An Introduction of Multi-Step K-means Clustering Applied to Rice Microarray Data

[†]Daehoon Park, [†]Yountae Kim, [††]Ok-Kyung Ham, [†]Sungshin Kim, and [††]Choon-Hwan Lee

[†]School of Electrical and Computer Engineering, Pusan National University, Busan, Korea
Pusan National University 30 Changjeon-dong, Keumjeong-ku, Busan 609-735, Korea
E-mail: dhsmile@pusan.ac.kr, dream0561@pusan.ac.kr, sskim@pusan.ac.kr

[††]School of Molecular Biology, Pusan National University, Busan, Korea
Pusan National University 30 Changjeon-dong, Keumjeong-ku, Busan 609-735, Korea
E-mail: 82dhrrud@pusan.ac.kr, chlee@pusan.ac.kr

Abstract – Long gene sequences and their products have been studied by many methods. The use of DNA(Deoxyribonucleic acid) microarray technology has resulted in an enormous amount of data, which has been difficult to analyze using typical research methods. This paper proposes that mass data be analyzed using division clustering with the *K*-means clustering algorithm. To demonstrate the superiority of the proposed method, it was used to analyze the microarray data from rice DNA. The results were compared to those of the existing *K*-means method establishing that the proposed method is more useful in spite of the effective reduction of performance time.

Keywords: K-means clustering, Microarray, Rice

1. Introduction

Since the inception of biotechnology, large amounts of new genetic information have been amassed for organisms varying from bacteria to human beings. This information has provided many clues for solving biological problems. However, new technologies are urgently needed because most genetics engineering methods have limitations. One method that has been developed to overcome the problems of the existing methods is searching genetic material using a DNA chip. The DNA chip is divided into a cDNA chip and oligonucleotide chip depending on the size of the genetic material. A large amount of genetic information can be obtained by using this DNA chip analysis technique [1]. The development of an efficient clustering algorithm for DNA microarray data will contribute to research in several important fields such as functional genomics and genetic networks. Furthermore this information can be analyzed by various data mining methods, and these results can be evaluated by many methods. A literature review of former data mining methods reveals a data clustering algorithm using graph theory and an algorithm by Hartuy and Ben-Dor et al. [2] [3], and Tamayo et al. developed the Self-Organizing Maps (SOM) algorithm [4]. And Eisen et al. proposed and developed a method using hierarchical clustering [5].

The most typical method of analysis for microarray is clustering. Clustering analysis classifies a large amount of genetic information into several groups that have similar properties, so it is effective for analyzing data. The hard clustering method has been especially useful because it is intuitive and its ease of use.

In this paper division clustering using a two step structure *K*-means clustering method is proposed in order to treat process the thousands of pieces of microarray data effectively. The proposed method processes enormous amounts of microarray data faster because it classifies and clusters data twice. The reliability of these clusters has been evaluated by comparing these results with those of the earlier *K*-means clustering method.

2. Microarry

People usually think that enormous number of genes and their products of a living organism create the mystery of life. Most molecular biological methods are used to study one gene during one experiment. Consequently, the DNA microarray method has attracted wide-spread attention from biologists to monitor the whole genome on a single chip so that researchers can better monitor the results of the simultaneous interactions of the numerous genes [6]. The DNA microarray technique is comprised of the cDNA microarray and oligonucleotide microarray techniques.

Microarray data from an experiment on the manifestations of a gene can be classified using the data clustering method. Thus results can be used in many fields including the development of drugs and toxicological research. In this paper we used the microarray data from rice genes consisting of 17,000 units.

3. Clustering Algorithm

3.1 Several Clustering Algorithms

The clustering algorithm classifies a whole data set into several clusters that have similar characteristics. This algorithm has been used in a variety of fields such as pattern analysis and classification, grouping, decision making, machine-learning situations, and data mining [7]. This clustering algorithm has been developed by experts from various scientific fields such as statistics, computer science and biology. Now the application of the clustering algorithm is more actively studied than the algorithm itself. The clustering algorithm is largely divided into a hierarchical clustering algorithm and a partitional clustering algorithm [7].

In the clustering algorithm, it is important to define the similarity of two clusters. Similarity was determined by Euclidian distance and measured using equation (1).

