The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

A Robust Clustering Technique for Grouping Biological Data: an Illustrative Study in Gene Expression Data*

Xuemei Ning[1,†] Shihua Zhang[2,‡]

[1]College of Science, Beijing Forestry University, Beijing 100083
[2]Academy of Mathematics and Systems Science, CAS, Beijing 100190

Abstract Clustering data based on a measure of similarity (or dissimilarity) is a critical step in scientific data analysis and especially in current bioinformatics field. A typical example is, with the advent of DNA Microarrays, clustering analysis becomes a powerful way to explore the expression profiles of all genes in the genome. And many algorithms have been developed for this problem. Here, we aim to introduce a more robust clustering technique (neural gas algorithm) for this problem and hope it can be applied to other problems in bioinformatics.

Keywords Gene expression profiles; Neural Gas; K-means; Clustering analysis

1 Introduction

Clustering analysis of data based on a measure of similarity is a critical and basic step for many problems in the field of computational biology. The common approach is to use original data to learn a set of centers such that the sum of error residues between data points and their closest centers is as small as possible. A typical example in computational biology is clustering analysis of Microarray data which can provide a compact global view of the whole expression profiles of all genes in the genome. To analyze the large number of genes in the complicated biological systems, researchers usually group the genes with similar expression profiles. It has been observed that genes with the same function or involved in the same biological process are likely to be co-expressed, hence grouping genes' expression profiles provides a means for understanding gene function, gene regulation, and cellular processes [1].

Clustering analysis has become a indispensable exploratory technique for many problems in bioinformatics, especially for analyzing gene expression data. Many clustering algorithms have been proposed for gene expression data analysis. For example, Eisen *et al.* [10] applied a variant of the hierarchical average-link clustering algorithm to identify groups of co-expression yeast genes. Tamayo *et al.* [21] used SOM to identify clusters

*This work is partially supported by the National Natural Science Foundation of China under grant No.60803099 and Beijing Forestry University Young Scientist Fund 200-8108.
†ningxuemei@bjfu.edu.cn
‡zsh@amss.ac.cn

in the yeast cell cycle and human hematopoietic differentiation data sets. Tavazoie *et al.* [16] used k-means method to analyze Microarray data generated from studies of the yeast cell cycle. More methods have been found in some review papers ([1], [2]).

However, a particular kind of vector quantization method, proposed by Martinetz and Schlten [9], the Neural Gas (NG) network method is ignored by researchers for clustering gene expression data. The NG algorithm has been successfully applied to vector clustering, pattern recognition and topology representation, etc. ([17], [18]). Actually, it is a more robust method than classical clustering methods like k-means (or k-medoid). In this paper, we introduce NG algorithm to the biological data analysis and as an illustrative study, we apply it to gene expression data. We compare it with the classical k-means method [15] using three popolar validation measures proposed by Datta *et al.* [4] on two well-known publicly available Microarray data sets.

2 Materials and Methods

2.1 Sporulation data

Data analyzed here were collected on the transcriptional program of sporulation in budding yeast [3]. It was obtained at http://cmgm.stanford.edu/pbrown/sporulation. There are 6118 protein-encoding genes in the yeast genome, and the mRNA levels were measured at seven time points during the sproulation process [3].

The ratio of each gene's mRNA levels to its mRNA level in vegetative cells just before transfering to sporulation medium is measured, more than 1000 genes whose root mean square of $\log_2$-transformed ratios at a given time point was greater than 1.13, showed significant changes in mRNA levels during sporulation [3]. About half of these genes were induced, and half were repressed during the process. Overall, there are 522 positively expressed genes for further analysis. The data are then summarized by a 522×7 matrix $X = (X_{ij})$, where X_{ij} denotes the expression ratio for gene i in time point j. There are no missing values and the data were normalized as described below.

