A COMBINED K-MEANS AND HIERARCHICAL CLUSTERING METHOD FOR IMPROVING THE CLUSTERING EFFICIENCY OF MICROARRAY

Tung-Shou Chen[1], Tzu-Hsin Tsai[2], Yi-Tzu Chen[3], Chin-Chiang Lin[4], Rong-Chang Chen[5], Shuan-Yow Li[2] and Hsin-Yi Chen[1]*

[1.] Dept. of Information Management, National Taichung Insitute of Technology
[2.] Institute of Medicine, Chung Shan Medical University
[3.] History of Art, Commonwealth Open University
[4.] Graduate Institute of Computer Sience and Information Technology, National Taichung Insitute of Technology
[5.] Dept. of Logistics Engineering and Management, National Taichung Insitute of Technology
*dalamar.shalafi@msa.hinet.net
No. 129, Sec. 3, Sanmin Rd., Taichung, Taiwan 404, ROC

ABSTRACT

Among the microarray data analysis clustering methods, K-means and hierarchical clustering are researchers' favorable tools today. However, each of these traditional clustering methods has its limitations. In this study, we introduce a new method, hierarchical K-means regulating divisive or agglomerative approach. The hierarchical K-means firstly employs K-means' algorithm in each cluster to determine K cluster while operating and then employs it on hierarchical clustering technique to shorten merging clusters time while generating a tree-like dendrogram. We apply this method in two original microarray datasets. The result indicates divisive hierarchical K-means is superior to hierarchical clustering on cluster quality and is superior to K-means clustering on computational speed. Our conclusion is that divisive hierarchical K-means establishes a better clustering algorithm satisfying researchers' demand.
Keywords: clustering, hierarchical, K-means, divisive

1. INTRODUCTION

Advances in microarray technology have made it possible to simultaneously monitor the expression of thousands of genes in genomes. The challenge is the technology should effectively analyze this large volume of information and accurately interpret the data. Clustering algorithms select similar genes, which should be considered members of a cluster, so as to help researchers interpret the gene relationship. However, an outstanding clustering should ensure that genes are similar to each other in the same clusters and different from other genes in other clusters. Until recently, many cluster-evaluation tools, such as hierarchical clustering and K-means, have been intended to formalize these situations. Notwithstanding the two are essentially excellent, researchers still long for alternative programs [1-3].

In K-means clustering, the method selects initial predetermined K cluster centroids and calculates the proximities from each point to all K centroids. When each datum is assigned to the K cluster members, the data are reallocated to one of the new clusters. The problem comes out that the iterative process will stop if the reassignment satisfies the criteria set by initial points. Different starting points may result in different clustering partitions. That is, the K-means algorithm may only find local optimum rather than global one. Consequently, this method suffers from the defeat that different runs of K-means on the same input data might produce different solutions [4-7].

The traditional hierarchical clustering method is an agglomerative approach, which organizes similar branch points into a cluster based on the choice of the distance measure and, therefore, results in a tree-like dendrogram. Usually this method does not guarantee that the within-dendrogram similarity is maximized because each cluster may consist of several different sub-clusters. The shortcoming is originated from visual inspection rather than standard criterion or algorithm for choosing a cut-off point for dendrogram; probably, this leads to the non-uniqueness of the dendrogram [8, 9].

Researchers have the thinking that there is no single best clustering method for all datasets and no single best way to evaluate a clustering method, so complementary methods may be helpful in analyzing datasets. On the other hand, these algorithms which are usually computationally expensive impede the wide application of them in practice such as in gene expression data analysis [10-12]. To tackle the addressed problems, we propose a new method, hierarchical K-means, which intends to improve the clustering quality and to shorten the merging time.

0-7803-9266-3/05/$20.00 ©2005 IEEE.

In the remainder of this paper, we first describe the hierarchical K-means algorithm. Experimental results are also presented to support the validity of the proposed approach. We conclude this paper with some remarks.

