



## Class discovery and classification of tumor samples using mixture modeling of gene expression data—a unified approach

Roxana Alexandridis, Shili Lin\* and Mark Irwin

Department of Statistics, Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA

Received on September 11, 2003; revised on March 4, 2004; accepted on April 19, 2004

Advance Access publication April 29, 2004

### ABSTRACT

**Motivation:** The DNA microarray technology has been increasingly used in cancer research. In the literature, discovery of putative classes and classification to known classes based on gene expression data have been largely treated as separate problems. This paper offers a unified approach to class discovery and classification, which we believe is more appropriate, and has greater applicability, in practical situations.

**Results:** We model the gene expression profile of a tumor sample as from a finite mixture distribution, with each component characterizing the gene expression levels in a class. The proposed method was applied to a leukemia dataset, and good results are obtained. With appropriate choices of genes and preprocessing method, the number of leukemia types and subtypes is correctly inferred, and all the tumor samples are correctly classified into their respective type/subtype. Further evaluation of the method was carried out on other variants of the leukemia data and a colon dataset.

**Contact:** shili@stat.ohio-state.edu

**Supplementary information:** The program implementing the method and additional details and figures are at <http://www.stat.ohio-state.edu/~statgen/PAPERS/DNC-MIX.html>.

### INTRODUCTION

Accurate classification of tumor samples is an essential tool for efficient cancer treatment. For many cancers, such as acute adult leukemias or non-Hodgkin's lymphomas, different subtypes show very different responses to therapy, although they have very similar morphological and histopathological appearance, reflecting the fact that they are molecularly distinct entities (Golub *et al.*, 1999). The DNA microarray technology has been increasingly used in cancer research, which enables classification of tissue samples based only on gene expression data, without prior and often subjective biological knowledge (Golub *et al.*, 1999; Dudoit *et al.*, 2002).

A considerable amount of research involving microarray data analysis is focused on the discovery of putative types and subtypes of cancers using gene expression profiles of disease samples. Unsupervised learning approaches, techniques commonly used for this problem, have the advantage of being impartial to currently accepted classes, but they may reveal a structure that is not biologically significant. Most of the recent publications on this issue utilize cluster analysis techniques to group tumor samples and/or genes, using techniques such as self-organizing maps (SOMs) (e.g. Golub *et al.*, 1999) and hierarchical clustering (e.g. Alon *et al.*, 1999).

In addition to class discovery, an equally important problem is to classify test samples into known classes, with the help of a training set containing samples whose classes are known. Numerous approaches based on gene expression data have been proposed for classifying test samples into known classes, without allowing them to belong to new classes. Some of these are applicable only to binary classification, such as the weighted voting scheme of Golub *et al.* (1999), whereas others can handle multitype classification problems. These approaches range from traditional methods, such as Fisher's linear and quadratic discriminant analysis, to more modern machine learning techniques, such as classification trees or aggregation of classifiers by bagging or boosting (for a review see Dudoit *et al.*, 2002). There are also approaches, which are able to identify test samples that do not belong to any of the known classes by imposing thresholds on the prediction strength (e.g. Golub *et al.*, 1999; Lee and Lee, 2002). However, they were not able to place these samples into new putative classes.

This paper proposes a unified approach to class discovery, classification into known classes, and the joint analysis of classification and class discovery. The method proposed is an extension of Lin and Alexandridis (2003), and is based on modeling the distribution of the gene expression profile of a test sample as a finite mixture of an unknown number of distributions, with each mixture component characterizing the gene expression levels within a class. The distributional

\*To whom correspondence should be addressed.

assumptions made here are the same as those in diagonal quadratic discriminant analysis (Dudoit *et al.*, 2002), but both the training samples (if they exist) and the test samples are used to estimate the parameters of the model in our formulation. We applied the method proposed to the leukemia data of Golub *et al.* (1999) and a number of resampled datasets based on it. Further evaluation of the method was carried out on the colon cancer data of Alon *et al.* (1999). We use several measures for gene selection, and we explore the sensitivity of the class discovery and class prediction results on the number of genes in a classifier.

## METHODS

### Mixture modeling of test samples

Let  $K$  be the number of known classes, which is zero in the absence of training samples. Let  $\mathbf{y}_{ki} = (y_{1ki}, \dots, y_{Gki})'$  denote the  $i$ -th training sample from class  $k$ ,  $k = 1, \dots, K$ ,  $i = 1, \dots, N_k$ . The length of the vector,  $G$ , is the number of genes used for class discovery and classification, and is referred to as the classifier size. Hence, the class labels of all training samples are known. Furthermore, we use  $\mathbf{x}_i = (x_{1i}, \dots, x_{Gi})'$ ,  $i = 1, \dots, T$ , to denote the  $i$ -th test sample, and we assume that the test samples can come from the  $K$  known classes as well as from  $U$  putative classes. However, there may not be any test samples belonging to some (or all) of the known classes. Consequently, the distribution of the test samples is modeled as a mixture of distributions of the  $M = K + U$  components as follows:

