Clustering Analysis of Stock Volume and Price Relationship based on Gaussian Mixture Model

Yaohui Bai, Jianwu Dang
Cloud Computation and Big Data Research Center
Jiangxi University of Finance and Economics
Nanchang, 330013, China
e-mail: byhnpu@163.com, dangjianwu@126.com

Abstract—The research of stock price volatility is very important. Traditionally, the stock price is usually processed as a time series, and don't consider the influence of stock trading volume. In this paper, we use Gaussian Mixture Model method to the clustering analysis of stock price volatility based on stock trading volume. The method is used to analysis the real data of Wuliangye stock in Shenzhen stock market of China. The experimental results show that it is possible to get the better clustering results by considering the influence of daily trading volume to stock closing price.

Keywords-Clustering analysis; Gaussian Mixture Model;Stock trading volume; Stock price

I. INTRODUCTION

The research of stock price volatility not only has important academic significance, but also has important practical significance. Traditionally, the stock price is usually processed as a time series. The modelling of such time series is extremely important and vital, and has been attracting the attention of both practitioners and researchers. However, it is also considered a rather difficult problem, due to the many complex features frequently present in stock price series , such as irregularities, volatility, trends and noise, and so on. A number of techniques have been developed in an attempt to model stock price series based on their present and past behavior.

Traditional time series modelling technologies, such as autoregressive integrated moving average(ARIMA)[1], exponential smoothing[2], decomposition[3], etc., have been widely and successfully used. More recently a number of machine learning techniques, such as neural networks[4], fuzzy systems[5], genetic algorithm[6], and SVM[7] are becoming promising directions in this fields. Some showed improvement compared to traditional models.

Among modern intelligent methods, cluster analysis method is an important technology. Cluster analysis[8] classifies a set of observations into two or more mutually exclusive unknown groups based on combinations of interval variables. The goal is that the objects within a group be similar to one another and different from the objects in other groups, and is to discover the share properties of the groups. The greater the similarity within a group and the greater the difference between groups, the better or more distinct the clustering. Clustering algorithms can be divided into partitioning, hierarchical, density-based, and grid-based algorithms.

Gaussian Mixture Models (GMMs)[9] are among the most statistically mature methods for clustering. The Gaussian Mixture Models are parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in data. The parameters of GMMs are usually estimated by the Expectation-Maximization (EM) algorithm[10]. This work uses the GMMs with EM algorithm to analysis the stock daily closing price based on the influence of stock daily trading volume.

II. METHODOLOGY

A. Gaussian Mixture Model

The hypothesis of Gaussian Mixture Model (GMM) is very simple, and has the assumption that the data obey the Gauss distribution. In other words, the data can be seen as generated from a number of Gaussian Distribution. In fact, GMM and k-means is actually very similar, the only difference between k-means and GMM is that the probability is introduced in the GMM. The GMM usually estimates probability density distribution of the sample, and the model estimated is a weighted sum of several Gaussian model. However, each Gaussian model is on behalf of a cluster. The probability of each cluster will be obtained when data samples are projected on several Gaussian model. Then we can choose the cluster with maximum probability as the result. In addition, Mixture Model itself actually can become arbitrarily complex, by increasing the number of Model, we can arbitrarily approximate any continuous probability density distribution. Each of the GMM is composed of K Gaussian distribution, that each Gaussian distribution is called a "Component", and the linear addition of these "Component" together constitute the probability density function of GMM, by,

$$\Pr(x) = \sum_{k=1}^{k} \pi_k N(x; u_k, \sum_k) \qquad \textbf{(1)}$$

where,

$$N(x;u_k,\sum_k)=\frac{1}{\sqrt{2\pi|\sum|}}\exp[-\frac{1}{2}(x-u)^T\sum^{-1}(x-u)] \quad \textbf{(2)}$$

x is a column vector of dimension *d*, *u* is the model expectation, and Σ is model variance.

When Gaussian mixture models are used for data clustering, clusters are assigned by selecting the component that maximizes the posterior probability. Usually, we want to find a set of parameters θ, so that the probability of generating these data points is maximum. The probability is,

$$L(\theta|x)=\prod_{i=1}^{N}\Pr(x_i;\theta) \quad (3)$$

which is called likelihood function. For the convenience of calculation, we take the logarithm of both sides of the equation(3),

$$\ln L(\theta|x)=\sum_{i=1}^{N}\ln\Pr(x_i;\theta)$$

$$=\sum_{i=1}^{N}\ln\left\{\sum_{k=1}^{k}\pi_k N(x;u_k,\sum_k)\right\} \quad (4)$$

which is called log-likelihood function.