2.2 Combined yeast data

An aggregation of data from experiments on the budding yeast *Saccharomyces cerevisiae*, including time courses of the mitotic cell division cycle [13], sproulation [3], the diauxic shift [14] and temperature and reducing shocks, was described in [10] and was obtained at http://rana.stanford.edu/clustering. This study produced gene-expression data for 2467 genes in 79 samples. The gene expression data are summarized by a 2467×79 matrix $X = (X_{ij})$, where X_{ij} denotes the base-2 logarithm of the Cy5/Cy3 fluorescence ratio for gene i in sample j. For each gene X_i with missing data, we simply estimate the missing value using the average of observations on X_i.

2.3 Normalization

A gene expression data set from a Microarray experiment can be represented by a matrix, where the rows represent genes, the columns represent the expression profiles of samples (e.g., single timepoints or conditions), and each cell is the measured expression level of a gene in a corresponding sample.

For the gene expression matrix, data normalization is indispensable before clustering analysis. Here, we normalize the $\log_2$-transformed expression ratios so that their mean

is zero and their variance is one before proceeding with the actual cluster algorithm [8]. Let X_{ij} denotes the $\log_2$-transformed expression ratio for gene X_i in sample j, then we can normalize the data by subtracting its mean across the time points from the expression level of each gene, and dividing by the standard deviation across the time points:

$$Y_{ij} = \frac{X_{ij} - \bar{X}_i}{\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2}},$$

where $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$ is the mean over observations on X_i.

2.4 Neural gas algorithm

As one of the partitioning clustering algorithm, the Neural Gas (NG) network algorithm has been successfully applied to vector quantization. Given a pre-specified number K, the first step in the algorithm is to randomly select K reference vectors $C_1, C_2, ..., C_K$. When an input vector X is presented, the reference vectors are ranked according its Euclidean distance to the input vector. After that the reference vectors are updated with different weight according their positions in the 'neighborhood ranking' list.

Let $C_{i_0}, C_{i_1}, ..., C_{i_k}, ..., C_{i_{K-1}}$ denote the neighborhood ranking of the reference vectors, where C_{i_0} represents the nearest reference vector to X, C_{i_k} is the k nearest reference vector to X. The ranking index associated with C_i is denoted by $k_i(X,C)$, then C_i is adjusted by

$$C_i = C_i + \alpha(t) \cdot H_\lambda(k_i(X,C)) \cdot (X - C_i), \quad i = 1, 2, ..., K,$$

where $\alpha(t) \in [0,1]$ is the learning rate, which describes the overall extent of modification form as $\alpha(t) = \alpha_0 \cdot (\alpha_f/\alpha_0)^{t/tmax}$, α_0 (0.5 by default) and α_f (0.005 by default) are the initial and final learning rate, t and $tmax$ represent the iteration step t and the maximum number of iterations. $H_\lambda(k) = exp(-k/\lambda) \in [0,1]$ is the Competitive Hebbian Learning scheme [9]. The parameter λ determines the number of reference vectors significantly changing their positions in the updating steps, and decreases gradually from a large initial to a small final value. For the iteration step t, $\lambda(t) = \lambda_0 \cdot (\lambda_f/\lambda_0)^{t/tmax}$, λ_0 ($K/2$ by default) and λ_f (0.01 by default) are used to set the rate at which learning rate α converges. In order to obtain a good convergence rate and stable model, the neural gas model requires fine tuning of these parameters. The detailed implementation of NG algorithm is presented in [20].

By using this deterministic annealing like strategy, the neural gas network is a good alternative to traditional partitioning clustering method: k-means clustering. One of the appealing features of NG algorithm is that it is seldom sensitive to different initializations due to the sequential learning scheme and use of neighborhood cooperation rule. And the NG algorithm converges to lower distortion error than that resulting from k-means clustering [9].

2.5 Validation

In order to compare NG algorithm with the k-means method, we use three validation measures, each of which can be used for an objective basis of checking the consistency of the grouping produced by a clustering algorithm [4]. The basic idea is that an algorithm

should be rewarded for consistency or stability. We compared the results of clustering with the full data and the reduced data after reducing the expression profiles by one unit.

