An improved *K*-Means Algorithm Based on Association Rules

Gang Liu, Shaobin Huang, Caixia Lu, and Yudan Du

***Abstract*—With the rapid development of clustering analysis technology, there have been many application-specific clustering algorithms, such as text clustering. *K*-Means algorithm, as one of the classic algorithms of clustering algorithms, and a textual document clustering algorithms commonly used in the analysis process, is widely used because of its simple and low complexity. This article in view of two big limitations that the *K*-Means algorithm has, namely requirements that users give the anticipated variety beforehand integer *K* and random selection of initial variety center, proposed *K*-Means improved algorithm based on the association rules technology. This method proposed the concept of the smallest rule covering set .It has relieved these two big limitations of *K*-Means algorithm effectively. It is used for the audit monitor target discovery and extraction process in social security domain basic old-age insurance audit methods. Thus it can provide better reference value and guiding sense for auditors.**

***Index Terms*—Clustering, text clustering, association rules, *K*-means algorithm, monitoring indicators.**

I. Introduction

Clustering analysis as one of the data mining primary missions has very important actual value and use significance in many domains. So far people have proposed the massive clustering algorithms. *K*-Means algorithm as classical clustering algorithm, because of its relatively expandable and high efficiency characteristics, is often treated as basis for improvement of the clustering algorithm by researchers, and successfully applied to various fields. However, the *K*-Means algorithm also has his limitations: Request the user to give anticipated variety beforehand integer *K* and random selection of initial variety center.

In this paper, the *K*-means algorithm given in advance of the expected cluster number *k* and the initial cluster centers randomly selected two major limitations, through the application of classical *K*-means clustering algorithm, and with association rule analysis technology effectively improves the two major limitations of the *K*-means clustering algorithm. This paper proposes *K*-means algorithm improved method based on association rules, and applies it to the audit of the basic pension insurance in the social security field in audit monitoring indicators discovery and extraction process, so that it can provide better reference value and guiding significance for auditing workers avoiding the audit monitoring indicators rely on audit manual extraction.

Manuscript received September 25, 2013; revised November 20, 2013. This work is sponsored by the Postdoctoral Science Funds of China under grant number 2013M541345, the Fundamental Research Funds for the Central Universities of China under grant number HEUCF100603 and HEUCFZ1212, the National Science & Technology Pillar Program under grant number 2009BAH42B02 and 2012BAH08B02.

The authors are with the College of Computer Science and Technology at Harbin Engineering University, China. (e-mail: liugang@hrbeu.edu.cn, huangshaobin@hrbeu.edu.cn, s311060106@hrbeu.edu.cn, duyudan@hrbeu.edu.cn).

II. Related Work

Our research, mainly relates to the theory knowledge of cluster analysis technology in the field of data mining. Text clustering as a means of data mining, the main task is gathering similar text together, so that the similarity between the similar texts becomes larger, and Similarity between the different types of texts is smaller.

A. Limitations of K-Means Clustering Algorithm

Cluster analysis is an important technology in the data mining technology. Clustering is gathering some things together into one class according to certain attributes. So that similarity between classes becomes as small as possible, similarity within one class as large as possible. Clustering is an unsupervised learning process. Differences between it and classification are: it is necessary to know in advance what is the basis of the data characteristics for classification, and clustering is to find out the data characteristics. Therefore, in many applications, the cluster analysis as a data preprocessing procedure is the basis for further analysis and data processing. For example, in business, cluster analysis can help market analysts find a different customer base from the client library. And buying patterns depict different features of the customer base.

Clustering analysis methods commonly used are: method based on classification, method based on hierarchical, method based on density, method based on grid and so on.

The *K*-means clustering algorithm (*K*-means clustering) [1] is one kind of classical clustering algorithm which is proposed by Mac Queen. Realization of this algorithm is simple; the complexity is low. It obtains extremely widespread use, and becomes improvement object or the foundation for many other algorithms. The *K*-means algorithm main steps are as follows:

Input: Data set D; the *K*-means clustering counts *K*
Output: clustering results C*
Step1：Selecting *k* points as initial central point
Step 2：Repeat
Assigning each point to the nearest the center, forming a *k*-th cluster. Recalculate the center of each cluster.
Until Center point does not change
Step3：Returning clustering results C*

K-means algorithm as a classical clustering algorithm although has relatively scalable and efficiency advantages. However, because of the limitations of the algorithm itself, there are still defects as follows:

