A Changing Window Approach to Exploring Gene Expression Patterns

Qiang Wang, Yunming Ye
Shenzhen Graduate School
Harbin Institute of Technology
Xili, Shenzhen 518055, China
mikewq@yahoo.cn
yeyunming@hit.edu.cn

Joshua Zhexue Huang
E-Business Technology Institute
The University of Hong Kong
Pokfulam Road, Hong Kong
jhuang@eti.hku.hk

Abstract

This paper presents a changing window approach to exploring gene expression patterns in "snapshot windows". A snapshot window is a sub-matrix of co-expressed microarray data representing certain expression pattern. In this approach, we use a feature weighting k-means subspace clustering algorithm to generate a set of clusters and each cluster defines a set of "snapshot windows" which are characterized by different sets of ordered sample weights that were assigned by the clustering algorithm. We define an accumulated weighting threshold (AWT) as the sum of weights of samples in the "snapshot window". Given a cluster, different "snapshot windows" can be obtained by changing AWT to explore all possible local expression patterns in the cluster. Experiment results have shown our approach is effective and flexible in exploring various expression patterns and identifying novel ones.

1 Introduction

Microarray is a revolutionary new technology which provides an opportunity to obtain the "global" view of the cell [1]. However, identifying patterns from subsets of genes co-expressed under subsets of samples poses great challenges to microarray data analysis [2]. Subspace clustering is an effective technique that can identify clusters of objects in different subsets of features in the dataset [3], which is appropriate in the context of microarray data analysis. In this paper, we investigate use of the Entropy Weighting *K*-Means (EWKM) subspace clustering algorithm [4] in microarray data analysis and propose a changing window approach to exploring gene expression patterns in different "snapshot windows".

Our approach is characterized in the following features. Firstly, EWKM is used to generate a set of clusters, each defining a set of "snapshot windows" by the same set of genes. Secondly, we define an accumulated weighting threshold (AWT) as the sum of weights of samples in the "snapshot window". For a given cluster, different "snapshot windows" can be obtained by changing AWT so that different local expression patterns can be explored. Thirdly, the sample weighting mechanism in the EWKM subspace clustering algorithm works on the hypothesis that different samples make different contributions to different genes clusters. Such contribution of a sample is represented as a weight that can be treated as the degree of the sample contribution to the cluster. Finally, the weight distribution is controlled by parameter γ of the EWKM algorithm. A large γ results in clusters with more evenly distributed sample weights. Therefore, changing γ can generate different clusters with different characteristics of "snapshot windows".

The rest of this paper is organized as follows. Section 2 presents related work of subspace clustering in microarray data analysis. Section 3 describes the EWKM subspace clustering algorithm and the changing window approach to exploring clusters in different snapshot windows. Section 4 presents experiment results on real microarray data. Section 5 summaries this work.

2 Related work

Clustering techniques are widely used in microarray data exploration and analysis. Traditional clustering algorithms, such as hierarchical clustering [5], *k*-means [6], self-organizing map (SOM) [7] generate clusters from microarray data across all samples. However, many gene co-expression patterns occur in subsets of samples. As such, these global clustering algorithms are inadequate for revealing all gene expression patterns.

Subspace clustering techniques cluster objects based on subsets of features in data. In microarray data analysis, subspace clustering is also referred to as bi-clustering, co-clustering, or two-mode clustering, which allows simultaneously clustering of rows and columns of a data matrix.

978-0-7695-3452-7/08 $25.00 © 2008 IEEE
DOI 10.1109/BIBM.2008.39

Each cluster represents a set of genes identified by a subset of samples and different clusters are represented in different subsets of samples. Recently, a few subspace clustering algorithms have been successfully applied to microarray data, including coupled two-way clustering (CTWC) [8], plaid models [9], δ-cluster [10], and biclustering [11]. These algorithms require predefinition of bicluster models which describe specific characteristics of clusters to be discovered, and search for the models from the data. Only a few models can be defined and the performance of a biclustering algorithm is highly constrained by the completeness and appropriateness of the definition of the bicluster model. Search for cluster models with these algorithms is also NP-hard.

In contrast, the changing window approach we propose has two major advantages. It is flexible to generate clusters and explore gene expression patterns in "snapshot windows". It is efficient for large microarray data because it is essentially a k-means clustering algorithm.

