Summer 6-19-2015

An Improved K-means Algorithm and Its Application for Assessment of Culture Industry Listed Companies

Haibo Sun
Business School of Jilin University, Changchun, Jilin, China;Economic Simulation Research Institute Jilin University of Finance and Economics Changchun, Jilin, China

Limin Wang
School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun, Jilin, China, wlm_new@163.com

Xuming Han
Changchun University of Technology,College of Computer Science and Technology , Changchun, Jilin, China

An Improved K-means Algorithm and Its Application for Assessment of Culture Industry Listed Companies

Haibo Sun[1, 2], Limin Wang[3], Xuming Han[4]*

1.Business School of Jilin University, Changchun, Jilin, China;

2. Economic Simulation Research Institute Jilin University of Finance and Economics Changchun, Jilin, China

3.School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun, Jilin, China;

4.Changchun University of Technology,College of Computer Science and Technology , Changchun, Jilin, China;

Abstract: Owing to K-means algorithm has the shortcoming that it always neglects the influence of cluster size when the Euclidean distances between samples and cluster center is calculated. In order to overcome the lack, the influence of cluster size is introduced into K-means algorithm in this paper. Therefore an improved K-means algorithm based on gravity is proposed, namely GK-means algorithm. The experimental simulation results show that GK-means algorithm has better performance compared with K-means algorithm. So the GK-means algorithm is adopted for assessing the performance of culture industry listed companies in this paper. Furthermore some satisfactory results are also obtained.

Keywords: K-means algorithm, clustering, listed company, performance assessment

1. INTRODUCTION

Culture industry is getting more and more attention, which as an emerging industry in recent year. The development potential and investment value are increasingly prominent, it has great significance to Chinese future economic development and become another hot property after the real estate industry. Currently it has become a novel research subject to assess the performance of culture industry listed companies for government, economists, and many scholars. Making scientific and accurate evaluation of culture industry listed companies, which is not only conducive to government economic policy-making, but also provide reference and reduce investment risks for investors. Therefore, performance assessment of culture industry listed companies has important practical significance for the whole cultural industry development.

The listed company performance evaluation methods mainly include multivariate statistical analysis method. But the traditional statistical analysis method has huge computation, computational complexity, etc. The results are often not satisfactory. Recently clustering as a common method in the field of data mining has been the research focus of many scholars, and which has been successfully applied to the field of image segmentation, speech recognition, information retrieval, market research, decision support, etc [1, 3]. K-means algorithm is a simple clustering algorithm. In this paper The influence of cluster size is introduced into the K-means algorithm, and an improved K-means algorithm based on gravity is proposed, which named GK-means algorithm. The experimental simulation results show that the improved K-means algorithm based on gravity has the performance of fast convergence and high precision, which is better than K-means algorithm. Moreover the GK-means algorithm is applied in the field of performance assessment of culture industry listed companies in \

*Corresponding author, E-mail: wlm_new@163.com

this paper. It can provide a new reference basis for government and investors, which has potential applications in the financial field.

2. K-MEANS ALGORITHM

K-means algorithm is a kind of clustering algorithm and widely used in the field of data mining. The K-means algorithm is described as follows [4]:

a) Initial cluster centre;

b) Calculate Euclidean distances from samples to

the cluster center, and join them to the cluster with the shortest Euclidean distance. The Euclidean distance equation as illustrated in Eq. (1);

$$D = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{s}\left|X_{ij} - C_j\right|^2} \tag{1}$$

where D refers to the Euclidean distance; n is the number of samples; s represents the number of dimensions; X_{ij} indicates the j dimensionality of the i-th data; C_j is the j dimensionality of cluster center.

c) Recalculate the mean value for all samples in each cluster, and take them as new cluster centers, as illustrated in Eq. (2);

$$C_k = \frac{1}{N_k}\sum X \tag{2}$$

where C_k is the new cluster center; N_k is the number of the k-th cluster; X refers to the samples of the k-th cluster.

d) Repeat steps *a)* and *c)*, until each cluster center will not change.