$$f(\mathbf{x}_i | \boldsymbol{\psi}_M) = \sum_{m=1}^M \pi_m f_m(\mathbf{x}_i | \boldsymbol{\mu}_m, \boldsymbol{\sigma}_m^2), \quad i = 1, \dots, T,$$

where  $f_m$  is the probability density function of the  $m$ -th component of the mixture and is assumed to be normal. The parameter set  $\boldsymbol{\psi}_M = (\pi_1, \dots, \pi_M, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_M, \boldsymbol{\sigma}_1^2, \dots, \boldsymbol{\sigma}_M^2)$  then contains the mixture coefficients  $\pi_m$  ( $\sum_{m=1}^M \pi_m = 1$ ), the mean vectors  $\boldsymbol{\mu}_m$ , and the vectors  $\boldsymbol{\sigma}_m^2$  of the parameters of the variance–covariance matrices,  $m = 1, \dots, M$ . Note that  $M \geq \max\{1, K\}$  such that, if training samples do not exist, there is still at least one putative class for the test samples. We further assume a diagonal variance–covariance matrix for each component density, therefore,  $\boldsymbol{\sigma}_m^2$  is a vector of the variances on the diagonal.

### EM estimation of parameters

The maximum-likelihood estimates (MLEs) of the parameters,  $\hat{\boldsymbol{\psi}}_M$ , are obtained using the Expectation Maximization (EM) algorithm (McLachlan and Peel, 2000). Let  $Z_i$  denote the unknown class label, the missing component, of the  $i$ -th test sample, which takes values in the set  $\{1, \dots, M\}$ . Thus, the complete data are  $\{(\mathbf{x}_i, Z_i), i = 1, \dots, T\} \cup \{(\mathbf{y}_{ki}, k), k = 1, \dots, K, i = 1, \dots, N_k\}$ , and the corresponding complete

data likelihood is

$$L_c(\boldsymbol{\psi}_M) = \left\{ \prod_{k=1}^K \prod_{i=1}^{N_k} f(\mathbf{y}_{ki} | \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2) \right\} \times \left\{ \prod_{i=1}^T \prod_{m=1}^M [\pi_m f(\mathbf{x}_i | \boldsymbol{\mu}_m, \boldsymbol{\sigma}_m^2)]^{I(Z_i=m)} \right\},$$

where  $I(Z_i = m)$  is an indicator function taking the value of 1 if  $Z_i = m$ . Note that this likelihood is conditioned on the class labels of the training samples. The  $(t + 1)$ -th EM iteration of the estimates, for  $m = 1, \dots, M$ , are

$$\begin{aligned} \pi_m^{(t+1)} &= \frac{\sum_{i=1}^T \tau_m^{(t)}(\mathbf{x}_i)}{T}, \\ \boldsymbol{\mu}_m^{(t+1)} &= \frac{\sum_{i=1}^{N_m} \mathbf{y}_{mi} + \sum_{i=1}^T \tau_m^{(t)}(\mathbf{x}_i) \mathbf{x}_i}{N_m + \sum_{i=1}^T \tau_m^{(t)}(\mathbf{x}_i)}, \\ \boldsymbol{\sigma}_m^{2(t+1)} &= \frac{\sum_{i=1}^{N_m} (\mathbf{y}_{mi} - \boldsymbol{\mu}_m^{(t)})^{\odot 2} + \sum_{i=1}^T \tau_m^{(t)}(\mathbf{x}_i - \boldsymbol{\mu}_m^{(t)})^{\odot 2}}{N_m + \sum_{i=1}^T \tau_m^{(t)}(\mathbf{x}_i)}, \end{aligned}$$

where  $\tau_m^{(t)}(\mathbf{x}_i)$  is the posterior probability of  $\mathbf{x}_i$  having class label  $m$  at iteration  $t$ , and  $\odot 2$  denotes element-wise squares. Furthermore, the first terms in both the numerator and the denominators of  $\boldsymbol{\mu}_m^{(t+1)}$  and  $\boldsymbol{\sigma}_m^{2(t+1)}$  are zero if there is no training sample in the class, i.e. when  $m > K$ .

### Starting values for EM

Since the EM algorithm is known to be sensitive to starting values of parameters, we explore several sets of starting points for each mixture model with  $M$  components, although we always use  $1/M$  as the starting values for the mixture weights. For the rest of the parameters, we consider two approaches for generating multiple sets of starting values. One approach is based on clustering algorithms, such as  $K$ -means (Hastie *et al.*, 2001) and PAM (Kaufman and Rousseeuw, 1990), to classify the entire dataset (training and test samples combined) into  $M$  clusters, and to use the sample means and sample variances of the resulting clusters as the starting values for the parameters of the corresponding classes.