3 EWKM algorithm and snapshot windows

In this changing window approach, the EWKM algorithm is first used to generate a set of clusters from a microarray dataset and to assign a set of weights to samples in each cluster. In each cluster, a set of sample weights indicate the importance of the samples in forming the cluster and can be used to specify the "snapshot windows" in this cluster. To compute the sample weights automatically from data, the objective function of the EWKM algorithm is defined as follows:

$$F(W,Z,\Lambda)=\sum_{l=1}^{k}\Big[\sum_{j=1}^{n}\sum_{i=1}^{m}w_{lj}\lambda_{lj}(z_{li}-x_{ji})^2+\gamma\sum_{i=1}^{m}\lambda_{li}log\lambda_{li}\Big] \tag{1}$$

subject to

$$\begin{cases}\sum\limits_{l=1}^{k}w_{lj}=1, & 1\le j\le n, \quad 1\le l\le k, \quad w_{lj}\in\{0,1\}\\ \sum\limits_{i=1}^{m}\lambda_{li}=1, & 1\le l\le k, \quad 1\le i\le m, \quad 0\le\lambda_{li}\le 1\end{cases} \tag{2}$$

where W is a partition matrix that indicates the assignment of gene j to cluster l, Z is a set of k cluster centers and Λ is an $k\times m$ matrix in which each row l represents a set of m weights assigned to samples in cluster l.

In the clustering process, the EWKM algorithm simultaneously minimizes the sum of the within cluster dispersions and maximizes the negative weight entropy to make more samples to contribute to the identification of clusters. The positive parameter γ controls the strength of the incentive for clustering on more samples. The clustering process is carried out by iterating the following three steps:

Step 1: Given Z and Λ, compute W by

$$w_{lj}=\begin{cases}1, & \text{if } \sum\limits_{i=1}^{m}\lambda_{li}(z_{li}-x_{ji})^2\le\sum\limits_{i=1}^{m}\lambda_{li}(z_{ri}-x_{ji})^2\\ & 1\le r\le k\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

Step 2: Given W and Λ, compute Z by

$$z_{li}=\frac{\sum\limits_{j=1}^{n}w_{lj}\,x_{ji}}{\sum\limits_{j=1}^{n}w_{lj}}, \qquad \text{for } 1\le l\le k \text{ and } 1\le i\le m \tag{4}$$

Step 3: Given W and Z, compute Λ by

$$\lambda_{lt}=\frac{exp\Big(\frac{-D_{lt}}{\gamma}\Big)}{\sum\limits_{i=1}^{m}exp\Big(\frac{-D_{li}}{\gamma}\Big)} \tag{5}$$

where

$$D_{lt}=\sum_{j=1}^{n}w_{lj}\Big(z_{lt}-x_{jt}\Big)^2 \tag{6}$$

The EWKM algorithm is summarized in Table 1. More details can be found in [4] and [12].

Parameter γ controls the distribution of the sample weights as follows. A large γ will result in more evenly distributed sample weights, while a small γ generates clusters with a few samples having large weights. Therefore, by using different γ, we can generate different sets of clusters with different weight distributions for exploring interesting gene expression patterns.

After a set of k clusters are generated by EWKM, each cluster l has a set of weights ordered as $\lambda_{l1}\ge\lambda_{l2}\ge\ldots\ge\lambda_{lm}$, where m is the number of samples in the dataset. Define the accumulated weighting threshold of cluster l, AWT(l) as

$$\sum_{t=1}^{p}\lambda_{lt}\le AWT(l)\le\sum_{t=1}^{p+1}\lambda_{lt} \tag{7}$$

where $1\le p<m$ and $0<AWT(l)\le 1$. For a given AWT, the "snapshot window" is defined by the size of p samples and the gene expression pattern under these p samples can be revealed in the window. By changing AWT, we can explore different "snapshot windows" in a cluster.

4 Experiments

We conducted a series of experiments on two real microarray datasets to investigate the changing window approach in exploring novel gene expression patterns. For each dataset, we used different γ values to generate different sets of clusters. For each cluster, we explored different "snapshot windows" by changing the AWT value. The datasets and experiment results are discussed below.

Table 1. The EWKM subspace clustering algorithm

Input	The data matrix, the number of clusters k, and parameter γ.
Initialization	Randomly choose k cluster centers and set all initial weights to 1/m.
Iteration	Compute the partition matrix W by (3); Compute the cluster centers Z by (4); Compute the sample weights Λ by (5).
Until	The objective function obtains its local minimum value.