3. THE IMPROVED K-MEANS ALGORITHM BASED ON GRAVITY

K-means algorithm can not focus on the influence of cluster size when Euclidean distance between samples and cluster center is calculated. In order to overcome this shortage, the thinking of gravity is introduced into K-means algorithm [5], and an improved K-means algorithm based on gravity is proposed in this paper. The cluster size is introduced into the Euclidean distance equation, as shown in Eq. (3).

$$D = \sqrt{\frac{1}{N_k}\sum_{i=1}^{n}\sum_{j=1}^{s}\left|X_{ij} - C_j\right|^2} \tag{3}$$

where D refers to the Euclidean distance; N_k is the number of the k-th cluster; n is the number of samples; s represents the number of dimensions; X_{ij} indicates the j dimensionality of the i-th data; C_j is the j dimensionality of cluster center.

4. EXPERIMENTAL SIMULATION

In order to verify the validity of the GK - means algorithm, 3 real word datasets from UCI machine repository (Wine, glass, balance-scale) are chosen to do experiments. Do five times of experiments of K-means algorithm and GK-means algorithm. The following are the comparison of average accuracy and the comparison of average iteration. Standard deviation is calculated for illustrating algorithm stability in this paper. The experimental simulation results as shown in Fig. 1, Fig. 2, Fig. 3 and Fig. 4.

From the experimental simulation results we can see that GK-means algorithm average of accuracy is higher than K-means algorithm and GK-means algorithm average of iteration is lower than K-means algorithm. As can be seen from accuracy standard deviation and clustering iteration standard deviation, the GK-means algorithm is more stable than K-means algorithm. The above shows that the performance of GK-means algorithm is better than K-means algorithm.

