Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 13 (9): 1499-1503, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.1499.1503

Dynamic Merge Clustering Algorithm and its Application in Evaluation of the Regional Scientific and Technological Innovation Capability

[1]Huang Zhong-Dong and [2]Tian Xue-Mei
[1]School of Management, Xuzhou Institute of Technology, Jiangsu Xuzhou, China
[2]Faculty of Higher Education, Swinburne University of Technology, Melbourne, Australia

Abstract: Cluster analysis is an important part of study and application in data mining and hierarchical clustering is currently the most widely used clustering method. A dynamic clustering algorithm named DCMA was proposed based on the defects of hierarchical clustering method. The irreversibility and the indispensable process ending condition of specifying the desired number of clusters and threshold adopts clusters diversity to automatically merge and divide the clusters. And clustering analysis and comprehensive evaluation were conducted to test the scientific and technological innovation capacity of 13 prefecture-level cities of Jiangsu Province. Results verified the feasibility and effectiveness of the proposed method. Data processing results show that this method can provide a scientific quantitative decision-making evaluation model for the relevant administrative departments.

Key words: Hierarchical clustering, dynamic clustering, diversity, jiangsu province, technological innovation, innovation capability

INTRODUCTION

Scientific and technological innovation capability is a measurement of the development strength of a nation and region. The national “12th Five-Year Plan” and that of Jiangsu province both put strengthening scientific and technological innovation ability as the key to enhance the comprehensive strength of science and technology (Xinhua News Agency, 2011; JIDP, 2011). China Science and Technology Development Research Reports puts forward that the evaluation index of scientific and technological innovation ability consists of the following five aspects, namely, the technology innovation environment, technology innovation input, technological innovative ability, innovation in economic performance and comprehensive ability of science and technology. The evaluation index system is based on the above five areas in this study, meanwhile, the evaluation index system of literature is also used for reference (Wang, 2009).

The literatures on technological innovation capability are relatively rich, but few of them provide evaluation methods of scientific quantitative decision-making or the comparison between the evaluation methods. Cluster analysis is a quantitative method that studies multi-factor classification problem, which can explain the complex relationship between objects, features and objects and features. In addition, it can provide scientific reference model for quantifying the comprehensive evaluation.

Among the methods of cluster analysis, hierarchical clustering is the most widely used clustering techniques. Although hierarchical clustering is extensively used, it is still very difficult to select the appropriate merging or splitting point. If a decision of choosing a merge or split point in a step is not well made, it may lead to the restriction of clustering quality. In addition, the user must decide when to stop clustering in the process of hierarchical clustering, so as to obtain the classification of a certain number, otherwise, the output of the algorithm is always a clustering (Xu and Wunsch, 2009). Aiming at the defects of hierarchical clustering, Dynamic-merge Cluster Algorithm (DMCA) which takes clusters diversity as a norm of automatic merging and splitting is proposed in this study. The algorithm does not require pre-set the clustering threshold to divide the clusters dynamically, in contrast, it automatically determines the merging and dividing process of the clusters, ultimately finding the optimal clustering. Moreover, by taking scientific and technological innovation capability index value of 13 cities in Jiangsu Province as experimental data, clustering analysis and comprehensive evaluation were conducted to test the scientific and technological innovation capacity of Jiangsu Province.

Corresponding Author: Huang Zhong-Dong, School of Management, Xuzhou Institute of Technology, Jiangsu Xuzhou, China

MATERIALS AND METHODS

Hierarchical clustering theory: Hierarchical clustering is a clustering method that organizes the data into certain groups to form a corresponding tree (Sambasivam and Theodosopoulos, 2006) and based on the clustering tree graph form, hierarchical clustering method can be divided into two types, one is top-down which is called split algorithm and the other is bottom-up named as merge algorithm. Since the specific implementation process of merged hierarchical clustering is more simple and useful, most of the hierarchical clustering method is merge-type (Davidson and Ravi, 2009). The basic idea of the method is to adopt bottom-up strategy. First, take each object as a cluster and then merge them step by step on the basis of distance criterion so as to reduce the number of clusters, until all the objects are in one cluster or a certain termination condition is met.

Relevant definitions

Definition 1 euclidean distance: Set the points in the p-dimensional space $X = (x_1, x_2,..., x_p)'$ and $Y = (y_1, y_2,..., y_p)'$, euclidean distance between two points is defined as:

$$d_{xy} = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \tag{1}$$

Euclidean distance is a common similarity measure method in clustering analysis and it can be used to express the proximity degree of the sample points. The sample points with closer distance are similar in nature and the farther ones are greatly different.

