

A Research of Exploration on College Students' Organization Form Based on Clustering Algorithm

Zhong Ping

China University of Geosciences(Wuhan)
Wuhan, China
e-mail: slxg@cug.edu.cn

Qin Zhimeng

China University of Geosciences(Wuhan)
Wuhan, China
e-mail: imqinzhimeng@163.com

Liu Yong*

China University of Geosciences(Wuhan)
Wuhan, China
e-mail: cugly@qq.com
* Corresponding author

Feng Shuai

China University of Geosciences(Wuhan)
Wuhan, China
e-mail: imfengshuai@163.com

Abstract—In this paper, K-means clustering algorithm is applied to the classification of college students, by which studies college students' organization form and then discovers new college students' organization form. By means of defining students' attributes, the whole text quantities and normalizes those attributes. After calculating weights and filtering those attributes, all data points are classified to K groups for the first time according to their intervals to initial centers which feature in the largest intervals to avoid K-means clustering algorithm's shortcomings. Using each group's average as new center, execute classification until stop standard is reached. Analyzing the classified groups by results of classification, common personal characters are easily to be figured out in each group. This research is considered to provide references and suggestions about college students' organization form newly discovered. How this classification results perform in the long run may need further study.

Keywords—Clustering algorithm; College students' organization form; Discoveries

I. INTRODUCTION

With a profound reform of our country's economy and society, all kinds of social associations and informal organizations emerge constantly in colleges and universities. Besides, inside colleges or universities, lots of new situations and problems present in all kinds of students' organizations, not only in their ways of activities but in the depth and range of both influence and cohesion that act on students, due to a thriving reform of systems of teaching, management and logistics. Meanwhile, along with a growing development of computer science and artificial intelligence technology, applying intelligent methods to the assistant analyses of the

existing problems will bring a total new scope to a scientific student affairs' management. Under that circumstance, how to use a modernized scientific method to explore student's organizations accurately effectively and studying organization form are of great importance both theoretically and practically. Students' organizations refer to the groups, associations or classes those are mutually related and structured in the course of students' social interactions and their academic activities. We stressed whether exists some groups whose internal relationship is close while whose external relationship is distant via students' relation organizations except for administrative organization (classes, departments) or not. Clustering points to individuals of a group are classified into several categories according to some sort of relevance, which makes similarity of the individuals of the same category big while makes similarity of the individuals of different categories small. Clustering method can make a collection of samples (data or variables) classify autonomously according to their attributes without prior knowledge, effectively conquers more randomness that relies on experience during classification in the past.

In colleges, mutual relationships like students' social interactions, academic activities and so on can be treated as a social network. Student organization is the grouping of every node in the network. The process of discovering student organization in the network can be regarded as a process of pelletization for net, and discovering student organization can expose the effects of student individuals in the network too. For instance, nodes at the edge of organizations are the important linkmen among them, but nodes in clustering center take effect on stabilizing associations.

II. CLUSTERING METHOD AND ITS PROCESS

K-means clustering algorithm is now the most widely used sort of clustering algorithms; its basic idea is seek a division method of k clusters by iteration. It uses the mean values of k clusters to represent the minimum of total error obtained from correspondent of all kinds of samples, thus that algorithm is simple and the rate of its convergence is rapid. The specific steps of k-means clustering algorithm, as follows. Choosing k objects as initial k class center of masses, classifying them according to center of mass and the distances of other objects, a new object will be added each time during the course and then center of mass will be correspondingly modified according to its core. The entire process is repeated constantly till square error converges.

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2 \quad (1)$$

In the above equation, J is the square error sum of all the objects x_i represents data points, c_j is the center of mass of the j cluster. The biggest drawback of k-means clustering algorithm is that different options of initial points will directly affect the classification accuracy of categories. As the initial points are chosen, the principle that needs to be obeyed is the maximum not similarity. The more not similar center of mass that is chosen at the beginning is, the more accurate subsequent division will be, besides, choosing initial points appropriately can speed up the rate of convergence of the algorithm. The proposal in this text is choose k initial points and make sum of square of distance between the two of those k initial points maximum.

$$L = \sum_{i=1}^k \sum_{j=1}^k (x_i - x_j)^2 \quad (2)$$

Among the whole proposals, choose k nodes that meet a demand for making L biggest as initial points.