An alternative approach is based on systematic generations of random starting values (McLachlan and Peel, 2000). For each known class, we use the vector of sample means and sample variances from the training samples in that class as starting values for the parameters of the corresponding component density. On the other hand, for each gene in each of the unknown classes, we randomly generate a number from the uniform distribution, defined on the range of the expression levels of the gene among the test samples, as a starting value for the mean. For the variance parameter, we use half of the sample variance of the expression levels of the gene as the starting value.

## Model selection with BIC/AIC

To determine the number of classes,  $M$ , which best explains the observed expression levels of the samples, we let  $M$  take integer values in the interval  $[\max\{1, K\}, M_{\max}]$ , where integer  $M_{\max}$  is chosen to be sufficiently large for each specific problem. Then for each value  $M$  and the corresponding mixture model, we obtain the MLEs ( $\hat{\psi}_M$ ) of the parameters (across all sets of starting points), and the corresponding log-likelihood,  $\log[L(\hat{\psi}_M)]$ . We explore two model selection criteria (Hastie *et al.*, 2001), the Bayesian Information Criterion (BIC), and the Akaike Information Criterion (AIC), for determining the best  $M$ . Specifically, for selection based on BIC, we calculate, for  $M = \max\{1, K\}, \dots, M_{\max}$ ,

$$\text{BIC}(M) = -2 \log[L(\hat{\psi}_M)] + \log(n)p(M),$$

where  $p(M)$  is the number of parameters in the mixture model with  $M$  components, and  $n = \sum_{k=1}^K N_k + T$  is the total number of samples. Then

$$\hat{M} = \operatorname{argmin}_M \{\text{BIC}(M), \max\{1, K\} \leq M \leq M_{\max}\},$$

is our estimate of the number of classes. Selection based on AIC can be carried out similarly, with the AIC defined as

$$\text{AIC}(M) = -2 \log[L(\hat{\psi}_M)] + 2p(M).$$

## A classification procedure

Once a model with  $\hat{M}$  components is selected, we use it to predict the class membership of each test sample. For each sample, the class label that achieves the maximum posterior probability, given  $\hat{M}$  and the corresponding MLEs  $\hat{\psi}_{\hat{M}}$ , is taken to be its predicted class, corresponding to the Bayes rule with 0–1 loss.

## APPLICATION TO THE LEUKEMIA DATA

### Datasets, preprocessing, gene selection and the set-up

The leukemia data of Golub *et al.* (1999) are analyzed in several ways to evaluate the proposed method for class discovery and classification, especially on the effect of the number of unknown classes. This dataset comprises the expression levels of 7129 genes, in three types of acute leukemia, AML, ALL-B and ALL-T. It consists of a training set with 38 samples and a test set with 34 additional samples. From this dataset, we reallocate some of the samples in the training set to the test set in several ways such that the new training set consists of samples from either 0, 1 or 2 types of leukemia, resulting in 7 variants. Together with the original dataset in which the training set contains samples from all three leukemia types, they can be divided into four groups, named UK<sub>0</sub>, UK<sub>1</sub>, UK<sub>2</sub> and UK<sub>3</sub>, with the subscript denoting the number of ‘unknown’ classes in the test set. The top half of Table 1 shows the composition of the training and test sets for each of these variants. The

class(es) named under UK<sub>1</sub> or UK<sub>2</sub> refer to the class(es) that are considered ‘unknown’ in the corresponding scheme. Note that UK<sub>0</sub> is the original dataset without any unknown class, while UK<sub>3</sub> does not have a training set at all.

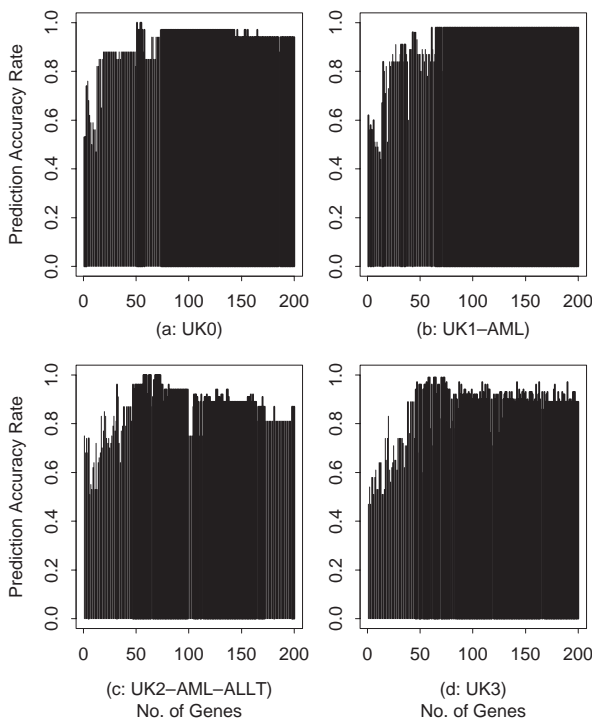
Three preprocessing procedures are explored for their effects on the outcome: (A) the method proposed by Dudoit *et al.* (2002), (B) the procedure used by Lee and Lee (2002) and (C) the preprocessing method employed by Golub *et al.* (1999). Note that the leukemia data, downloaded from [www.genome.wi.mit.edu/MPR](http://www.genome.wi.mit.edu/MPR), has been preprocessed under (C), and the additional preprocessing steps in (A) and (B) are based on this preprocessed data. Moreover, we can assume that any gene with missing expression level in any of the samples will be excluded by the preprocessing procedures.