The first criteria is the average proportion of non-overlap measure:

$$V_1(K) = \frac{1}{Ml}\sum_{g=1}^{M}\sum_{i=1}^{l}\left(1 - \frac{n(C^{g,i}\bigcap C^{g,0})}{n(C^{g,0})}\right),$$

where K is the number of clusters , l is the number of samples, M is the total number of genes. For each gene $1 \leqslant g \leqslant M$, $C^{g,i}$ denotes the cluster containing gene g in the clustering based on the data set with time point i deleted, $C^{g,0}$ denotes the cluster in the original data containing gene g, $n(C)$ is the cardinality of cluster C. This measure computes the (average) proportion of genes that are not put in the same cluster by the clustering method under consideration on the basis of the full data and the data obtained by deleting the expression levels at one time point at a time.

The second criteria is the average distance between means measure:

$$V_2(K) = \frac{1}{Ml}\sum_{g=1}^{M}\sum_{i=1}^{l} d(\bar{x}_{C^{g,i}}, \bar{x}_{C^{g,0}}),$$

where $\bar{x}_{C^{g,0}}$ denotes the average expression profile for genes across cluster $C^{g,0}$ and $\bar{x}_{C^{g,i}}$ denotes the average expression profile for genes across cluster $C^{g,i}$. This measure computes the (average) Euclidean distance between the mean $\log_2$-expression ratios of all genes that are put in the same cluster by the clustering method under consideration on the basis of the full data and the data obtained by deleting the expression levels at one time point at a time.

The third criteria is the average distance measure:

$$V_3(K) = \frac{1}{Ml}\sum_{g=1}^{M}\sum_{i=1}^{l}\frac{1}{n(C^{g,0})n(C^{g,i})} \times \sum_{g\in C^{g,0}, g'\in C^{g,i}} d(x_g, x_{g'}),$$

where $d(x_g, x_{g'})$ is a Euclidean distance between the expression profiles of genes g and g'. This measure computes the average distance between the $\log_2$-expression levels of all genes that are put in the same cluster by the clustering method under consideration on the basis of the full data and the data obtained by deleting the expression levels at one time point at a time.

For a good clustering algorithm, these values should be as small as possible [4].

3 Results

Here, we test NG algorithm to group gene expression data and compare it with the classic k-means method. We evaluate their performance based on the validation measures on two well-known publicly available Microarray data sets. In this paper, we use Euclidean distance to measure the distance between two vectors.

First we consider 522 significant sporulation genes based on their 7 dimensional expression profiles [3]. Chu *et al.* (1998) aimed to cluster the expressed genes into seven