Definition 2: The shortest distance between classes: the merger between class and class is involved in the process of clustering, so distance measurement between classes should be considered. The following four distance measurement methods between classes are widely used: the minimum distance method, the maximum distance method, the class-average distance method and the centroid method. The minimum distance method is adopted in this study, which means the minimum distance between classes is taken as their merging norm. Let A and B be two clusters, thus the minimum distance between them is defined as:

$$D_{min}(A,B) = \min\{d(x_A, x_B)\} \quad x_A \in A, x_B \in B \tag{2}$$

where, $d(x_A, x_B)$ stands for the Euclidean distance between sample x_A of class A and sample x_B of class B; d_{min} (A, B) stands for the minimum distance between all the samples of class A and class B. If class C is merged from class A and class B, that is, $C = A \cup B$, then the minimum distance between class C and another class D is:

$$D_{min}(C,D) = \min\{d_{AD}, d_{BD}\} \tag{3}$$

Definition 3: The average distance of the intra-class: Let class C contain c clusters (C_1, C_2,..., C_c) each cluster C_i contains n_i samples, i = 1, 2,.., c, then the average distance of the intra-class of class X is defined as:

$$R_i = \frac{1}{n}\sum_{j=1}^{n}\| x - \bar{r} \|^2, \bar{r}^{(j)} = \frac{1}{n_i}\sum_{i=1}^{n_i} R_i^{(j)} \tag{4}$$

DYNAMIC-MERGE CLUSTER ALGORITHMS (DMCA)

Algorithm thinking: Hierarchical clustering calculates the degree of difference through the different characteristics index value of the samples and variable data. The variables or samples is recombined and classified on the basics of difference degree between them, resulting in a more efficient class. However, hierarchical clustering method is irreversible, once the two clusters are merged, it is impossible to get back to the initial state. Moreover, the user needs to specify the desired number of clusters and threshold as the process ending condition, which is very difficult to prejudge in advance.

Based on merge-type hierarchy clustering, a Dynamic-merge Cluster Algorithm (DMCA) is proposed. The core idea of the algorithm is: The two sub-clusters' merging or not is determined by the relative degree of proximity and that of interconnection between the clusters, the latter of which is defined as cluster diversity in this study. Meanwhile, compare the minimum distance with the average distance of the intra-class between two clusters to decide whether to merge these two classes. By taking clusters diversity as a norm of automatic merge and split, the defects that hierarchical clustering method is irreversible and the threshold need to be pre-set can be overcome. Instead of simply adopting the shortest original distance between classes as clusters merging criterion, the introduction of a new measurement basis helps realize clustering without having to foresee the number of clusters and achieve automatic cluster analysis of data set without having to know the classification information of clusters.

Merge criterion: Let two clusters be C_i and C_j, their shortest distance between classes is D_{min} (C_i, C_j) according to Eq. 1 and 2, their average distance of the intra-class is $R(C_i)$ and $R(C_j)$ according to Eq. 4, thus the diversity represented by σ_{ij} between C_i and C_j is defined as:

$$\sigma_{ij} = \min\{(D_{\min}(C_i, C_j) - R(C_i))(D_{\min}(C_i, C_j) - R(C_j))\} \quad (5)$$

Merge criterion: if $\sigma_{ij} \leq 0$, it means the two clusters are very close and are in a high degree of interconnection, then merge class C_i and C_j into one class C_{ij}; if $\sigma_{ij} > 0$, it suggests the shortest distance between the two clusters is greater than their respective average distance of the intra-class, then divide class C_i and C_j as two different classes.

Algorithm description:

- **Algorithm:** Dynamic-merge Cluster Algorithm (DMCA)
- **Input:** Input the data set containing N objects
- **Output:** Output the automatically merged cluster results
- **Step 1:** N initial data samples are sui generis and calculate the distance between different classes (different samples) according to Eq. 1, getting the initialized distance matrix
- **Step 2:** Quick sort the N(N-1)/2 elements in distance matrix by distance in an order from small to large and store them in the one-dimensional array D
- **Step 3:** About the current element D_{ij} of D, judge whether class C_i and C_j have been merged into the class, if not, calculate the diversity σ_{ij} between class C_i and C_j
- **Step 4:** Judge σ_{ij}, if $\sigma_{ij} \leq 0$, then merge class C_i and C_j into one class C_{ij} and replace C_i and C_j with C_{ij}, otherwise, turn to Step 5
- **Step 5:** Take the next element of array D, repeat step 2 to 4, until there are no clusters that can be merged in the cluster sequence
- **Step 6:** Output the merged clustering results

APPLICATIONS OF DCMA IN SCIENTIFIC AND TECHNOLOGICAL INNOVATION ABILITY EVALUATION OF THE CITIES IN JIANGSU PROVINCE

Jiangsu Province, a total of 13 prefecture-level cities under its jurisdiction, can be divided into three regions according to the level of economic development, namely, South Jiangsu, Central Jiangsu and North Jiangsu. South Jiangsu is the developed area of Jiangsu Province, Central Jiangsu is less developed areas and North Jiangsu is underdeveloped area.