An additional step after preprocessing is executed further to reduce the number of genes for our analysis. For each of the datasets in which there are at least two classes of training samples (UK<sub>0</sub> and UK<sub>1</sub>), the BW measure (Dudoit *et al.*, 2002) is used to select 1–200 genes that are deemed most ‘significant’ among the training samples. For datasets in which there are at most one known class of training samples (UK<sub>2</sub> and UK<sub>3</sub>), BW is no longer applicable, therefore, we use the Inter-Quartile-Range (IQR) measure to order the genes across all samples (training and tests combined), with higher IQR considered as more informative for classification and hence more ‘significant’.

For each set of genes ( $1 \leq G \leq 200$ ) selected, we fit the mixture model to the dataset comprising the samples with the expressions of the genes selected. In other words, we fit the model with 200 different classifier sizes, for each  $M$  that ranges from  $\max\{1, K\}$  to  $M_{\max}$ , with  $M_{\max}$  taken to be 10 for these applications. In addition to the two sets of starting points generated from  $K$ -means and PAM, we also use 10 sets of starting points generated from the random starting scheme. We choose  $G$  up to 200 following Lin and Alexandridis (2003), as they found that to be sufficiently large for classification problems.

## Results

Figure 1 shows the estimated number of classes using BIC and the corresponding prediction accuracy rate for 4 representative variants, for each of the 200 classifier sizes. The bottom half of Table 1 records the best prediction results with the smallest classifier among the 200 classifier sizes, for each of the 8 data variants. In general, we can see that the method performs well in all the variants with moderate classifier sizes, especially for UK<sub>0</sub> and the three UK<sub>1</sub> variants. For example, for UK<sub>0</sub>, the best prediction accuracy rate of 1.0 is achieved for the classifiers with 51, 54 and 55 genes. That is, there is no prediction error for a small number of classifier sizes. We also observe that, for datasets with a moderate number of genes, say  $80 \leq G \leq 140$ , BIC identifies the correct number of classes, with only one mispredicted sample, corresponding to an accuracy rate of 0.97. For the three UK<sub>2</sub> variants and UK<sub>3</sub>,



**Fig. 1.** BIC model selection results and the corresponding prediction accuracy rates for 200 classifiers, for a selected number of datasets: (a) UK<sub>0</sub>, (b) UK<sub>1</sub>-AML, (c) UK<sub>2</sub>-AML-ALLT and (d) UK<sub>3</sub>. In each of the four plots, for each classifier with  $G$  ( $1 \leq G \leq 200$ ) genes, the result is shown by a vertical line, with its height representing the prediction accuracy rate, and with its gray level denoting the model being correctly estimated (black;  $M = 3$ ), underestimated (light gray;  $M < 3$ ) or overestimated (dark gray;  $M > 3$ ).

the problems are harder, and the results can be more sensitive to the classifier sizes. Despite the lack of training samples for most/all of the classes, good results are still obtained, with at most two misclassified samples for appropriate classifier sizes.

For the task of model selection, unlike BIC, AIC does not perform well, usually finding too many classes, as it does not penalize sufficiently for overfitting. Since we observe that BIC consistently outperforms AIC, the results presented in the paper are based on BIC. Similarly, we only present results based on preprocessing method (A), as it is found to be superior to (B) and (C) for most of the datasets. More detailed descriptions and color plots of all results can be found from our website that provides Supplementary information.

## A RESAMPLING STUDY

### Study design

This study is carried out to evaluate further the performance of the proposed class discovery and classification procedure. For each of the 8 variants of the leukemia data, we resample 100 datasets of the same variant based on the original samples. For the variants under UK<sub>0</sub>, UK<sub>1</sub> and UK<sub>2</sub>, two-third of all the samples belonging to the ‘known’ class(es) are randomly selected as training samples, while the remaining samples are assigned to the test set, making up one resampled dataset. For UK<sub>3</sub>, since no training samples are available, 80% of the samples from each class are randomly selected to make up a resampled dataset consisting of test samples only. Alternatively, one could use an unstratified sampling scheme, but we chose to use the current one so that our procedure could be better tested with samples from each of the known classes.

