An Approach to Data Mining in Healthcare: Improved K-means Algorithm

Dinh Thuan Nguyen, Gia Toan Nguyen, and Vu Tuan Nguyen Lam
Faculty of Information Technology, University of Information Technology, VNU-HCMC, VietNam
Email: {thuannd, toanng, tuanvnl}@uit.edu.vn

***Abstract*—Nowadays, the application of data mining in the healthcare industry is necessary. Data mining brings a set of tools and techniques that can be applied to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions. In more detail, clustering the patients that have the same status helps discovering new disease, but the suitable number of clusters is not often obvious. This paper first reviews existing methods for selecting the number of clusters for the algorithm. Then, an improved algorithm is presented for learning k while clustering. Finally, we evaluate the algorithm, apply to dataset of patients and results show its efficiency.**

***Index Terms*—clustering, fuzzy c-means, selecting the number of clusters**

I. Introduction

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of data objects that are similar between themselves and dissimilar to objects of other groups. Dissimilarities are assessed based on the attribute values describing the objects. Often, distance measures are used. Representing data by fewer clusters necessarily lose certain fine details, but achieves simplification.

K-means is a popular clustering method, but it also has disadvantages. One is the fixed number of clusters must be specified as an input to the algorithm. Moreover, the initial randomly choice data points as cluster means can result in different final clusters. That means each rerun will produce a different result.

Determining the number of clusters is usually so hard to achieve a good clustering result. A number of researchers used method based on information obtained during the K-means clustering operation itself to select the number of clusters, K.

One of these methods was addressed by D. T. Pham, S. S. Dimov and C. D. Nguyen [1], proposed Building of Measure Function. Yin Z., Tang Y., Sun F. and Sun Z. [2] proposed Fuzzy Clustering with Separable Criterion, M. V. B. T. Santhi, V. R. N. Sai Leela, P. U. Anitha, and D. Nagamalleswari [3] proposed method finding the better initial centroids and provides an efficient way of assigning the data points to the suitable clusters. Mohammad F. E. and Wesam M. A., [4] proposed to solve the problems generated from randomly initialized k-means algorithm, it depends on initializing prototypes according to statistical information calculated from the data, it initiates prototypes as points located on a surface of a hypersphere centered on the mean of the sample. Haizhou W. and Mingzhou S. [5] develop an exact solution to 1-D clustering in a practical amount of time, as an alternative to heuristic k-means algorithms.

The paper proposed a way to improve the traditional K-means, based on the two-step method of Ming-Y. S., Jar-Wen J. and Lien-Fu L. [6], and the method for selecting the number of clusters of Dinh Thuan N. and Huan D. [7].

The remainder of the paper consists of four sections. Section 2 reviews the existing methods mentioned above, which are used in this research. Section 3 builds an improved K-means algorithm. Section 4 presents the results and evaluates the new algorithm. Section 5 concludes the paper.

II. Reference Works

In the healthcare industry, determining the number of groups of patients is an important problem and it requires high precision. One way to find out the appropriate number of clusters *k* is running the algorithm with various k, then choose k that the result clusters are the best. But first, we need to overcome one weakness of K-means, the results are not the same after each rerun.

To replace the randomly selecting individual objects as cluster centers, the two-step method proposed in [4]. Specifies that using agglomerative hierarchical clustering in the first step to cluster the original dataset into some subsets, which will be the initial set of clusters in K-means clustering algorithm. Then calculate the centroid of every formed cluster, and apply K-means algorithm again to regroup formed cluster into desired k groups. It is believed that this strategy will be a better solution than a random selection individual data.

Since the final clusters are not changed after each rerun, now we can compare the results achieved in order to find the suitable number of clusters.

There are many approaches for choosing the right number of clusters *k*, and one effective way is select k based on information obtained during the K-means clustering operation itself. The method was addressed in

Manuscript received January 7, 2013; revised April 24, 2013.