Five scientific and technological innovation capability index data of thirteen prefecture-level cities in Jiangsu province are selected according to Jiangsu Province Statistical Yearbook 2011(Statistics Bureau of Jiangsu Province, 2011), which is shown in Table 1. It includes technology innovation environment, technology innovation input, technological innovative ability, innovation in economic performance and comprehensive ability of science and technology.

DMCA algorithm is applied to cluster analysis of the data above and the results are as shown in Table 2 and comparison of the cluster results are as shown in Fig. 1. From Table 2 can see that the algorithm proposed in this study can merge the cluster results into three classes without pre-setting the threshold, which fits the actual development situation in Jiangsu province. K-means algorithm and hierarchical clustering algorithm obtain the similar cluster results when the cluster number is 4, but that does not tally with the actual situation in Jiangsu province. When the cluster number is 3, the cluster results of the third class is the same after adopting three kinds of clustering algorithms and the first and the second are different when the algorithm changes. K-means algorithm classify Suzhou as a separate class and isolated point appears, which affects the cluster results; the difference of cluster results between hierarchical clustering algorithm and the proposed algorithm in this study is to classify Changzhou as the first class or the second, it can be seen

Table 1: Index data of scientific and technological innovation capability of Jiangsu Province

City	Technology innovation environment	Technology innovation input	Technological innovative ability	Innovation in economic performance	Comprehensive ability of science and technology
Nanjing	1.6625947	1.4997727	1.2668179	0.5569964	0.4294016
Wuxi	0.8950966	0.5943143	1.3176259	0.9835620	0.9152497
Xuzhou	-0.6517200	-0.1121360	-0.7721080	-0.5445370	-0.5930840
Changzhou	0.4216844	0.7613168	0.3849960	-0.0186580	0.1956543
Suzhou	1.1692531	0.7691429	2.1540502	2.1585670	1.8816988
Nantong	0.0025128	-0.2657240	-0.2657240	-0.2678640	-0.0285110
Lianyungang	-0.7492490	-0.7495540	-0.7495540	-0.4164490	-0.5123270
Huaian	-0.8247520	-0.8324280	-0.8324280	-0.8176680	-0.6356750
Yancheng	-0.8363280	-0.8046080	-0.8046080	-0.5053470	-0.5051180
Yangzhou	-0.0132120	0.2384570	-0.2384570	-0.4687810	-0.2012430
Zhenjiang	0.1715065	0.0015495	0.0015495	-0.2328320	-0.0362910
Taizhou	-0.3980850	0.4343210	-0.4343210	-0.0786290	-0.2220750
Suqian	-0.8493020	-1.7236020	-1.0278410	-0.3483610	-0.5876800

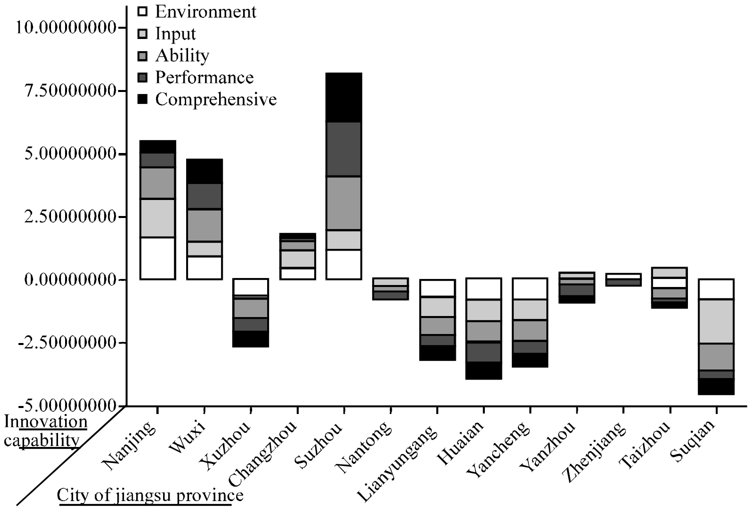

Fig. 1: Analysis results of dynamic-merge cluster algorithm (DMCA)