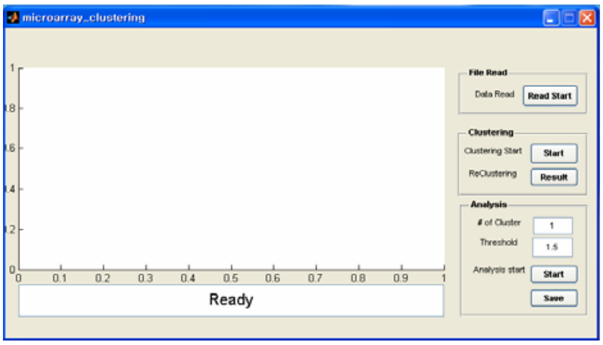
$$d(x_i, x_j) = \left\{ \sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2 \right\}^{\frac{1}{2}} \tag{1}$$

3.2 *K*-means Clustering Algorithm

The *K*-means clustering algorithm used in this paper is the most frequently used of the partitional clustering algorithms. *K*-means clustering differs from hierarchical clustering in that the number of clusters, *k*, needs to be determined at the outset. The goal of the *K*-means clustering algorithm is to divide the objects into *k* clusters such that some metric relative to the centroids of the clusters is minimized. Two procedures are available to search for the optimum set of clusters. The first assigns each object to a cluster and the second sets initial positions for the cluster centroids. The *K*-means algorithm consists of the *K*-means method and K-medoid method. In this paper we used the average value as the center of the cluster.

4. Implementation of the Division System Using the Two-Step Structure K-means Algorithm

In this paper, we used Matlab 7.1 and GUI is shown in Figure 1. GUI is divided into three parts: input, clustering, and results. On the left-side of GUI, the results are shown in the graph window and character window. On the right-side of the GUI, the button that are used to operate the algorithm in the order they are needed.



Ready

# of Cluster 1

Threshold 1.5

Fig. 1. GUI of the divided clustering system.

Figure 2 shows the flow chart of the program used in this paper, as can be seen, the rice gene microarray data has been normalized and clustered twice. The clustering step ends when the data has been clustered for the second time then the data from the center value of each cluster can be gathered, and the results saved.

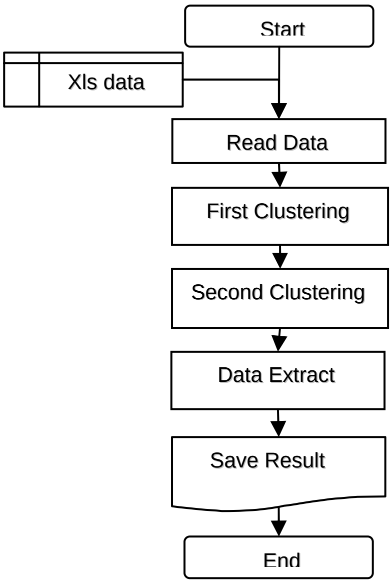


Fig. 2. Flow chart of the proposed algorithm.

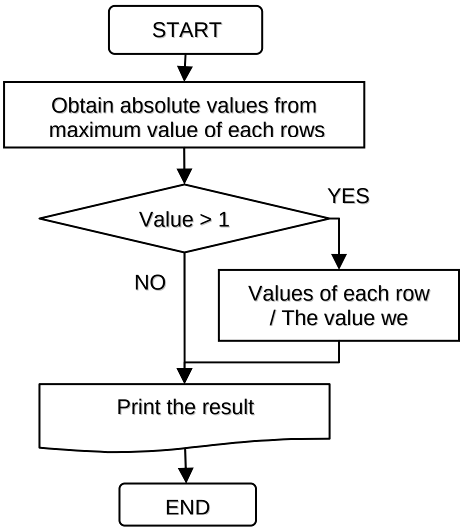

START
Obtain absolute values from maximum value of each rows
Value > 1
YES
NO
Values of each row / The value we
Print the result
END

Fig. 3. Normalization flow chart of normalization.

4.1 Normalization

The microarray data used in this paper has a value between -1 and 1, but a few datum are out of this range. The clustering algorithm is based on the similarities between the data, so if the data is outside of the normal range, that value determines its total similarity. Therefore this data needs to be adjusted so that it is between -1 and 1. Consequently it is divided by the maximum or minimum value to normalization. Figure 3 shows the normalization flow chart.

