An Ensemble Method of Discovering Sample Classes Using Gene Expression Profiling

Dechang Chen¹, Zhe Zhang², Zhenqiu Liu³, and Xiuzhen Cheng⁴

¹ Department of Preventive Medicine and Biometrics
Uniformed Services University of the Health Sciences
4301 Jones Bridge Road, Bethesda, MD 20814, USA
dchen@usuhs.mil

² Department of Biomedical Engineering
University of North Carolina
Chapel Hill, North Carolina 27599, USA
zhangz@email.unc.edu

³ Bioinformatics Cell, TATRC
110 North Market Street, Frederick, MD 21703, USA
liu@bioanalysis.org

⁴ Department of Computer Science
The George Washington University
801 22nd St. NW, Washington, DC 20052, USA
cheng@gwu.edu

Summary. Cluster methods have been successfully applied in gene expression data analysis to address tumor classification. Central to cluster analysis is the notion of dissimilarity between the individual samples. In clustering microarray data, dissimilarity measures are often subjective and predefined prior to the use of clustering techniques. In this paper, we present an ensemble method to define the dissimilarity measure through combining assignments of observations from a sequence of data partitions produced by multiple clusterings. This dissimilarity measure is then subjective and data dependent. We present our algorithm of hierarchical clustering based on this dissimilarity. Experiments on gene expression data are used to illustrate the application of the ensemble method to discovering sample classes.

Key words: Cluster analysis, Dissimilarity measure, Gene expression

1 Introduction

Microarrays provide a very effective approach to interrogate hundreds or thousands of genes simultaneously. Such high throughput capability poses great challenges in terms of analyzing the data and transforming the data into useful information. As an exploratory data analysis tool, clustering has become

a useful technique for identifying different and previously unknown cell types [3, 4, 9, 10].

Among many clustering methods applied to cluster samples, hierarchical clustering and its variations have received a special attention [2, 4, 6, 8]. This is mainly because a hierarchical method can produce a dendrogram, which provides a useful graphical summary of the data. However, dendrograms depend on the measures of dissimilarity for each pair of observations. A dissimilarity measure is amplified graphically by means of a dendrogram. In microarray data analysis, dissimilarity measures are commonly based on Pearson correlation. Such measures are restrictive, since Pearson correlation coefficient only describes the linear relationship between the observations on two variables. Therefore, it is difficult to see how one might be capable of making valid biological inferences. In this paper, we present an ensemble method to define the dissimilarity measure. This method derives dissimilarity by combining assignments of observations from a sequence of data partitions produced by multiple clusterings. Thus, the dissimilarity measure is subjective and data dependent. We then present our algorithm of hierarchical clustering based on this dissimilarity. Experimental results show that the ensemble method is efficient in discovering sample classes using gene expression profiles.

2 Methods

Assume that there are $k(\geq 2)$ distinct tumor tissue classes for the problem under consideration, and that there are p genes (inputs) and n tumor mRNA samples (observations). Suppose x_{li} is the measurement of the expression level of the l th gene from the i th sample for $l = 1, \dots, p$ and $i = 1, \dots, n$. Let $\mathbf{G} = (x_{li})_{p \times n}$ denote the corresponding gene expression matrix. Note that the columns and rows of the expression matrix $\mathbf{G}$ correspond to samples and genes, respectively. The i th column may be written as a vector $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{pi})^T$, where T represents the transpose operation. We consider clustering the data $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$. Clustering may be viewed as a process of finding natural groupings of the observations $\mathbf{x}_i$. A key issue related to this type of groupings is how one measures the dissimilarity between data points. Most clustering algorithms presume a measure of dissimilarity. For example, K -means clustering uses Euclidean distance as a dissimilarity measure and Hierarchical clustering often uses correlation based dissimilarity measures. In this section, we briefly review K -means method and linkage methods, special cases of Hierarchical clustering techniques. Then we present an ensemble clustering algorithm using K -means and linkage methods.

