BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 22 2005, pages 4148–4154
doi:10.1093/bioinformatics/bti681

Gene expression

Classification of microarrays to nearest centroids

Alan R. Dabney

Department of Biostatistics, University of Washington, Seattle 98195, USA

Received on August 15, 2005; revised on September 15, 2005; accepted on September 17, 2005
Advance Access publication September 20, 2005

ABSTRACT

Motivation: Classification of biological samples by microarrays is a topic of much interest. A number of methods have been proposed and successfully applied to this problem. It has recently been shown that classification by nearest centroids provides an accurate predictor that may outperform much more complicated methods. The 'Prediction Analysis of Microarrays' (PAM) approach is one such example, which the authors strongly motivate by its simplicity and interpretability. In this spirit, I seek to assess the performance of classifiers simpler than even PAM.

Results: I surprisingly show that the modified *t*-statistics and shrunken centroids employed by PAM tend to increase misclassification error when compared with their simpler counterparts. Based on these observations, I propose a classification method called 'Classification to Nearest Centroids' (ClaNC). ClaNC ranks genes by standard t-statistics, does not shrink centroids and uses a class-specific gene-selection procedure. Because of these modifications, ClaNC is arguably simpler and easier to interpret than PAM, and it can be viewed as a traditional nearest centroid classifier that uses specially selected genes. I demonstrate that ClaNC error rates tend to be significantly less than those for PAM, for a given number of active genes.

Availability: Point-and-click software is freely available at http://students.washington.edu/adabney/clanc

Contact: adabney@u.washington.edu

Supplementary Information: http://students.washington.edu/adabney/clanc/supplement.pdf

INTRODUCTION

Gene-expression microarrays (Schena *et al*., 1995) can be used to discriminate between multiple clinical or biological classes. A classifier is built from training data, consisting of expression profiles for samples of known class. Many methods exist for building classifiers, but the overall goal is the same: (1) find characteristics that define each class, and (2) build a function that compares the expression profile of an unknown sample with each class on the basis of these defining characteristics. The unknown sample is assigned to the class to which it is most similar.

Candidate classifiers are evaluated on the basis of (1) accuracy, (2) interpretability, and (3) practicality. High accuracy corresponds to low misclassification error. Interpretability may or may not be important, depending on the application. Often, it is desirable to learn something about the underlying processes at play in addition to classifying accurately. It may be very difficult to characterize the contribution of a single gene to the overall classifier (in turn making it difficult to learn about the biological functionality of that gene) if a complicated method is used. A complicated classifier may then be rejected for a simpler alternative, even if the simpler alternative does not perform as well. Finally, it must be practical to implement the classifier. Because of the resources required to form a microarray, it is desirable to base the classifier on the fewest genes possible.

Linear Discriminant Analysis (LDA) (Mardia *et al*., 1979) is a classical method of prediction that is simple to understand and has been shown to perform well with microarrays (Dudoit *et al*., 2002; Lee *et al*., 2005). Each class is characterized by its vector of means or 'centroid.' An unknown sample is evaluated by computing the scaled distance between its expression profile and each class centroid. The unknown is assigned to the class to which it is nearest. Thus, LDA can be thought of as a 'nearest centroid classifier.'

To identify the fewest genes necessary for discrimination, a feature-selection step can be added. Note that the goal of this step will depend on the setting. For example, with unlimited resources, we may wish to use *all* relevant genes in the classifier; although, we may be able to find a subset of genes that classify just as well as (or even better than) the complete set. In other settings, it may be necessary to make tradeoffs between accuracy and practicality.

An obvious way to incorporate feature-selection into LDA is to compute test statistics for each gene that measure that gene's ability to distinguish the classes, rank those statistics and choose only the top genes from this list to base the classifier on. In early work, Dudoit *et al*. (2000, 2002) used *F*-statistics for ranking genes. The Prediction Analysis of Microarrays (PAM) method (Tibshirani *et al*., 2002) uses what are essentially *t*-statistics. However, instead of the usual *t*-statistic, PAM adds a 'fudge-factor' to each statistic's denominator. This prevents genes with large *t*-statistics but small mean differences (numerators) from being selected. A further level of complexity is added to PAM by shrinking the class centroids toward their overall mean. Thus, PAM can be thought of as a 'nearest shrunken centroid classifier.'

