BIOINFORMATICS

Vol. 20 no. 12 2004, pages 1905–1913
doi:10.1093/bioinformatics/bth177

Supervised cluster analysis for microarray data based on multivariate Gaussian mixture

*Yi Qu and Shizhong Xu**

Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

Received on May 25, 2003, revised on November 26, 2003; accepted on 29 January 2004
Advance Access publication March 25, 2004

ABSTRACT

Motivation: Grouping genes having similar expression patterns is called gene clustering, which has been proved to be a useful tool for extracting underlying biological information of gene expression data. Many clustering procedures have shown success in microarray gene clustering; most of them belong to the family of heuristic clustering algorithms. Model-based algorithms are alternative clustering algorithms, which are based on the assumption that the whole set of microarray data is a finite mixture of a certain type of distributions with different parameters. Application of the model-based algorithms to unsupervised clustering has been reported. Here, for the first time, we demonstrated the use of the model-based algorithm in supervised clustering of microarray data.

Results: We applied the proposed methods to real gene expression data and simulated data. We showed that the supervised model-based algorithm is superior over the unsupervised method and the support vector machines (SVM) method.

Availability: The program written in the SAS language implementing methods I–III in this report is available upon request. The software of SVMs is available in the website http://svm.sdsc.edu/cgi-bin/nph-SVMsubmit.cgi

Contact: xu@genetics.ucr.edu

INTRODUCTION

DNA microarray experiments allow us to measure the expression levels of thousands of genes simultaneously under various conditions. Gene expression profiles provide some clue to the functions of individual genes. This is because genes with similar functions are likely to show similar expression patterns under various conditions (Carr *et al*., 1997; Cho *et al*., 1998; Hughes *et al*., 2000; Szabo *et al*., 2002). By comparing the expression patterns of unknown genes to those of known functions, one can predict the functions of unknown genes. This is the primary objective of the supervised cluster analysis of gene expression data.

Many clustering techniques are available. The commonly used methods in microarray data analysis include hierarchical clustering (Carr *et al*., 1997), K-means (Tavazoie *et al*., 1999), self-organizing maps (SOMs) (Herrero *et al*., 2001) and support vector machines (SVMs) (Brown *et al*., 2000). These algorithms are largely, heuristically motivated and they do not require any underlying statistical models. One possible alternative to these 'heuristic' algorithms is the model-based clustering method (Dasgupta and Raftery, 1998) which is based on the assumption that the whole set of microarray data is a finite mixture of the same type of distribution each with a different set of parameters, such as the finite mixture of multivariate Gaussian distributions. One obvious advantage of the model-based clustering algorithm over the 'heuristic' algorithms is that with the underlying assumption, the choice of the optimal number of clusters and the models which fit the data best can be done by using some objective statistical criteria, i.e. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), whereas for the 'heuristic' algorithms, choosing the 'correct' number of clusters and the best clustering method is still a question open to discussion. Successful application of the model-based clustering to microarray data has been reported (Yeung *et al*., 2001; Mclachlan *et al*., 2002; Ghosh and Chinnaiyan, 2002).

In addition to heuristic clustering and model-based clustering, we may also categorize clustering methods into unsupervised and supervised clustering, according to the characteristics of the sampled genes. The supervised clustering method uses the expression profiles of genes with known functions as training samples. Unsupervised clustering, on the other hand, classifies all genes according to the same criteria, regardless of the functions of the genes. The supervised clustering method is obviously advantageous over the unsupervised one because the former utilizes additional information from the functional genes as the prior knowledge. Several supervised clustering techniques have been applied to microarray data, i.e. the SVMs (Brown *et al*., 2000) and multilayer perceptrons (Mateos *et al*., 2002). These algorithms belong to the category of 'heuristic' algorithms. To the best of our knowledge, a model-based supervised clustering method on microarray data has not been investigated.

An intuitive method to implement the model-based supervised clustering algorithm is to (a) estimate the parameters

*To whom correspondence should be addressed.

from the training sample and (b) use the estimated values of the parameters to classify the genes in the test dataset. A drawback of this method is that only the training set is used to estimate the parameters, and information from the test dataset is completely ignored. To tackle this problem, we propose to utilize both the training sample and the test dataset to estimate the model parameters using the expectation–maximization (EM) algorithm (Dempster *et al*., 1977). In fact, a majority of the information comes from the test dataset rather than the training sample because the latter usually accounts for only a small proportion of the entire dataset.