2. METHOD

The proposed hierarchical K-means aims on using K-means method to decide K clusters before clustering. This new approach generates two functions, divisive hierarchical K-means (divisive HK) and agglomerative hierarchical K-means (agglomerative HK). Divisive HK follows a top-down approach because it works by splitting large clusters into smaller ones; namely, Divisive HK divides the K cluster dataset into K+1 clusters using K-means method. This function helps pick up the two elements that are furthest from each other in this cluster, so as to divide this distance between the two into 3 equivalent parts to produce one more new cluster. Agglomerative HK follows a bottom-up approach because it works by putting smaller clusters together. This function intends to merge K cluster into K-1 cluster. The goal is to determine the most similar clusters in order to merge the two clusters into a new one.

To verify the efficiency of the methods, we use gene expression profiling as inputs to run hierarchical clustering, K-means, divisive HK and agglomerative HK. We apply these methods to compare the performance of each method to two recently published microarray studies form Stanford Microarray Database (SMD). Data 1 is the prostate cancer cells dataset of Deprimo et al. (2002), which contains 19 samples represented by the expression values of 382 genes [13]. Data 2 is B-cell lymphoma of Alizadeh et al. (2000), which consists of 47 samples described by the expression levels of 148 genes [14]. Then we compare the total within-cluster variation and computational time under different clustering methods.

3. RESULT

Our program was coded in JAVA on Windows XP. Figure 1 shows the system flow. We use hierarchical clustering, K-means, divisive HK and agglomerative HK to cluster data. Figure 2 shows the result of the total within-cluster variation on the same number of clusters of Data 1. Hierarchical clustering has the largest value among the four methods. That is, the quality of the hierarchical clustering is inferior to the other three methods. Figure 3 compares the clustering time cost of divisive HK and K-means on the same number of clustering of Data 2. Divisive HK-means proceeds to the result more quickly.

4. CONCLUSIONS

Due to the limitations of various clustering methods, it is important to effectively display clustered data in a manner that allows researchers to examine the variation of different clustering algorithms. Multiple statistical methods, including the within-cluster variation method used in this study, have been developed for assessing the quality of clusters produced by different algorithms [3,12]. In this study, we consider 4 clustering algorithms modified from hierarchical and K-means clustering and evaluate their performance on two well-known public available microarray data on SMD. To sum up, we find divisive HK method leads to the best result on the clustering quality and computational time.

In general, the top-down clustering tends to be faster than bottom-up clustering but the outcome from the former tends to reflect less accurately. For example, traditional hierarchical clustering algorithms working by divisive method require less computational time and are therefore faster but may produce inferior results than the agglomerative method [1,9]. In our design, we combine hierarchical and K-means method in order to strengthen hierarchical clustering's quality but preserve its merit on time cost.

5. REFERENCES

[1] Sultan, M., Wigle, D.A., Cumbaa, C. A., Maziarz, M., Glasgow, J., Tsao, M. S. and Jurisica, I., "Binary tree-structured vector quantization approach to clustering and visualizing microarray data," *Bioinformatics*, Vol.18, Supp1 1, pp. 111-9, (2002).

[2] Kerr, M.K., Churchill, G.A., "Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments," *Proc Natl Acad Sci* U S A, Vol.98, No.16 pp. 8961-5, (2001).

[3] Datta, S. and Datta, S., "Comparisons and validation of statistical clustering techniques for microarray gene expression data," *Bioinformatics*, Vol.19, No.4, pp. 459-66, (2003).

[4] Kanungo, T., Mount, D., Piatko, C., Silverman, R., Wu, A., "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.24, No.7, pp. 887-892, (2002).

[5] Krishna, K. and Murty, M. Narasimha, "Genetic k-means algorithm," *IEEE Transactions on Systems Man And Cybernetics-Part B: Cybernetics*, Vol.29, No.3, pp. 433-439, (1999)

[6] Camastra, F. and Verri, A., "A novel kernel method for clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.27, No.5, pp. 801-804, (2005).

[7] Cheung, Y.M., "K-means: a new generalized k-means clustering algorithm," *Pattern Recognition Letters*, Vol.24, No.15, pp. 2883-2893, (2003).

[8] Tsai, C.A., Lee, T.C., Ho, I.C., Yang, U.C., Chen, C.H., Chen, J.J., "Multi-class clustering and prediction in the analysis of microarray data," *Mathematical Biosciences,* Vol.193, Issue 1, pp. 79-100, (2005).

