RESEARCH ON A CLUSTERING ANALYSIS ALGORITHM FOR OPTIMAL ALLOCATION OF HUMAN RESOURCES IN COLLEGES AND UNIVERSITIES

Gangmin Weng[*]
Jingyu Zhang
School of Economics and Management
Yanshan University
066004, Qinhuangdao City, Hebei Province
China
zhangjinyu_zjy001@126.com

Abstract. With the development of digitalization, more and more colleges and universities have applied information management technology into the optimal allocation of human resources. How to manage the growing number of human resource management data and mine their potential laws to realize the optimal allocation of human resources in colleges and universities has become a top priority. In this paper, an improved k-means algorithm was introduced and its superiority was verified through an experiment by comparing the results before and after improvement. The result showed that the calculation time and complexity of the improved algorithm decreased greatly, suggesting that it could be applied for the optimal allocation of human resources in colleges and universities.

Keywords: university human resources, optimal configuration management, k-means algorithm.

Introduction

Nowadays, with the rapid development of computer technology, information management technology has been applied to the university's human resources management, resulting in a large number of data in university databases [1]. Without effective data mining technologies, these data can not play their roles. As an effective data mining algorithm, the clustering analysis algorithm has been studied by scholars all over the world. Xu et al. [2] applied the k-means algorithm based on Web user log data to perform clustering of the Web users, studied their historical Web usage data and behavioral characteristics and found that the algorithm was feasible and effective in data mining and could provide useful knowledge for Web user cluster. RJ Kuo et al. [3] integrated particle swarm optimization algorithm and k-means algorithm to cluster data and found that the particle swarm optimization algorithm could be applied to find the cluster centroid with user specified number. Besides, they used four data sets

[*]. Corresponding author

to evaluate the proposed particle swarm optimization algorithm and found that the algorithm had great potential and could complete the mining of data. In this paper, an improved k-means algorithm was introduced and its superiority was verified through an experiment by comparing the results before and after improvement. The result showed that the calculation time and complexity of the improved algorithm decreased greatly, suggesting that it could be applied for the optimal allocation of human resources in colleges and universities, which provided some reference for the application of clustering analysis algorithm in the optimization of human resources in colleges and universities.

1. Clustering analysis algorithm

1.1 Clustering analysis

Clustering analysis refers to the grouping of a collection of physical or abstract objects into a number of classes that consist of similar objects, whose aim is to collect data on a similar basis for classification [4-5]. As there are many types of clustering analysis algorithms, appropriate ones should be chosen based on the specific data type and clustering purpose [6]. The common clustering algorithms are as follows:

1.1.1 Classification method

A database containing j objects is divided into k categories, with each category representing a cluster, where all the objects are similar, $k \leq j$. Two conditions must be met using this method. Firstly, each object can belong to only one group, rather than multiple groups. Secondly, there must be one object in each group. The method is carried out as follows: the group number k is given; iteration positioning is used to move the objects between partitions and divide them; make the objects within the same group to be as similar as possible while those of different groups to be as diverse as possible. Currently, the k-means algorithm is a popular heuristic method, where each group can be represented with the average value of the objects in the group. In this paper, we use the improved k-means algorithm [7] to excavate human resource data.

1.1.2 Hierarchy method

The hierarchy method is to perform hierarchical decomposition of objects, which includes two types. One is the coagulation method, which includes all the similar objects into one group until all the groups are merged into one. The other is the splitting method, which splits s a big group into small groups until each object is in a separate group [8].

1.1.3 Density method

The density method means that the clustering behavior is continued as long as the number of objects within a neighbor region exceeds a threshold value [9].

1.1.4 Grid method

In the grid method, the space where the objects lie is divided into several grids, where clustering of the objects is performed so as to improve the clustering speed. STING is one of the commonly used grid methods [10].

1.2 k-means algorithm

The k-means algorithm divides j objects into k groups based on similarity, with the average value of the objects in each group as its center and significant differences between groups. The detailed algorithm is as follows:

(1) k objects are selected from j objects to be the center of k groups.

(2) Repeat step (1).

(3) According to the distance between each object to the center object of each group, they are divided into corresponding groups.

(4) Recalculate the average value of each group.

(5) Take the renewed average value of each group as new centers.

(6) Repeat step (3) and (4) until the group centers do not change any longer.

Generally, square error criterion is taken as distance calculation function, with its formula as follows:

Where is the average value of group and a is a data in group.