1) Requiring the user in advance to give the desired cluster number k

Clustering algorithm based on *K*-means clustering algorithm is the most classic and the most commonly used algorithm. The clustering method requires users in the process of cluster analysis to input expected clustering number *k*. For example, when clustering computer audit methods of the basic endowment insurance, users who has many years society guarantee work experience will set the expected clustering number as 5 according to his society guarantees experience and knowledge, anticipated bunch of number *k* supposes will be 5 (i.e. basic old-age insurance synthesis auditing methods, basic old-age insurance collection- pay auditing methods, basic old-age insurance management auditing methods, basic old-age insurance payment auditing methods and basic old-age insurance finance auditing methods), but those who do not have any society guarantees experience and knowledge will stochastically set the expected clustering number *k*. Because the clustering results are extremely sensitive to regarding the input parameter, different input can obtain entirely different clustering results. Consequently, *K*-means algorithm requires the user in advance to give expected clustering number *k*. This deficiency not only increases the users 'burden, but also made it difficult to control the clustering results 'quality.

2) Random selection of initial cluster centers

Selecting the appropriate initial cluster center is a key step in the traditional *K*-means algorithm process. However, it often leads to the final local optimization results that the traditional *K*-means algorithm always randomly selects initial cluster centers. It can be explained by data set in literature [2]. The data set consists of two clusters: that is four types of data composition. Within each cluster is closer, while distance between clusters is larger. If setting two initial centers for each cluster, final cluster centers will not change, even if the two initial centers are assigned to one cluster, with the iteration of the algorithm, the cluster center will re-distribution. If a cluster is only assigned one initial centers, another cluster three, after several iterations, two clusters originally belong to one cluster are divided, and two clusters originally did not belong to one cluster are merged.

Thus, in the *K*-means clustering process, different initialized clustering center can produce different results. Furthermore, it can discover: So long as two initial central points of one cluster fall on the internal cluster, no matter which position falls on, the optimal cluster can be obtained. That is because the multiple iteration of algorithm will redistribute the cluster center, and finally each cluster will has one cluster center. However, with clustering objects increase, possibilities that central point of cluster become more or less than two also gradually increase. Because distance between clusters is larger, center point cannot the redistributed. Therefore, only local optimum can be obtained.

B. Researches in Clustering Technology Based on Association Rules

Along with the development of association rules and clustering- two mining technologies, research in clustering technology based on association rules has also become more and more. Firstly, researchers had many improvements in the similarity computing methods mainly through the association rules technology. Literature [3] has given a new association rule method. It measures the distance between the rules by commodity information classification information. The entire process scanned primitive data sets only once, thus it saves time. Literature [4] proposes one similarity computing algorithm based on the words" relational degree". This algorithm obtains good clustering results. In addition, the frequent item set is the foundation of association rules, so clustering technology based on the frequent item set had many improvements. Literature [5] has improved text clustering method based on frequent item-set in WEB documents through the cross link chart instead of traditional calculating methods obtaining the frequent item-set.

To solve the two limitations *K*-means algorithm has, Longhao, Fengjianlin, et propose R-means algorithm [6].

C. Clustering Effect Evaluation Index

This article accurately evaluates clustering results, according to some effective expert category messages. For example, information entropy (*Entropy*) and purity (*Purity*) [7]. Definitions of the two targets are as follows:

$$Entropy = \sum_{i=1}^{k} \frac{n_i}{N} \left(-\frac{1}{\log q} \sum_{j=1}^{q} \frac{n_i^j}{n_i} \log \frac{n_i^j}{n_i}\right) \quad (1)$$

$$Purity = \sum_{i=1}^{k} \frac{1}{N} \max_j (n_i^j) \quad (2)$$

q is the number of the original data class. n_i^j is the number of the i-th data cluster got in the originally part of the j-th class data. The Smaller entropy, the larger purity and the better clustering. The ideal case is entropy = 0.0, purity = 1.0.

III. Improved *K*-Means Algorithm Based on Association Rules Technology

To solve the two limitations *K*-means algorithm has, we propose improved *K*-means algorithm based on minimum cover set [8].

Algorithm 1: Seeking the minimum rules set

Input: Frequent closed item sets FCI; Minimum confidence minconf;
Output: minimum rules set MRS
For each item *K* in FCI {
Finding all subset s in *K*;
For each subset s in *K* {
If (*s*->*K* ¢ MRS){
Confidence=support(I*K*)/SUPPORT(*s*);
If(Confidence>=minconf)
Add *s*->*K* to MRS;
}
}
}

Algorithm 2: Seeking the minimum rules covering set

Input: minimum rules set MRS

Output: minimum rules covering set MCS

1) Obtain their corresponding data set *D*' according to the minimum rules set MRS, form a new set *Q*, each item in the set consists of {rules, data set}
2) obtain its subset *Q*1 according to *Q*;
3) Obtain the intersection of *Q*1 is *Q*2 -subset of primitive data set *D*;
4) Get subset *Q*3 having the least rules number in *Q*2;
5) If *Q*3 is not unique, then obtain the overlapped rate smallest subset, namely MCS.