SYSTEMS AND METHODS

Multivariate Gaussian mixture

The finite mixture of multivariate normal distributions has been used to fit microarray data by a number of authors (Mclachlan *et al*., 2002; Ghosh and Chinnaiyan, 2002). Yeung *et al*. (2001) applied this model to fit the yeast cell cycle data and showed good fitness of the model, as judged by the BIC value (Ghosh and Chinnaiyan, 2002). First, we assume that the expression levels of all genes have been appropriately normalized (Yeung *et al*., 2001). Therefore, each observation in the data is already a processed data point and analysis can be directly performed. As usual, the data are arranged in an $n \times m$ matrix denoted by $\boldsymbol{Y}$, where n is the number of genes and m is the level of treatment (cases or time points). Let y_{ij} be the expression level of the i-th gene in the j-th treatment, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Let $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{im}]^{\mathrm{T}}$ be the i-th column of matrix $\boldsymbol{Y}^{\mathrm{T}}$, i.e. an $m \times 1$ vector for the expression data of gene i under all treatments. The values of y_{ij} across all the m treatments represent the expression profile of the i-th gene. With the finite multivariate Gaussian mixture model, each $\boldsymbol{y}_i$ is assumed to follow an m-dimensional mixture of normal distributions. Mathematically, the mixture distribution for C clusters is expressed as

$$f(\boldsymbol{y}_i) = \sum_{k=1}^{C} \pi_k f_k(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1}$$

where π_k, with $\sum_{k=1}^{C} \pi_k = 1$, is the mixing proportion of cluster k, and

$$f_k(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-m/2} |\boldsymbol{\Sigma}_k|^{-1/2} \times \exp\left[-\tfrac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)\right] \tag{2}$$

is the probability density of the k-th normal distribution with a mean vector $\boldsymbol{\mu}_k$ (an $m \times 1$ vector) and a variance–covariance matrix $\boldsymbol{\Sigma}_k$ (an $m \times m$ matrix). The mixing proportion π_k is defined as the proportion of genes that belong to the k-th cluster.

Unsupervised clustering algorithm (method I)

The model-based supervised clustering algorithm is developed based on the unsupervised clustering algorithm. Therefore, we first review the unsupervised method and then, in the next section, modify this algorithm to incorporate information from the training sample for the supervised algorithm.

The model-based unsupervised clustering algorithm assigns each gene to one of C clusters with a certain probability. Let us denote the probability that the i-th gene is assigned to the k-th cluster by p_{ik}. A gene will be assigned to the k-th cluster if p_{ik} is greater than a certain pre-determined value. This probability may be obtained via the EM algorithm (Dempster *et al*., 1977). With the EM algorithm, we can also estimate other model parameters, π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, for $k = 1, \ldots, C$. The number of clusters, C, can also be treated as an unknown parameter and inferred with the BIC or AIC test (Schwarz, 1978; Akaike, 1974). If a Bayesian approach is taken, C may be estimated via the reversible jump Markov chain Monte Carlo (Green, 1995).

The EM algorithm starts with some initial values of all unknowns and iteratively updates each parameter conditional on the parameter values in the previous round of the iteration. Without any prior knowledge, each gene may be assigned an equal probability to each cluster. The EM iteration is described in the following steps:

(0) Initialize prior probabilities of cluster assignment and the mixing proportions,

$$p_{ik}^{(0)} = 1/C \quad \forall i = 1, \ldots, n; \;\; k = 1, \ldots, C \quad \text{and}$$
$$\pi_k^{(0)} = 1/C \quad \forall k = 1, \ldots, C. \tag{3}$$

(1) Update the mean vectors,

$$\boldsymbol{\mu}_k^{(t)} = \left[n\pi_k^{(t-1)}\right]^{-1} \sum_{i=1}^{n} p_{ik}^{(t-1)} y_i \quad \forall k = 1, \ldots, C. \tag{4}$$

(2) Update the variance–covariance matrices,

$$\boldsymbol{\Sigma}_k^{(t)} = \left[n\pi_k^{(t-1)}\right]^{-1} \sum_{i=1}^{n} p_{ik}^{(t-1)} [\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(t)}][\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(t)}]^{\mathrm{T}} \quad \forall k = 1, \ldots, C. \tag{5}$$

(3) Update the posterior probabilities of cluster assignment,

$$p_{ik}^{(t)} = \frac{\pi_k^{(t-1)} f_k[\boldsymbol{y}_i | \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}]}{\sum_{k'=1}^{C} \pi_{k'}^{(t-1)} f_{k'}[\boldsymbol{y}_i | \boldsymbol{\mu}_{k'}^{(t)}, \boldsymbol{\Sigma}_{k'}^{(t)}]} \quad \forall i = 1, \ldots, n; \;\; k = 1, \ldots, C. \tag{6}$$

(4) Update the cluster proportions,

$$\pi_k^{(t)} = n^{-1} \sum_{i=1}^{n} p_{ik}^{(t)} \quad \forall k = 1, \ldots, C. \tag{7}$$

(5) Repeat (1)–(4) until convergence.

This EM iteration scheme is robust and well behaved. The convergence speed is also reasonably fast.

The number of clusters may be treated as another parameter and inferred from the data. The BIC or AIC test is used to estimate the optimal number of clusters (Akaike, 1974; Schwarz, 1978). The BIC is

$$\mathrm{BIC} = 2 \ln\ L(\hat{\Phi}) - p \ln(n), \tag{8}$$

where p is the number of parameters to be estimated in the model, $L(\hat{\Phi})$ is the likelihood value evaluated at $\hat{\Phi}$, the vector of maximum likelihood estimate of the parameters and n is the size of the dataset. The number of clusters (C) that has the maximum BIC value is the estimated C. Note that C is fixed in the supervised analysis because it is determined by the number of functional groups in the training sample.