This algorithm calculates the square error of the objects, according to which j objects are divided. If the sum of the squared differences between them is large, group centers must be redefined to continue clustering until the sum of the square errors reaches the minimum. When the data is relatively large, the algorithm can efficiently complete the data mining work. Besides, most cases using the algorithm for data mining are ended with local optimization.

Because the algorithm needs to divide the starting center of each group, it is necessary to determine the number of groups and the initial center of each group firstly. Therefore, the k-means algorithm does not apply to groups where the size of the objects in the database is too different. Also, it is susceptible to isolated point data, which can exert great impact on the clustering analysis results.

1.3 Improved k-means algorithm

In order to avoid the impact of extreme differences, we improved the k-means algorithm by removing x maximum values and x minimum values respectively. The specific algorithm is as follows:

(1) Rank the j objects from large to small, removing x maximum values and x minimum values.

(2) Calculate the average value F of all the remaining $j - 2x$ objects and take $(0 - 2)$ times that of the value of F to be the initial center of each group.

(3) Repeat the above step.

(4) According to the distance between each object and each group center, they are divided into corresponding groups.

(5) Recalculate the average value of each group.

(6) Take the recalculated average value as the new group center.

(7) Repeat step (5) and (6) until the group centers do not change any longer.

2. Clustering analysis of human resource management in colleges and universities

2.1 Application of clustering analysis in post setting in colleges and universities

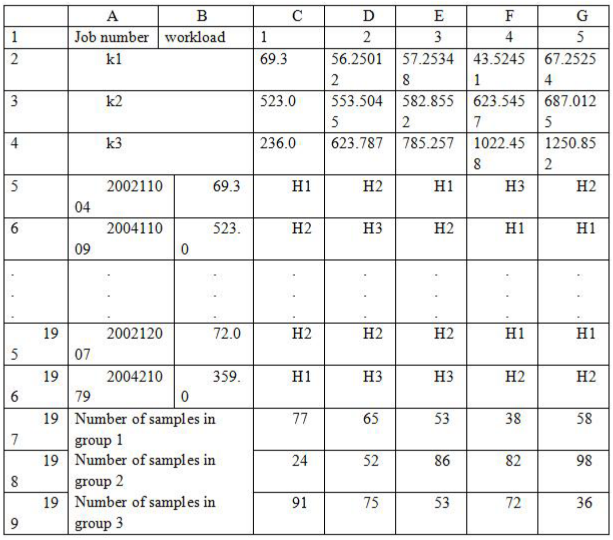
In this paper, both the k-means algorithm before and after improvement were applied to extract the human resource data of School of Economics and Management, Yanshan University, which was compared to verify the significance of clustering analysis algorithm in human resource data mining. Before the clustering analysis, the preliminary statistics on the human resources data in the university was performed and the mining objects were determined.

This paper studied the university professional and technical personnel database. The class information table and scientific research information table of teachers between 2015-2016 were obtained from the office of academic affairs, including 1,524 records on in-service staff appointment time, standard class hours and student ratings. As the research objects are the teachers, 855 related records were selected while others were omitted. Then, after screening based on the deputy senior title, 192 records were kept for clustering analysis.

2.2 Application of k-means algorithm in human resource management in colleges and universities

In this design, the 192 included records were divided into three groups, i.e., $k = 3$. Firstly, 3 objects were taken as the center of clustering. Then, according to Euclidean distance, each object was assigned to the group to which it is close in its average value. Besides, the mean vector of these objects to each cluster point was calculated and the total mean value was used as the center to perform clustering again. All the screened data were stored in an Excel table and the Excel built-in functions were used to perform clustering analysis on the data.

In the table, lines 5 to 196 are the 192 records and lines 2-4 are the new average values. A and B stand for the job number and workload respectively. G5 to G196 are the groups each point was assigned to through calculation. H1, H2 and H3 represent three groups; C1-G1 are numbers of iterations; lines 194-199 are the numbers of objects in each group after each iteration.