I present here an alternative LDA-based classifier that I call ClaNC, for Classification to Nearest Centroids. ClaNC (1) does not shrink centroids, (2) uses unmodified *t*-statistics to select genes, (3) carries out class-specific feature selection, and (4) allows each gene to be active in at most one class. I first individually evaluate the four proposed changes on simulated datasets (Table 1). Shrinkage, fudge factors and class-nonspecific gene selection can all increase misclassification error, depending on the situation. The error rates for ClaNC are substantially lower than those for PAM at all considered numbers of active genes. Furthermore, ClaNC error estimates are consistently less variable than their PAM analogs. I then compare PAM and ClaNC on four previously published datasets (Tables 2–5 and Fig. 1). I again evaluate each of the four proposed changes individually. Overall, shrinkage, fudge factors

Table 1. Comparison of error rates on three simulated examples (standard deviations in parentheses)

	Active genes					
	4	8	12	20	40	60
PAM[a]	0.41 (0.13)	0.31 (0.14)	0.26 (0.13)	0.17 (0.10)	0.07 (0.06)	0.03 (0.03)
I	0.38 (0.12)	0.29 (0.14)	0.23 (0.10)	0.12 (0.07)	0.03 (0.03)	0.01 (0.01)
II	0.14 (0.14)	0.04 (0.08)	0.02 (0.06)	0.01 (0.03)	0.00 (0.00)	0.00 (0.00)
III	0.35 (0.14)	0.29 (0.15)	0.26 (0.13)	0.17 (0.10)	0.05 (0.04)	0.02 (0.02)
IV	0.41 (0.13)	0.31 (0.14)	0.26 (0.13)	0.17 (0.10)	0.07 (0.06)	0.03 (0.03)
ClaNC	0.02 (0.03)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
PAM[b]	0.53 (0.04)	0.51 (0.04)	0.50 (0.04)	0.43 (0.07)	0.34 (0.10)	0.25 (0.08)
I	0.52 (0.04)	0.50 (0.04)	0.48 (0.04)	0.43 (0.07)	0.31 (0.11)	0.21 (0.09)
II	0.52 (0.05)	0.49 (0.06)	0.45 (0.08)	0.39 (0.08)	0.28 (0.08)	0.22 (0.08)
III	0.30 (0.07)	0.21 (0.05)	0.16 (0.06)	0.08 (0.03)	0.03 (0.02)	0.01 (0.01)
IV	0.53 (0.04)	0.51 (0.04)	0.50 (0.04)	0.43 (0.07)	0.34 (0.10)	0.25 (0.08)
ClaNC	0.25 (0.06)	0.12 (0.04)	0.07 (0.04)	0.03 (0.02)	0.00 (0.01)	0.00 (0.00)
PAM[c]	0.31 (0.10)	0.22 (0.07)	0.19 (0.03)	0.17 (0.04)	0.14 (0.05)	0.12 (0.05)
I	0.24 (0.11)	0.11 (0.07)	0.05 (0.03)	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)
II	0.33 (0.12)	0.23 (0.08)	0.18 (0.04)	0.16 (0.04)	0.13 (0.05)	0.10 (0.05)
III	0.18 (0.07)	0.08 (0.05)	0.05 (0.04)	0.02 (0.02)	0.00 (0.01)	0.00 (0.00)
IV	0.31 (0.10)	0.22 (0.07)	0.19 (0.03)	0.17 (0.04)	0.14 (0.05)	0.12 (0.05)
ClaNC	0.10 (0.04)	0.03 (0.02)	0.01 (0.01)	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)

[a]Simulation one: noisy data.
[b]Simulation two: class one easier to distinguish than the others.
[c]Simulation three: class one harder to distinguish than the others.
I PAM minus shrinkage.
II PAM minus fudge factors.
III PAM with class-specific selection.
IV PAM with unique features.