Supervised clustering algorithm (method II)

In the supervised cluster analysis, we know the functions of genes in the training sample, and thus know which gene belongs to which cluster. Let n_1 and n_2 be the number of genes in the training sample and the test dataset, respectively, and the total number of genes in the microarray experiment be $n_1 + n_2 = n$. In addition, we know the number of genes that belong to each of the k clusters in the training sample, denoted by n_{1k}, for $\sum_{k=1}^{C} n_{1k} = n_1$. The most intuitive method for the model-based supervised cluster analysis is to use the training sample to estimate the cluster means and variance–covariance matrices, denoted by $\hat{\boldsymbol{\mu}}_k$ and $\hat{\boldsymbol{\Sigma}}_k$ for all $k = 1, \ldots, C$. For each of the n_2 genes in the test dataset, indexed from $n_1 + 1$ to $n_1 + n_2$, we need to calculate the posterior probability that the i-th gene belongs to the k-th cluster. The EM iteration is described below:

(0) Initializing the prior probabilities of cluster assignment,

$$p_{ik}^{(0)} = 1/C \quad \forall i = n_1 + 1, \ldots, n; \ \ k = 1, \ldots, C.$$

(1) Updating the cluster proportions,

$$\pi_k^{(t)} = n^{-1}\Big[n_{1k} + \sum_{i=n_1+1}^{n} p_{ik}^{(t-1)}\Big] \quad \forall k = 1, \ldots, C.$$

(2) Updating the posterior probabilities of cluster assignment,

$$p_{ik}^{(t)} = \frac{\pi_k^{(t-1)} f_k(\boldsymbol{y}_i | \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)}{\sum_{k'=1}^{C} \pi_{k'}^{(t-1)} f_{k'}(\boldsymbol{y}_i | \hat{\boldsymbol{\mu}}_{k'}, \hat{\boldsymbol{\Sigma}}_{k'})}$$
$$\forall i = n_1 + 1, \ldots, n; \ \ k = 1, \ldots, C.$$

(3) Repeating steps (1) and (2) until convergence.

Note that $\hat{\boldsymbol{\mu}}_k$ and $\hat{\boldsymbol{\Sigma}}_k$ for all $k = 1, \ldots, C$ are estimated from the training sample and they are not updated in the iteration process. This intuitive method is similar to Fisher's discriminate analysis (Fisher, 1936), except that this method can handle more than two clusters.

Supervised clustering algorithm (method III)

The simple and intuitive supervised method given above usually performs well if the number of genes within each known cluster in the training sample is sufficiently large. The large sample requirement is to ensure high accuracy of the estimates of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$. However, for small training samples, these estimates are subject to large errors. Sometimes, the estimated variance–covariance matrices may not even be positive definite. This may happen if the number of genes within a cluster is smaller than the number of treatments (m). Furthermore, much information from the test dataset has not been fully utilized. The test dataset is usually much larger than the training sample, implying that the unutilized information may be substantially more than that contained in the training sample. It is our intention to incorporate this additional information into the algorithm, leading to the new supervised algorithm (method III). The EM iterations are performed based on the following steps:

(0) Initialize prior probabilities of cluster assignment and the mixing proportions,

$$p_{ik}^{(0)} = 1/C \quad \forall i = n_1 + 1, \ldots, n; \ \ k = 1, \ldots, C$$

and

$$\pi_k^{(0)} = 1/C \quad \forall k = 1, \ldots, C.$$

(1) Update the mean vectors,

$$\boldsymbol{\mu}_k^{(t)} = \Big[n_{1k} + \sum_{i=n_1+1}^{n} p_{ik}^{(t-1)}\Big]^{-1} \times \Big[n_{1k}\hat{\boldsymbol{\mu}}_k + \sum_{i=n_1+1}^{n} p_{ik}^{(t-1)} y_i\Big] \quad \forall k = 1, \ldots, C.$$

(2) Update the variance–covariance matrices,

$$\boldsymbol{\Sigma}_k^{(t)} = \left[n_{1k} + \sum_{i=n_1+1}^{n} p_{ik}^{(t-1)}\right]^{-1} \times \left[n_{1k}\hat{\boldsymbol{\Sigma}}_k + \sum_{i=n_1+1}^{n} p_{ik}^{(t-1)} [\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(t)}][\boldsymbol{y}_i - \boldsymbol{\mu}_k^{(t)}]^{\mathrm{T}}\right]$$
$$\forall k = 1, \ldots, C.$$

(3) Update the posterior probabilities of cluster assignment,

$$p_{ik}^{(t)} = \frac{\pi_k^{(t-1)} f_k[\boldsymbol{y}_i | \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}]}{\sum_{k'=1}^{C} \pi_{k'}^{(t-1)} f_{k'}[\boldsymbol{y}_i | \boldsymbol{\mu}_{k'}^{(t)}, \boldsymbol{\Sigma}_{k'}^{(t)}]}$$
$$\forall i = n_1 + 1, \ldots, n; \ \ k = 1, \ldots, C.$$

(4) Update the cluster proportions,

$$\pi_k^{(t)} = n^{-1}\left[n_{1k} + \sum_{i=n_1+1}^{n} p_{ik}^{(t)}\right] \quad \forall k = 1, \ldots, C.$$

(5) Repeat (1)–(4) until convergence.

