Journal of Computer Science 7 (7): 954-957, 2011
ISSN 1549-3636

An Efficient Unified K-Means Clustering Technique for Microarray Gene Expression Data

[1]P. Valarmathie, [2]K. Dinakaran and [3]T. Ravichandran
[1]Department of Computer Science and Engineering,
Saveetha Engineering College, Chennai, India
[2]Department of Computer Science and Engineering,
RMD Engineering College, Chennai, India
[3]Department of Computer Science and Engineering,
Hindustan Institute of Technology, Coimbatore, India

Abstract: Problem statement: Using microarray techniques one could monitor the expressions levels of thousands of genes simultaneously. One challenge was how to derive meaningful insights into expressed data. This might be carried out by clustering techniques such as hierarchical and k-means, but most of the clustering techniques were largely heuristic in nature and are associated with some unresolved issues like how to fix the precise number of clusters and how to visualize the results in a pictorial form. **Approach:** Determine accurate number of clusters from gene expression data and validate the results using correctness ratio and sum of squares criteria. A new approach suggested to addresses the primary issue of k-means clustering algorithm that predefining number of clusters. This approach provides accurate number of clusters by minimizing the squared error function and maximizing the correctness ratio value. **Results:** The experimental results have shown the efficiency of our method by calculating and comparing the sum of squares with different k values. It was concluded that the number of clusters were accurate with minimum sum of squares value and maximum value of correctness ratio. **Conclusion:** The results showed that the quality of clusters and performance of this new approach is improved.

Key words: Microarray, expectation maximization, clustering technique, squared error function

INTRODUCTION

The advent of microarray technology made it possible to monitor the expression levels of thousands of genes concurrently whereas in traditional approaches one can focus local examination and collection of data on single gene (Wilkin and Huang, 2007; Chen *et al*., 2005). Microarray may be used to measure gene expression in many ways, but one of the most popular applications is to compare expression of a set of genes from a cell maintained in a particular 'condition A' to the same set of genes from a reference cell maintained under normal 'condition B'. The process data, after the normalization procedure, can be represented in the form of matrix. Each row in the matrix corresponds to a particular gene and each column could either correspond to an experimental condition or to a specific time point at which expression of genes has been measured. Huge volume of data generated by microarray techniques are collected and stored in massive databases. Traditional techniques and tools are not adequate to deal with this data and obtain the desired results (jiang *et al*., 2004; Eisen *et al*., 1998; Ali *et al*., 2009). The challenge is to effectively analyze and interpret such a huge volume of information. Two statistical operations commonly applied to microarray data are classification and clustering (Suresh *et al*., 2009; Kumar, 2009). Classification technique is a supervised one in which objects is classified by known class label, whereas clustering is an unsupervised technique requiring no predefined class labels. As we have little knowledge of the complete data set, we have favored unsupervised methods (Eisen *et al*., 1998). The patterns within the groups are similar to one another and dissimilar to the patterns in different groups. Many tools that cluster microarray data employ methods such as hierarchical clustering, k-means clustering and self organizing maps to analyze and interpret the data. As each technique has its own disadvantages, a new approach is required to overcome them.

Corresponding Author: P. Valarmathie, Department of Computer Science and Engineering, Saveetha Engineering College, Chennai, India

K means clustering adopts a non-hierarchical approach to cluster N objects into K partitions where 0<K<N. It randomly selects k of the objects, each of which initially represents a cluster means then calculates mean value for each of the remaining object to which it is the most similar, based on the distance between the object and the cluster mean (Chen *et al*., 2005; Jaradat *et al*., 2009). Very common measures include the sum of distances or sum of squared Euclidean distances from the mean of each cluster. It then re-computes the mean value for each cluster, this process being repeated until no more reassignment occur (Han and Kamber, 2001). The objective of k-means is to minimize total intra-cluster variance, or the squared error function. The mathematical formula for squared error function is

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

This algorithm is sensitive to initial value of k; hence it may produce different results for different k values and it may find only local optimum rather than global one (Al-Zoubi *et al*., 2010). Also, it is sensitive to noise and outlier objects since a small number of such objects can substantially influence the mean value.