**Table 1.** Composition of the training and test sets as well as minimal prediction errors for the test samples for each of the eight variants of the Leukemia data

Class	UK <sub>0</sub>		UK <sub>1</sub>		ALLT		ALLB		AML-ALLB		UK <sub>2</sub>		ALLB-ALLT		UK <sub>3</sub>	
	Tr	Ts	Tr	Ts	Tr	Ts	Tr	Ts	Tr	Ts	Tr	Ts	Tr	Ts	Tr	Ts
ALLB	19	19	19	19	19	19	—	38	—	38	19	19	—	38	—	38
ALLT	8	1	8	1	—	9	8	1	8	1	—	9	—	9	—	9
AML	11	14	—	25	11	14	11	14	—	25	—	25	11	14	—	25
Total <sup>a</sup>	38	34	27	45	30	42	19	53	8	64	19	53	11	61	—	72
ALLB	—	—	—	—	—	—	—	—	—	17 <sup>c</sup>	—	—	—	17 <sup>d</sup>	—	17 <sup>d</sup>
ALLT	—	—	—	67 <sup>c</sup>	—	—	—	67 <sup>e</sup>	—	67 <sup>c</sup>	—	—	—	0	—	—
AML	—	—	—	—	—	66 <sup>d</sup>	—	66 <sup>e</sup>	—	—	—	—	—	0	—	—
Pred <sup>b</sup>	51	0	61	1	65	1	48	2	48	2	57	0	57	1	58	1

<sup>a</sup>For the first half of the table, the numbers under the headings ‘Tr’/‘Ts’ are the number of training/test samples of the identified class on the left. ‘—’ denotes unavailability of training samples in the class. Note that, in UK<sub>3</sub>, the class memberships of all the samples are unknown, representing a class discovery problem.

<sup>b</sup>The second half of the table gives the best prediction results. For each data variant, the first number in the row ‘Pred’ reports the smallest classifier size for which best prediction occurs, while the second number is the corresponding total number of prediction errors. The specific indices (as given by Golub *et al.*) of the samples misclassified are given under ‘Ts’, with ‘—’ indicates no sample in that class is misclassified.

<sup>c</sup>The sample is misclassified as AML.

<sup>d</sup>The sample is misclassified as ALLB.

<sup>e</sup>The sample is misclassified as ALLT.

Each resampled dataset is first preprocessed using method (A) and gene selection is done by the BW measure (for UK<sub>0</sub> and UK<sub>1</sub> based on the training samples) or the IQR measure (for UK<sub>2</sub> and UK<sub>3</sub>) to build the classifiers, with sizes ranging from 1 to 200, as with the leukemia variants. To keep the computing time within a reasonable limit, we let the number of models  $M$  taking integers in the range  $[\max\{1, K\}, (K + 2)I\{K > 0\} + 5I\{K = 0\}]$ , where the two indicator functions are defined in the usual way. Since our experience with the leukemia data indicates that models exceeding the set limits are rarely picked for reasonably large classifier sizes, we doubt that this limitation on the class size would lead to a significant information loss.