In traditional K-means clustering algorithm, all attributes own the same weight, which inconsistent with the actual situation. To reflect the importance of each attribute, text adopts a method based on weighted distance. Specific process is as follows:

Assume that s samples are known, which can form a matrix M

$$M = \begin{bmatrix} x'_{11} & x'_{12} & L & x'_{1s} \\ x'_{21} & x'_{22} & L & x'_{2s} \\ M & M & O & M \\ x'_{m1} & x'_{m2} & L & x'_{ms} \end{bmatrix} \quad (3)$$

By calculating the product of matrix M and its transpose matrix M' , we can obtain an $m * m$ matrix N :

$$N = \begin{bmatrix} y_{11} & y_{12} & L & y_{1m} \\ y_{21} & y_{22} & L & y_{2m} \\ M & M & O & M \\ y_{m1} & y_{m2} & L & y_{mm} \end{bmatrix} \quad (4)$$

The weight for the i_{th} attribute can be described as follows:

$$l_i = \left(\sum_{j=1}^m y_{ij}^2 \right)^{\frac{1}{2}} \quad (5)$$

Normalization is needed for the weight

$$l'_i = \frac{l_i}{\sum_{i=1}^m l_i} \quad (6)$$

Using k-means clustering algorithm to discover the basic process of student organization structure as the figure presents. In the first place, quantify students' information including some mutual relations of students such as students' discipline orientations, students' hobbies etc. In the second place, mapping students' quantified information onto a vector in the vector space followed by choosing measure methods and clustering method of similarity to classify students. In the end, analyzing and assessing results according to situations of classification and discovering new organization forms or organizational ways of students.



Figure 1. The Process of Discovering Students' Organizations by Clustering Method

The realization process of k-means clustering algorithm.

Input: k numbers of clusters, n numbers of students' information.

Output: k numbers of students organizations meeting the requirements of the smallest distance.

- 1) To calculate vectors of all the students.
- 2) To calculate initial k vectors according to formula and to remove k vectors from N .

- 3) To choose anyone node j from N and to calculate $d(G, i_n, j)$.
- 4) To choose the minimum $d(G, i_n, j)$ and to classify J into i_n , belonging to the minimum d , then to remove J from N . If there are two or more minimums, the node is left to be classified later.
- 5) To calculate center of mass of i_n again.
- 6) To repeat the steps from the third to the fifth and to make sense that N is null.
- 7) Output k clusters.

III. EXPERIMENT RESULTS

In order to validate the clustering effect of this arithmetic and find the abilities of discovery and organization, the students from Institute of mathematics and physics in grade two, majoring mathematics, are regarded as this research objects. In the first place, simplify the students' attributes. In the process of collecting materials, I have collected the students' feature attributes, such as, Name, gender, the source of students, the scores of university entrance exam, rewards and punishments to help credit, credit point, student affairs adviser, class, personality, and hobbies. For the classification of students' organizations, we need to simplify students' attributes. According to the analytic survey, we should remove the secondary factor attributes for the later clustering process. And meanwhile simplify the students' attributes to only four kinds, gender, credit points, student affairs guidance, and hobbies. And the second is the students' attribute quantification processing. In gender, the female is 0, and male is 1. Credit point data samples are standardized processing, mapping to $[0, 1]$ interval. Using the normalization formula is:

$$l' = \frac{l - l_{\min}}{l_{\max} - l_{\min}} \quad (7)$$

In (7), l_{\max} , l_{\min} is respectively the maximum and minimum boundaries of the samples according to the actual engineering situation.

Using this formula, the credit points can be quantified to an interval $[0, 1]$. The conditions of qualifications and titles of the student affairs guidance teachers are sorted and then averagely normalized to an interval $[0, 1]$. They are assigned to (0, 0.2, 0.4, 0.6, 0.8, 1). According to the static or the dynamic types, students' hobbies are assigned to a number from 0 to 1. For example, reading assignment is 0 and bungee jumping and extreme sports assignment is 1. According to the specific circumstances of each student, assign them numbers.