Table 2. Real world microarray datasets

Datasets	Genes	Samples	Clusters
Iyer	517	12 (time point)	10
Golub	7129	38 (27ALL, 11AML)	2

4.1 Datasets

In the experiments, we used two public microarray datasets, Iyer's dataset and Golub's dataset as shown in Table 2. More details about these two data sets can be found in [13] and [14]. Each dataset is represented as an $m \times n$ matrix of real-valued expression levels $Y = y_{ij}$, where genes were represented as rows and samples as columns. All sample columns were standardized to zero means and one standard deviation to eliminate the scale difference.

4.2 Experiment analysis

The purpose of the experiments was to explore novel expression patterns in different "snapshot windows" by changing AWT. For each data set, we ran the EWKM algorithm 10 times. In each run, the number of clusters was set to 10 and the same initial cluster centers were used. Parameter γ was changed from 0.5 to 9.5 with increment of 1. Altogether, 100 different clusters were generated from each data set.

Figure 1 shows gene expression profiles in different "snapshot windows" of a cluster from Iyer's data. In each window, the horizontal axis represents the twelve time points (i.e., 15min, 30min, 1hr, 2hr, 4hr, 6hr, 8hr, 12hr, 16hr, 20hr, 24hr, and UNSYN), and the vertical axis represents the gene expression level. Each thin line represents one gene profile in the cluster. The red line with circle dots represents the average expression level of all genes in the cluster. The circle dots represent the significant samples that were included in the "snapshot window".

To visualize different possible expression patterns in a cluster, we changed "snapshot windows" by moving the

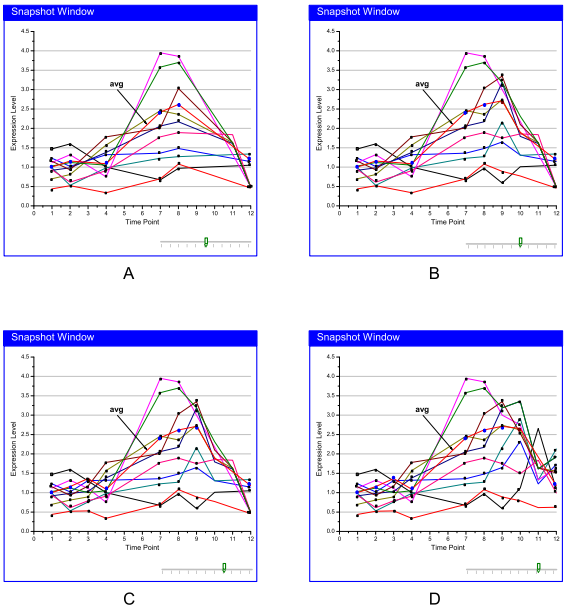

Figure 1. Reveal expression details in different snapshot windows by increasing AWT

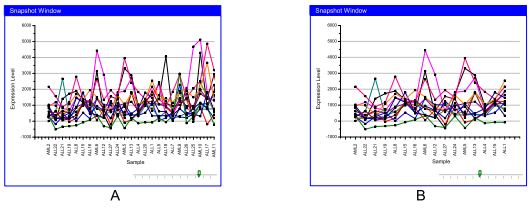

Figure 2. Discover specific expression patterns by decreasing AWT

changing bar to increase or decrease AWT. When AWT was increased, more samples were included in the "snapshot window" and the order of samples to be included was determined by the sample weights which indicate the significance of the samples in expressing the genes. For example, the snapshot window A of Figure 1 was generated with AWT=0.5 and it contains a gene expression pattern in six samples with comparatively large weights. This pattern represents a six-point time course (i.e., 15min, 30min, 2 hr, 8 hr, 12 hr, and UNSYN), which shows that most genes had low expression levels in the first 2 hours. The expression levels increased significantly afterwards and arrived at peak levels after 8 hours. The expression level reduced after 12 hours and reached the initial level at UNSYN. These six sample points revealed the general expression pattern of the

genes in this cluster.