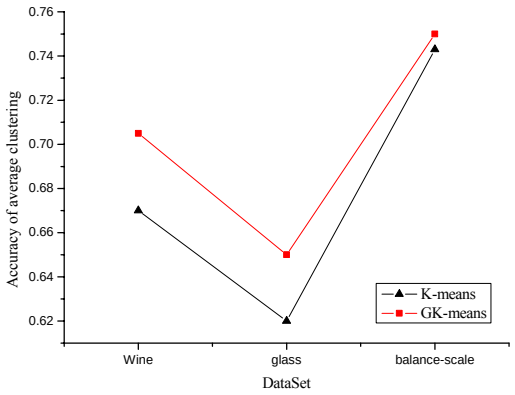

Fig.1 Comparison of average clustering accuracy

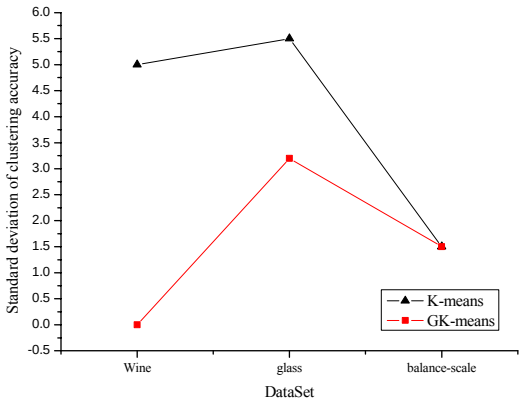

Fig.2 Comparison of average accuracy standard deviation

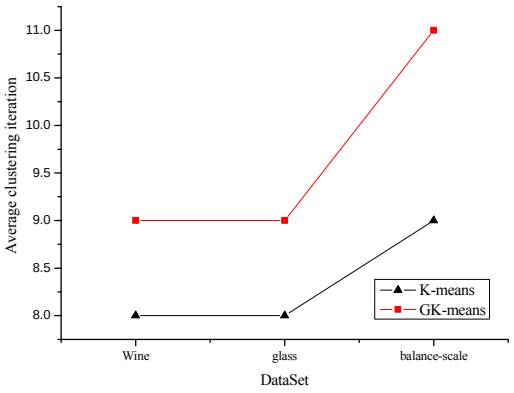

Fig.3 Comparison of average clustering iteration

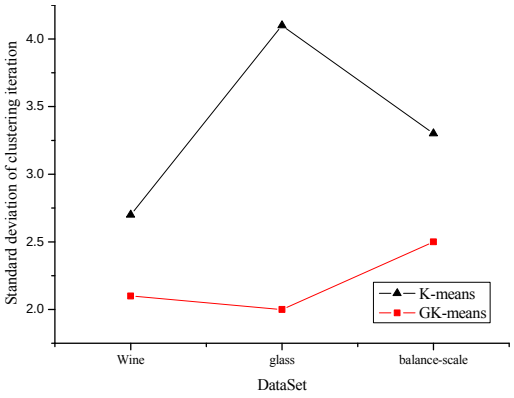

Fig.4 Comparison of average clustering iteration standard deviation

5. THE APPLICATION OF GK-MEANS ALGORITHM TO PERFOEMANCE ASSEMENT OF CLUTURE INDUSTRY LISTED COMPANIES

In the view of GK-means algorithm has good clustering performance, it is applied in the performance assessment of culture industry listed companies. This paper select 30 cultural industry listed companies in Shanghai and Shenzhen, and the financial report data (December 31, 2011) of 30 companies is chosen as the data sample. The assessment indicators are profitability (main camp service profit margin, rate of return on common stockholders' equity, earnings per share), operational capacity (receivable turnover, inventory turnover, total asset turnover), debt-paying ability (current ratio, quick ratio, asset-liability ratios), development ability (growth rate of gross operating income, net profit growth rate, growth rate of total assets), expanding capability (undistributed profits per share, accumulation fund per share, net asset value per share). Using the GK-means algorithm to assess the performance of culture industry listed companies[7-11].

5.1 Data preprocessing

In order to minimize the impact of the unit, the original data should be standardized before the cluster analysis, which means the conversion of dimensionless value. The standardized formulas are presented in Eq.(4), Eq. (5) and Eq. (6).

$$x_{ij} = \frac{(x_{ij} - \hat{x}_j)}{S_j} \tag{4}$$

$$\hat{x}_j = \frac{1}{n}\sum x_{ij} \tag{5}$$

$$S_j = \sqrt{\frac{1}{n}\sum (x_{ij} - \hat{x}_j)^2} \tag{6}$$

Where the i of x_{ij} is the number of website; j refers to the evaluation index of website; x_{ij} is the standardized data; $\hat{x}_j$ means the value of the index j; S_j represents the variance of the index j.

5.2 Clustering result and analysis

This paper divides 30 cultural industry listed companies into 5 classes, the clustering result is shown in table I.

Table 1. Clustering result

Class	Stock code				
1	（300291）	（601801）	（300148）	（002238）	（300027）
	（601098）	（300133）	（600551）	（601928）	（600637）
	（002181）	（600386）	（601999）	（000793）	（600880）
2	（300336）	（600757）	（300071）	（600373）	（300226）
	（601929）				
3	（300251）	（300235）			
4	（000917）	（600831）	（600681）	（300104）	
5	（600088）	（600037）	（000504）		

The listed companies in the third class take the leading position in cultural industry. The third class consists of ENLIGHT MEDIA and KINGSUN. ENLIGHT MEDIA has the highest earnings per share and net asset value per share, KINGSUN has the lowest asset-liability ratio in the 30 cultural industry listed companies. The earnings per share of ENLIGHT MEDIA increased 1.37 yuan in 2010 to 1.88 yuan in 2011, which has increased 37.23%. The net asset value per share of ENLIGHT MEDIA increased 2.82 yuan in 2010 to 16.32 yuan in 2011, which has increased 478.72%, the asset-liability ratio reduced 87.15% than last year. The average asset-liability ratio of listed companies in the fourth class is highest. For example, asset-liability ratio of Letv increased 9.01% in 2010 to 40.42% in 2011, which has increased 348.61%. The average rate of return on common stockholders' equity, average growth rate of gross operating income and average net profit growth rate in the fifth class are negative. The companies in the forth class and fifth class have poorer performance the other three class. The GK-means algorithm is adopted to assess performance of culture industry listed companies, and the results obtained are consistent with the actual situation of companies.