According to the above method, we can get sixty students' attributes $D_i = \{A_i, B_i, C_i, D_i\}$. Some data are as follows:

TABLE I. STUDENT ATTRIBUTES

1	0.27	1	0.25
1	0.37	0	0.75
0	0.91	1	0.5
1	0.35	0	0.5
0	0.82	1	0.5
1	0.34	1	0
1	0.78	1	0.75
0	0.18	0	0.75
1	0.69	1	0.5
1	0.71	0.6	1
1	0.69	0.8	0.75
...
1	0.59	0.6	0.25
0	0.31	0.4	0.25
1	0.29	1	0.5
0	0.66	1	0.75
1	0.67	0	0.75
1	0.68	1	0.5
1	0.52	0.8	0.5
1	0.10	1	0.5

The distance formula of any two nodes, i and j , is

$$D_{ij} = \{(A_i - A_j)^2 * w_1 + (B_i - B_j)^2 * w_2 + (C_i - C_j)^2 * w_3 + (D_i - D_j)^2 * w_4\}^{1/2} \quad (8)$$

The numbers of cluster represent the numbers of student organization. If the organization quantity is too little, it is not beneficial for the exploration of organizations. But if the numbers of organization are too much, it is easy to cause the misclassification. This paper selects the $k=4$, finally clustering to the formation of the four student organizations.

With the formula (2) and through the calculation, select the four points which can make L the biggest and they are

$$D_5 = \{0, 0.82, 1, 0.5\}$$

$$D_{18} = \{0, 0.92, 0.2, 0.25\}$$

$$D_{59} = \{1, 0.12, 1, 1\}$$

$$D_{108} = \{1, 0.42, 0.6, 0.75\}$$

According to the previous algorithm process, classify the rest 136 nodes. The results achieved are as follows.

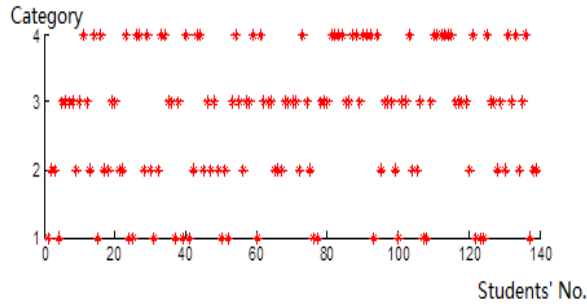


Figure 2. The Category of the 140 students

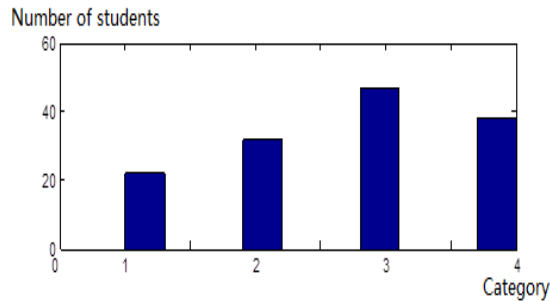


Figure 3. The Histogram of the Student Numbers in Four Categories

IV. CONCLUSION ANALYSIS

Usually application of algorithm into practice needs some evaluation method to evaluate and assess the effects and performances. But, for the fields in this paper, that is to say, using the clustering algorithm to classify students is later to find that new students' organizations are difficult to be quantitatively evaluated for the algorithm. However, in the practical operation, I find a lot of interesting phenomena and practical conclusions.

Firstly, the cadres of class and departments are easy to be classified into one group. Generally speaking, it would be thought that the generation of student cadres has some randomness. However, through our clustering algorithm it can be proved that the generation of student cadres and their own attributes cannot be separated.

In our classification, most of G3 group is the student cadres. And the ones who are not student cadres have certain potentials to become the student cadres.

Secondly, gather the students who have been classified into one group to organize one activity. And then we will find that such kind of organization forms is better and more cooperative. In our classifications, G4 group has less common places in our cognitions. After an investigation, we know that the students in this group recognize each other, but have no communications in the daily life. For such students, we organize one interesting basketball game (considering the group has girls) according to their hobbies. Result shows that the students both in cooperation and interaction are obviously better than the randomly distributed group. What kind of role they play for this clustered team needs us to do further research.

ACKNOWLEDGMENT

This research was funded by Hubei Province, humanities and social science key research base of College Students' development and innovation of the scientific research project of Open Education Research Center (No.DXS20140024), and Central Colleges of basic scientific research projects special fund operating expenses (No.2012269070).