For example, data set *D* is as follows Table I. Set R consists of rules on this data set. Table II shows minimum rules set P getting from R and object N meeting rules. Then we can find out rules subset of P is *r*1= {rules 1, rules 2, rules 3}, *r*2= {rules 1, rules 3, rules 4}, and so on. Union of objects in these rules is *D*, but it is not the rules set covering the least rules number. There is rules subset *r*={rules 3, rules 4},which meets union of objects in *r* is *D*, and *r* contains the least rules number 2. Here has only one such set. Therefore, rules set *r* is the minimum rules covering set of data set *D*.

TABLE I: DATA SET *D*

Objects N	Items
1	ABCD
2	ABC
3	ACD
4	BCD

TABLE II: RULES SET *R* AND OBJECTS MEETING RULES

Rules set R	Objects
1：BD=>C	1,4
2：AB=>C	1,2
3：AC=>D	1,3
4：B=>C	1,2,4
5：AB=>CD	1
6：BD=>AC	1

TABLE III: CLUSTERING RESULTS OF EXPERIMENT 1

Cluster	Elements of cluster(text n)
cluster 1	2、3、4、10、13、14
cluster 2	8、9、12、23、24
cluster 3	6、7、11、16、17
cluster 4	1、5、15、18、19、20、21、22

To solve the limitation (requiring the user in advance to give the desired cluster number *k*) of the traditional *K*-means algorithm, we set *K* as *N* which is the number of elements in minimum rules covering set *r*.

Meanwhile, get *K* initial cluster according to the corresponding object each of the rules has in the minimum covering set *r*. Compute the average value of the object in each cluster. Then obtaining *K* initial cluster centers has laid the good foundation for the later *K*-means clustering process.

IV. EXPERIMENTS

This experiment is on real data sets about basic old-age insurance in the social security area audit methods. Text clustering is carried on through *N*=24 basic old-age insurance audit methods as the data set.

Experiment 1 uses the classic *K*-means algorithm for clustering on the above data set. Specific clustering results are shown in Table III.

In order to solve the limitation of *K*-means algorithm stochastic initialization, experiment two proposes the association rules-based *K*-means algorithm using association rules technology. Obtain the initial cluster result according to the documents that the keywords frequently together presents must be similar. Farther obtain cluster number *K* and initial cluster centers, initialization *K*-means algorithm.

Analysis of the association rules is run on data mining software weka. Association rules analysis is done: select weight ranking the top seven Keywords in each document on behalf of this document. Analysis results are shown in Table IV.

TABLE IV: RESULTS OF ASSOCIATION RULES ANALYSIS

The minimum covering set	Elements of cluster（text *n*）	Initial clustering centers
Keyword 2= credit side=> Keyword 5= expenditure	18、19、20、21、22、23	18
Keyword 3=account=> Keyword 6= individual	8、9、10、12、14、15	8
Keyword 4= one-off payment=> Keyword 6= individual	13、14、15	13
Keyword 6= business accounting=> Keyword 7= period dealing with subordinate	1、5、24	1
Keyword 7= payment=> Keyword 2= unit	6、3、2	6

Then we can get *k*=5, and initial clustering centers are text 18, text 8, text 13, text 1, and text 6. The results on the same data set clustering are shown in Table V.

TABLE V: RESULTS OF EXPERIMENT TWO CLUSTERING

Cluster	Elements of cluster(text n)
cluster 1	18、19、20、21、22
cluster 2	8、9、12、23
cluster 3	10、13、14
cluster 4	1、4、5、11、15
cluster 5	6、3、2、7、16、17、24

TABLE VI: COMPARING CLUSTERING RESULTS

	Purity	Entropy
Experiment1	0.3167	0.7253
Experiment2	0.6667	0.3456
reference value	1.0	0.0

The clustering effect comparing the improved algorithm with standard *K*-means algorithm after the experiment is shown in Table VI. Table VI shows that comparing with the standard *K*-means algorithm; the *K*-means algorithm based on association rules obviously has the enhanced purity. Simultaneously the information entropy is also reduced. Thus

from another side it confirms the *K*-means algorithm based on association rules validity.

Audit monitoring indicators, can be understood as indicators, for audit workers, have a good reference value and significance or should be considered or influencing factors during the audit work. For example, when we analyze the stability of basic old-age insurance transfers, we find that throughout the whole analysis process we should consider citizens' average wage as one key monitoring indicator.