	A	B	C	D	E	F	G
1	Job number	workload	1	2	3	4	5
2	k1		69.3	56.25012	57.25348	43.52451	67.25254
3	k2		523.0	553.5045	582.8552	623.5457	687.0125
4	k3		236.0	623.787	785.257	1022.458	1250.852
5	200211004	69.3	H1	H2	H1	H3	H2
6	200411009	523.0	H2	H3	H2	H1	H1
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
195	200212007	72.0	H2	H2	H2	H1	H1
196	200421079	359.0	H1	H3	H3	H2	H2
197	Number of samples in group 1		77	65	53	38	58
198	Number of samples in group 2		24	52	86	82	98
199	Number of samples in group 3		91	75	53	72	36

Table 1. Clustering analysis in universities and colleges

2.2.1 Determination of iteration group center

Based on the k-means algorithm, k objects were randomly taken as the initial group centers, which are C2-C4 in table 1. Afterwards, new average values were generated after each iteration as new group centers, noted by D2, whose function is D2 = AVERAGEIF (C 5 : C 196, "H1", B 5: B 196). As an average function, AVERAGEIF is mainly used to calculate the average value of multiple table cells. In column D, the function was applied to calculate the average value of the workload data which meet the H1 condition between D5-D196.

2.2.2 Grouping of records after each iteration

In the data area, this paper made an iterative grouping of all the probabilities, and registered the value of each record and the European distance of the center of three groups [11]. According to the distance, they were assigned to the neighbor groups. For example, the function of D5 is: suppose $(\$B5 - D\$2)^2$ to be m, $(\$B5 - D\$3)$ to be n, and $(\$B5 - D\$4)^2$ to be p, then: $D5 = if((m \leq n, ifa \leq c,$ $"H1", "H2"))$, if $(n \leq p, "H2", "H3")$.

Where m refers to the European distance between record D and group 1; n refers to the European distance between record D and group 2; p refers to the European distance between record D and group 3.

When $m \leq n$, if $m \leq p$, then D belongs to group 1, denoted by $H1$; if $m > p$, then D belongs to group 3, denoted by $H3$. When $m > n$, if $n \leq p$, then D belongs to group 2, denoted by $H2$; if $n > p$, then D belongs to group 3, denoted by $H3$.

2.2.3 Calculation of the number of samples in each group

In the data area, the number of samples in each group after each iteration is recorded. For example, the function of C197 is:

C197 = COUNTIF (C5: C197, "H1")

COUNTIF is the function to calculate the number of records which meet the H1 condition.

2.2.4 Iteration

After one iteration is completed, the content in D2-D196 is used to continue the following iterations.

2.2.5 End

The iteration is stopped when the group centers and the number of samples in each group do not change any longer.

2.3 Improved k-means algorithm

As mentioned in section 2.2, we set the number of groups to be 3, i.e., $k = 3$. Then, the 192 records were ranked, the maximum vale 487.0 and minimum value 3.4 were selected and the average value of the remaining samples was calculated, denoted by Q. Afterwards, 0.5Q and 1.5Q were used as the initial center of the clustering and each object was assigned to the group to which it was close in its average value. Finally, the mean vectors of each object to each clustering point were calculated and the whole mean value was taken as the new center to perform the clustering again. The Excel table used in this section was the same as table 1, with the specific steps as follows:

2.3.1 Determination of the iteration group center

Using the improved algorithm, the ranked samples were calculated. After removing the maximum value and the minimum value, the average value Q was calculated. Then, $0.5Q$ and $1.5Q$ was used as the initial group center of the iteration and the value of $C2 - 4$, with the calculation formula as follows:

$C2 = AVERAGE(B\$6 : B\$195) * 0.5$

$C3 = AVERAGE(B\$6 : B\$195)$

$C4 = AVERAGE(B\$6 : B\$195) * 1.5$

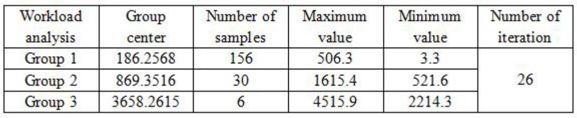
After each iteration, the average value of the group of the previous iteration was calculated. For example, the function of D2 is as follows:

$D2 = AVERAGEIF(C\$5 : C\$196, "H1"\$B\$5 : \$B\$196)$

The following steps were as the same as mentioned in the above sections. When the group centers and the number of samples in groups did not change any longer, the iteration ended.

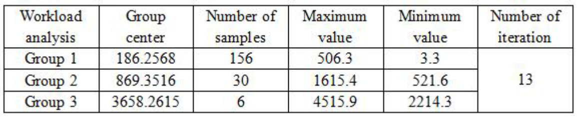
3. Results

The clustering analysis on the data of scientific research work of university teachers was performed using the above methods, with the results shown in table 2 and 3.