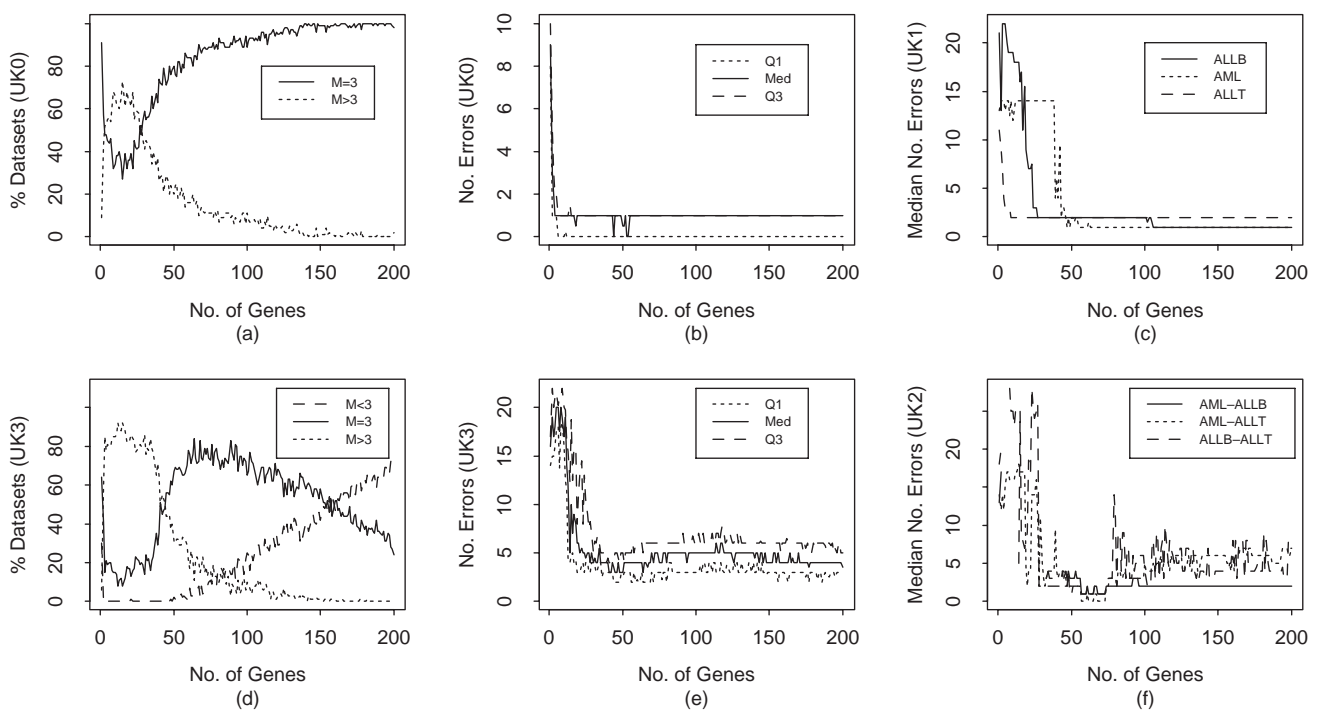
## Results

**No unknown class—UK<sub>0</sub>** For UK<sub>0</sub>, the results about the discovery of the number of classes are plotted in Figure 2a, which shows, for each classifier size, the percentages of the resampled datasets for which the models with  $M$  classes are selected by BIC. The correct number of classes,  $M = 3$ , is always picked more often than the other alternatives, for decent sized classifiers ( $G > 30$ ). In fact, for classifiers with  $G > 100$ , the correct number of classes is selected in  $>90\%$

of the resampled datasets. Summary statistics of the classification results, when the correct number of classes is selected, are plotted in Figure 2b. As can be seen from this plot, for classifiers whose sizes are not too small, the first quartiles of the numbers of prediction errors are all zero, indicating that there are no prediction errors in  $>25\%$  of the replications, when the correct model is selected. Furthermore, the third quartiles of 1's indicate that  $<25\%$  of the replicates have more than one prediction error.

**One unknown class—UK<sub>1</sub>** The performance of BIC for model selection for all three UK<sub>1</sub> datasets is similar to that for UK<sub>0</sub> as depicted in Figure 2a. Namely, for classifiers that are not too small, the model with the correct number of classes,  $M = 3$ , is picked most often, and among classifiers with more than 100 genes,  $M = 3$  is selected at a very high rate. The median errors, among the replicates with the correct model selected, are plotted in Figure 2c for all three variants. We observe consistent results for a large range of classifier sizes ( $G > 100$ ). For UK<sub>1</sub>–ALLB and UK<sub>1</sub>–AML, there is only one error, while there is one more error for UK<sub>1</sub>–ALLT.

**Two unknown classes—UK<sub>2</sub>** For UK<sub>2</sub>–AML–ALLB, as with all the UK<sub>0</sub> and UK<sub>1</sub> variants, for classifiers with decent sizes, the correct model is identified most often, with the



**Fig. 2.** BIC model selection results and summary statistics for the number of prediction errors among the resampled replicates in which  $M = 3$  is selected, for each of 200 classifiers: (a) and (d) display the percentages of resampled replicates in which the model is correctly estimated ( $M = 3$ ), underestimated ( $M < 3$ ) or overestimated ( $M > 3$ ), for UK<sub>0</sub> and UK<sub>3</sub>, respectively; (b) and (e) show several summary statistics for the number of prediction errors for UK<sub>0</sub> and UK<sub>3</sub>, respectively; and (c) and (f) plot the median errors for the three UK<sub>1</sub> variants and the three UK<sub>2</sub> variants, respectively.

percentage of correct identification reaching 100% for most of the classifiers with more than 100 genes. However, for both UK<sub>2</sub>-AML-ALLT and UK<sub>2</sub>-ALLB-ALLT, as with the corresponding original data variants, the classifier size has a greater effect on the performance of BIC. The results are similar, qualitatively, to that plotted in Figure 2d for UK<sub>3</sub>, showing that too large a classifier may lead to underestimation of the number of classes, although the percentages for the correct number of classes are generally much higher than that shown in Figure 2d. The medians for all three UK<sub>2</sub> variants, plotted in Figure 2f, show clearly that the number of errors for UK<sub>2</sub>-AML-ALLT and UK<sub>2</sub>-ALLB-ALLT are generally higher than for most of the other variants considered, especially when the classifier sizes are large. However, we note that in a small range of classifier sizes ( $65 \leq G \leq 75$ ), the correct numbers of classes are identified >95% of the time, and among these datasets, the median numbers of errors are zero and one, for UK<sub>2</sub>-AML-ALLT and UK<sub>2</sub>-ALLB-ALLT, respectively.

*Three unknown classes—UK<sub>3</sub>* The results for UK<sub>3</sub>, in which there is no training sample, are plotted in Figure 2d and e. In terms of selection with BIC, the results show that not only too few genes can make it hard to estimate the correct number of classes, but too many genes ( $G > 150$ ) can also lead to underestimation of the number of true classes. For moderate classifier sizes ( $60 \leq G \leq 100$ ), however, BIC identifies the correct number of classes in more than two-third of the resampled replicates. Among them, the median number of misclassified test samples varies between 4 and 5 (out of 57 test samples), giving a median prediction accuracy rate of between 91 and 93%.

