Application of an Improved K-means Algorithm in Gene Expression Data Analysis

Qian Ren
School of Science
Beijing University of Posts and Communications
Beijing ,China
renqian-86517@163.com

Xinjian Zhuo
School of Science
Beijing University of Posts and Communications
Beijing ,China
zhuoxj@bupt.edu.cn

***Abstract*-K-means algorithm is one of the most classic partition algorithms in clustering algorithms. The result obtained by K-means algorithm varies with the choice of the initial clustering centers. Motivated by this, an improved K-means algorithm is proposed based on the Kruskal algorithm, which is famous in graph theory. The procedure of this algorithm is shown as follows: Firstly, the minimum spanning tree (MST) of the clustered objects is obtained by using Kruskal algorithm. Then K-1 edges are deleted based on weights in a descending order. At last, the average values of the objects contained by the k-connected graphs resulting from last two steps are regarded as the initial clustering centers to cluster.Make the improved K-means algorithm used in gene expression data analysis, simulation experiment shows that the improved K-means algorithm has a better clustering effect and higher efficiency than the traditional one.**

Keywords-Clustering; K-means Algorithm; Kruskal Algorithm; MST; Gene expression data

I. INTRODUCTION

As the physical basis of heredity, DNA determines the synthesis of protein. Gene is able to produce a specific protein fragment of DNA sequence, which is the basic unit of human genetics. To detect the gene expression level, in 1990s, people developed the DNA microarray gene chips, that is, gene chip. It can simultaneously measure thousands of gene expression data which are called DNA microarray data or large-scale gene expression profiling. How to organize, analyse and deal with these vast amounts of gene expression data to extract the effective biological and medical information has attracted attention from people. Because of the large, high dimensional and heterogeneous features of the gene expression data, produced massive gene expression data should be dealt with an effective analytical technique. However, clustering methods in data mining can meet this end, for it tends to classify each object to a different subset according to some similarity measure while making the objects of different subsets different and the objects of the same subset similar.

Traditional clustering methods generally include followings: (1) method based on partition; (2) method based on level; (3) method based on density; (4) method based on grid; (5) method based on model. In addition, fuzzy clustering method, kernel clustering method and spectral clustering method also belong to them[1]. Until now, many clustering algorithms have already been used for cluster analysis of gene expression data, such as hierarchical clustering, K-Means clustering, self organizing map algorithm (SOMs), PCA clustering algorithm, methods based on density and grid. All these clustering algorithms have been widely used in earlier studies, such as hierarchical clustering algorithm used by Eisen to identify co-regulated yeast genes[2], self-organizing map algorithm (SOMs) used by Tamayo to identify yeast cell cycle and differentiation of human hematopoietic cell data[3]. And at present, the fuzzy clustering algorithm, clustering algorithm based on artificial neural network, intelligent algorithms are also widely applied in gene clustering analysis. For example, genetic algorithm (GA) which is one of the intelligent algorithms has been widely applied in the analysis of gene expression data clustering in the late 90s of last century. Among those pioneers are Lixin Tang, who put the combination of K-means and GA applied to the analysis of gene expression data[4] and Fangxiang Wu, who put forward with a hybrid algorithm GWKMA of weighted K-means algorithm and genetic algorithm in 2007[5].

As a typical partition method, K-means clustering algorithm has the advantages of easy and fast. However, this algorithm depends on the choice of initial clustering centers, because quite different results would be gotten to different initial values. Thus, the algorithm is easy to converge to a local minimum value if the initial clustering centers have been improperly chosen. To solve this problem, many improved and optimized algorithms have been proposed in recent years, for example clustering algorithm based on sampling by PaulS.Bradley[6], density estimation based on division by Moh'd B Al-Daoud, and Stuart A. Roberts[7], estimate the initial cluster centers of class by the local density of data by Kaufman[8], K-means algorithm that select the initial cluster centers by connected branches based on graph theory by Haiyan Zhou[9], algorithm based on Euclidean distance to determine the farthest distance between two data points, the data set are divided into k segments, then look for the cluster centers by Yuanyuan Bu[10], k-means algorithm based on greedy algorithm by Weiping Li[11], K-means algorithm based on partition by Jinqi Su[12]. In this paper, an

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

improved K-means algorithm based on the Kruskal algorithm in graph theory to select the cluster centers is given. By applying the improved algorithm into gene expression data analysis, the results of experiment show that this method lessens its dependence on initial cluster centers than traditional K-means algorithm, and increases the stability and accuracy of clustering results.