4.2 Division Clustering Algorithm

The data consisted of 17,000 units, so it took too long to cluster the entire data set at one time. First, the 17,000 units were divided into 17 groups consisting of 1,000 units, and then each group was clustered. In this paper, *k* was set at 36. The clustering algorithm was performed a second time using the typical values of the first clustering. Thus the data was processed twice, and 17,000 units of data were clustered.

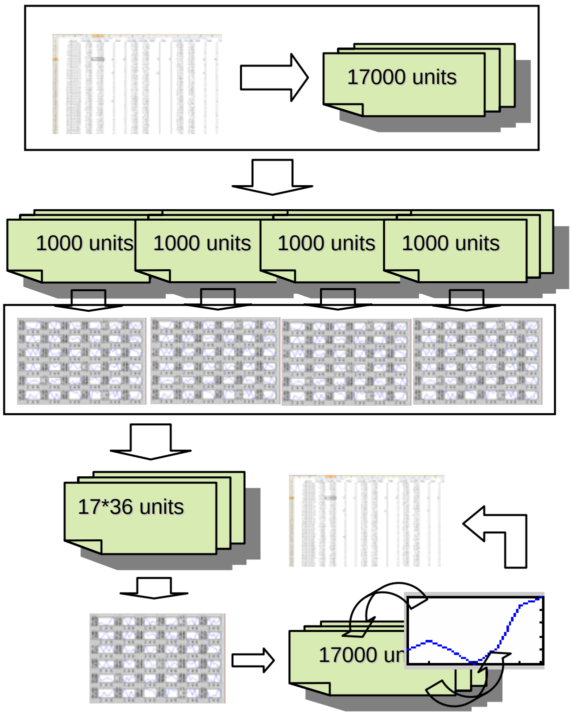


Fig. 4. Process of dividing the clustering.

4.3 Data Analysis and Saved Data

The results of the clustering determined the data that was saved. The graph file of typical values and the text file of the gene numbers were saved from the clustering results.

5. Simulation and Results

In this paper the clustering method for rice microarray data was analyzed and a new clustering method for a large amount of data like that in the rice microarray data was proposed. Figure 5 shows the results of the clustering of the proposed algorithm.

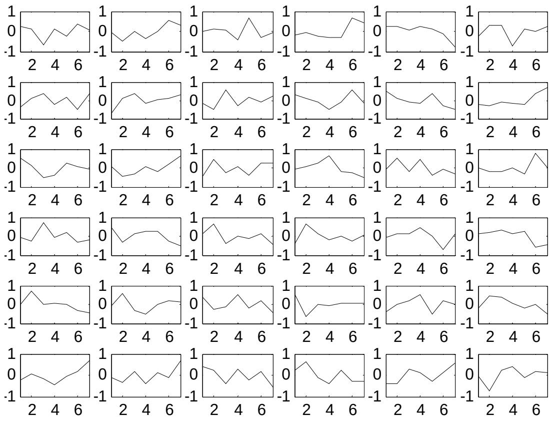


Fig. 5. The clustering results using the proposed method.

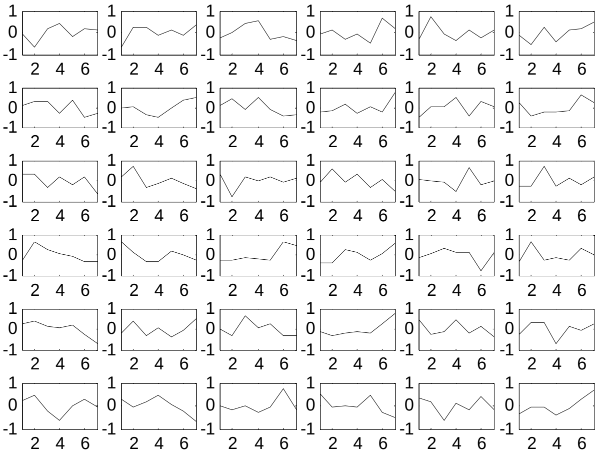


Fig. 6. The clustering results using the previous method.

Figure 6 shows the result of the previous method that clustered the entire data set of 17,000 units at one time. Table 1 compares the results of both method using two values. First, the time each method needed to cluster each data set was compared, and the shorter time is considered to be better. And second, the standard deviation between the typical values resulting from each algorithm was compared, and the larger standard deviation is an indication of a cleaver classification of the data. A computer with a 3GHz CPU and 512 MB RAM was used to run the algorithm. The previous method required 251 seconds to cluster the data and the proposed method, 47.76 seconds. Therefore, the proposed method sharply reduced the clustering time and is better than the previous method. The standard deviation of the previous method is 0.7702, and the standard deviation is proposed method, 0.7987. Consequently it can be concluded that the proposed method is better.