This work was supported in part by the University of Information Technology, VNU-HCM under Grant B2012-26-03 and 15/HĐ-ĐHCNTT.

doi: 10.12720/jiii.1.1.14-18

[7], proposed using intra-class coefficient α and inter-class coefficient β that indicate the appropriate number of clusters. Those coefficients are calculated as follows:

$$\alpha = \frac{d_{\max}}{d_{avr}}$$

where α is the intra-class coefficient, dmax is the largest distance among all objects within a cluster, and davr is the average distance among all objects within a cluster. When the cluster is only one object, it accepts α = 0. If a certain cluster has large α, it means this cluster needs to be splitted.

$$\beta = \frac{\phi_{\min}}{\phi_{avr}}$$

where β is the inter-class coefficient, $\phi_{\min}$ is the smallest distance from center of the cluster to different cluster, and ϕ_{avr} is the average distance from center of the cluster to different cluster. When it is only one cluster, it accepts β = 0. If β is close to 1, it means some clusters should be grouped.

With the above definition of α when the α coefficient of a cluster becomes larger, the more unbalanced cluster. In an unbalanced cluster the similarity among objects in the cluster is not high. According to clustering theory, unbalanced clusters need to separate clusters.

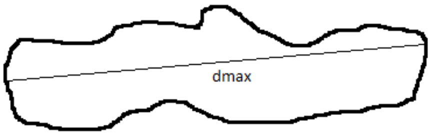

Figure 1. Unbalanced cluster

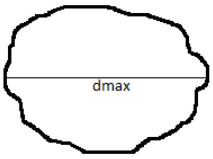

Figure 2. Balanced cluster

Call $\alpha_{\max}$ is the largest coefficient of the α coefficients of all the clusters after clustering.

After clustering if a certain cluster has large α, it means $\alpha_{\max}$ is large (on the basis of what is called great, we'll find a value standard for comparison in section 4.A), so this cluster needs to be separated clusters.

With the above definition of β, cluster 2 will be quite small and flat when the β is close 1 (β = 1 when cluster 2 is only one object) (Fig. 3). Thus the coefficient β is close 1, cluster 1 and cluster 2 have to be coupled.

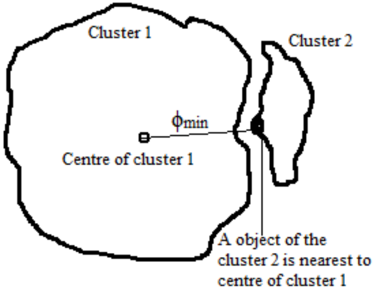

Figure 3. Two clusters tend to fit together

Call $\beta_{\max}$ is the largest coefficient of the β coefficients after clustering.

After clustering if a certain β is close 1, it means $\beta_{\max}$ is close 1. Then the two clusters need to be coupled into a cluster.

III. Improved K-Means Algorithm

A. Some Concepts about Cluster

Distance between two objects in a data set is signed d.

The large distance between all objects within a cluster is signed dmax.

The average distance among all objects within a cluster is signed d_{avr}.

The smallest distance from center of the cluster to different cluster is signed $\phi_{\min}$.

The average distance from the center of the cluster to different cluster is signed ϕ_{avr}.

Goal of clustering algorithms is minimization of the distortion of clusters and maximization of the obvious division among clusters. A good clustering result makes balanced clusters. With the definitions above; $d_{\max}$, d_{avr} obviously are factors reflecting of inside distortion of cluster. Also $\phi_{\min}$, ϕ_{avr} are factors reflecting of distance between the boundary of the cluster and a center of different cluster.

B. Algorithm K-means++

The improved K-means algorithm, combining of traditional K-means and agglomerative hierarchical clustering, is described as follow:

1) Input n objects.
2) Input fuzzy parameter *m*>1, input epsilon small enough.
3) Input weighted vector W:

$$\sum_{l=1}^{k} w_l = 1$$

where *k* is the number of dimensions of x_i.