Note that the gain in efficiency of the new supervised method over the simple supervised method comes from the more precise estimates of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$. Both parameter sets are functions of $\hat{\boldsymbol{\mu}}_k$ and $\hat{\boldsymbol{\Sigma}}_k$ obtained from the training sample and data $\boldsymbol{y}_i, \forall i > n_1$, contained in the test dataset. In other words, the sample size for estimating $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ has been increased from n_{1k} in the simple supervised method to $n_{1k} + \hat{n}_{2k}$ in the new supervised method, where $\hat{n}_{2k} = \sum_{i=n_1+1}^{n} p_{ik}^t$ is the estimated number of genes in the test dataset that belong to cluster k.

IMPLEMENTATION

Yeast cell cycle data

The yeast cell cycle data were published by Cho *et al*. (1998). The data contained the expression profiles of 6220 genes over 17 time points (treatments) taken at 10-min intervals, covering nearly two cell cycles. This set of data has been analyzed by many authors, e.g. Lukashin and Fuchs (2001), Yeung *et al*. (2001) and Tamayo *et al*. (1999). The entire dataset (raw data) is available at http://cellcycle-www.stanford.edu. In the study by Yeung *et al*. (2001), a subset of 384 genes was used ($n = 384$). These genes had expression levels peaking at different times corresponding to the five ($C = 5$) phases of the cell cycle (Fig. 1). This subset of the data is available at http://www.cs.washington.edu/homes/kayee/model. For preprocessing, we removed the data corresponding to the 90- and 100-min time points, because these two time points were reported to be unreliable (Tavazoie *et al*., 1999). After the deletion, the total number of treatments became 15 ($m = 15$). We then standardized each gene expression profile by subtracting the mean expression from the original value and dividing the difference by the SD so that the transformed expression level has mean 0 and variance 1. All the 384 genes were assigned to one of the five clusters by the original investigators (Cho *et al*., 1998; Yeung *et al*., 2001). Therefore, we can use this dataset to test the performances of the clustering algorithms developed here and compare them with the performance of existing methods.

Four methods were compared using the same dataset: the unsupervised method (method I), the simple supervised method (method II), the new supervised method (method III) and the SVMs algorithm (method IV). Methods I and IV were previously developed by other authors (Yeung *et al*., 2001; Brown *et al*., 2000). Methods II and III were developed in this study. Methods II–IV are all supervised clustering algorithms. The SVM algorithm (method IV) is the only other supervised clustering method used for comparison. There are other supervised clustering methods, e.g. linear, quadratic, mixture and the functional discrimination analyses (Hastie and Tibshirani, 1996; James and Hastie, 2001). These discrimination analysis procedures share similar features with our Gaussian mixture model. Therefore, we only compare our methods with the SVM, a heuristic approach. We adopted the commonly used three-fold cross-validation experiments to test the three supervised methods (Brown *et al*., 2000), where we randomly divided the total number of genes into three groups and used genes from two groups as the training sample and the genes from the third group as the test data. There were three possible ways to combine two out of three groups. A detailed description of the method can be found in Brown *et al*. (2000). Overall, we did three separate analyses to complete one cross-validation experiment for each method. For the unsupervised method (method I), we analyzed all the 384 genes simultaneously without regrouping the genes.

Table 1. Comparison of the clustering results of various classification methods on the yeast cell cycle microarray data

Cell division phase	Methods	FP	FN	TP	TN
Early G_1	I	50	12	55	267
(67 genes)	II	21	17	50	296
	III	21	21	46	296
	IV	38	10	57	279
Late G_1	I	28	40	95	221
(135 genes)	II	22	38	97	227
	III	24	35	100	225
	IV	43	10	125	206
S	I	33	49	26	276
(75 genes)	II	41	28	47	268
	III	37	36	39	272
	IV	72	18	57	237
G_2	I	28	41	11	304
(52 genes)	II	6	38	14	326
	III	18	29	23	314
	IV	46	5	47	286
M	I	38	42	13	291
(55 genes)	II	9	28	29	320
	III	19	8	47	310
	IV	47	2	53	282

After the test, each gene had four possible outcomes: false positive (FP), false negative (FN), true positive and true negative. The total error rate was defined as FP + FN. Table 1 summarizes the results of the three-fold cross-validation experiments for each of the five clusters. For methods I–III, we assigned a gene to a cluster if the probability of the gene belonging to that cluster exceeded 0.8. For method IV, a gene was assigned to a cluster depending on whether it was a member of the cluster or not. Table 2 summarizes the total error rate of the four algorithms for all the methods. We can see that methods II and III, the two model-based training methods developed in this study, have similar performance in terms of