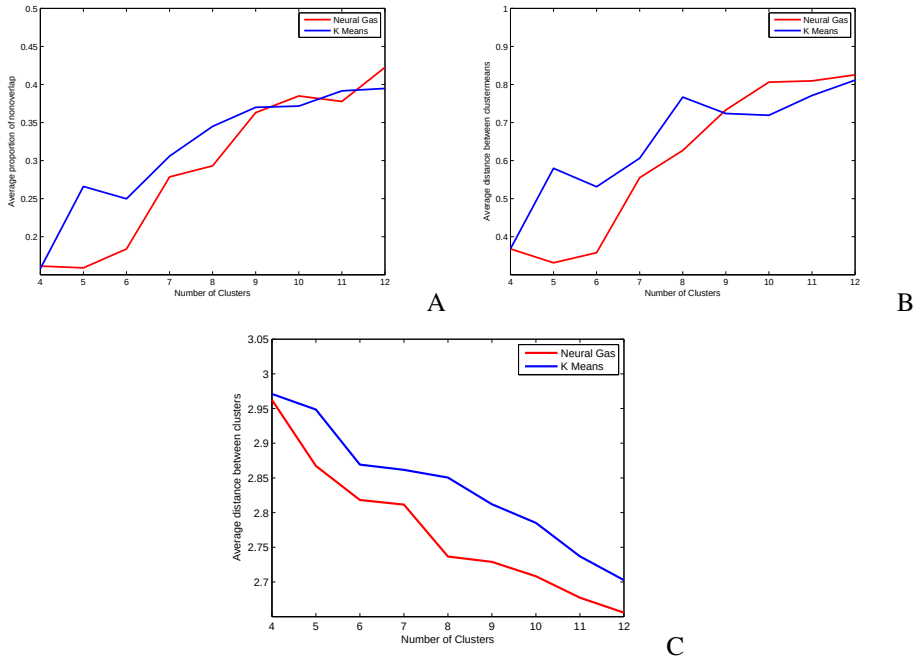

Figure 1: The comparative results for the two clustering algorithms with three validation measures respectively. In each plot, the red line represents the results of NG algorithm and the blue line represents the result of k-means method respectively. (A) a panel plot for the overall 'proportion of non-overlap' validation measure; (B) a panel plot for the average distance between means measure; (C) a panel plot for the average distance measure;

temporal classes from biological view. For each of the two clustering algorithms under consideration, we compute the three validation measures over a range of k values between four and twelve like in [4]. Considering the randomicity of the two clustering algorithms, we run each algorithm repeatedly (50 times) with each k value, and compare the best clustering results about the three validation measures of each clustering algorithm. Figure 1 displays the corresponding results. The three plots show the NG has consistent better performance than k-means for most cases. We can say that NG algorithm can converge to a 'better' partition than k-means.

We further analyze the 'best' clustering results of the two clustering algorithms with $k = 7$. In Chu's research [3], sporulation in yeast is characterized by sequential transcription of seven sets of genes-rapid including transient induction (metabolic), early I induction, early II induction, early-middle induction, middle induction, mid-late induction and late induction. In the 522 significant sporulation genes, 52 genes are marked metabolic, 61 genes are marked early I, 45 genes are marked early II, 93 genes are marked early-mid, 157 genes are marked middle, 261 genes are marked mid-late, 5 genes are marked late and 48 genes are unmarked. We use these prior knowledge to label clusters by hyper-

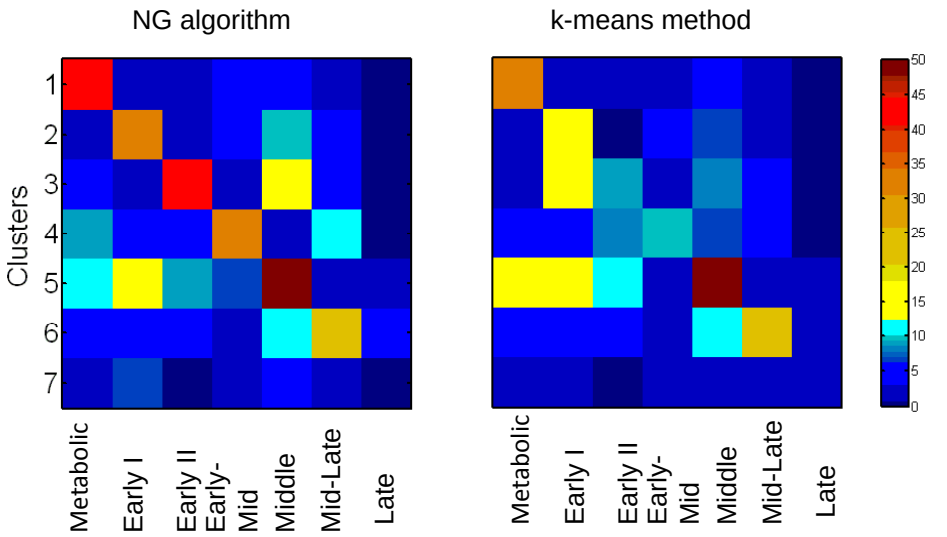

Figure 2: The enrichment analysis of clusters with respect to the known gene annotations.