Workload analysis	Group center	Number of samples	Maximum value	Minimum value	Number of iteration
Group 1	186.2568	156	506.3	3.3	26
Group 2	869.3516	30	1615.4	521.6	
Group 3	3658.2615	6	4515.9	2214.3	

Table 2. k-means algorithm analysis results



Workload analysis	Group center	Number of samples	Maximum value	Minimum value	Number of iteration
Group 1	186.2568	156	506.3	3.3	13
Group 2	869.3516	30	1615.4	521.6	
Group 3	3658.2615	6	4515.9	2214.3	

Table 3. Improved k-means algorithm analysis results

As shown in table 2, the center of the three groups was 186.2568, 869.3516 and 3658.2615 respectively, with great differences between groups and small differences within each group. The minimum values of the groups are the boundary points of university teacher recruitment, i.e., the minimum standards of recruitment. Based on the workload, we can assign the recruitment conditions. The results suggest that the k-means algorithm can adapt to the excavation of human resource data in colleges and universities and complete the optimal allocation of human resources.

By comparing table 2 with table 3, it can be seen that the group centers, number of samples in each group, the maximum values and minimum values of groups were the same in the two tables, suggesting that the improved k-means algorithm also adapted to the excavation of human resource data in colleges and universities. Moreover, the number of iteration of the improved algorithm was 13 times less than that before improvement, which reduced the calculation complexity to a large degree. Therefore, the improved k-means algorithm was more convenient and can realize better optimal allocation of university human resources.

4. Conclusion

As one of clustering analysis methods, the k-means algorithm can well complete the clustering analysis of human resource data [12]. HM Hussain et al. [13] proposed a highly parallel hardware design that accelerates the k-means clustering of microarray data by implementing the k-means algorithm in a field programmable gate array. Q Ren et al. [14] improved the k-means algorithm using the kruskai algorithm, and obtained the minimum spanning tree of the clustering object by kruskai algorithm and proved that the improved algorithm was more efficient than the traditional algorithm through an experiment. In this paper, the k-means algorithm before and after improvement were both applied to carry out clustering analysis on university human resource data and the results showed that the improved algorithm was more convenient, which provide references for the application of clustering analysis in the optimization of human resources in colleges and universities.

References

[1] Y. Liu, *Analysis on the effective integration of information technology and personnel management in colleges and universities*, Creative Education, 6 (2015), 785-789.

[2] Jin Hua Xu, Hong Liu, *Web user clustering analysis based on K means algorithm*, International Conference on Information NETWORKING and Automation, IEEE, 2010.

[3] R.J. Kuo, M.J. Wang, T.W. Huang, *An application of particle swarm optimization algorithm to clustering analysis*, Soft Computing, 15 (2011), 533-542.

[4] K. Xia, Y. Wu, X. Ren et al., *Research in clustering algorithm for diseases analysis*, Journal of Networks, 8 (2013), 1632-1639.

[5] S. Cheng, Y. Shi, Q. Qin et al., *Solution clustering analysis in brain storm optimization algorithm*, Swarm Intelligence, IEEE, 2013, 111-118.

[6] R. Bala, S. Sikka, J. Singh, *A Comparative analysis of clustering algorithms*, International Journal of Computer Applications, 100 (2014), 35-39.

[7] K. Singh, D. Malik, N. Sharma, *Evolving limitations in K-means algorithm in data mining and their removal*, International Journal of Computational Engineering & Management, 2011, 2230-7893.

[8] F. Murtagh, P. Legendre, *Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?*, Journal of Classification, 31 (2014), 274-295.

[9] Y. Fan, Y. Rao, *A density-based path clustering algorithm*, International Conference on Intelligent Computation and Bio-Medical Instrumentation, IEEE, 2011, 224-227.

[10] S. Krinidis, V. Chatzis, *A robust fuzzy local information C-means clustering algorithm*, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 19 (2010), 1328-1337.

[11] L. Liberti, C. Lavor, N. Maculan et al., *Euclidean distance geometry and applications*, Quantitative Biology, 56 (2012), 3-69.

[12] J. Zhu, H. Wang, *An improved K-means clustering algorithm,* Journal of Networks, 9 (2014), 44-46.

[13] H.M. Hussain, K. Benkrid, H. Seker et al, *FPGA implementation of K-means algorithm for bioinformatics application: an accelerated approach to clustering microarray data*, Adaptive Hardware and Systems, IEEE, 2011, 248-255.

[14] Q. Ren, X. Zhuo, *Application of an improved K-means algorithm in gene expression data analysis*, IEEE International Conference on Systems Biology, IEEE, 2011, 87-91.

Accepted: 21.01.2016