K -means

Cluster analysis aims at partitioning the observations into clusters (or groups) so that observations within the same cluster are more closely related to each

other than those assigned to different clusters. Partitioning is one of the major clustering approaches. A partitioning method constructs a partition of the data into clusters that optimizes the chosen partitioning criterion. The K -means is one of the most popular partitioning methods. This method uses Euclidean distance as the dissimilarity measure. It starts with a given assignment and proceeds to assign an observation to the cluster whose mean is closest. The process is repeated until the assignments do not change.

Linkage Methods

Commonly used linkage methods include single linkage (SL), complete linkage (CL), and average linkage (AL). They are special cases of agglomerative clustering techniques and follow the same procedure: beginning with the individual observations, at each intermediate step two least dissimilar clusters are merged into a single cluster, producing one less cluster at the next higher level [7]. The difference among the linkage methods lies in the dissimilarity measures between two clusters, which are used to merge clusters. SL, CL, and AL define the dissimilarity between two clusters to be the minimum distance between the two clusters, the maximum distance between the two clusters, and the average distance between the two clusters, respectively. Specifically, suppose we have n observations $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$. Let $d(\mathbf{x}_i, \mathbf{x}_j)$ denote the predefined dissimilarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. Given two clusters G_1 and G_2 , containing n_1 and n_2 observations, respectively, the dissimilarity between G_1 and G_2 defined by SL, CL, and AL are

$$\begin{aligned} d_{SL}(G_1, G_2) &= \min_{\mathbf{x}_i \in G_1, \mathbf{x}_j \in G_2} d(\mathbf{x}_i, \mathbf{x}_j), \\ d_{CL}(G_1, G_2) &= \max_{\mathbf{x}_i \in G_1, \mathbf{x}_j \in G_2} d(\mathbf{x}_i, \mathbf{x}_j), \\ d_{AL}(G_1, G_2) &= \frac{1}{n_1 n_2} \sum_{\mathbf{x}_i \in G_1, \mathbf{x}_j \in G_2} d(\mathbf{x}_i, \mathbf{x}_j). \end{aligned}$$

Hierarchical Clustering by Ensemble Procedure

Learning Dissimilarity from Data

The clustering procedure partitions the data into clusters so that observations in one cluster are more like one another than like observations in other clusters. This procedure requires a measure of “closeness” or “similarity” between two observations. Such a measure can be made by using any metric $d(\cdot, \cdot)$, which, for any observations $\mathbf{x}_i, \mathbf{x}_j$, and $\mathbf{x}_k$, satisfies the following four properties: (i) (nonnegativity) $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$; (ii) (reflexivity) $d(\mathbf{x}_i, \mathbf{x}_j) = 0$ if and only if $\mathbf{x}_i = \mathbf{x}_j$; (iii) (symmetry) $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$; (iv) (triangle inequality) $d(\mathbf{x}_i, \mathbf{x}_j) + d(\mathbf{x}_j, \mathbf{x}_k) \geq d(\mathbf{x}_i, \mathbf{x}_k)$. The smaller the measure $d(\mathbf{x}_i, \mathbf{x}_j)$ is, the

more similar the two observations. As an example, the Euclidean distance $d(\mathbf{x}_i, \mathbf{x}_j) = (\sum_{l=1}^p (x_{li} - x_{lj})^2)^{1/2}$ is a metric.

Clustering methods also accept other measures of “closeness” that may not meet the reflexivity or triangle inequality. In general, one can introduce a dissimilarity function $dis(\mathbf{x}_i, \mathbf{x}_j)$ to measure ‘closeness’ or ‘similarity’ between $\mathbf{x}_i$ and $\mathbf{x}_j$. A dissimilarity function is a function that satisfies nonnegativity and symmetry. For example, $dis(\mathbf{x}_i, \mathbf{x}_j) = 1 - \cos(\mathbf{x}_i, \mathbf{x}_j)$ is a dissimilarity function, where $\cos(\mathbf{x}_i, \mathbf{x}_j)$ refers to the cosine of the angle between $\mathbf{x}_i$ and $\mathbf{x}_j$. (It is easy to show that neither reflexivity nor the triangle inequality is satisfied by this measure.)