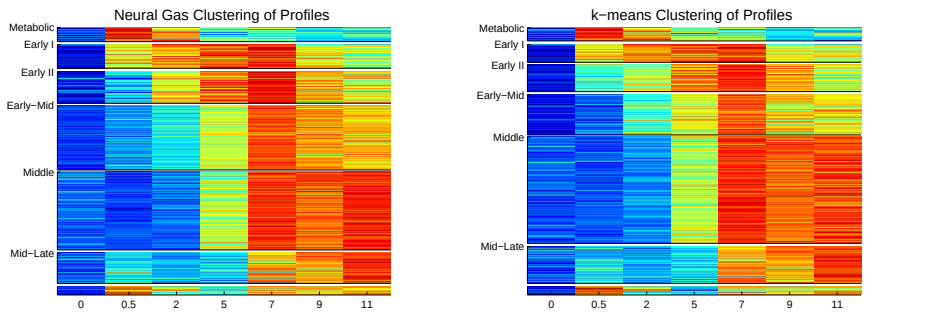

Figure 3: The colored images of the two clustering results. In each colored image, a single row of colored boxes represents one gene's $\log_2$-expression ratio, a single column represents one time point. Seven separate clusters are indicated by colored bars and by identical temporal class labels. (A) The colored image of NG algorithm; (B) The colored image of k-means method.

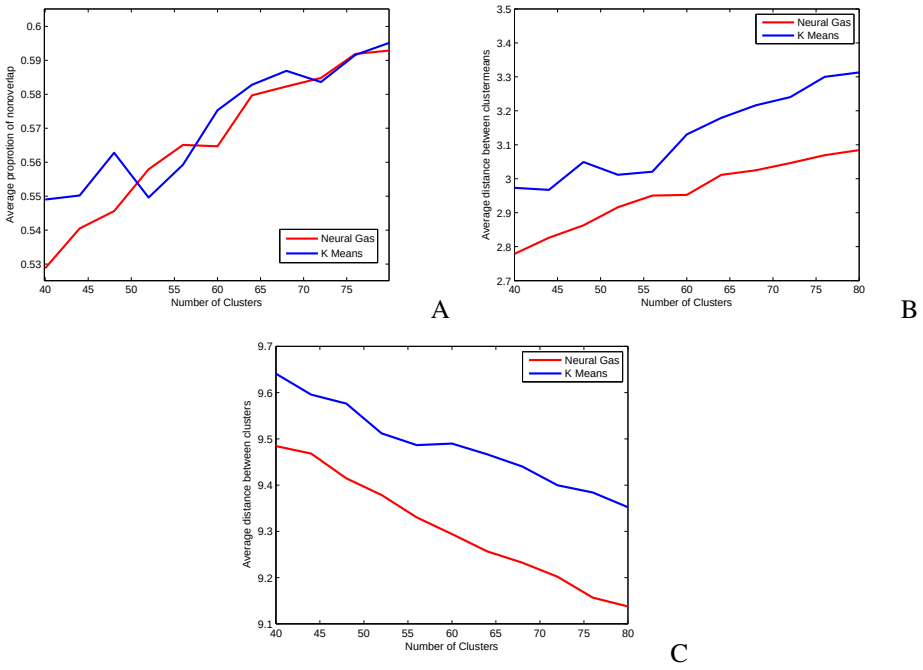

Figure 4: The Panel plots for the two clustering algorithms on combined yeast dataset. (A) a panel plot for the overall "proportion of non-overlap" validation measure; (B) a panel plot for the average distance between means measure; (C) a panel plot for the average distance measure;

geometric test. For each cluster obtained by the two clustering methods, we compute the statistical significance of each cluster with each set of marked genes, and assign each temporal class to the cluster with the highest significance. Figure 2 shows the value of $-log10(P)$ where P is the hyper-geometric P-value. We can easily find that the clusters detected by NG algorithm are more significant than that of k-means. Six clusters correspond to the known temporal condition well. Three clusters of k-means are significant in 'Early I' stage, and the fifth cluster of k-means are significant in three different stages. These show the clusters of NG algorithm are more biological relevant. The colored image of each cluster are given in Figure 3. In the colored image, a single row of colored boxes represents one gene's $\log_2$-expression ratio, a single column represents one time point. Seven separate clusters are indicated by colored bars and by identical temporal class labels.

We also test all the above analysis on the combined yeast data [10]. We compute the results with k ranging from 40 to 80. Figure 4 provides the results of the three validation measures for NG algorithm and k-means algorithm. It is clear that NG algorithm has much better performance than the classical k-means (Here limited by the space, we just show the results of the three measures).