Table 2: Comparison of analysis results of three kinds of clustering algorithms

Class	K-means algorithm	Hierarchical clustering algorithm	DMCA
3 classes			
1st class	Suzhou	Suzhou, Nanjing, Wuxi	Suzhou, Nanjing, Wuxi, Changzhou
2nd class	Nanjing, Wuxi, Changzhou, Nantong, Yangzhou, Zhenjiang, Taizhou	Changzhou, Nantong, Yangzhou, Zhenjiang, Taizhou	Nantong, Yangzhou, Zhenjiang, Taizhou
3rd class	Xuzhou, Lianyungang, Huaian, Yancheng, Suqian	Xuzhou, Lianyungang, Huaian, Yancheng, Suqian	Xuzhou, Lianyungang, Huaian, Yancheng Suqian
4 classes			
1st class	Suzhou	Suzhou	Algorithm is over and the cluster results are merged into 3 classes automatically
2nd class	Nanjing, Wuxi	Nanjing, Wuxi	
3rd class	Changzhou, Nantong, Yangzhou, Zhenjiang, Taizhou	Changzhou, Nantong, Yangzhou, Zhenjiang, Taizhou	
4th class	Xuzhou, Lianyungang, Huaian, Yancheng, Suqian	Xuzhou, Lianyungang, Huaian, Yancheng, Suqian	

*DMCA is the algorithm in this study, which is the short of dynamic-merge cluster algorithm

clearly from Fig. 1 that it's better to classify Changzhou, Suzhou, Wuxi and Nanjing as one class. According to the analysis above, the advantage of the DMCA can be seen clearly, with which not only the quality of clustering is improved but the clustering results are more practical, forming higher reference value.

According to the cluster results comparison shown in Fig. 1, the science and technology innovation ability of Suzhou, Wuxi, Nanjing and Changzhou respectively rank the top four cities in Jiangsu Province. These cities generally have the following characteristics: as compared to the area with weak science and technology innovation ability, these cities have relatively better scientific and technological base, more foreign investment, especially Suzhou, which has become the champion city in attracting foreign investment. It has not only driven the development of high and new technology industry but also enhanced the comprehensive competitive strength of scientific and technical innovation. The comprehensive rank of Nantong, Yangzhou, Zhenjiang and Taizhou of Central Jiangsu is in the middle level and the last five cities are Huaian, Suqian, Yancheng and Lianyungang of North Jiangsu. It can be seen that the distribution of the science and technology innovation ability of each prefecture-level cities in Jiangsu Province is not balanced. The cities in South Jiangsu have obvious advantages in science and technology innovation and that in Central Jiangsu cities need to be improved and the North part are relatively weak. Therefore, massive investment in technological innovation needs to be enhanced and appropriate policies and measures need to be made to promote the development of scientific and technological innovation ability.

CONCLUSION

On the basis of the idea of merged hierarchical clustering, a dynamic clustering algorithm named DCMA that adopts clusters diversity to automatically merge and divide clusters was expounded. This algorithm overcomes the defects of hierarchical clustering method such as the irreversibility and the need of pre-establishing the threshold. According to the practice, the algorithm has been applied to evaluate the science and technology innovation ability of Jiangsu Province. And it has provided scientific quantitative decision-making evaluation for 13 prefecture-level cities of Jiangsu Province. The feasibility and effectiveness of the algorithm is verified. Compared with other clustering methods, cluster results of DCMA are more in line with the objective reality, which provides reference for analysis of science and technology innovation ability of different regions.

REFERENCES

Davidson, I. and S.S. Ravi, 2009. Using instance-level constraints in agglomerative hierarchical clustering: Theoretical and empirical results. Data Min. Knowl. Discov., 18: 257-282.

JIDP, 2011. Outline of twelfth five-year plan for the economic and social development of Jiangsu province. Jiangsu Institute of Development Planning, Nanjing, China, July 6, 2011.

Sambasivam, S. and N. Theodosopoulos, 2006. Advanced data clustering methods of mining Web documents. Inform. Sci. Inform. Technol., 3: 564-579.

Statistics Bureau of Jiangsu Province, 2011. Jiangsu Statistical Yearbook 2011. China Statistical Press, Beijing, China.

Wang, F., 2009. The assessments and strategic solutions of scientific and technological innovation capability in Jiangsu province. J. Sci. Technol. Econ. Market, 7: 63-64.

Xinhua News Agency, 2011. Outline of the twelfth five-year plan for the national economic and social development of the People's Republic of China. Xinhua News Agency, Beijing, March 16, 2011. http://news.xinhuanet.com/politics/2011-03/16/c_121193916.htm

Xu, R. and D. Wunsch, 2009. Clustering. John Wiley and Sons, New York.