Table 2. Comparison of error rates on SRBCT data

	Active genes					
	4	8	12	20	40	60
PAM	0.46	0.43	0.27	0.06	0.05	0.07
I	0.29	0.17	0.14	0.06	0.05	0.04
II	0.29	0.20	0.11	0.05	0.07	0.04
III	0.47	0.30	0.19	0.04	0.02	0.00
IV	0.46	0.42	0.29	0.10	0.05	0.04
ClaNC	0.13	0.07	0.04	0.05	0.06	0.04

See Table 1 footnote for definitions of classifiers I–IV.

Table 3. Comparison of error rates on lymphoma data

	Active genes					
	3	6	9	15	30	45
PAM	0.28	0.34	0.33	0.24	0.14	0.33
I	0.17	0.19	0.17	0.21	0.12	0.24
II	0.30	0.21	0.16	0.14	0.17	0.21
III	0.22	0.05	0.12	0.26	0.03	0.00
IV	0.28	0.28	0.26	0.24	0.12	0.12
ClaNC	0.10	0.09	0.03	0.02	0.03	0.02

See Table 1 footnote for definitions of classifiers I–IV.

Table 4. Comparison of error rates on NCI data

	Active genes					
	8	16	24	40	80	120
PAM	0.75	0.67	0.65	0.49	0.40	0.40
I	0.75	0.61	0.54	0.44	0.40	0.28
II	0.72	0.72	0.65	0.60	0.49	0.39
III	0.58	0.47	0.44	0.32	0.07	0.04
IV	0.75	0.70	0.63	0.58	0.40	0.30
ClaNC	0.51	0.44	0.32	0.35	0.32	0.35

See Table 1 footnote for definitions of classifiers I–IV.

Table 5. Comparison of error rates on leukemia data

	Active genes					
	2	4	6	10	20	30
PAM	0.47	0.32	0.18	0.11	0.05	0.05
I	0.32	0.26	0.21	0.05	0.05	0.05
II	0.45	0.26	0.16	0.16	0.16	0.11
III	0.18	0.18	0.16	0.13	0.08	0.08
IV	0.29	0.26	0.11	0.11	0.08	0.08
ClaNC	0.18	0.11	0.11	0.05	0.05	0.03

See Table 1 footnote for definitions of classifiers I–IV.

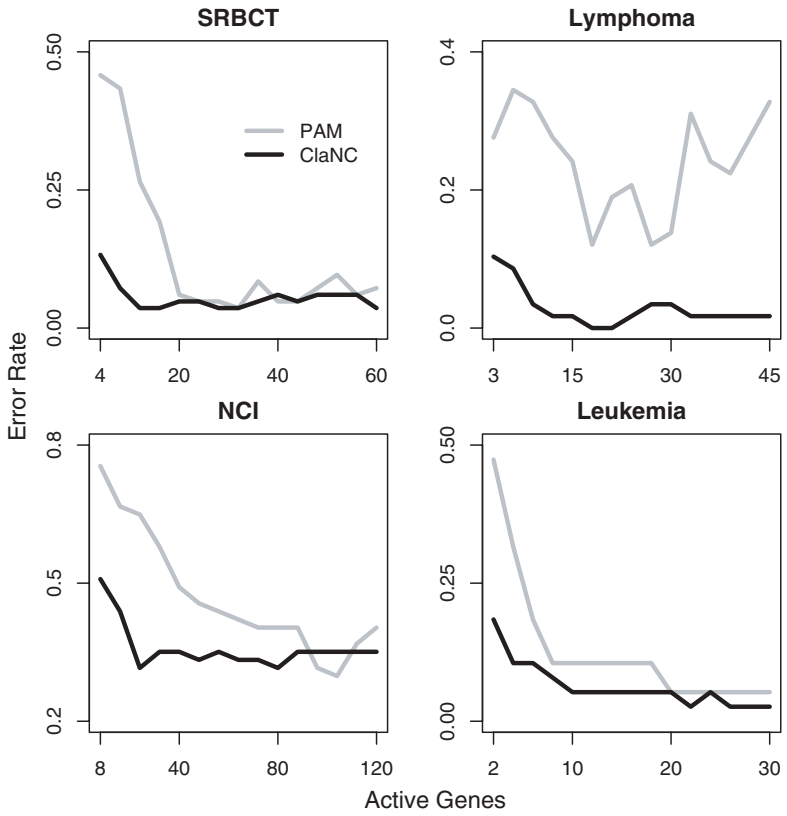

Fig. 1. Comparison of cross-validated misclassification error rates for PAM and ClaNC on four real datasets.

and class-nonspecific gene selection tend to increase error in the real examples. There is some evidence that repeat representations of a single gene can also increase error. ClaNC error rates again tend to be smaller than those for PAM. Surprisingly, then, LDA-based classifiers that are even simpler than PAM can perform very well.