The notion of dissimilarity is central to cluster analysis. Different choices of dissimilarity functions can lead to quite different results. Prior knowledge is often helpful in selecting an appropriate dissimilarity measure for a given problem. However, it is possible to learn a dissimilarity function from the data. We describe such a procedure as follows.

Partitioning methods are usually not stable in the sense that the final results often depend on initial assignments. However, if two observations are assigned to the same cluster by a high percentage of the times of use of the same partitioning method, it is then very likely that these two observations come from a common “hidden” group. This heuristic implies that the “actual” dissimilarity between two observations may be derived by combining the various clustering results from repeated use of the same partitioning technique. Here we formalize this combining process using K -means partitioning method.

For the data $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, we can select K centroids and then run K -means technique to partition the data into K clusters. It is known that the final assignment usually depends on the initial reallocation. Now we run K -means N times. Each time a number K is randomly picked from a given interval $[K_1, K_2]$. By doing this, we may end up with N possibly different final assignments. Given observations (samples) $\mathbf{x}_i$ and $\mathbf{x}_j$, let p_{ij} denote the probability that they are not placed into the same cluster by the final assignment of a run of K -means clustering. This probability p_{ij} can be estimated by using the results of repeated K -means clustering method. Define $\delta_m(i, j) = 1$ if the m th use of the K -means algorithm does not assign samples $\mathbf{x}_i$ and $\mathbf{x}_j$ into the same cluster; and $\delta_m(i, j) = 0$ otherwise. Then $\delta_1(i, j), \delta_2(i, j), \dots, \delta_N(i, j)$ are iid Bernoulli(p_{ij}). It is well known that the best unbiased estimator of p_{ij} is $\sum_{m=1}^N \delta_m(i, j)/N$. This estimate will be used as the dissimilarity measure between $\mathbf{x}_i$ and $\mathbf{x}_j$, i.e.,

$$dis(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{m=1}^N \delta_m(i, j)}{N}. \quad (1)$$

A smaller value of $dis(\mathbf{x}_i, \mathbf{x}_j)$ is expected to imply a bigger chance that samples $\mathbf{x}_i$ and $\mathbf{x}_j$ come from the same “hidden” group.

We now state our ensemble method using K -means technique and linkage methods. Here the word ensemble refers to the sequence of K -means procedures involved in the method.

Algorithm

1. Given N , K_1 , and K_2 , run the K -means clustering method N times with each K randomly chosen from $[K_1, K_2]$.
2. Construct the pairwise dissimilarity measure $dis(\mathbf{x}_i, \mathbf{x}_j)$ between the observations by using the equation (1).
3. Cluster the n samples (observations) by applying a linkage method and the dissimilarity measure $dis(\mathbf{x}_i, \mathbf{x}_j)$ learnt in Step 2.

3 Results and Discussion

Datasets

We considered two gene expression datasets: COLON [1] and OVARIAN [11]. The COLON dataset consists of expression profiles of 2000 genes from 22 normal tissues and 40 tumor samples. The OVARIAN dataset contains expression profiles of 7129 genes from 5 normal tissues, 27 epithelial ovarian tumor samples, and 4 malignant epithelial ovarian cell lines.

Standardization and Gene Selection

Following Dudoit and Fridlyand [3], we standardized the gene expression data so that the samples have mean 0 and variance 1 across genes. This simple standardization procedure achieves a location and scale normalization of different arrays.

Each gene dataset considered here contains several thousand genes. Genes showing almost constant expression levels across samples are not expected to be useful for clustering. Thus in our analysis, we used $p = 100$ genes which correspond to 100 largest variances of the gene expression levels across the samples.