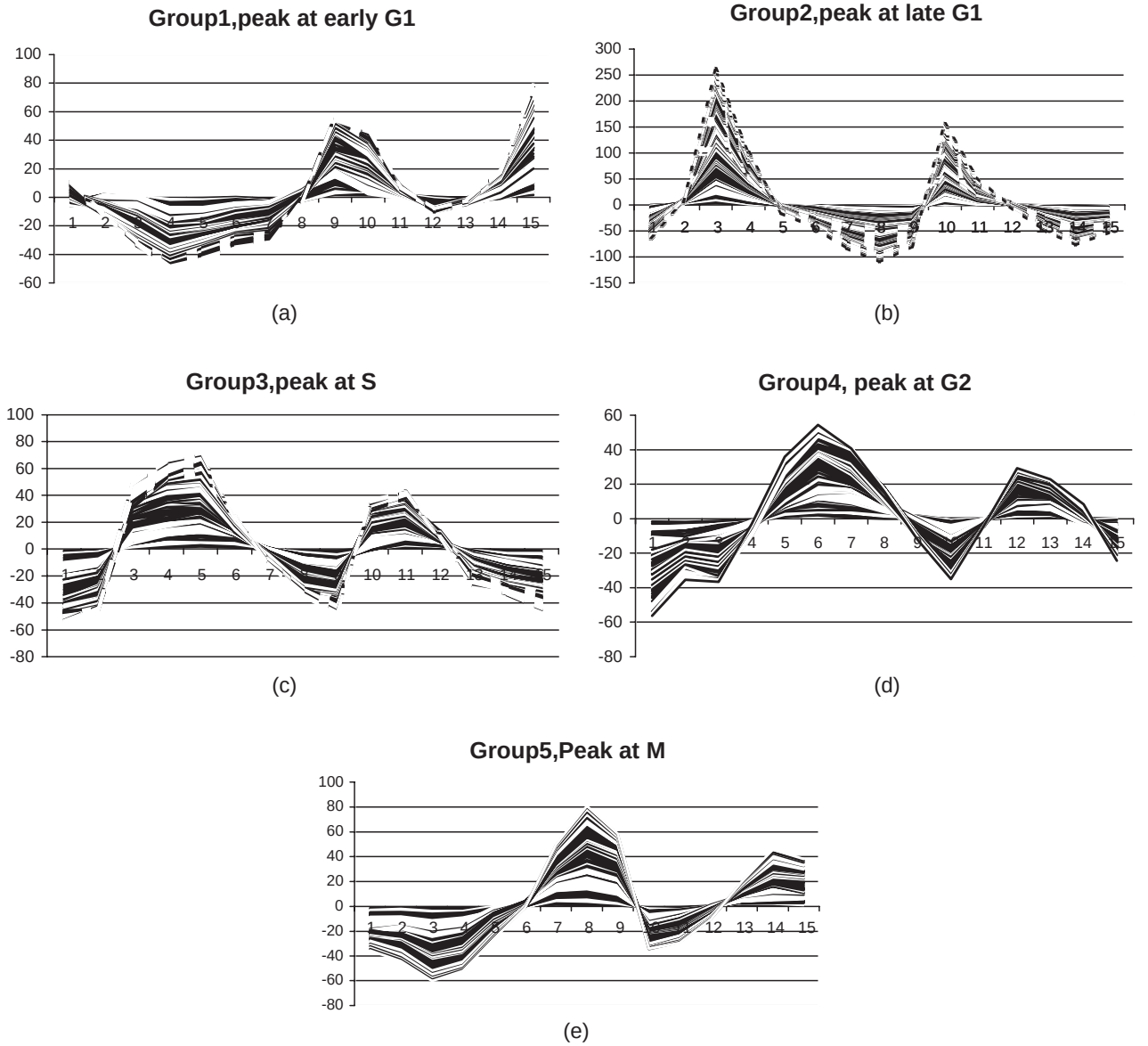

Fig. 1. The five groups of genes whose expression levels peak at different phases of the cell cycle (a–e). The numbers 1–15 in the x-axis correspond to the 15 data points for each gene. The expression profile was measured for 160 min at 10-min intervals, while the data points corresponding to 90 and 100 min were deleted because they were unreliable. These 15 data points cover almost two cell cycles.

Table 2. Comparison of the overall error rates of four clustering algorithms on the yeast cell cycle microarray data

Methods	FP	FN	FP+FN
I	177	184	361
II	99	149	248
III	119	129	248
IV	246	45	291

the total number of errors, FP + FN. The unsupervised method (method I) produced significantly larger number of both FP and FN due to the poor recognition of the last two groups. This demonstrates the advantage of supervised clustering over unsupervised clustering based on the Gaussian mixture model.