After increasing AWT to 0.6, another relatively more important sample corresponding to time point 16hr was included in window B. We can see that a new peak appeared at the 16 hour time point. However, since this new peak was only slightly higher than the peak at the 12 hour point in window A and the sample weight of this new peak was smaller than the sample weight of the peak at the 12 hour time point, the previous peak was stronger than the new peak. This can be observed from window B where most genes were peaked at the 12 hour point and the expression levels reduced afterwards. A few genes were peaked at the 16 hour point.

Further increasing AWT to 0.7 and 0.8, more samples were included in windows C and D. A small peak occurred at the 1 hour point in window C, which may indicate that serum started effective after one hour. An important sample occurred at the 20 hour point in window D. From this time point, the expression level reduced.

Figure 2 shows another example of gene expression profiles of a cluster from Golub's leukemia data set. This data set contained two types of samples of acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Different gene expression patterns in ALL and AML could be explored in "snapshot windows". Window A corresponded to AWT=0.8 and window B to AWT=0.5. In window A, many samples in both ALL and AML were included. It was difficult to judge whether this pattern was corresponding to ALL or AML. When AWT was reduced, many samples with small weights were removed from window A. The remaining samples with relatively bigger weights showed a strong pattern that was related to ALL because most samples were ALL samples.

The above examples show that this changing window approach is flexible to explore gene expression patterns at different details. In other subspace clustering algorithms, only one pattern is explored in a cluster.

5 Conclusions

In this paper, we have presented a changing window approach to exploring gene expression patterns in clusters generated with the EWKM subspace clustering algorithm from microarray data. In this approach, importance of samples in a cluster is ordered by the sample weights generated in the EWKM clustering process. A "snapshot window" is defined to visualize a gene expression pattern of a cluster in a subset of samples. By changing the accumulated weighting threshold (AWT), different expression patterns with different details can be explored in a cluster. Therefore, this approach offers biologist a more flexible tool to explore more expression patterns from microarray data.

In our future work, we will investigate post-processing methods to remove noise genes from "snapshot windows" and develop modeling techniques to model subspace cluster patterns revealed in snapshot windows as in [8]. We will also develop a tool for gene expression visualization.

Acknowledgments

The authors would like to thank Iyer *et al.* and Golub *et al.* for their contribution of the fibroblast and leukemia dataset respectively. This research is supported in part by NSFC under grant No.60603066 and China National High-tech Program under grants No.2007AA01Z436 and No. 2006AA01A124. Part of Joshua Huang's research was supported by the 863 project matching fund from The University of Hong Kong.

References

[1] J.L. DeRisi, V.R. Iyer, and P.O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale In *Science*, 278:680-686, 1997.

[2] H. Wang, F. Chu, W. Fan, P.S. Yu, and J. Pei. A Fast Algorithm for Subspace Clustering by Pattern Similarity. *In SSDBM*, 51-60, 2004.

[3] L. Parsons, E. Haque, and H. Liu. Subspace Clustering for High Dimensional Data: A Review. *In SIGKDD Explorations*, 6(1): 90-105, 2004.

[4] L. Jing, M.K. Ng, and J.Z. Huang. An Entropy Weighting K-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data. *IEEE TKDE*, 19(8):1-16, 2007.

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

[6] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nat. Genet.*, 22: 281-285, 1999.

[7] P. Tamayo et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *PNAS*, 96:2907-2912, 1999.

[8] G. Getz, E. Levine, and E. Domany. Coupled Two-way Clustering Analysis of Gene Microarray Data. *Proc. Natl. Acad. Sci. USA*, 97(22): 12079-12084, Oct. 2002.

[9] L. Lazzeroni, and A. Owen. Plaid Models for Gene Expression Data. *Statistica Sinica*, 12(1):61-86, 2002.

[10] J. Yang, W. Wang, H. Wang, and P.S. Yu. delta-cluster: Capturing Subspace Correlation in a Large Data Set. *In ICDE 2002*, 517-528, 2002.

[11] Y. Chen, and G.M. Church. Biclustering of Expression Data. *Proceedings of ISMB'00*, 93-103, 1999.

[12] J.Z. Huang, M.K. Ng, H. Rong, and Z. Li. Automated Variable Weighting in k-Means Type Clustering. *IEEE Trans. PAMI*, 27(5): 1-12, May 2005.

[13] V.R. Iyer et al. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83-87, 1999.

[14] T.R. Golub et al. Molecular classification of cancer: Class discovery and class prediction by gene expression. *JBCB*, 286(5439):531-537, 1999.