II. K-MEANS ALGORITHM

A. The basic principle of K-means algorithm

It is assumed that there were n clustered objects to get a sample set, that is, $X = \{x_1, x_2, ..., x_n\}$. By using K-means algorithm, n sample objects are grouped into K clusters to ensure the similarities among samples in the same cluster and the differences among samples in different clusters. Specific procedure is as follows:

1) Randomly select K objects as initial cluster centers as following: $c_1, c_2, ..., c_K$

2) According to the minimum distance principle, that is,

$$D_j = \min\left\|X - c_j\right\|, X = \{x_1, x_2, \ldots, x_n\}, j = 1,2, \ldots K \quad (1)$$

each sample object is assigned to one of K clusters

3) Take the average values of objects of each cluster as new clustering centers, average values can be get by

$$C_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i, \; j = 1,2, \ldots, K \quad (2)$$

n_j is the number of objects in the cluster j

4) If the cluster centers have changed, repeat 2), 3) steps until the cluster centers do not change. As a result, clustering criterion function can be converged

$$J_c = \sum_{j=1}^{K}\sum_{i=1}^{n_j}\left\|X_i^{(j)} - c_j\right\|^2, X_i^{(j)} \in S_j \quad (3)$$

C_j is the clustering center of cluster S_j

The flow of K-means algorithm is shown in Fig. 1.

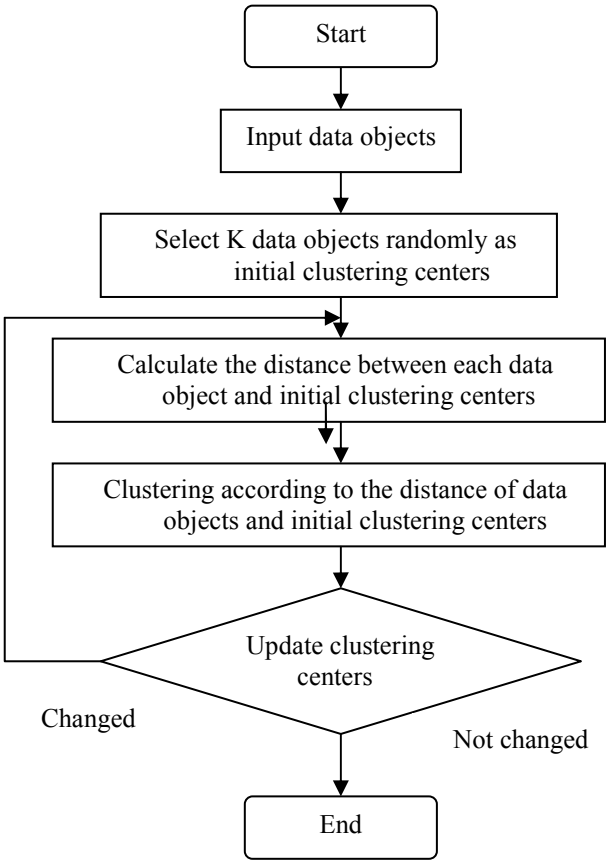
B. The shortcomings of K-means algorithm

K-means algorithm has two shortcomings:

(1) Different initial clustering centers may result in different clustering results;

(2) The algorithm is optimization algorithm based on the objective function, usually by gradient method to solve the extremum, so the algorithm is easy to converge to the local extremum result[13].

This paper aims at solving the first shortcoming, proposed an improved K-means algorithm based on Kruskal algorithm in graph theory to select the initial clustering centers.



Figuer 1. Flow of K-means algorithm

III. IMPROVED K-MEANS ALGORITHM OF CHOOSING THE CLUSTERING CENTERS BASED ON KRUSKAL ALGORITHM

A. The introduction of Kruskal algorithm

Assuming undirected connected graph G=(V, E), T=(U,H)is the minimum spanning tree of G, the initial values are U=V, H= $\varnothing$, so that the vertices of T formed a connected branch respectively. And then select edges from E according to the weights of edges in a descending order, if the vertices of the edge belong to two connected branches of T, add this edge to H and combine the two connected branches into one; If the vertices of the edge belong to the same connected branch, then delete the edge to avoid loop. As this goes on, when the number of connected branches come to 1, this connected branch is a minimum spanning tree of G.