The yeast cell cycle data have been analyzed by many authors, but all took the unsupervised approach. Tamayo *et al*. (1999) analyzed 828 genes with the SOM and found 30 clusters. Lukashin and Fuchs (2001) identified five clusters among the 1306 genes that passed the variation filter with the simulated annealing algorithm. Because these authors did not

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 6 \\ 5 \\ 5.5 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1.5 & 0.3 \\ 0.5 & 0.3 & 1 \end{bmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} -4 \\ 4 \\ 4.5 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.5 & 0.8 & 1.3 \\ 0.8 & 2 & 0 \\ 1.3 & 0 & 3 \end{bmatrix}$$

$$\boldsymbol{\mu}_3 = \begin{bmatrix} -5 \\ -6 \\ -3 \end{bmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{bmatrix} 2 & 1 & 0.8 \\ 1 & 1.5 & 1.3 \\ 0.8 & 1.3 & 2 \end{bmatrix}$$

$$\boldsymbol{\mu}_4 = \begin{bmatrix} 2.5 \\ -7 \\ -1 \end{bmatrix}, \quad \boldsymbol{\Sigma}_4 = \begin{bmatrix} 2 & 0.1 & 1 \\ 0.1 & 3 & 1.2 \\ 1 & 1.2 & 1.5 \end{bmatrix}$$

$$\boldsymbol{\mu}_5 = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_5 = \begin{bmatrix} 10 & 4.0 & 3.0 \\ 4.0 & 8.0 & 5.0 \\ 3.0 & 5.0 & 9.0 \end{bmatrix}$$

Fig. 2. The mean vectors and variance–covariance matrices for the five clusters of the simulated data.

use exactly the same set of genes as those analyzed in this study, and their purposes were to identify the optimal number of clusters (different from ours), their findings are not comparable with our results. Yeung *et al*. (2001) analyzed the same set of genes as ours (384 genes) with the unsupervised model-based algorithm (method I), but their purpose was, again, to identify the optimal number of clusters, and thus did not provide the error rates. However, the results of method I given in Table 1 of this study should be identical to the error rates of the analysis by Yeung *et al*. (2001), if such information were provided.

Simulated data

In the yeast cell cycle data analysis, the FP + FN error rates may reflect the confounding errors of the methods and the human-made clusters. Therefore, we conducted a simulation experiment in which the clusters are known exactly without human-made errors. In this case, the FP + FN error rates reflect the true errors due to different methods. We tested our algorithms on different simulated multivariate normal mixture distributions. We found that, most of the times, the unsupervised method performed equally well as the supervised methods (data not shown), but it usually showed a significant reduction in the ability to recognize small groups from large overlapping background. We then simulated a dataset with 2400 observations from five (clusters) multivariate normal distributions ($C = 5$). The first four clusters each had 100 observations with smaller variances, while the fifth cluster had 2000 observations with larger variances. The mean vectors and variance–covariance matrices of the five clusters are given in Figure 2. The mean vectors of the four small groups are close to the center of the large group.

Table 3. Comparison of the clustering results of various classification methods on the simulated data

Cluster	Methods	FP	FN	TP	TN
1 (100 genes)	I	1	100	0	2299
	II	6	30	70	2294
	III	6	34	66	2294
	IV	831	0	100	1469
2 (100 genes)	I	56	0	100	2244
	II	8	22	78	2292
	III	10	19	81	2290
	IV	557	0	100	1743
3 (100 genes)	I	23	99	1	2277
	II	3	46	54	2297
	III	3	45	55	2297
	IV	321	1	99	1979
4 (100 genes)	I	6	16	84	2294
	II	5	21	79	2295
	III	5	18	82	2295
	IV	152	2	98	2148

Table 4. Comparison of the overall error rates of four algorithms on the simulated data

Methods	FP	FN	FP+FN
I	86	215	301
II	22	119	141
III	24	116	140
IV	1861	3	1864

We analyzed the simulated dataset with the four methods as we did in the yeast cell cycle data analysis. This time, we performed the two-fold cross-validation experiments. We randomly divided the data into two groups. The classifiers were trained with one group and were tested with the other group. For all the model-based methods (methods I–III), we assigned an observation to a cluster if the probability of the observation belonging to that cluster exceeded 0.8. Because cluster five represented a widely distributed background, it was excluded from the analysis for error calculation. The results are summarized in Table 3 for the test values for individual clusters, and in Table 4 for the consensus test values of all the clusters. In terms of the small error rates, FP, FN or FP + FN, the performances of methods II and III were superior over the other two methods, indicating that both supervised clustering methods performed well above the other methods in the situation where a Gaussian mixture model applies. The unsupervised method (method I) performed well in clusters

Table 5. Comparison of the clustering results of methods II and III with different training sample sizes on the simulated data

Training sample size	Methods	Cluster (no. of genes)	FP	FN	TP	TN	Total FP	Total FN	Total FP+FN
10%	II	1 (90)	3	45	45	2067	11	178	189
(240 genes)		2 (89)	3	61	28	2068			
		3 (92)	2	55	37	2066			
		4 (92)	3	17	75	2065			
	III	1 (90)	5	30	60	2065	27	106	133
		2 (89)	14	16	73	2057			
		3 (92)	3	43	49	2065			
		4 (92)	5	17	75	2063			
15%	II	1 (83)	5	27	56	1952	18	151	169
(360 genes)		2 (88)	8	49	39	1944			
		3 (85)	4	48	37	1951			
		4 (83)	1	30	53	1956			
	III	1 (83)	5	29	54	1952	24	105	129
		2 (88)	12	18	70	1940			
		3 (85)	3	40	45	1952			
		4 (83)	4	18	65	1953			

two and four, but poorly in clusters one, three and five. This was due to the fact that many of the data simulated from clusters one, three and five were assigned back to these three clusters but with a high chance of incorrect classification. With our criterion of assigning an observation to a cluster (0.8), these genes could not be assigned to any of the clusters. For the first four clusters, the SVM method had a high power for identifying true positives, but at the cost of a high false positive rate.