[9] Qin, J., Lewis, D.P., Noble, W.S., "Kernel hierarchical gene clustering from microarray expression data," *Bioinformatics*, Vol. 19, No.16, pp. 2097-2104, (2003).

[10] Garai, G. and Chaudhuri, B. B., "A novel genetic algorithm for automatic clustering," *Pattern Recognition Letters,* Vol.25, Issue2, pp. 173-187, (2004).

[11] Ushizawa, K., Herath, C.B., Kaneyama, K., Shiojima, S., Hirasawa, A., Takahashi, T., Imai, K., Ochiai, K., Tokunaga, T., Tsunoda, Y., Tsujimoto, G., Hashizume, K., "cDNA microarray analysis of bovine embryo gene expression profiles during the pre-implantation period," *Reproductive Biology and Endocrinology*, Vol.2, No.77, (2004).

[12] Bolshakova, N., Azuaje, F., and Cunningham, P., "An integrated tool for microarray data clustering and cluster validity assessment," *Bioinformatics*, Vol.21, No.4 pp. 451-455, (2005).

[13] DePrimo, S. E., Diehn, M., Nelson, J. B., Reiter, R. E., Matese, J., Fero, M., Tibshirani, R., Brown, P. O., Brooks, J. D., "Transcriptional programs activated by exposure of human prostate cancer cells to androgen," *Genome Biology*, Vol.3, No.7, Research0032, pp. 1-12, (2002).

[14] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, Vol.403, pp503-511, (2000)

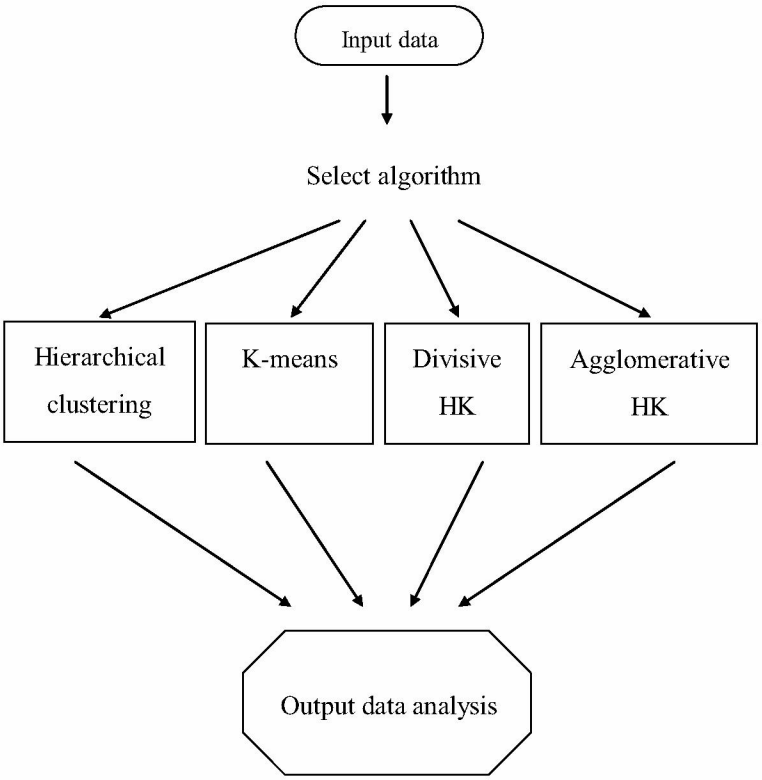

Figure 1: System flow.