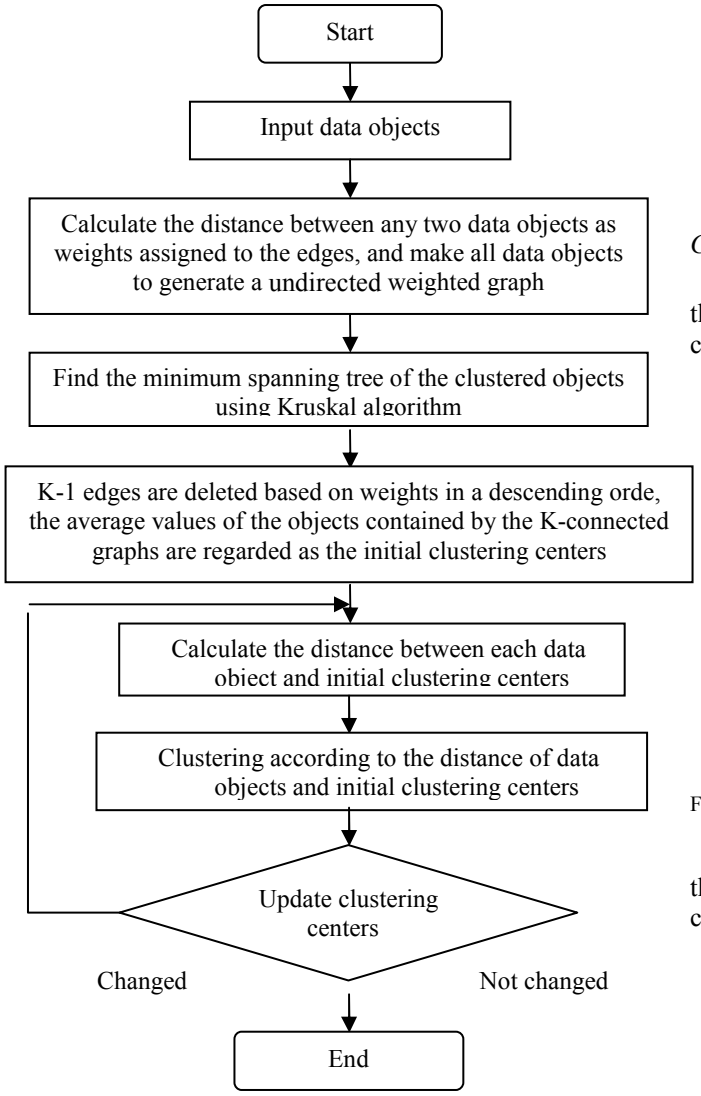
B. The thought of improved K-means algorithm

For a given data set $X = \{x_1, x_2, ..., x_n\}$, our task is to cluster the data objects of X. Firstly, decide the distance

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

values between the data objects. As Euclidean distance has been adopted in this paper, calculate the distance values between any two objects as weight is given to the edges connecting the two objects, so an undirected connected graph G(X) has been gotten. Secondly, obtain the minimum spanning tree (MST) of G(X) by using Kruskal algorithm, then K-1 edges are deleted according to weights in a descending order. At last, the average values of the objects contained by the k-connected graphs resulting from last two steps are regarded as the initial clustering centers. At last, use traditional K-means to do clustering analysis.

The flow of improved K-means algorithm is shown in Fig. 2.



Figuer 2. Flow of improved K-means algorithm

IV. SIMULATION AND ANALYSIS OF THE RUSULTS

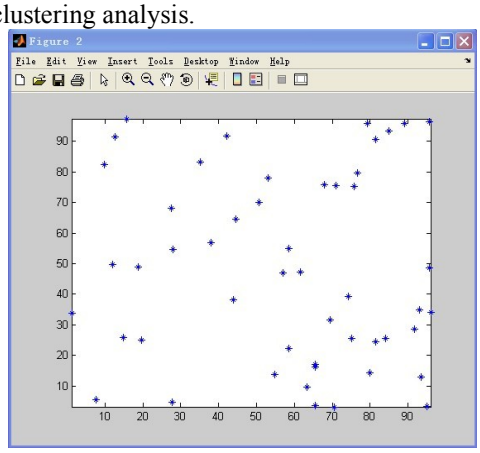
A. *Experimental environment*

Windows XP operating system, MATLAB7.1 programming language

B. *Experimental data*

Randomly generated 50, 100 and 200 two-dimensional data objects respectively among which 50 random two-dimensional data objects are shown in Fig. 3.