When the size of the training sample was large, say one-half or two-third of the whole set of the data, the two supervised methods (methods II and III) proposed in this study performed equally well, although both were better than the unsupervised method and SVM. To compare the performance of the two methods with a smaller training sample, we randomly selected 10 or 15% observations (240 or 360 observations) from the simulated data (2400 observations), and used the subset as the training sample to classify the rest of the data. The classification results of the two methods are summarized in Table 5. From this table, we can see that in terms of the small total error rate, FP + FN, method III was better than method II. The difference between the two methods came from the way the parameters were estimated. Method II used the training sample to estimate the parameters, whereas method III used both the training sample and the test dataset. Therefore, the additional information from the test dataset indeed improved the performance of the clustering. In fact, the simulation experiment was replicated several times. The results are all consistent with the one reported (data not shown).

The time complexity of the algorithms is roughly linear on the number of genes, but not linear on the levels of the treatment and the number of clusters. For the simulated data, the slowest algorithm (method I) took about 40 min to converge on a Pentium IV PC.

DISCUSSION

We developed two algorithms for supervised model-based clustering analysis of microarray data. These two algorithms were tested and compared with existing methods using both real and simulated data. We found that the supervised methods (methods II and III), were superior over the other two methods evaluated for both datasets. When the size of the training sample was small (10 or 15% of the whole dataset), the simple supervised model-based method (method II), which uses only the training sample to estimate the parameters and then uses the estimated value of the parameters to classify the test dataset, performed not as well as method III, which utilizes both the training and test datasets into a single EM algorithm to simultaneously estimate parameters and perform classification. This advantage has been demonstrated in our classification results. The unsupervised method (I) produced similar classification results to the supervised methods (II and III) in many datasets we simulated (data not shown), but we found that it had some problems in identifying small groups from big overlapping background, as shown in Tables 3 and 4. The SVM method has been proved to be a successful knowledge-based 'heuristic' clustering method (Brown *et al*., 2000; Mateos *et al*., 2002). Here, we used SVM to classify the data and compared the results with our model-based methods. We found that the proposed new method (III) performed better than the SVM on both datasets tested. The SVM was likely to produce a larger number of FP but a smaller number of FN than the new EM method. One possible reason may be that the SVM identifies members in a one-versus-rest fashion. This binary

SVM was likely to include more members in a cluster than it should, i.e. produce a large number of FP, and thus it may not be optimal for the multiclass problem. An extension has been made by Lee and Lee (2003) as multicategory support vector machines (MSVM). Further investigation may be necessary to compare the performance of the model-based algorithm with the MSVM. The binary SVM (Brown *et al*., 2000), however, has a user friendly web-based program available for general use. Therefore, we compared our EM algorithm only with the binary SVM. The SVM is an approach completely different from the model-based method. Normally, a model-based method is always better than a heuristic approach, except that the former is always more time consuming. Therefore, we do not expect the MSVM method to be better than our model-based method. The model-based algorithm is not heuristic and will guarantee finding the optimal clusters if the sample size is sufficiently large. This property is called consistency in statistics. However, the property of consistency may not be shared by all heuristic approaches. The model-based method depends on a probability model. The probability model itself is usually proposed based on experience and the feasibility of the model. The Gaussian mixture model is a convenient choice and also quite robust. Furthermore, the model-based method provides substantial information than the heuristic method, e.g. the posterior probabilities of gene classifications and the parameters of each cluster.

In this study, we assigned a gene to a cluster if the posterior probability was greater than 0.8. This criterion was chosen in an arbitrarily manner and it may affect the results of classification. We also tried the criteria of 0.5 and 0.9 with method III. We found that for the yeast cell cycle data, criterion 0.5 produced a slightly larger FP (125) but a slightly smaller FN (124), while criterion 0.9 produced a slightly smaller FP (113) but a slightly larger FN (135). The total error rates FP + FN were almost the same for all these three criteria. For the simulated data, criterion 0.5 produced a larger FP (64) and a smaller FN (44). The total error rate reduced to 108. Criterion 0.9 produced a smaller FP (2) and a larger FN (214). The total error rate increased to 216. We think that the 0.8 criterion is the reasonable choice.