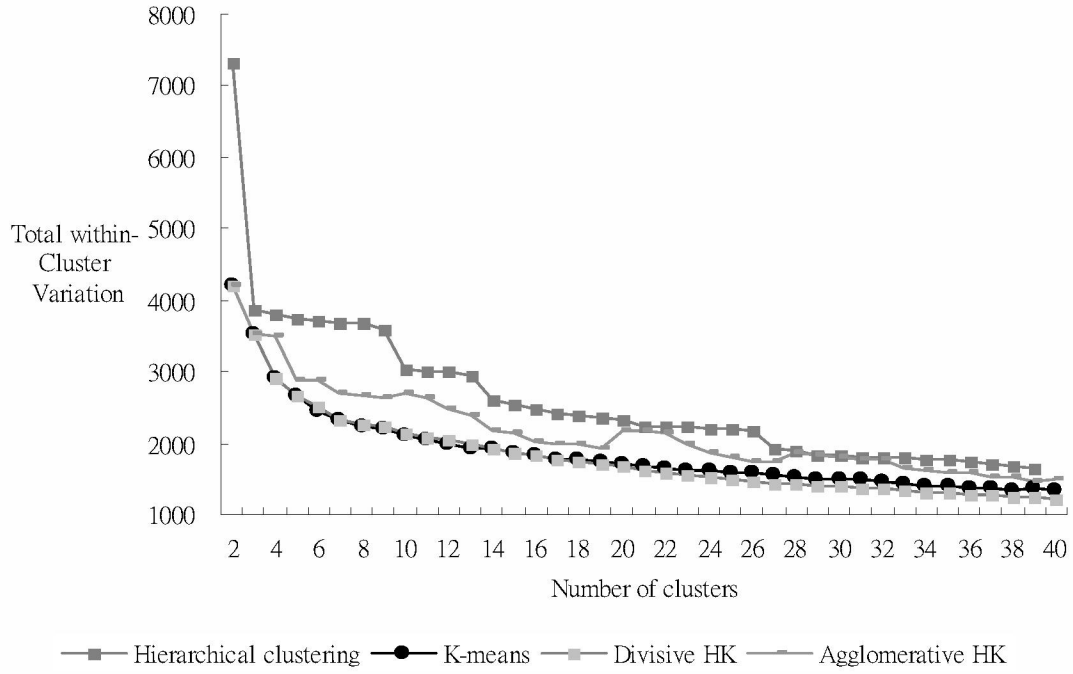

Figure 2: Total within-cluster variation of the four clustering methods on the same number of clusters.

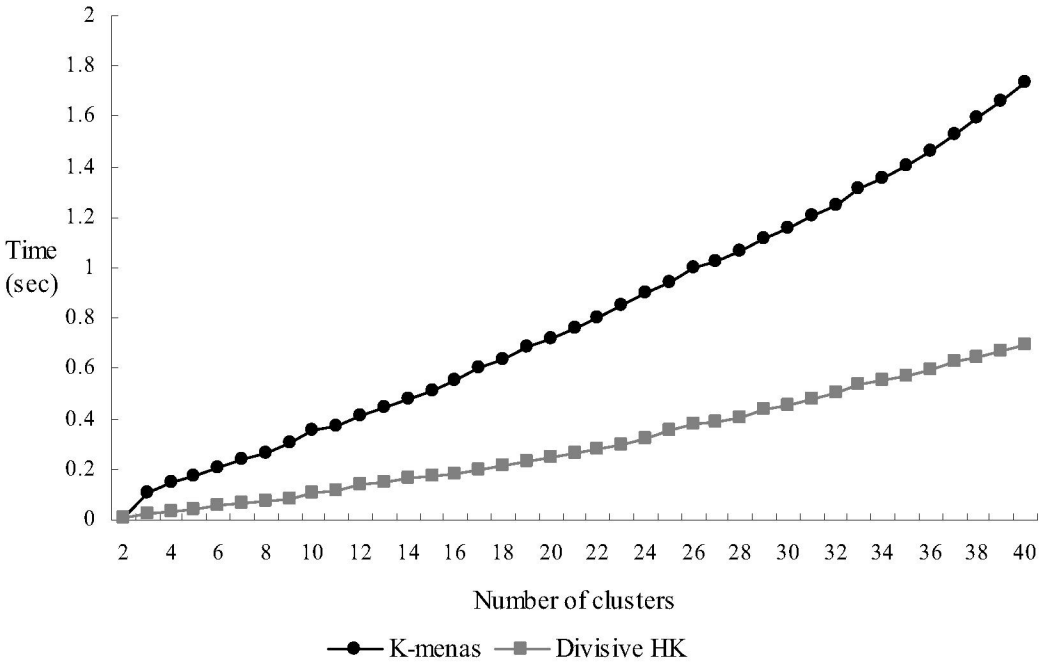

Figure 3: Time cost of divisive K-means and hierarchical K-means on the same number of clusters.