REFERENCES

- [1] Chen An, Chen Ning, Zhou Long Xiang. "data mining technology and its application" [J]. 2006.
- [2] Leon Danon, Jordi Duch, Albert Diaz-Guilera, Alex Arenas. "Comparing community structure identification" [J]. *Journal of Statistical Mechanics: Theory and Experiment*, 2005,9008.
- [3] Li Bin. "On K-Means Clustering Algorithm With Global Optimization Ability" [J]. *Journal of Southwest China Normal University (Natural Science Edition)*, 2014,07:36-40.
- [4] Sun Defen. "University student organization functions and management research" [J] *education and occupation*, 2007 (30): 50-51.
- [5] Sun Ke, Liu Jie, Wang Xueying. "Improvement of K-Means clustering algorithm for initial center choice" [J] . *Journal of Shenyang Normal University (Natural Science Edition)*, 2009,04:448-450.
- [6] Wang Jun, Wang Chuanyu, Zhou Mingzheng. "Improved semi - supervised K - Means clustering algorithm" [J]. *Computer Engineering and Applications*, 2009,28:137-139.
- [7] Xie Juanying, Jiang Shuai, Wang Chunxia, Xie Weixin. "An improved globalK-means clustering algorithm" [J]. *Journal of Shanxi Normal University (Natural Science Edition)*, 2010,02:18-22.
- [8] Gou Guoqi. "THINKING ABOUT THE CONSTRUCTION OF STUDENT ORGANIZATION IN COLLEGES AND UNIVERSITIES" [J]. *Journal of China West Normal University (PHILOSOPHY AND SOCIAL SCIENCES EDITION)*, 2005,05:129-131
- [9] Hu Zhikui, Tang Ping, Zhang Yi, Chen Songling, Tang Cheng. "Research of clustering algorithm based on K-means for identification function bunch in neural biopsy dyeing image" [J]. *Image Processing and Multimedia Technology*, 2012,03:42-45+49.
- [10] Meng Zijian, Ma Jianghong. Improved K-Means algorithm with optional initial cluster centers [J]. *Statistics and Decision*, 2014,12:12-14.
- [11] R.S.Burt. "Positions in networks" [J]. *Social Forces* 55, 1976,93122.
- [12] Song Xiujing, Niu Zhiqiang, Tao Fei. "A comparative study of college student organizations" [J]. *SCIENCE & CHNOLOGY INFORMATION*, 2010,19:172+193
- [13] Tao Xinmin, Xu Jing, Yang Libiao, Liu Yu. "Improved Cluster Algorithm Based on K-Means and Particle Swarm Optimization" [J]. *Journal of Electronics & Information Technology*, 2010,01:92-97.
- [14] Wang Xuzhao, Wang Yadong, Zhan Yan, Yuan Fang. "Optimization of K-means Clustering by Feature Weight Learning" [J]. *OURNAL OF COMPUTER RESEARCH AND DEVELOPMENT*, 2003,06:869-873.
- [15] Xie Juanying, Guo Wenjuan, Xie Weixin, Gao Xinbo. "K-means clustering algorithm based on optimal initial centers related to pattern distribution of samples in space" [J]. *Application Research of Computers*, 2012,03:888-892.
- [16] Xiong Bing. "Study on Construction mode of College Students' Association" [J]. *Education and Vocation*, 2012 (29):39-40.
- [17] Xu Yifeng, Chen Chunming, Xu Yunqing. "AN IMPROVED CLUSTERING ALGORITHM FOR K-MEANS" [J]. *ComputerApplications and Software*, 2008,03:275-277.

- [18] Zhang Li, Sun Gang, Guo Jun. "Unsupervised Feature Selection Method Based on K-means Clustering" [J]. *Application Research Of Computers*, 2005(3):23-24.
- [19] Zhang Wenjuan, Gu Xingfa, Chen Liangfu, Yu Tao, Xu Hua. "An Algorithm for Initializing of K-Means Clustering Based on Mean-standard Deviation" [J]. *JOURNAL OF REMOTE SENSING*, 2006,05:715-721.
- [20] Zhang Xinhua. "Improvement and Innovation of Student Work organization model" [J] *Jilin Engineering and Technical Teachers College*, 2007 (11): 58-60.
- [21] Zhou Haiyan, Bai Xiaolin. "K - Means clustering method based on graph initial cluster centers in select" [J]. *Computer Measurement & Control*, 2010 18(9):2167-2169.
- [22] Zhou Shibing, Xu Zhenyuan, Tang Xuqing. "New method for determining optimal number of clusters in K-means clustering algorithm" [J]. *Computer Engineering and Applications*, 2010,46(16):27-31. .