B. EM Algorithm

EM（Expectation-Maximization）algorithm is to find the maximum likelihood estimation algorithm of parameter in the probability model, in which probabilistic models rely on hidden variables that cannot be observed. Expectation-Maximization algorithm calculating by two steps alternately, the first step is to calculate the expectation (E), in which the hidden variables is contained as can be observed, to calculate the maximum likelihood expectation; the second step is to maximize (M), in which maximum likelihood expectation found in the E step is maximized to calculate the maximum likelihood estimates of the parameters. This process continues alternately.

1) E step

A latent variable Z is introduced in model. Under the assumption that the model parameters is known, it is to seek the expectations that hidden variable Z take z1, z2, ..., respectively. It is to seek the probability of data points generated by each component in GMM, by

$$\gamma(i,k)=\alpha_k pr(z_k|x_i;\pi,u,\sum) \quad (5)$$

where weight factor αk represents the frequency of data points of the training set belonging to the category zk. So

$$\gamma(i,k)=\frac{\pi_k N(x_i;u_k,\sum_k)}{\sum_{j=1}^{k}\pi_j N(x_i;u_j,\sum_j)} \quad (6)$$

2) M step

It is to use the maximum likelihood method to obtain model parameters. Now we think that $\gamma(i,k)$ is the probability of a data point xi generated by the component k obtained in the previous step. Therefore, it can be obtained by,

$$N_k=\sum_{i=1}^{N}\gamma(i,k) \quad \textbf{(7)}$$

$$u_k=\frac{1}{N_k}\sum_{i=1}^{N}\gamma(i,k)x_i \quad \textbf{(8)}$$

$$\sum_k=\frac{1}{N_k}\sum_{i=1}^{N}\gamma(i,k)(x_i-u_k)(x_i-u_k)^T \quad (9)$$

III. EXPERIMENTS AND RESULTS

In this paper, we use the data of Wuliangye stock(2013/7/1-2014/6/20) in Shenzhen stock market of China to test our proposed method. The testing process mainly choose daily closing price and daily stock trading volume of Wuliangye stock as test data. The data chosen is totally 238 groups, we choose the first 233 groups as training data, and the remaining 5 groups as test data. The reason that only five data is selected as test data, is that the volatility of stock price leads to its long-term forecasting poorly, and the user is more concerned about the short-term predictions. The selected daily closing price data of Wuliangye stock is shown in Fig .1, and the corresponding daily trading volume is shown in Fig .2.

In the clustering process, it is need to calculate similarity distance between the data. However, the data of stock price and stock trading volume has the different scaling factors. In order to eliminate the influence of different scale factor to the similarity distance, we first normalize the values in the data before calculating the distance information. The data processed is shown in Fig .3. Otherwise, before performing Gaussian mixture model clustering analysis, we must first determine the number of Gaussian mixture model components, namely the number of clusters. The AIC information of models with different number of components is calculated, and the model with lowest AIC information is selected to get the number of clusters which is 4. The information of the number of clusters is shown in table 1. The probability density contour of the training data is shown in Fig .4, and the fitted results of the training data is shown in Fig .5. The posterior probability of corresponding to component 1 of the training data is shown in Fig .6, in which the color blue indicates that posterior probability is small, and the color red indicates that the posterior probability is big. The fitted results of the test data is shown in Fig .7, and Table 2, and it is shown clearly the clustering results of the test data fitted by GMM. The results shows that the combination of stock price and stock trading volume get a better clustering results.