Parameter Setting

To run our algorithm, we chose parameters as follows. The choice of N depends on the rate at which dis in (1) converges to p_{ij} . A large number should be chosen for N . For obtaining quick results, we set $N = 1000$. Since the number of true clusters for each dataset is at least 2, we set $K_1 = 2$. A large K_2 is required to guarantee that $\delta_1(i, j), \delta_2(i, j), \dots, \delta_N(i, j)$ are iid Bernoulli(p_{ij}). However, in practice, the algorithm will not work for very large values of K_2 . For example, $K_2 > n$ is clearly impractical. In our examples, we used $K_2 = 30$.

Cluster Analysis

For both gene expression datasets, we used single linkage in our algorithm. Cluster analysis for COLON and OVARIAN is shown in Fig. 1 and Fig. 2, respectively.

Fig. 1. Dendrogram from clustering of COLON data.

The dendrogram for COLON clearly shows that all the samples are clustered into two classes with 6 misclassified samples (NORMAL08, NORMAL18, NORMAL20, TUMOR23, TUMOR27, and TUMOR34). Alon et al. [1] conducted a two-way clustering analysis based on 20 genes with the most significant difference between tumor and normal tissues. They found two clusters with 8 misclassified samples: 5 tumor samples clustered with normal tissues and 3 normal tissues clustered with tumor samples.

From the dendrogram for OVARIAN, our algorithm shows that all the samples are clustered into three classes: the class for normal tissues, the class for malignant epithelial ovarian cell lines, and the class for tumor samples. The figure indicates that only one sample BCELL is misclassified. Clearly,

Fig. 2. Dendrogram from clustering of OVARIAN data.

our result is better than that of Welsh et al. [11], where the samples are mixed together.

The above clustering analysis shows that our algorithm could be employed to effectively discover sample classes using gene expression profiling. From the above two examples, we have found that the dendrograms do not show any significant change when increasing N or p or both, where p is the number of genes used in the analysis. This robustness will be further studied both theoretically and experimentally. SL was used in our algorithm for the above two gene expression datasets. The effect of linkage methods other than SL will be investigated in our future work. Our algorithm uses K -means to define dissimilarity. In our future work, we will also investigate the effect of replacing K -means by other partitioning methods. The ensemble method of clustering presented here was discussed in [5] from the perspective of evidence accumulation.

Acknowledgments

D. Chen was supported by the USUHS grant R087RH-01 and the National Science Foundation grant CCR-0311252.

References

1. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.* 96, 6745-6750.
2. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, X., Ben-Dor, A., Dougherty, E., Wang E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N., and Trent, J. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406, 536-540.
3. Dudoit, S. and Fridlyand, J. (2002). A prediction based resampling method to estimate the number of clusters in a dataset. *Genome Biology*, 3(7), 0036.1-0036.21.
4. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* 95, 14863-14868.
5. Fred, A. and Jain, A. K. (2002). Data clustering using evidence accumulation. *Proc. of the 16th International Conference on Pattern Recognition, ICPR 2002*, 276-280.
6. Goldstein, D. R., Ghosh, D., and Conlon, E. M. (2002). Statistical issues in the clustering of gene expression data. *Statistica Sinica*, 12, 219-240.
7. Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.
8. Li, W., Fan, M., and Xiong, M. (2003). SamCluster: An integrated scheme for automatic discovery of sample classes using gene expression profile. *Bioinformatics*, 19(7), 811-817.
9. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci.* 96, 2907-2912.
10. Tavazoie, S., Hughes, J., Campbell, M., Cho, R., and Church, G. (1999). Systematic determination of genetic network architecture. *Nature Genetics*, 22, 281-285.
11. Welsh, J. B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Burger, R. A., Monk, B. J., and Hampton, G. M. (2001). Analysis of gene expression in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl. Acad. Sci.* 98, 1176-1181.