4) Input number of clusters k ($1 \le k \le n$).
5) Applying agglomerative hierarchical clustering. Place each object in its own cluster. Calculate the distance of each cluster to another cluster, then find the closest distance. The two clusters that have that closest distance will be merged into a larger cluster.
6) Continue merge these clusters, until all of the objects are in k clusters.
7) From now on, applying K-means algorithm. Computes mean of the objects in the cluster. For each of the remaining objects, an object is reassigned to the cluster to which it is the most similar, based on the distance between the object and the cluster center. It then computes the new mean for each cluster.
8) Repeat the above step until no change.
9) Calculate $d_{\max}$, d_{avr}, α, $\alpha_{\max}$:

$$d_{\max} = \underset{q}{Max}\sqrt{\sum_{l=1}^{k} w_l\left(x_{il} - x_{jl}\right)^2}$$

and

$$d_{avr} = \frac{\sum_{q} \sqrt{\sum_{l=1}^{k} w_l \left(x_{il} - x_{jl} \right)^2}}{q}$$

where *j*=1,..,*p*, i = 1,..,p, i # j, *p* is number of objects in the each cluster, k is number of demensions, *q* is number of distances between objects in the each cluster.

$$\alpha = \frac{d_{\max}}{d_{avr}}$$

and $$\alpha_{\max} = \underset{c}{Max}(\alpha)$$

where c is number of clusters.

10) Calculate $\phi_{\min}$, ϕ_{avr}, β, β_{max}:

$$\phi_{\min} = \underset{p}{Min} \sqrt{\sum_{l=1}^{k} w_l \left(x_{il} - c_{jl} \right)^2}$$

$$\phi_{avr} = \frac{\sum_{p} \sqrt{\sum_{l=1}^{k} w_l \left(x_{il} - c_{jl} \right)^2}}{p}$$

where C_j is a center of cluster *j*, *j* = 1,..,*c*, *i* = 1,..,*p*, where *c* is number of clusters, *p* is number of objects in cluster x, *k* is number of dimensions.

$$\beta = \frac{\phi_{\min}}{\phi_{avr}}$$

and $$\beta_{\max} = \underset{c(c-1)}{Max}(\beta)$$

where *c* is number of cluster.

11) Based on the results of calculations α_{max}, β_{max}, if the number of clusters is unsuitable then it returns to step 2 to adjust the number of clusters according to indicating of α_{max}, β_{max}.

Otherwise the algorithm ends.

IV. Experiment Result

A. Analysis of Data on the Data Sets

Set:

$$f(c) = \underset{c}{Max}(\alpha) = \alpha_{\max}$$

$$g(c) = \underset{c(c-1)}{Max}(\beta) = \beta_{\max}$$

where c is number of cluster.

Running the algorithm with the number of clusters c from 1 to 9 with input data is a set including 91 records and a set including 695 records. The α_{max}, β_{max} obtained in each run is presented in Table I

Looking at the statistics table (Table I) and the graph of f(c), it predicts that the location of the number of appropriate clusters is in the neighborhood of the point at which the graph intends to go horizontally.

TABLE I. Statistic Table of A_{MAX}, B_{MAX}

Number of clusters c	Set including 91 records		Set including 695 records	
	α_{max} f(c)	β_{max} g(c)	α_{max} f(c)	β_{max} g(c)
1	15.45	0	42.84	0
2	7.78	0.79	14.13	0.70
3	4.82	0.93	7.05	0.93
4	3.45	0.98	5.14	0.98
5	3.52	1	4.46	0.98
6	3.87	1	4.04	0.99
7	3.87	1	4.17	1
8	3.89	1	3.95	1
9	3.85	1	3.82	1

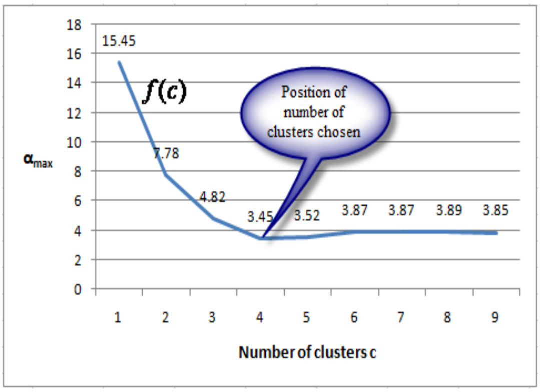

Figure 4. The graph shows the variation of f(c) and position for selecting of appropriate number of clusters on a set of 91 customers