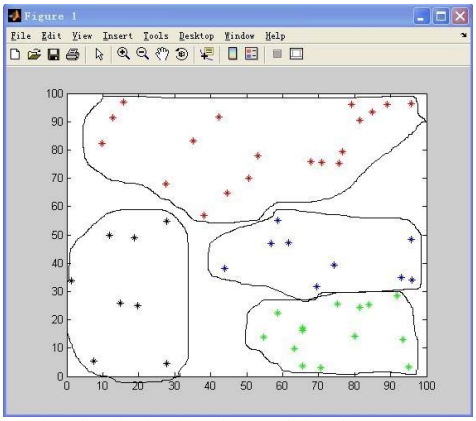
K, the number of clustered data objects, should be input before the clustering analysis.



Figuer 3. Distribution of 50 random two-dimensional data objects

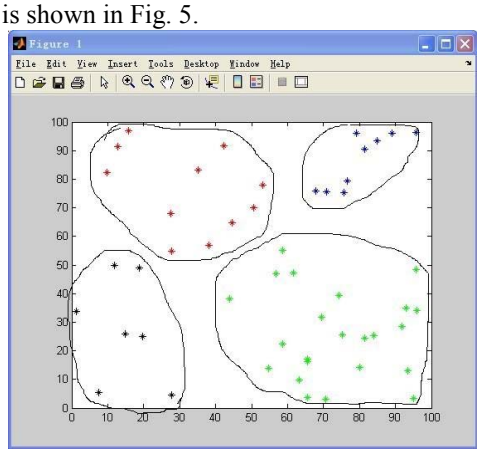
C. *Analysis*

By the traditional K-means clustering algorithm (K=4), the result of 50 random two-dimensional data objects clustered is shown in Fig. 4.



Figuer 4. Clustering distribution based on traditional K-means algorithm

By the improved K-means clustering algorithm (K=4), the result of 50 random two-dimensional data objects clustered is shown in Fig. 5.



978-1-4577-1666-9/11/$26.00 ©2011 IEEE

Figuer 5. Clustering distribution based on improved K-means algorithm

Take the randomly generated 50, 100 and 200 two-dimensional data objects which are grouped according to different categories to do 5 experiments using traditional K-means algorithm, the improved algorithm and algorithm mentioned in references[11] (select two farthest data objects to form a set U. Then select data whose distance with the data objects of U are maximum in the remaining n-2 data objects to add into U. Continue this until the number of the data centers in U equals to K. At last, take these K data objects as initial cluster centers to cluster) respectively. The objective function values are shown in TABLE Ⅰ, TABLE Ⅱ,TABLE Ⅲ.

TABLE I. FUNCTION VALUES OF 50 RANDOM TWO-DIMENSIONAL DATA OBJECTS BASED ON K IS 2,3,4,5 RESPECTIVELY

Number	K-means algorithm	Improved algorithm	Algorithm mentioned in[11]
K=2	57439	47887	47840
K=3	35197	28877	28521
K=4	28122	18165	18258
K=5	26289	14580	14918

TABLE II. FUNCTION VALUES OF 100 RANDOM TWO-DIMENSIONAL DATA OBJECTS BASED ON K IS 2,3,4,5 RESPECTIVELY

Number	K-means algorithm	Improved algorithm	Algorithm mentioned in[11]
K=2	113725	104957	104473
K=3	74689	62043	61775
K=4	49001	38045	38968
K=5	33643	29500	29851

TABLE III. FUNCTION VALUES OF 200 RANDOM TWO-DIMENSIONAL DATA OBJECTS BASED ON K IS 2,3,4,5 RESPECTIVELY

Number	K-means algorithm	Improved algorithm	Algorithm mentioned in[11]
K=2	226900	216040	215940
K=3	143628	131230	130525
K=4	100452	88687	89453
K=5	80159	72887	73991

From the Fig. 4 and Fig. 5, we can know that as improved K-means algorithm has gotten the initial centers by Kruskal algorithm, the clustering results are stable and the overall distribution are more uniform, it improves clustering analysis results and reduces the dependence on the initial cluster centers.

Because the initial values are randomly selected by K-means algorithm, the objective function values are different in each experiment and the results are volatile. Here, we just take the average values of 5 experiments. From TABLE Ⅰ, and TABLE Ⅱ,TABLE Ⅲ, the objective function values gotten by the improved algorithm are much smaller than by the traditional K-means algorithm, in accordance with the requirements of objective function value reached to a minimum. Compare with the results gotten by the algorithm mentioned in references [11], data in the TABLE shows that when K is large, the objective function value is smaller by using of this improved algorithm for clustering.