4 Discussion

Clustering analysis has become powerful tool for analyzing biological data in the past decade. A huge number of methods have been developed for these problems ([10], [11], [12], [16], [21]) . The classical problem of Microarray data clustering analysis have made it to be typical test sets for evaluating the performance of different algorithms ([4], [5], [6], [7]).

Discovering patterns hidden in gene expression data offers a tremendous potential for advanced investigation in computational biology. Because of the large number of genes and the complexity of biological systems, clustering is an necessary exploratory technique for further analysis.

Here, we introduced the robust NG algorithm to the bioinformatics field for its efficiency. An illustrative study has been done on gene expression data and compare it with the most common partition method: k-means algorithm [15]. We evaluate their performances with three validation measures [4] on two well-known publicly available Microarray data sets. Considering gene expression data is noisy and has a significant number of outliers, and the neural learning process makes NG more robust than k-means algorithm to noisy data. From the results we can say that NG algorithm can group gene expression data into more stable clusters than k-means method. Thus, we hope that the NG algorithm can be a useful complementary to the computational biology and can be applied to other biological problems.

References

[1] Jiang D, Tang C, and Zhang A: Clustering analysis for gene expression data: a survey. IEEE Trans Knowledge and Data Eng 2004, 16: 1370-1386.

[2] Belacel N, Wang Q, and Cuperlovic-culf M: Clustering methods for microarray gene expression data. OMICS: A journal of Integrative Biology 2006, 10(4):507-531.

[3] Chu S, DeRisi J, Eisen M, *et al.* The transcriptional program of sporulation in budding yeast. Science 1998, 282: 699-705.

[4] Datta S, Datta S: Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 2003, 19: 459-466.

[5] Datta S, Datta S: Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. BMC Bioinformatics 2006, 7:397.

[6] Datta S, Datta S: Evaluation of clustering algorithms for gene expression data. BMC Bioinformatic 2006, 7(Suppl4):S17.

[7] Priness I, Mainom O, and Ben-Gal I: Evaluation of gene-expression clustering via mutual information distance measure. BMC Bioinformatics 2007, 8:111.

[8] Smet FD, Mathys J, Marchal K, *et al.*: Adaptive quality-based clustering of gene expression profiles. Bioinformatics 2002, 18:735-746.

[9] Martinetz TM, Berkovich SG and Schulten KJ: "Neural-gas" network for vector quantization and its application to time-series prediction. IEEE Transactions on Neural Networks, 1993, 4(4): 558-569.

[10] Eisen MB, Spellman PT, Brown PO and Botstein D: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 1998, 95(25):14863-14868.

[11] Carmona-Saez P, Pascual-Marqui RD, Tirado F, *et al.*: Biclustering of gene expression data by non-smooth non-negative matrix factorization. BMC Bioinformatics 2006, 7:78.

[12] Fu L and Medico E F: A novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics 2007, 8:3.

[13] Spellman PT, Sherlock G, Iyer VR, *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell 1998, 9: 3273-3297.

[14] DeRisi JL , Iyer VR and Brown PO: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997, 278: 680-686.

[15] McQueen, JB: Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, 1: 281-297.

[16] Tavazoie S, Jason D, Hughes JD, Campbell RJ, Raymond JS et al. Systematic determination of genetic network architecture. Nat Genet 1999, 22(3): 281-285.

[17] Atukorale AS, Downs T, and Suganthan PN: Boosting the HONG network. Neurocomputing 2003, 51: 75-86.

[18] Winter M, Metta G, and Sandini G: Neural-gas for function approximation: a heuristic for minimizing the local estimation error. Proceeding of 2000 international joint conference on neural network (IJCNN00) 2000, 4: 535-538, Italy.

[19] Kohonen T, (2001). Self-organizing maps (3rd ed.). Berlin: Springer.

[20] Fritzke, B. (1997). Some competitive learning methods. Technique report. Institute for Neural Computation, Ruhr-University, Bochum.

[21] Tamayo P, Slonim D, Mesirov J, *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl Acad. Sci. USA 1999, 96: 2907-2912.