## DISCUSSION

In this paper, we propose a method for the discovery of putative types/subtypes of cancers and classification of the interrogated tumor samples into the identified classes. This method enables us to handle datasets with no training samples (class discovery problems), datasets where training samples exist and all the test samples belong to the known classes (classification), as well as datasets in which training data exist but some of the test samples do not belong to any of the known classes (joint analysis of class discovery and classification). This unified approach to class discovery and classification is achieved by modeling the gene expression profile of a test sample as from a mixture of an unknown number of distributions, each characterizing the gene expression levels of genes within a class. Model selection criterion BIC clearly outperforms its competitor AIC, as the correct number of classes is much more frequently inferred with BIC than with AIC, which tends to lead to overestimation, due to its insufficient penalty for overfitting. Note that mixture models have been used by others to cluster genes or samples based on gene expression data (e.g. Broet *et al.*, 2002; Ghosh and Chinnaiyan, 2002;

McLachlan *et al.*, 2002), although their settings all differ from the problems that we are addressing in the current paper.

We apply our method to the leukemia data of Golub *et al.* (1999) and its seven variants through reallocating the 'known' and 'unknown' classes. It is perhaps most sensible to compare the results on UK<sub>0</sub> to those from other approaches for classification, as UK<sub>0</sub> is the original dataset in which all test samples belong to some 'known' classes. However, we note that the problem is more difficult within our context as we allow the possibility of new classes, which is more realistic in practice. Compared with Lin and Alexandridis (2003), which followed the same mixture formulation as the current paper but did not allow for extra new classes, our procedure does not perform as well. This is expected as certain information for classification is lost in an effort to select the correct number of classes as well. Nevertheless, our results are still as good as those from several other approaches in the literature. With our procedure, the best prediction accuracy rate of 1.0 is achieved for several classifier sizes, and there is only one prediction error for a large number of med-size classifiers, with preprocessing procedure (A). With preprocessing procedures (B) and (C), even better results are achieved, with no prediction errors for a large number of classifiers. Keller *et al.* (2000) also yielded no misclassification for a small number of classifiers, while Lee and Lee (2002) did not achieve perfect prediction with the limited number of classifier sizes studied.

For the class discovery problem without training samples, our results compare favorably with those from Golub *et al.* (1999) using SOMs. In particular, their four-cluster SOM with the collapsing of B3 and B4 into a single cluster representing ALL-B seems to be the most appropriate for comparison. Their results, based on the 38 training samples only, show that there are two misclassifications, giving a prediction accuracy rate of 0.95. Whereas our method gives a prediction accuracy rate of about 0.99, with only one error. The EMMIX-GENE approach of McLachlan *et al.* (2002) also yielded one misclassification, but the result was obtained in the simpler context without a model selection component.

For the variants in which there are test samples belonging to the known classes as well as samples belonging to unknown classes, we are not aware of comparable procedures in the literature in this context. Although the approaches of Golub *et al.* (1999) and Lee and Lee (2002) have a prediction strength measure to weed out samples that cannot be confidently classified to one of the known classes, they do not place these samples into any putative class(es). Therefore, our discussion here is based on our observations only. For two of the UK<sub>1</sub> variants, because there are two known classes in the training samples, the gene selection measure BW is capable of selecting genes that best separate these two samples into the classifiers, yielding good results, with the correct number of classes inferred, and with only one prediction error for a large number of classifiers. For UK<sub>1</sub>-ALLB, there is one more prediction error for most of the classifiers, which may reflect

the fact that there are fewer training samples in this dataset than in its two UK<sub>1</sub> counterparts. For the UK<sub>2</sub> variants, the results are not as good as for UK<sub>1</sub>, as expected, as all training samples belonging to a single known class does not allow for selection of genes that can discriminate among known classes. In two of the three cases, the number of genes used is a far more important variable, as having too many genes tend to underestimate the true number of classes.

To investigate further whether the patterns revealed by the Golub *et al.* (1999) data are mainly due to the particular division of ‘training’ and ‘test’ samples, we carry out a study by analyzing 100 resampled datasets for each of the 8 variants of the Golub *et al.* data. Our analyses confirm that best results are indeed achieved with the UK<sub>0</sub> data. Furthermore, the UK<sub>1</sub> data are easier to handle than the UK<sub>2</sub> and UK<sub>3</sub>. In terms of prediction accuracy rates for the replicates whose number of classes are correctly inferred, the results are reasonable, although we need to point out that prediction accuracies are likely to suffer when the number of classes is mis-estimated in the first place. Also, the performance of the method is rather insensitive to classifier sizes for UK<sub>0</sub> and UK<sub>1</sub>, whereas in UK<sub>2</sub> and UK<sub>3</sub>, a med-size classifier is essential, otherwise, the number of classes may be over-estimated, or under-estimated. This is not surprising, as increasing the number of genes in the classifier would eventually introduce genes with little or no discriminatory power, which should lead to masking of the signals in the important genes. Furthermore, this effect should be more profound with the IQR gene selection measure, as it does not allow for specific targeting of genes that have the greatest discriminatory power among known classes.