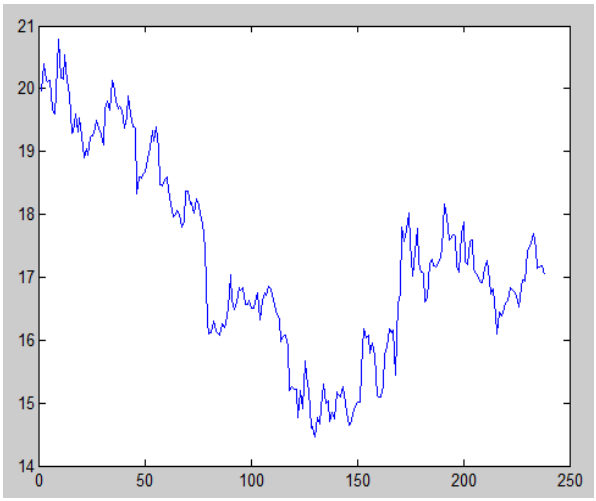

Figure 1. The daily closing price of Wuliangye stock

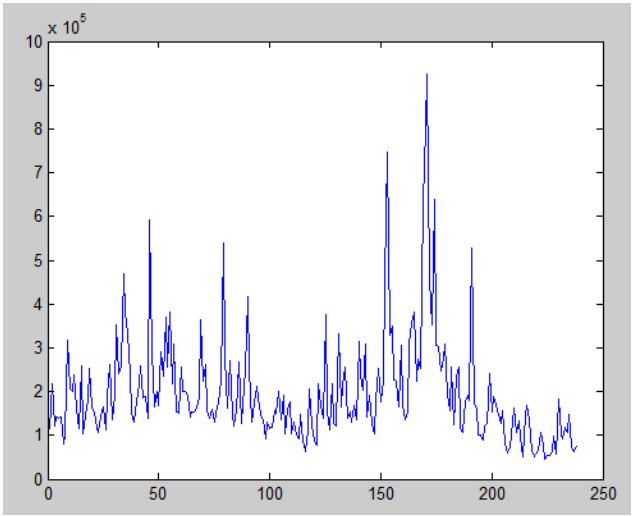

Figure 2. The daily trading volume of Wuliangye stock

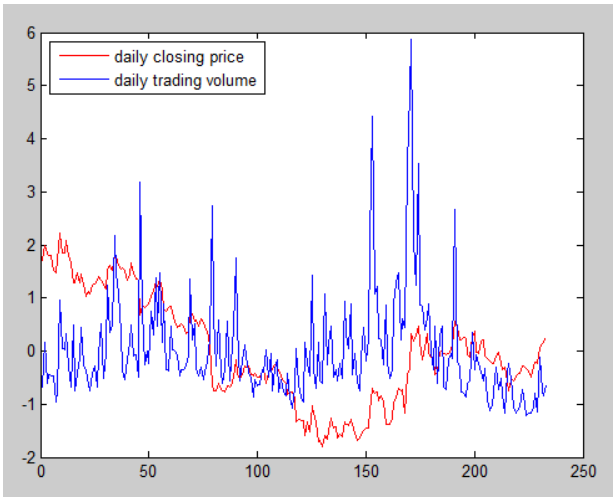

Figure 3. Normalized data of daily closing price and daily trading volume of Wuliangye stock

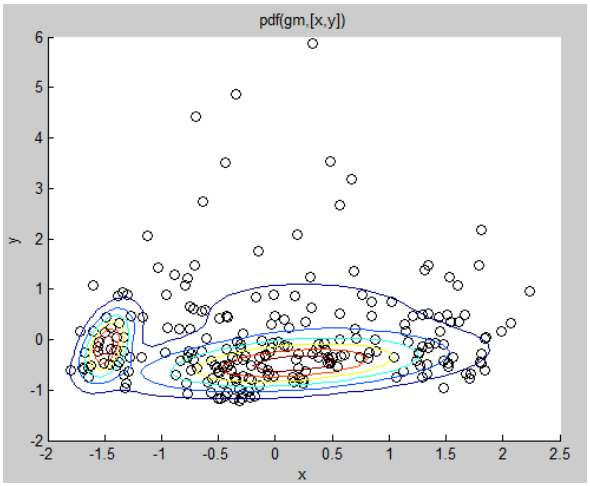

Figure 4. The probability density contour of the training data

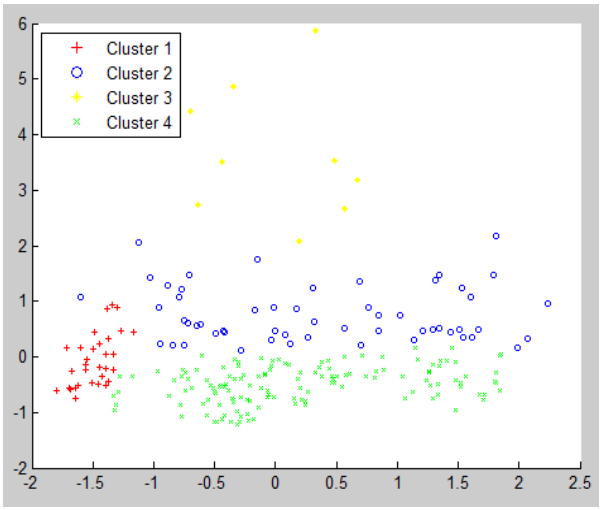

Figure 5. Fitted cluster results of the training data

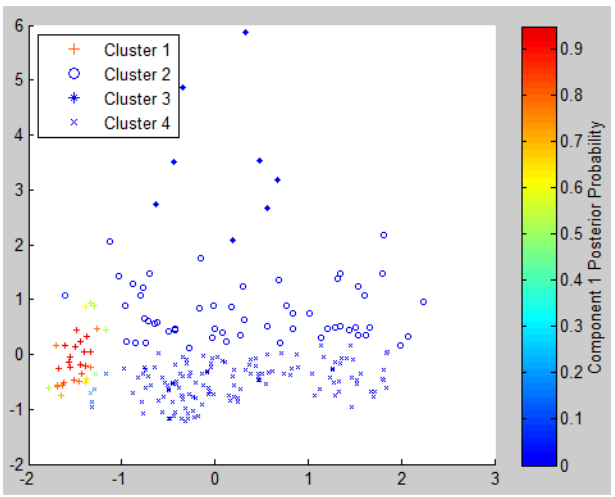

Figure 6. The posterior probability of corresponding to component 1 of the training data