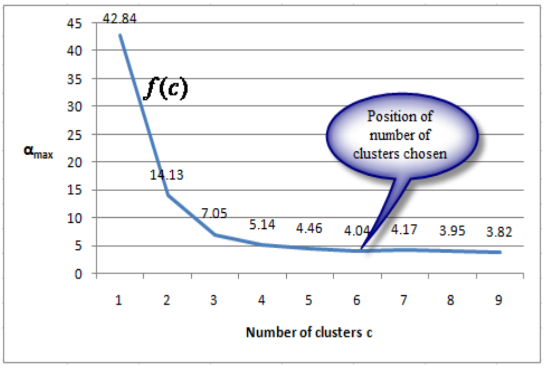

Figure 5. The graph shows the variation of f(c) and position for selecting of appropriate number of clusters on a set of 695 customers

Similarity, the graph of g(c) predicts that the location of the number of appropriate clusters is in the neighborhood of the point at which the graph increases approximately to 1, begining of tendency going across.

B. Experiment Result

In this session, we present the results of applying the improved K-means algorithm on data of the approximately 1000 patient records from MQIC database that were used to develop the Health Visualizer. Every object has 4 attributes:

Age, Diab, Hypertension and BMI. The distance measure used is the Euclidean distance.

Running the K-means++ with the number of clusters k from 2 to 10 with input data is a set including 500 records. The results are presented in Table II and Table III. We also provide the Davies–Bouldin index to evaluate the clustering results. Since algorithms that produce clusters with high intra-cluster similarity and low inter-cluster similarity will have a small Davies–Bouldin index.

TABLE II. STATISTIC TABLE AFTER RUNING TRADITIONAL K-MEANS WITH 500 RECORDS

No of clusters	α_{max}	β_{max}	Davies–Bouldin index
2	1.6560927129	0.679889904391	0.50886175404
3	1.6690286418	0.940263266254	0.41914956304
4	1.6474335185	0.967347184540	0.90272767337
5	1.6474335185	0.990146068766	0.30162327378
6	1.6604463350	0.988635121911	1.0596422255
7	1.7191484342	0.988635121911	1.32129589587
8	1.7032789487	0.988635121911	1.01053352555
9	1.6654647193	0.988635121911	1.69919300282
10	1.671040357	0.9886351219	2.21630151249

TABLE III. STATISTIC TABLE AFTER RUNNING K-MEANS++ WITH 500 RECORDS

No of clusters	α_{max}	β_{max}	Davies–Bouldin index
2	1.67103955531	0.910034156	0.6134555359549
3	1.6710395553	0.97910313	0.351749353276
4	1.6710395553	0.99003896	0.288680606141
5	1.6710395553	1	0.359118830370
6	1.6712867283	1	0.390487998676
7	1.6629464080	1	0.468652709143
8	1.6311254393	1	0.47130436473
9	1.6311254393	1	0.508404795624
10	1.6327367664	1	0.602087192374

Look at Table II, since using the traditional K-means, it is quite difficult to choose the suitable number of clusters k. We may base on the smallest Davies–Bouldin index to choose k (k = 3), and α_{max}, β_{max} are also small enough (α_{max} = 1.66 and β_{max} = 0.94). If selecting k that has β_{max} = 1, the size of the clusters will be too small. If selecting k that has large α_{max}, the size of clusters will be large, then the similarity of the objects in the cluster is not high.

In Table III, after applying K-means++, it is easier to determine k based on two coefficients. α_{max} is already small, so we should choose k = 3 with β_{max} = 0.97.

Running the improved algorithm again, with 1000 records. The result is shown in Table IV. Here we choose k = 4. The similarity of the data objects in the each cluster is rather good. Also, the Davies–Bouldin index is smallest.