This article is based on the above clustering. For each cluster extract several key words on behalf of the cluster. Use TF/IDF features extracting algorithm (here TF refers to term frequencies of some word in all articles in this cluster, but is not the number of times presented in the article. IDF is inverse document frequency presented in all articles) and the method of extracting keyword, combining the domain knowledge, then choose the appropriate keyword as the monitor target. Finally, joining into the monitor target storehouse.

Here, steps extracting keywords for each cluster are as follows:

1) Give the segmentation in all documents of each cluster, then input the frequency of each word in a dictionary
2) Traverse each word. Get the values that IDF of each word in all documents multiply its frequency presented in the cluster (TF);
3) Save all of the information in a dictionary (key is one word. Value is weight TF*IDF) .Sort the information by value. Finally take the top weight of words as keywords.

The larger weights keywords in each cluster are shown in Table VII.

TABLE VII: KEYWORDS BASED ON CLUSTERS

Cluster	Keywords 1	Keywords 2	Keywords 3
cluster 1	insurance unit	unit payment	insurance expenses
cluster 2	credit side	accounting	GLVch
cluster 3	disposable	one-off payment	one-time insurance
cluster 4	payment details	time quantum	audit
cluster 5	in-service staff	retired people	subscription

Table VII is keywords that have larger weight in each cluster. From Table VII, we can obviously see that the cluster 3 is related to the old-age insurance management audit methods according to the keywords of the cluster 3 and exiting audit methods, while "the disposable payment" can be used as the core word of this method, considering it as the monitor target and joining it into the monitor target storehouse, which provides the basis and reference value for the auditing workers.

V. CONCLUSION

To solve the two big limitations the traditional *K*-means algorithm has that it needs users to give anticipated bunch of integer *K* and the initial bunch of random, we regard the smallest rules covering collection as basis. We propose *K*-means algorithm based on the association rules. And we use it in discovering audit monitor targets and the extraction process in society guarantees in the domain the basic old-age insurance audit methods. Through many times tests, we can see that the *K*-means algorithm based on the association rules is more effective than the traditional *K*-means algorithm. The monitor targets obtained by the *K*-means algorithm based on the association rules have provided the very good instruction and the reference value for the later period auditing work.

REFERENCES

[1] F. Weng, L. Chen, and Q. Jiang, "Clustering ensemble based on the knn algorithm," *Journal of Computer Research and Development*, vol. 44, no. 4, pp. 187-191, 2007.
[2] M. Fan and H. Fan, *Introduction to Data Mining*, Beijing: Posts and Telecom Press, pp. 313-316, 2006.
[3] B. Ruan and Y. Zhu, "Association rule clustering based on taxonomy information," *Journal of Computer Research and Development*, vol. 2, pp. 352-360, 2004.
[4] S. Qu, Q. Wang, Y. Zou, and Q. Zhu, "Research on text clustering algorithm based on association rule," *Application Research of Computers*, vol. 4, pp. 986-988, 2008.
[5] J. Wang, L. Wang, and C. Han, "Improvement of the application of web documents clustering based on frequent itemsets," *Modern Computer*, vol. 10, pp. 11-13, 2009.
[6] H. Long, J. Feng, and Q. Li, "R-means: exploiting association rules as means for text clustering," *Computer Science*, vol. 32, no. 9, 2005.
[7] M. Edwin and T. Raymond, "A unified notion of outliers: properties and computation," *Knowledge Discovery in Data*, pp. 219-222, 1997.
[8] G. Ma and R. Cui, "Mining the smallest association rule set based on cover operations," *Computer Engineering & Science*, no. 6, pp. 65-69, 2005.

Gang Liu was born in 1976. He is an associate professor with the College of Computer Science and Technology at Harbin Engineering University, China. He received his Ph.D. from Harbin Engineering University in 2008. His research interests include grid computing, distributed computing and simulation, and policy analysis.

Shaobin Huang was born in 1965. He is a professor with the College of Computer Science and Technology at Harbin Engineering University, China. He received his Ph.D. from Harbin Engineering University in 2004. His research interests include grid computing, distributed computing and simulation, and model checking.

Caixia Lu was born in 1986. She is a postgraduate student at the College of Computer Science and Technology at Harbin Engineering University, China. She received her B.D. from Henan Polytechnic University in 2011. Her research interests include distributed computing and simulation, and policy analysis.

Yudan Du was born in 1989. She is presently a MSc with the College of Computer Science and Technology, Harbin Engineering University, China. She received her Bachelor from Zhongyuan University of Technology, China in 2012. Her research interests include ontology and rule-based reasoning.