We also apply our method to the colon cancer data of Alon *et al.* (1999). We first analyze the data excluding the five likely contaminated samples, as in Soukup and Lee (2003), so that our result can be compared with their results. We use the split of training and test sets that has led to the best result obtained by Soukup and Lee (2003), which would give advantage to their method for comparison purpose. BIC selects the correct number of classes for most of the classifiers explored, with a prediction accuracy rate of 1.0 for almost all classifiers with 19–200 genes. With only two genes, we misclassify a tumor sample as a normal sample. Since the best split for Soukup and Lee (2003) is unlikely to be the best split for our method, we believe that our results are comparable with those obtained by Soukup and Lee (2003) for classification. We then re-analyze the entire dataset containing all 62 samples so that our result can be compared with those obtained by other authors for the problem of class discovery. Our best result misclassifies five samples, including, surprisingly, only one of the five samples often misclassified using other methods (Alon *et al.*, 1999; McLachlan *et al.*, 2002 and references therein). The number of samples misclassified is one less than the best results obtained by McLachlan *et al.* (2002), while the analysis by Alon *et al.* (1999) misses eight samples. We also analyze the data under the variants of having training samples from at least one of

the two classes (cancer or normal). The full results, with a more detailed description of the analyses, are available at our website. In summary, for the more difficult colon dataset, our method performs well compared with other methods for the pure classification or class discovery problems. Furthermore, encouraging results are also obtained for the situations with partial training sets, which are not being tackled by the other methods.

Selection of classifier sizes can be done through leave-one-out cross-validation of the training samples, as carried out in Lin and Alexandridis (2003), although we note that such a procedure will not be feasible with class discovery problems without training samples. In the absence of a good method for choosing classifier sizes, our results based on the leukemia data led to the recommendation of using mid-size (50–120 genes) classifiers, although we caution that this recommendation needs to be studied further as it is based on a limited number of results. We also note that this recommendation includes classifier sizes that are much larger than the recommendation of Golub *et al.* (1999) with 50 genes, which, as stated in their paper, was picked rather arbitrarily.

Since EM is known to be sensitive to starting points, it is important to explore multiple sets of them, as exemplified in our study. First, we observe that although *K*-means seems to outperform PAM, in the sense that starting values generated from *K*-means usually lead to higher likelihoods than those from PAM, the reverse can also happen with a non-negligible frequency. In addition to the starting points from these two, and perhaps other, clustering procedures, we strongly recommend the use of random starting points as well, as one of these can frequently lead to the highest likelihood, as we have seen in our study.

In our mixture modeling formulation, we assume that each class has a multivariate normal density with diagonal variance–covariance matrix. This assumption, made for analytical tractability, and also used by other recent classification papers (e.g. Keller *et al.*, 2000; Dudoit *et al.*, 2002), works well for the data that we have analyzed. However, this is not an inherent assumption, as other distributions, and/or some levels of dependencies among expression levels of genes, can be used. If the EM iterates become intractable, other estimation procedures, such as the versatile class of Markov chain Monte Carlo methods, can be entertained.

## ACKNOWLEDGEMENTS

The authors would like to thank the two referees for helpful comments. This work was supported in part by NSF grants DMS-9971770 and DMS-0306800, and NIH grant 1R01HG002657-01A1.

## REFERENCES

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., USA*, **96**, 6745–6750.
- Broet,P., Richardson,S. and Radvanyi,F. (2002) Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *J. Comput. Biol.*, **9**, 671–683.
- Dudoit,S., Fridlyand,J. and Speed,T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.
- Ghosh,D. and Chinnaiyan,A.M. (2002) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18**, 275–286.
- Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,C., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
- Hastie,T., Tibshirani,R. and Friedman,J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer–Verlag, New York.
- Kaufman,L. and Rousseeuw,P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
- Keller,A.D., Schummer,M., Hood,L. and Ruzzo,W.L. (2000) Bayesian Classification of DNA Array Expression Data. *Technical Report UW-CSE-2000-08-01*, Department of Computer Science and Engineering, University of Washington, WA.
- Lee,Y. and Lee,C.-K. (2003) Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, **19**, 1132–1139.
- Lin,S. and Alexandridis,R. (2003). Classification of tissue samples using mixture modeling of microarray gene expression data. *IMS Lecture Notes-Monogr. Ser.*, **40**, 419–435.
- McLachlan,G. and Peel,D. (2000) *Finite Mixture Models*. Wiley, New York.
- 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.
- Soukup,M. and Lee,J.K. (2003) Developing optimal prediction models for cancer classification using gene expression data. *J. Bioinform. Comput. Biol.* **1**, 681–694.