In the past, hierarchical and k-means methods have been the primary clustering tools employed to perform the task of clustering microarray data. The major limitation of these methods is their inability to determine the number of clusters (Mar and McLachlan, 2003). Model based clustering has become an essential one in microarray gene expression data in order to determine the number of clusters and provides a statistical framework to model the cluster structure of gene expression data. In this approach the data is generated by a finite mixture of underlying probability distributions in which each component represents a different cluster (Yeung *et al*., 2001). For a fixed number of components G, the model parameters can be estimated using the EM algorithm. It is a general approach to maximum likelihood in the presence of incomplete data. Let the dataset be $y_i=(x_i,z_i)$, where $z_i=(z_{i1},\ldots\ldots z_{iG)}$. The EM algorithm iterates between E-step in which the values of Z_{ik} are computed from the data with the current parameter estimates. In M-step, model parameters are estimated so as to maximize the likelihood of complete data for the given estimated Z_{ik} parameters. Each data object is assigned to the component with the maximum conditional probability when the algorithm converges (Suresh *et al*., 2009; Fraley and Raftery, 1998). In order to ascertain the number of clusters represented by the model based method, we calculated the correctness ratio (Arima and Hanai, 2003). In this study, we suggest a new approach to solve the problems not addressed in the conventional methods.

MATERIALS AND METHODS

Most of the clustering algorithms that have been employed in the literature are heuristic and have the disadvantage of requiring beforehand the precise number of clusters. The present work focuses on K-means clustering algorithm where the number of clusters k has to be defined by the user arbitrarily in advance. This may not help the researchers to achieve the desired aim; hence drawing inference of biological significance becomes difficult for them. The present method helps them avoid this arbitrariness by automatically suggesting the correct number of clusters that is obtained by applying the results of the model based algorithm to k-means clustering. The sample dataset is downloaded from the machine learning database in order to examine the performance of the proposed method.

Algorithm: AUKCA clustering algorithm
Input: Data objects $X=\{x_1\ldots.x_n\}$ and model structure $M = \{m_1\ldots.m_k\}$.

Output: K clusters with maximum sum of square.

Step: 1 Estimate the no. of components K using EM algorithm.
Step: 2 Select K object as the initial cluster centers from step 1.

Repeat:

Step: 3 Assign each object to the cluster to which the object is the most
similar based on the cenroid value.
Step: 4 Update the cluster centroids using any similarity metric
Until Centroid values remain unchanged or else goto step 2.

RESULTS AND DISCUSSION

The result of model based clustering is shown Fig. 1. This figure shows the best model EEV, the highest point in the plot provides four components (clusters). In order to ascertain the number of clusters represented by the model, we have calculated and compared the correctness ratios for different k values. It is confirmed that the number of components provided by the EEV model is correct with respect to ratio value 0.135, corresponding to the value of k = 4 as given in Table 1.

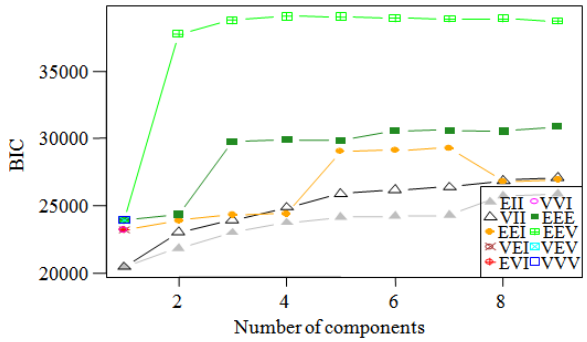

Fig. 1: The best model EEV is the highest point in the plot. The no. of components (clusters) is four

Table 1: Correctness ratio corresponding to k values

K=3	K = 4	K = 5	K = 6
0.10	0.135	0.083	0.073

Table 2: Sum of squares corresponding to K values

K = 4	K = 5	K = 6
61.00	64.14	72.18

To minimize the squared error function in k-means clustering algorithm, we calculated the sum of squares for different clusters in Table 2. From the obtained values, one can conclude that the number of clusters four is optimum with minimum sum of squares value.