TABLE IV. STATISTIC TABLE AFTER RUNNING K-MEANS++ WITH 1000 RECORDS

No of clusters	α_{max}	β_{max}	Davies–Bouldin index
2	1.690734561	0.9167388531	0.3716078492
3	1.730427069	0.9816915768	0.3987139406
4	1.730427069	0.9897063678	0.3179442254
5	1.730427069	1	0.3373329941
6	1.727844717	1	0.3587848233
7	1.700482257	1	0.4304007094
8	1.700482257	1	0.4778225755
9	1.700482257	1	0.4387431591
10	1.700482257	1	0.4631025678

The weakness of the new method is its speed is slower than the traditional one. So if the data is huge, it is better to choose the algorithm based on speed or ease of use.

The choice of algorithms depend much on the collected data. Although the improved algorithm proposed in this paper provides a criterion to select number of clusters, its speed is not good, and need to be optimized. If you do not really care about the accuracy, may be it is a good way to choose tradition K-means, or another clustering algorithm.

V. CONCLUSION

The improved algorithm, proposed by Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai, optimizes K-means, as well as agglomerative hierarchical clustering. It overcomes the weakness of traditional K-means, makes the result is more easily analyzable, but may reduces the execution speed.

An approach proposed by D.T.Nguyen and H.Doan mentioned in this paper is a good way to determine the number of clusters using the clustering information obtained in the clustering process. It also provides a new measure for selecting the number of clusters.

The success of data clustering often depends on good data, rather than good algorithms. If the dataset is huge and not clear, your choice of clustering algorithm might not really matter so much in terms of performance, so you should choose your algorithm based on speed or ease of use instead.

REFERENCES

[1] D. T. Pham, S. S. Dimov, and C. D. Nguyen, “Selection of K in K-means clustering,” in *Proc. IMechE Mechanical Engineering Science*, 2005, vol. 219, pp. 103-119.

[2] Z. Yin, Y. Tang, F. Sun, and Z. Sun, “Fuzzy clustering with novel separable criterion,” *Tsinghua Science and Technology*, pp. 50-53, 2006.

[3] M. V. B. T. Santhi, V. R. N. Sai Leela, P. U. Anitha, and D. Nagamalleswari, “Enhancing K-Means clustering algorithm,” *International Journal of Computer Science & Technology* , vol. 2, no. 4, pp. 73-77, Nov 2011.

[4] F. E. Mohammad and M. A. Wesam, “Initializing K-Means clustering algorithm using statistical information,” *In’t Journal of Computer Applications*, pp. 51-55, 2011.

[5] W. Haizhou and S. Mingzhou, “Optimal k-means clustering in one dimension by dynamic programming,” *The R Journal*, vol. 3, pp. 29-33, 2011.

[6] M. Y. Shih, J. W. Jheng, and L. F. Lai, “A Two-Step method for clustering mixed categorical and numeric data,” *Tamkang Journal of Science and Engineering,* vol. 13, no. 1, pp. 11–19, 2010.

[7] D. T. Nguyen and H. Doan. “An approach to determine the number of clusters for clustering algorithms,” in *Proc. 4th International Conference ICCCI,* Vietnam, LNAI, Springer-Verlag, Nov 2012, vol. 7653, pp. 485–494.

Dr. Dinh Thuan Nguyen received the Ph.D degree in Information Technology from Institute of Information Technology, Vietnamese Academy of Science and Technology, Viet Nam in April 2004, MSc degree in Information Technology from University of Natural, Vietnam National University of Ho Chi Minh City, Viet Nam in December, 1998, and BSc degree in Mathematics from Dalat University, Vietnam in August 1984. He is presently working as Dean of Faculty of Information Systems, University of Information Technology, VNU-HCM. He has published more number of research papers in journals, books, conferences, and workshops. His research interest include Data Mining and Business Intelligence.

Mr. Toan Nguyen Gia received the B.E degree in Information Systems from University of Information Technology, VNU-HCM. His area of interests includes Data Mining and Spatial Databases.

Mr. Vu Tuan Nguyen Lam received the B.E degree in Information Systems from University of Information Technology, VNU-HCM. His area of interests includes Data Mining and Business Intelligence.