One possible problem with the model-based method is that the correct number of clusters is needed for a good classification. We tested the supervised EM algorithm with the same dataset used by Brown *et al*. (2000). We divided the 2467 genes into six clusters based on their functions: TCA cycle, respiration, ribosome, protease, histones and others. Although we used two-third of the data as training sample, the method still could not classify the rest of the data correctly (data not shown). This was because the sixth cluster itself may be a mixture of many distributions. We actually treated it as a single distribution in our model. This indicates that a good estimation of the number of the clusters is very important for classification. As we mentioned earlier, an advantage of the model-based model is that we can use statistical criteria, such as AIC and BIC, to find the number of clusters that fits the data best. So far, finding the number of clusters using AIC or BIC only applies to unsupervised clustering analysis because, in the supervised clustering, the training sample already contains the fixed number of groups of genes and, thus, it is not necessary to find the number of clusters. It will be very informative to extend our method to combine the unsupervised and the supervised clustering method into a single analysis if the training sample does not contain all functional groups that exist in the test dataset. This requires redefining the number of clusters as $C_1 + C_2$, where C_1 is the number of clusters in the training sample and C_2 is the number of additional clusters contained in the test dataset. The BIC or AIC will play a role in determining the optimal C_2. The reversible jump MCMC which was originally developed for inferring the number of distributions in a mixture (Green, 1995; Richardson and Green, 1997) is an ideal tool for estimating C_2 if the problem is tackled from a Bayesian perspective.

Finally, data normalization is the prerequisite of microarray data analysis. It serves as a tool to remove systematic environmental effects so that comparisons are made on an equal basis. Therefore, normalization is a tool to centralize the data. In addition, normalization provides a way to reshape the distribution of the expression data. The raw data collected by the experimenters are rarely distributed in a normal fashion. Log or other form of data transformation is important to force the data to follow a normal distribution. The model-based method developed here depends on the normal assumption. Therefore, the method may be more sensitive to any departure from normality. Further investigations are necessary to address this problem.

ACKNOWLEDGEMENTS

The research was supported by the National Institute of Health Grant R01-GM55321 and the USDA National Research Initiative Competitive Grants Program 00-35300-9245 to S.X.

REFERENCES

Akaike,H. (1974) A new look at Statistical Model Identification. *IEEE Trans. Autom. Control*, **19**, 716–723.

Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci., USA*, **97**, 262–267.

Carr,D.B., Somogyi,R. and Michaels,G. (1997) Templates for looking at the gene expression clustering *Stat. Comput. Stat. Graph. Newslett.*, **8**, 20–29.

Cho,R.J., Campbell,M.J, Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. and Davis,R.W.A (1998) genome-wide transcriptional analysis of the mitotic cell cycle *Mol. Cell*, **2**, 65–78.

Dasgupta,A. and Raftery,A.E. (1998) Detecting features in spatial point processes with clutter via model-based clustering. *J. Am. Stat. Assoc.*, **93**, 294–302.

Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion) *J. R. Stat. Soc. B*, **39**, 1–38.

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., USA*, **95**, 14863–14868.

Fisher,R.A. (1936) Use of multiple measurements in taxonomic problem. *Ann. Eug.*, **7**, 179–184.

Fraley,C. and Raftery,A.E. (1998) How many clusters? Which clustering method?—answers via model-based cluster analysis. *Comput. J.*, **41**, 578–588.

Ghosh,D. and Chinnaiyan,A.M. (2002) Mixture modeling of gene expression data from microarray experiments. *Bioinformatics*, **18**, 275–286.

Green,P. (1995) Reversible jump Markov chan Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hastie,T. and Tibshirani,R. (1996) Discriminant analysis by Gaussian mixture. *J. R. Stat. Soc. B*, **58**, 267–288.

Herrero,J., Valencia,A. and Dopazo,J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.

Hughes,T.R., Marton,M.J., Jones,A.R., Roland Stoughton,C.J.R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D., Kidd,M.J. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

James,G. and Hastie,T. (2001) Functional linear discriminant analysis for irregularly sampled curves. *J. R. Stat. Soc. B*, **63**, 533–550.

Lee,Y. and Lee,C. (2003) Classification of multiple cancer types by multcategory support vector machines using gene expression data. *Bioinformatics*, **19**, 1132–1139.

Lukashin,A.V. and Fuchs,R. (2001) Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, **17**, 405–414.

McLachlan,G.J., Bean,R.W. and Peel,D. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

Mateos,A., Dopazo,J., Jansen,R., Tu,Y., Gerstein,M. and Stolovizky,G. (2002) Systemic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genom. Res.*, **12**, 1703–1715.

Richardson,S. and Green,P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components. *J. R. Stat. Soc. B*, **59**, 731–792.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 2907–2912.

Szabo,A., Boucher,K., Carroll,W.L., Klebanov,L.B., Tsodikov,A.D. and Yakovlev,A.Y. (2002) Variable selection and pattern recognition with gene expression data generated by the microarray technology. *Math. Biosci.*, **176**, 71–98.

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., USA*, **96**, 2907–2912.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 218–285.

Yeung,K.Y., Fraley,C., Murua,A., Raftery,A.E. and Ruzzo,W.L. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.