CONCLUSION

In this study, we have described a novel clustering approach for performing clustering of microarray gene expression data. We examined the results of model based clustering to obtain the precise k clusters and applied the same to k-means clustering. The results of clustering yeast data show the efficiency of the new method. The future work is to enhance the performance of this new algorithm that can be achieved by reducing the dimensionality of the dataset so that the outliers are removed and thereby increasing its efficiency and the accuracy.

REFERENCES

Ali, S.A., N. Sulaiman, A. Mustapha and N. Mustapha, 2009. K-means clustering to improve the accuracy of decision tree response classification. Inform. Technol. J., 8: 1256-1262. http://docsdrive.com/pdfs/ansinet/itj/2009/1256-1262.pdf

Al-Zoubi, M.B., A. Hudaib, A. Huneiti and B. Hammo, 2008. New efficient strategy to accelerate k-means clustering algorithm. Am. J. Applied Sci., 5: 1247-1250. DOI: 10.3844/ajassp.2008.1247.1250

Arima, C. and T. Hanai, 2003. Gene expression analysis using fuzzy k-means clustering. Genome Inform., 14: 334-335. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.63.5074&rep=rep1&type=pdf

Chen, T.-S., T.-H. Tsai, Y.-T. Chen, C.-C. Lin and R.-C. Chen *et al*., 2005. A combined k-means and hierarchical clustering method for improving the clustering efficiency of microarray. Proceedings of 2005 International Symposium on Intelligent Signal Processing and Communication Systems, Dec. 13-16, Hong Kong, pp: 405-408. DOI: 10.1109/ISPACS.2005.1595432

Eisen, M.B., P.T. Spellman, P.O. Brown and D. Botstein, 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA., 95: 14863-14868. PMID: 9843981

Fraley, C. and A.E. Raftery, 1998. How many clusters? Which clustering method? Answers via model based cluster analysis. Comput. J., 41: 578-588. DOI: 10.1093/comjnl/41.8.578

Han, J. and M. Kamber, 2006. Data Mining Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publishers, An Imprint of Elsevier, First Indian Reprint, ISBN: 1558609016, pp: 770.

Jaradat, A., R. Salleh and A. Abid, 2009. Imitating K-means to enhance data selection. J. Applied Sci., 9: 3569-3574. http://en.scientificcommons.org/51553806

Jiang, D., C. Tang and A. Zhang, 2004. Cluster analysis for gene expression data: A survey. IEEE Trans. Knowl. Data Eng., 16: 1370-1386. DOI: 10.1109/TKDE.2004.68

Kumar, R.M., 2009. The widely used diagnostics "DNA Microarray"-A review. Am. J. Infect. Dis., 5: 207-218. DOI: 10.3844/ajidsp.2009.207.218

Mar, J.C. and G.J. McLachlan, 2003. Model based clustering in gene expression microarrays: An application to breast cancer data. Proceedings of the Asia Pacific Bioinformatics Conference, Feb. 4-7. Australian Computer Society, Inc. Darlinghurst, Australia, Australia. ISBN: 0-909-92597-6

Suresh, R.M., K. Dinakaran and P. Valarmathie, 2009. Model based modified k-means clustering for microarray data. Proceedings of the International Conference on Information Management and Engineering, Apr. 03-05, Kuala Lumpur, Malaysia, pp: 271-273. http://doi.ieeecomputersociety.org/10.1109/ICIME.2009.53

Wilkin, G.A. and X. Huang, 2007. K-means clustering algorithms: Implementation and comparison. Proceedings of the 2nd International Multi Symposium on Computer and Computational Sciences, Aug. 13-15, The University of Iowa, Iowa City, Iowa, USA., pp: 133-136. http://doi.ieeecomputersociety.org/10.1109/IMSCCS.2007.51

Yeung, K.Y., C. Fraley, A. Murua, A.E. Raftery and W.L. Ruzzo, 2001. model-based clustering and data transformations for gene expression data. Bioinformatics, 17: 977-987. DOI: 10.1093/bioinformatics/17.10.977