6. CONCLUSIONS

This paper introduces the influence of cluster size into K-means algorithm and proposes an improved clustering algorithm, namely GK-means algorithm. By using GK-means algorithm for assessing the performance of listed companies, experiment simulation and numerical analysis show that GK-means algorithm has better clustering performance than the K-means algorithm. GK-means algorithm is feasible and effective on performance assessment of listed companies. It provides an innovative way for performance assessment of listed companies.

ACKNOWLEDGEMENT

The authors are grateful to the support of the National Science Foundation of China under Grant No. 61202306，61472049 and 61402193, Science-Technology Research Foundation of Jilin Province under Grant

No. 2012185, 2012189 and 2014159, the Social Science Research Foundation of Jilin Province under Grant No. 2014B166. The Application Basis Foundation of Jilin Provincial Science & Technology Department under Grant No. 20100507, 201215119, 20130522177JH and 20130101072JC, the Young Talent Funds of Jilin University of Finance and Economics, and the Education Science Programming Funds of Jilin Province under Grant No. GH14216 and GH13444.

REFERENCES

[1] Renato CordeirodeAmorim, BorisMirkin. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering [J]. Pattern Recognition, 2012, 45:1061-1075.

[2] S. H. Yue, J. S. Wang, G. Tao, H. X. Wang. An unsupervised grid-based approach for clustering analysis [J]. SCIENCE CHINA Information Sciences, 2010, (7) : 1345-1357.

[3] Waseem Ahmad,Ajit Narayanan.Feature Weighing For Efficient Clustering[C].6th International Conference on Advanced Information Management and Service, Seoul, 2010: 236-242.

[4] H. G. Rong, M. W. Li, L. J. Cai. An early recognition algorithm for BitTorrent traffic based on improved K-means [J]. J.Cent. South Univ. Technol, 2011, (18) : 2061-2067.

[5] S. Y. Jiang, Q. H. Li. Gravity-based clustering approach [J]. Computer Application, 2005, 25 (2) : 286-300.

[6] Y. M. He. L. M. Wang et al. A Novel Model based on SOM2W Network for Listed Reality Companies Analysis. 2009 International Conference on Financial Theory and Engineering, 2009, 195-198.

[7] Hao H, Wang Z, Xu H. The Evaluation of Listed Banks' Competitiveness Based on Principal Component Analysis[C]. International Conference on Management Science and Engineering, 2010, 5: 110-114.

[8] Liu H F, Wang J. Integrating Independent Component Analysis and Principal Component Analysis with Neural Network to Predict Chinese Stock Market[J]. Mathematical Problems in Engineering. DOI: 10.1155/2011/382659, 2011.

[9] Huang Y D, Du L B. Evaluation of Performance of Listed Companies Using Factor Analysis-Take Listed Companies in Luzhong Region as an Example[C]. International Conference on Regional Management Science and Engineering, 2010: 711-715.

[10] Gao C, Fan Z, Zhang J. New Energy Listed Companies Competitiveness Evaluation Based on Modified Data Envelopment Analysis Model[C]. International Conference on Intelligent Computing and Information Science, 2011, 135: 613-618.

[11] Chen J. The Analysis of the Investment Value of the Real Estate Listed Companies Based on the AHP[C]. International Conference on Engineering and Business Management, 2011: 2058-2062.