V. APPLICATION OF AN IMPROVED K-MEANS ALGORITHM IN GENE EXPRESSION DATA ANALYSIS

This paper adopts Yeast Cell Cycle data of Stanford University which have genes with known functions, because clustering of these data objects has an immediate significance to the evaluation of clustering algorithm's application in gene expression data. There are 130 genetic data of the Yeast Cell Cycle selected in this paper and each gene expression data is 76 dimensional, resulting from different experimental conditions and time point combination. Gene can be divided into 5 categories according to their RNA expression levels in the cell cycle time of the peak point, including 41 G1 type, 20 S type, 22 G2 type, 24 M type and 20 M/G1 type, G, S, M respectively represent the periods of cell's Growth, Synthetic and Mitosis.

By K-means algorithm and the improved algorithm, these data objects have been clustered respectively and the results are compared in TABLE Ⅳ, we can see that the accuracy by the improve algorithm is higher than by the traditional K-means algorithm, the objective function value and the number of iterations are less than by the traditional K-means algorithm. However, time of using improved algorithm is nearly 5 times than using the traditional K-means algorithm, because the process of obtaining the minimum spanning tree is relatively time-consuming.

TABLE IV. THE COMPARISON OF TWO ALGORITHM IN GENE EXPRESSION DATA ANALYSIS

Clustering algorithm	Accuracy	Objective function value	Iterations	Running time
K-means algorithm	65%	1805	11	47
Improved algorithm	72%	1780	7	210

VI. CONCLUSION

As the clustering results by using the traditional K-means clustering algorithm depends on the choice of the initial clustering centers, how to overcome this problem is the purpose of this paper. Thus, the paper here tries to obtain the minimum spanning tree (MST) of the clustered objects by using Kruskal algorithm in graph theory to get the initial clustering centers, which improves K-means clustering algorithm. Then make this improved K-means algorithm used in gene expression data analysis. From simulation experiments, this algorithm has successfully solved the problem that traditional K-means algorithm depend on the initial clustering centers, and it has a better accuracy and stability.

REFERENCES

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

[1] Han J,Kamber M. Data Mining: Concepts and Techniques [M].The 2nd Edition,Singapore:Elsevier Inc,2006.285-382.

[2] Eisen MB,Spellman PT,Brown PO,et a1.Cluster analysis and display of genome-wide expression patterns[J].Proc Natl Acad Sci USA,1998,95:14863-14868.

[3] Tamayo P,Slonim D,Mesirov J,et a1.Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation[J].Proc Natl Acad Sci USA,1 999,96:2907-2912.

[4] Lixin Tang, Zihou Yang. Improve the K-means algorithm using genetic algorithms [J]. Mathematical Statistics and Applied Probability,1997,12(4):350-356.

[5] Fangxiang Wu.A Genetic Weighted K-means Algorithm for Clustering Gene Expression Data[J].BMC Bioinformatics,2008,9(6):68-75.

[6] Bradley P S, Fayyad U M.Refining initial points for K-Means clustering [C]//Proc. of the 15th International Conf. on Machine Learning. San Francisco, CA: MorganKaufmann, 1998: 91-99.

[7] Moh'd B Al-Daoud, Stuart A Roberts. Newmethods for the initialization of clusters [J].Pattern Recognition Letters, 2001 (17) : 451- 455.

[8] Kaufman L,Rousseeuw P J. Finding groups in data:an introduction to cluster analysis[M]. NY:John Wiley &Sons ,1990.

[9] Haiyan Zhou,Xiaolin Bai. The choice of initial cluster centers based on Graph and K-means algorithm [J]. Computer Measurement & Control,2010,18(9):2167-2169.

[10] Yuanyuan Bu, Zhongren Guan. The research of K-Means clustering algorithm [J].The Nationalities of Guangxi University 2009:35(1):198-200.

[11] Weiping Li. The improvement of K-means clustering algorithm [J]. West China, 2010,9(24):49-50.

[12] Jinqi su,Huifeng Xue. K-means Initial Clustering Center Optimal Algorithm Based on Partitioning [J]. Microelectronics and Computer,2009,26(1):8-11.

[13] Guangqiu Huang,Xideng Wang. An Improved Artificial Fish-Swarm Algorithm Based on Gridding Method [J] Microelectronics and Computer,2007,24(7):83-86.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE