International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:1, No:4, 2007

Iterative Clustering Algorithm for Analyzing Temporal Patterns of Gene Expression

Seo Young Kim, Jae Won Lee, Jong Sung Bae

***Abstract*—**Microarray experiments are information rich; however, extensive data mining is required to identify the patterns that characterize the underlying mechanisms of action. For biologists, a key aim when analyzing microarray data is to group genes based on the temporal patterns of their expression levels. In this paper, we used an iterative clustering method to find temporal patterns of gene expression. We evaluated the performance of this method by applying it to real sporulation data and simulated data. The patterns obtained using the iterative clustering were found to be superior to those obtained using existing clustering algorithms.

***Keywords*—**Clustering, microarray experiment, temporal pattern of gene expression data.

I. Introduction

THE rapid development of microarray technologies has made it possible to monitor the expression levels of thousands of genes simultaneously [1]. These technologies have proved a boon in the biological and medical sciences, where they have assisted researchers in tackling such broad problems as tumor classification. Microarray experiments provide a wealth of information; however, extensive data mining is required to identify the patterns that characterize the underlying mechanisms of action. For biologists, a key aim when analyzing microarray data is to group genes based on the temporal patterns of their expression levels, which may provide insights into genetic capacities and their interactions. Indeed, microarray experiments in cellular contexts have shown that genes with similar functions often evince similar temporal patterns of co-regulation [2], [3]. Due to the large number of genes involved in these experiments and the complexity of biological processes in general, an effective clustering algorithm for grouping genes is crucial to such studies.

Clustering analysis is faced with two problems: how to determine the number of true clusters and how to evaluate the samples assigned to those clusters. Since the results of clustering analysis rely heavily on limited biological and medical information (i.e. tumor classification), they are not only sensitive to noise but they can also be prone to over-fitting, which is a major concern in the clustering analysis of gene expression data. Previous studies on the analysis of gene expression data have extensively explored the use of unsupervised clustering analysis to find temporal patterns [4]-[10]. Recently, resampling and cross-validation methods have been shown to be effective for evaluating the stability of clusters [11]-[14]. In particular, Monti et al. [15] proposed a consensus clustering method for class discovery based on a resampling method. Kim et al. [16] devised an extension of consensus clustering [15] that exploits a mixed clustering algorithm based on a mixed distance measure.

The iterative clustering procedures of Monti et al. [15] and Kim et al. [16] have been applied to the problem of discovering taxonomy, or distinct and non-overlapping sub-populations within a larger population in gene expression data analysis. We applied iterative clustering procedures to the problem of identifying temporal patterns of gene expression in time course microarray data spanning a small set of times.

Here, we introduce a new clustering method based on an iterative algorithm that measures the relative stabilities of clusters from cross-validation criteria. The performance of the proposed approach is compared with those of the more commonly used agglomerative and divisive hierarchical clustering methods. One important property of temporal gene expression data is that the data for a given gene at different times may be correlated. Furthermore, gene expression levels may vary markedly over time [8]. To reflect such time dependencies in observed data, we compare the stability and consistency of the results produced by deleting one set of temporal observations at a time. In addition, we compare the average expression patterns in each group with the model profiles obtained using our iterative algorithm and existing clustering algorithms.

II. Iterative Clustering Algorithm

A. Review of previous studies

The consensus clustering method is a type of resampling-based method [15]. In consensus clustering, the original data set is perturbed by subsampling iteratively, and then existing clustering methods are iteratively applied to the perturbed data set to construct a distance measure. The data are

Manuscript received January 19, 2005. This work was supported by grant R08-2003-000-10572-0 from the Basic Research Program of the Korea Science & Engineering Foundation. JW Lee was also supported by grant R14-2003-002-0102-0 from the Korea Science and Engineering Foundation.

S.Y. Kim is with the Research Institute for Basic Science, Chonnam National University, Gwangju, 500-757 Korea (corresponding author to provide phone: +82-62-530-0442; fax: +82-62-530-3449; e-mail: gong@chonnam.ac.kr).

J.W. Lee is Department of Statistics, Korea University, Seoul 136-701 Korea (e-mail: jael@korea.ac.kr)

J. S. Bae is with Department of Statistics, Chonnam National University, Gwangju, 500-757 Korea (e-mail: jsbae@chonnam.ac.kr)..

represented as a matrix, $X=[x_{gt_i}]$, where x_{gt_i} denotes the expression of the g th gene at time T_i, $1 \le i \le t$. First, a data resampling scheme and an initial clustering algorithm must be chosen. Then a similarity matrix is used to assign genes to the proper clusters obtained by applying the algorithm to the various perturbed data sets. The similarity matrix is defined as follows:

$$S(i,j)=\frac{\sum_h M^h(i,j)}{\sum_h I^h(i,j)},\ 1 \le i,j \le N,\ 1 \le h \le H \qquad (1)$$

where N is the number of genes and M^h is the matrix corresponding to the results obtained by applying the initially selected clustering algorithm to the hth perturbed data set. $M^h(i,j)$ equals 1 if observations i and j belong to the same cluster, and 0 otherwise. In (1), I represents the indicator matrix such that $I^h(i,j)$ equals 1 if observations i and j are in the h th perturbed data set, and 0 otherwise. The similarity matrix is symmetric, and hence $0 \le S \le 1$. When all the entries of S are close to 1 or 0, we can infer that the results have been well clustered. Here, the matrix $I-S$ represents a distance matrix.

When the consensus clustering method is applied to gene expression data to identify temporal patterns, the results obtained depend on the choice of the initial clustering method. To improve the consensus clustering, we propose mixing two similarity matrices, S_1 and S_2, for clustering gene expression data based on temporal patterns. Specifically, we take the mixed similarity matrix to be the average of two similarity matrices, i.e., $S_m = average(S_1, S_2)$.

B. Iterative clustering algorithm

We apply S and S_m as similarity matrices in clustering stages to find temporal patterns. The iterative clustering algorithm is as follows.

```
Do k=1 to K.
  Initialize matrices M and I.
    Do h=1 to H
     Resample D^h, Construct I^h
       Execute initial clustering algorithm
           Construct M^h
         Let M be the union of M and M^h
         Let I be the union of I and I^h
  End h
     Compute similarity matrix S from M and I
    Compute mixed similarity matrix S_m (if the matrix is used)
End k
Determine optimal k
Classify X into final optimal clusters based on S^k or S_m^k.
```

III. SYSTEMS AND METHODS

A. Initial clustering algorithm

Clustering is an exploratory tool for examining associations among gene expression data. Hierarchical clustering dendrograms allow us to visualize such data. The information provided by these methods can be used as the basis for hypotheses regarding the relationships between genes and classes.

Agglomerative hierarchical clustering: This algorithm initially takes each gene as a single cluster and then constructs progressively bigger clusters by grouping similar genes together until the entire data is contained in one final cluster. This algorithm is computationally simpler, and more available than non hierarchical method. It is more representative of the original data structure at the bottom levels than top levels of the dendrogram. This algorithm should be considered for use in problems requiring the identification of small clusters or many clusters.

Divisive hierarchical clustering: Divisive clustering first places all genes into a single cluster and then progressively splits this initial cluster into smaller and smaller subsets until each subset contains only a single gene. This algorithm retains the overall data structure, i.e., the upper levels of the dendrogram are very representative of the original data structure. Divisive clustering should be considered when searching for large clusters or a small number of clusters. The divisive algorithm is thus well suited to gene expression data clustering.

B. Assessing clusters

Two measures were used to assess and compare the performance of various clustering methods. The idea behind the validation method used is that an algorithm should be rewarded for consistency. We considered that expression data are observed over all the genes at t time points, say $T_1, T_2, \cdots, T_t$. For all time points, iterate the clustering algorithms for each of the t data sets obtained by deleting the observations at time T_i from the original data set.

The average proportion of non-overlap measure computes the average proportion of genes that are not placed in the same cluster by the clustering method under consideration on the basis of the entire data set and the data sets obtained by deleting the expression levels at one time point at a time [10].

$$VM_1(K)=\frac{1}{Nt}\sum_{g=1}^{N}\sum_{i=1}^{t}\left(1-\frac{n(C^{g,i}\cap C^{g,0})}{n(C^{g,0})}\right) \qquad (2)$$

where $C^{g,i}$ denotes the cluster containing gene g in the clustering based on the data set from which the observations at time T_i have been deleted, and $C^{g,0}$ denotes the original cluster containing gene g in the clustering based on the entire data set. A good algorithm is expected to yield a small value of $VM_1(K)$.

The average of the adjusted rand index computes the average degree of agreement between two partitions. Given a set of N observations $D=\{o_1, o_2, \cdots o_N\}$, suppose $U=\{u_1, u_2, \cdots, u_R\}$ and $V=\{v_1, v_2, \cdots, v_C\}$ represent two different partitions of the observations in D. Here, for $1 \le i \ne i' \le R$ and $1 \le j \ne j' \le C$, $\cup_{i=1}^{R} u_i = \cup_{j=1}^{C} v_j = D$ and N_{ij} is the number of observations that are in both classes u_i and v_j, and $N_{i.}$ and $N_{.j}$ are the

numbers of observations in classes u_i and v_j respectively. The adjusted rand index is as follows [17], [18]:

$$Rand = \frac{\sum_{ij} N_{ij} C_2 - \left[\sum_i N_{i.} C_2 \sum_j N_{.j} C_2\right] / {}_N C_2}{(1/2)\left[\sum_i N_{i.} C_2 + \sum_j N_{.j} C_2\right] - \left[\sum_i N_{i.} C_2 \sum_j N_{.j} C_2\right] / {}_N C_2} \quad (3)$$

We calculate the adjusted rand index in (3) for the clustering results of the entire data set and the clustering results based on the data set after deleting the observations at time T_i, and then we calculate t adjusted rand indices by deleting the observations at time T_i for all i. The average adjusted rand index for all time points is as follows: For a good clustering algorithm, we would expect these values to be high.

$$VM_2(K) = \sum_{i=1}^{t} Rand_i / t \; . \quad (4)$$

IV. Implementation and Results

We used the agglomerative clustering method UPGMA (Unweighted Pair Group Method with Arithmetic Mean) [10] and the divisive clustering method Diana [10] as initial clustering algorithms, and additionally applied iterative clustering with UPGMA (ITU), iterative clustering with Diana (ITD), and iterative mixed clustering with UPGMA and Diana (ITM). The clustering performance was tested on a real data set and a simulated data set. For each cluster we ran 20 resampling iterations. At each iteration, the perturbed dataset was obtained by sampling, without replacement, 80% of the observations in the original data set.

A. Results for gene expression data

We used the publicly available [19] gene expression data set on yeast sporulation obtained experimentally in [5]. The data set consists of expression levels of 6118 genes in the yeast genome measured at seven time points during the sporulation process (i.e., 0, 0.5, 2, 5, 7, 9 and 11.5 hours).

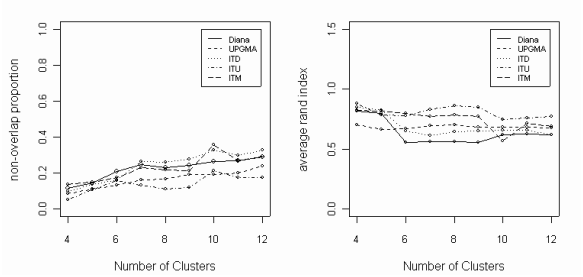

Fig. 1. The average proportion of non-overlap measures and the average of the adjusted rand index for the sporulation data.

For the five clustering algorithms under consideration, we computed the two cluster-assessment measures, $VM_1(K)$ and $VM_2(K)$, over a range of cluster numbers around 7, specifically, K = 4–12. The average proportion of non-overlap measure gave similar results for the five algorithms, although UPGMA and ITU appeared to be the best as judged by this measure (left of Fig. 1). The results for the average of the adjusted rand index (right of Fig. 1), on the other hand, indicated that ITU and ITM gave the best performance. A somewhat surprising finding is that the performance of Diana appears to be the worst as judged by the two measures. To classify yeast genes based on their expression levels, Chu et al. [5] hand picked seven small subsets of representative genes using their knowledge of the yeast genome. We used the same subsets to construct our model temporal profiles by averaging the log-expression ratio of all genes in each subset. Taking the model profiles obtained by [5] as a benchmark, inspection of the plots for the various algorithms suggests that the ITM plots are closest to the model profiles (Fig. 2) to the model profiles (Fig. 2).

B. Results for simulated data

We generated a simulated data set with the same distinct temporal patterns over 10 time points according to the method described in [20] and [10]. Independent random variables were added to these mean expression-ratio values so as to generate 50 genes around each of the nine patterns. Half of the total genes were generated from a normal distribution with mean 0 and standard deviation 1, and the remaining half were generated from an exponential distribution with location -0.2 and scale 0.2. The model profile is displayed in the plot in the bottom right hand corner of Fig. 4. Fig 3 shows the results obtained using the two clustering-assessment measures. According to both measures, ITU and ITM give the best performance. Comparison of the profiles generated using the five algorithms and the model profiles indicates that the profiles generated by ITM and ITD methods are perhaps the closest to the model profiles (Fig. 4).

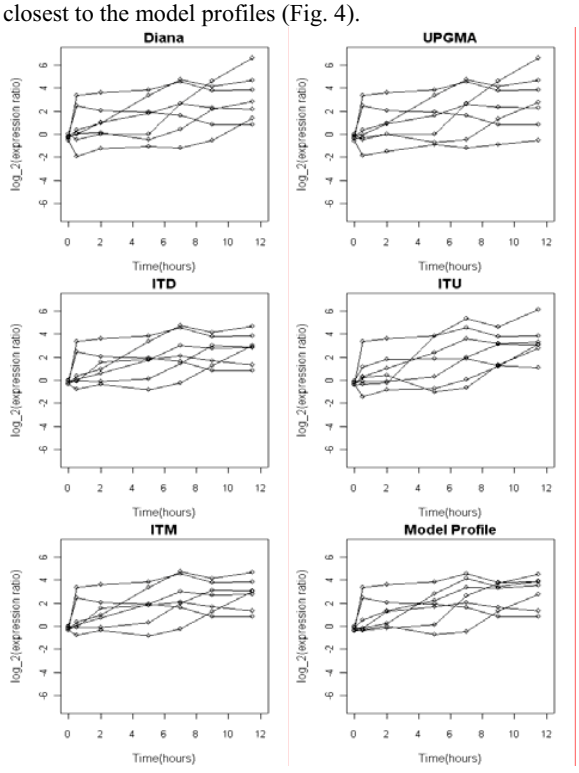

Fig. 2. Average temporal profiles of seven groups obtained using five clustering algorithms and the model profiles for the sporulation data.

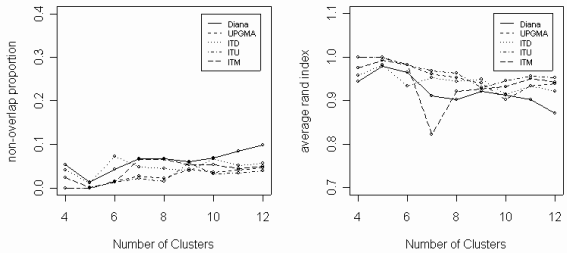

Fig. 3. The average proportion of non-overlap measures and the average of the adjusted rand index for the sporulation data.

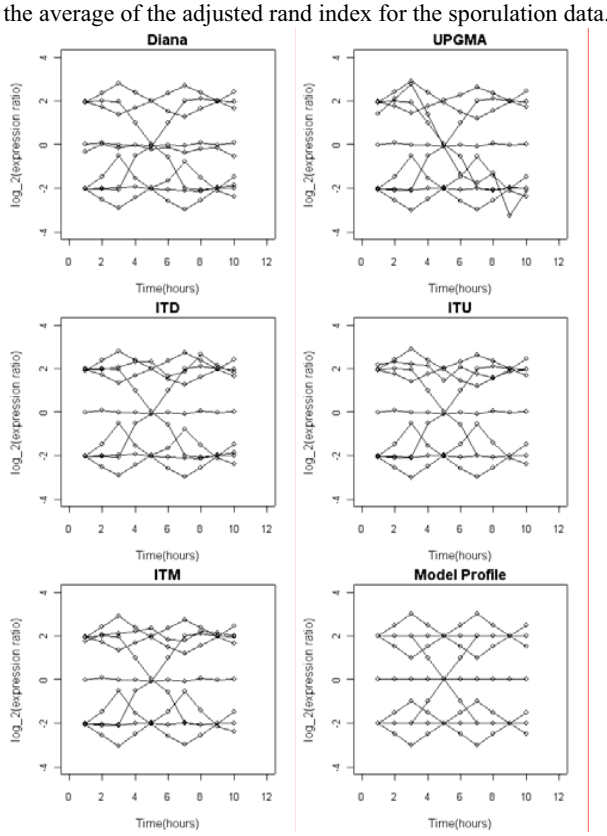

Fig. 4. Average temporal profiles of nine groups obtained using five clustering algorithms and the model profiles for the simulated data.

V. CONCLUSION AND DISCUSSION

We have compared the performance of 3 iterative clustering algorithms and 2 existing algorithms using temporal gene expression data and simulated data. The iterative algorithms were found to be more accurate and consistent than existing methods. Furthermore, the mixed iterative algorithm gave superior results to the other iterative algorithms tested. The present findings suggest that the mixed iterative algorithm overcomes the demerits of the agglomerative and divisive hierarchical clustering algorithms.

ACKNOWLEDGMENT

This work was supported by grant R08-2003-000-10572-0 from the Basic Research Program of the Korea Science & Engineering Foundation. JW Lee was also supported by grant R14-2003-002-0102-0 from the Korea Science and Engineering Foundation.

REFERENCES

[1] P.O. Brown, and D. Botstein, "Exploring the new world of the genome with DNA microarrays", *The chipping forecast, vol.* 21, 1999. pp. 33-37.
[2] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns", *Proceeding of the National Academy of Sciences*, vol. 95, 1998, pp. 14863-14868.
[3] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yest Saccharomyces cerevisiae by microarray hydridization", *Mol. Biol. Cell*, vol. 9, pp. 3273-3279.
[4] J.L. DeRisi, V.R. Iyer, and P.O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale", *Science*, vol. 278, 1997, pp. 680-686.
[5] S. Chu, and J.L. DeRisi et al., "The transcriptional program of sporulation in budding yeast", *Science*, vol. 282, 1998, pp. 699-705
[6] R.J. Cho, et al., "A genome-wide transcriptional analysis of the mitotic cell cycle", *Mol. Cell*., vol. 2, 1998, pp. 65-73.
[7] M.J.L. De Hoon, S. Imoto, and S. Miyano, "Statistical analysis of a small set of time-ordered gene expresion data using linear splines", *Bioinformatics*, vol. 18, 2002, pp. 1477-1485.
[8] Y. Luan, and H. Li, "Clustering of time-course gene expression data using a mixed-effects model with B-splines", *Bioinformatics*, vol. 19, 2003, pp. 474-482.
[9] S.D. Peddada, E.K. Lobenhofer, L. Li, et al., "Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference", *Bioinformatics*, vol. 19, 2003. pp. 834-841.
[10] S. Datta, and S. Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data", *Bioinformatics,* vol. 19, 2003, pp. 459-466.
[11] A. Bhattacharjee, W.G. Richards, and J. Staunton, J. et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinomas sub-class", *Proceeding of the National Academy of Sciences*, vol. 98, 2001, pp. 13790-13795
[12] S. Dudoit, J. Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset", *Genome Biology*, vol. 3, 2002, research0036.1-0036.21.
[13] A.K. Jain, and J. Moreau, "Bootstrap techniques in cluster analysis", *Pattern Recognition*, vol. 20, 1988, pp. 547-568.
[14] E. Levine, and E. Domany, "Resampling method for unsupervised estimation of cluster validity", *Neural Computation*, vol. 13, 2001, pp. 2573-2593.
[15] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus Clustering: A resampling based method for class discovery and visualization of gene expression microarray data", *Kluwer Academic Publishers, 2003.*
[16] S.Y. Kim, J.W. Lee, and T.M. Choi, "Ensemble clustering method based on the resampling similarity measure for gene expression data", 2004, Submitted.
[17] L. Huber, and P. Arabie, "Comparing partitions", *Journal of Classification*, vol. 2, 1985, pp. 193-218.
[18] K.Y. Yeung, and W.L. Ruzzo, "An empirical study on principal component analysis for clustering gene expression data", Technical Report 2000 UW-CSE-00-11-01, Department of Computer Science and Engineering, University of Washington.
[19] http://smgm.stanford.edu/pbrown/sporulation.
[20] J. Quackenbush, "Computional analysis of microarray expression data.", *Bioinformatics*, vol. 18, 2001, pp. 1-10.