indices [J]. Intelligent Data Analysis, 2015, 19(6): 1275-1296.

[32] Li L, He J, Sui X. Research on structural correlation of HS 300 stock index based on AR (n)-XARCH-Copula model[C]//Management Science & Engineering (ICMSE), 2014 International Conference on. IEEE, 2014: 1190-1194.

[33] Kapil S, Chawla M. Performance evaluation of K-means clustering algorithm with various distance metrics[C]//Power Electronics, Intelligent Control and Energy Systems (ICPEICES), IEEE International Conference on. IEEE, 2016: 1-4.

[34] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding[C]//Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007: 1027-1035.

[35] Zhu Z. A clustering method for high-dimensional data analysis in stock market [J]. RISTI (Revista Iberica de Sistemas e Tecnologias de Informacao), 2016 (17A): 116-125.

[36] Kumar K M, Reddy A R M. An Efficient k-Means Clustering Filtering Algorithm Using Density Based Initial Cluster Centers [J]. Information Sciences, 2017.

[37] Xiong C, Hua Z, Lv K, et al. An Improved K-means Text Clustering Algorithm by Optimizing Initial Cluster Centers[C]//Cloud Computing and Big Data (CCBD), 2016 7th International Conference on. IEEE, 2016: 265-268.

[38] Hoenig J M, Heisey D M. Use of a log-linear model with the EM algorithm to correct estimates of stock composition and to convert length to age [J]. Transactions of the American Fisheries Society, 1987, 116(2): 232-243.

[39] Shinozaki T, Ostendorf M. Cross-validation and aggregated EM training for robust parameter estimation [J]. Computer Speech & Language, 2008, 22(2): 185-195.

[40] Coakley K J. A cross-validation procedure for stopping the EM algorithm and deconvolution of neutron depth profiling spectra [J]. IEEE Transactions on Nuclear Science, 1991, 38(1): 9-15.

[41] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the royal statistical society. Series B (methodological), 1977: 1-38.

[42] Karlis D. An EM algorithm for multivariate Poisson distribution and related models [J]. Journal of Applied Statistics, 2003, 30(1): 63-77.

[43] Zhu D, Guralnik D P, Wang X, et al. Statistical properties of the single linkage hierarchical clustering estimator [J]. Journal of Statistical Planning and Inference, 2017, 185: 15-28.

[44] Großwendt A, Röglin H. Improved Analysis of Complete-Linkage Clustering [J]. Algorithmica, 2017, 78(4): 1131-1150.

[45] Gagolewski M, Bartoszuk M, Cena A. Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm [J]. Information Sciences, 2016, 363: 8-23.

[46] Srivastava R K, Leone R P, Shocker A D. Market structure analysis: hierarchical clustering of products based on substitution-in-use [J]. The Journal of Marketing, 1981: 38-48.

[47] Lahmiri S. Clustering of Casablanca stock market based on hurst exponent estimates [J]. Physica A: Statistical Mechanics and its Applications, 2016, 456: 310-318.

[48] Sharma S, Tiwari R. Canopy Clustering Based Multi Robot Area Exploration [J]. IFAC Proceedings Volumes, 2014, 47(1): 505-510.

[49] He H, Guo L, Geng Y. The Optimization of CMAC Neural Network Structure Based on Canopy-k-means Algorithm [J]. International Journal of Advancements in Computing Technology, 2012, 4(22).

[50] Wang Y F, Chuang Y L, Hsu M H, et al. A personalized recommender system for the cosmetic business [J]. Expert Systems with Applications, 2004, 26(3): 427-434.

[51] Shu G, Zeng B, Chen Y P, et al. Performance assessment of kernel density clustering for gene expression profile data[J]. Comparative and Functional Genomics, 2003, 4(3): 287-299.

[52] Kasturi J, Acharya R, Ramanathan M. An information theoretic approach for analyzing temporal patterns of gene expression [J]. Bioinformatics, 2003, 19(4): 449-458.

[53] Chen G, Jaradat S A, Banerjee N, et al. Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data [J]. Statistica Sinica, 2002: 241-262.

[54] Bezdek J C, Pal N R. Cluster validation with generalized Dunn's indices[C]//Artificial Neural Networks and Expert Systems, 1995. Proceedings., Second New Zealand International Two-Stream Conference on. IEEE, 1995: 190-193.

[55] Raghuvanshi A S, Tiwari S, Tripathi R, et al. GK clustering approach to determine optimal number of clusters for wireless sensor networks[C]//Wireless Communication and Sensor Networks (WCSN), 2009 Fifth IEEE Conference on. IEEE, 2009: 1-6.

[56] http://www.standardandpoors.com/en_US/web/guest/home?pagename=sp/Page/IndicesIndexPg&r=1&b=4&s=6&ig=51&l=EN&i=56&xcd=500

[57] http://www.cs.waikato.ac.nz/ml/weka/