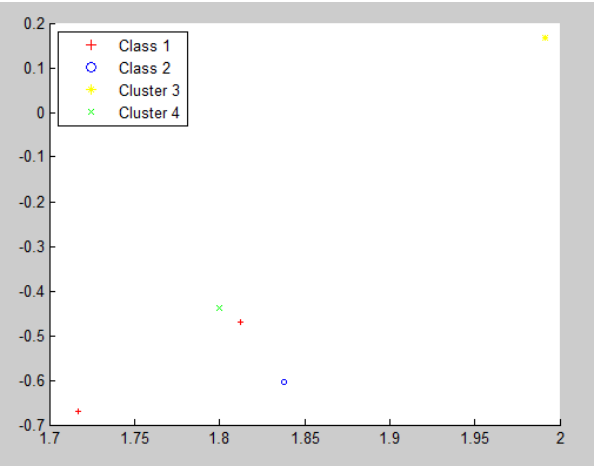

Figure 7. Fitted cluster results of the test data

TABLE I. THE INFORMATION OF THE NUMBER OF CLUSTERS

Component	Mixing proportion	Mean	
		Price	*Volume*
1	0.256	-0.3654	-0.6767
2	0.102	0.7467	2.0061
3	0.444	0.6419	-0.0737
4	0.198	-1.3387	0.1064

TABLE II. THE FITTED RESULTS OF THE TEST DATA

Test data	Price	Volume	Cluster results
1	0.1212	-0.7421	1
2	-0.0829	-0.3990	3
3	-0.0701	-0.9898	2
4	-0.0638	-1.0754	4
5	-0.1531	-0.9777	1

IV. CONCLUSIONS

In this paper, we use Gaussian Mixture Model method to the clustering analysis of stock price volatility. The premise of this problem is that the daily trading volume of stock can affect stock price. The method is used to analysis the real data of Wuliangye stock(2013/7/1-2014/6/20) in Shenzhen stock market of China. The experimental results show that it is possible to get the better clustering results by considering the influence of daily trading volume to stock closing price. Further work may use the cluster results to improve the performance of classifier.

ACKNOWLEDGMENT

This work is supported by Humanities and Social Sciences Planning Project of Chinese Ministry of Education (Project No. 07JA630090).

REFERENCES

[1] Box G, G Jenkins, and G C. Reinsel, Time series analysis: forecasting and control, 4th ed, John Wiley & Sons, Inc. 2008.

[2] James W. Taylora, and Ralph D. Snyderb, "Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing," Special Issue on Forecasting in Management Science, vol. 40(6), 2009, pp. 748–757, doi:10.1016/j.omega.2010.03.004.

[3] Theodosiou, Marina, "Forecasting monthly and quarterly time series using STL decomposition," International Journal of Forecasting, vol. 27(4), 2011, pp. 1178–1195, doi:10.1016/j.ijforecast.2010.11.002.

[4] Charles Wong and Massimiliano Versace, "CARTMAP: a neural network method for automated feature selection in financial time series forecasting," Neural Computing and Applications, vol. 21(5), 2012, pp. 969-977, doi:10.1007/s00521-012-0830-8.

[5] Ivette Luna and Rosangela Ballini, "Top-down strategies based on adaptive fuzzy rule-based systems for daily time series forecasting," International Journal of Forecasting, vol. 27(3), 2011, pp. 708–724, doi:10.1016/j.ijforecast.2010.09.006.

[6] Yi-Hui Liang, "Combining seasonal time series ARIMA method and neural networks with genetic algorithms for predicting the production value of the mechanical industry in Taiwan," Neural Computing and Applications, vol. 18(7), 2009, pp. 833-841, doi:10.1016/j.ijforecast.2010.09.006.

[7] Guo, ZQ, Wang, HQ, and Liu, Q, "Financial time series forecasting using LPP and SVM optimized by PSO," SOFT COMPUTING, vol. 17(5), 2013, pp. 805-818, doi:10.1007/s00500-012-0953-y.

[8] Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, 3rd ed, Elsevier Inc. 2011.

[9] Isabella Morlini, "A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model," Advances in Data Analysis and Classification, vol. 6(1), 2012, pp. 5-28, doi:10.1007/s11634-011-0101-z.

[10] Plechawska-Wojcik, M, "Application of Expectation-Maximization algorithm and Maximum Likelihood Rule to estimation of mixture models parameters," ACTUAL PROBLEMS OF ECONOMICS, vol. 120, 2011, pp. 346-353.