Table 1. Comparison of the performance of the previous method and proposed method.

clustering method evaluation standards	former method	proposed method
performance time (sec)	251	46.45
standard deviation	0.7702	0.7987

Figure 7 compares two typical values of the two methods when *k* is 36. As can be seen, the clustering results are similar. Figure 8 compares the methods when *k* is 16, and the result are also similar.

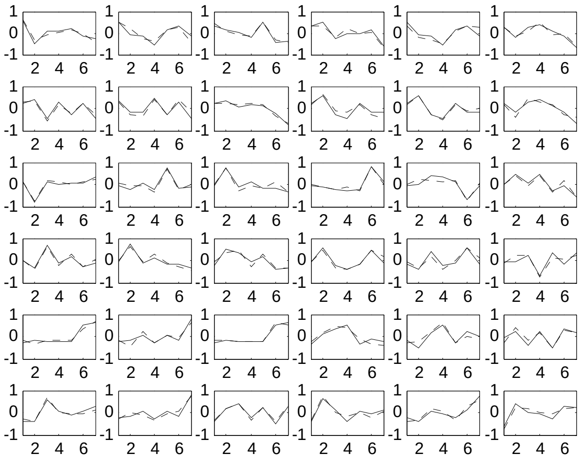


Fig. 7. Comparison of two typical values of the two methods using 17 groups, and a *k* of 36.

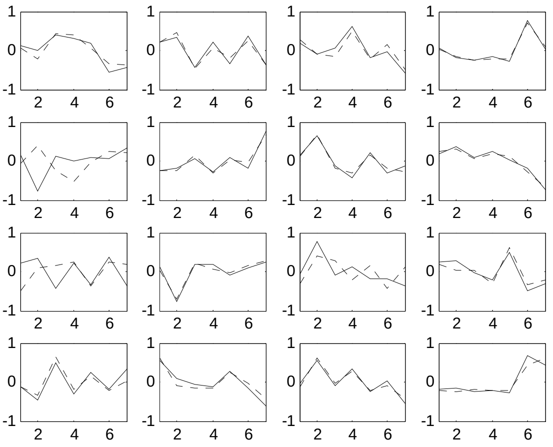


Fig. 8. Comparison of two typical values: two methods using 17 groups, and a *k* of 16.

6. Conclusion

Most of the study methods used previously have limits. For example, the hierarchical clustering methods cannot process mass data. The previous partitional clustering method can takes more time than proposed method to classify data. In this paper, clustering was divided into two steps using *K*-means clustering. This method allowed a large amount of data to be processed more quickly than the previous *K*-means method but with similar results. The speed of the proposed algorithm needs to be verified using other large data sets.

ACKNOWLEGEMENT

This work was supported by the Second-stage of the Brain Korea 21 project in 2006.

References

[1] Hwang, Seung Yong: DNA chip technology. *Korea Information Science Society*, 18(8).(2000), 23- 28 (in Korean)

[2] Hartuv, E., Schumitt, A., Lange, J., Meier-Ewert, S., Legrach, H. and Shamir, R.: An Algorithm for Clustering cDNAs for Gene Expression Analysis. *Proceed ings of the Third International Conference on Computational Molecular Biology (RECOMB 99)*, (1999), 188- 197

[3] Ben-Dor, A., Shamir, R., Yakhini, Z.: Clustering Gene Expression Patterns. *Journal of Computational Biology*, 6. (1999), 281- 297

[4] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R.: Interpreting patterns of gene expression with self-organizing maps : Methods and application to Hematopoeitic differentiation. *Proceedings of National Academy of Sciences of the USA* , 96. (1999), 2907- 2912

[5] Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D.: Cluster analysis and display of genome-wide expression patterns. *Proceedings of National Academy of Sciences of the USA* , 95. (1998), 14863- 14868

[6] http://gene-chips.com

[7] Jain, A. K., Murty, M. N. and Flynn, P.J.: Data clustering: A review. *ACM Computing Surveys*, 31(3). (1999), 264-323