I note briefly some of the many other classification methods besides those based on LDA. These include neural networks (Khan *et al*., 2001), support vector machines (Ramaswamy *et al*., 2001), CART (Breiman *et al*., 1984), random forests (Breiman, 2001), and methods based on generalized linear models (Zhu and Hastie, 2004; Nguyen and Rocke, 2004). LDA-based methods have been shown to perform well when compared with more complicated classifiers (Dudoit *et al*., 2002; Lee *et al*., 2005). Furthermore, the simplicity of LDA-based methods arguably makes them preferable to more complicated alternatives in settings where interpretation of the classifier is important.

SYSTEM AND METHODS

Linear Discriminant Analysis (LDA)

We would like to classify unknown samples into one of K classes. To build a classifier, we obtain n_k training samples per class, $k = 1, 2, \ldots, K$, with m genes on each microarray. For each training sample, we observe class membership Y and expression profile X. For simplicity, I will represent the classes by the numbers $1, 2, \ldots, K$. Note that each expression profile is a vector of length m. We assume that expression profiles from class k are distributed as $N(\mu_k, \Sigma)$, the multivariate normal distribution with mean vector μ_k and covariance matrix Σ. Call $L(\cdot; \mu_k, \Sigma)$ the corresponding probability density function. Finally, we agree upon prior probabilities π_k that an unknown sample comes from class k, $k = 1, 2, \ldots, K$.

Bayes' theorem states that the probability that a sample comes from class k, given that sample's expression profile, is proportional to the product of the class density and prior probability:

$$Pr(Y = k|X = x) \propto L(x; \mu_k, \Sigma) \times \pi_k. \quad (1)$$

We call Equation (1) the posterior probability that array x comes from sample k. LDA assigns the sample to the class with the largest posterior probability:

$$\hat{y} = \arg\max_k \{L(x; \mu_k, \Sigma) \times \pi_k\}. \quad (2)$$

This can be shown to be the rule that minimizes misclassification error (Mardia *et al*., 1979).

The innards of the right side of Equation (2) are proportional to

$$|\Sigma|^{-1/2} \exp\Big\{ -\tfrac{1}{2}(x - \mu_k)^{\mathrm{T}} \Sigma^{-1} (x - \mu_k) + \log(\pi_k) \Big\}. \quad (3)$$

Since the covariance matrix Σ is the same for all classes, only the exponential component of Equation (3) is relevant to classification. We can then rewrite Equation (2) as

$$\hat{y} = \arg\min_k \{(x - \mu_k)^{\mathrm{T}} \Sigma^{-1} (x - \mu_k) - 2\log(\pi_k)\}. \quad (4)$$

Thus, a sample is assigned to the class to which it is nearest, as measured by the metric $\|x-\mu\|^2 - 2\log(\pi)$, where $\|x-\mu\|^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$ is the square of the Mahalanobis distance between x and μ.

We can further simplify the problem by assuming independence between genes. This allows us to simplify the LDA classification rule (4) to

$$\hat{y} = \arg\min_k \left\{ \sum_{i=1}^{m} \left(\frac{x_i^* - \mu_{ik}}{\sigma_i} \right)^2 - 2\log(\pi_k) \right\}. \quad (5)$$

Then, in order to select genes that provide the most discriminatory information, we can form test statistics for each gene. For example, Dudoit *et al*. (2000, 2002) use *F*-statistics to measure the between-class sum-of-squares relative to the within-class sum-of-squares:

$$F_i = \frac{\text{BSS}}{\text{WSS}} = \frac{\sum_{j=1}^{n} \sum_{k=1}^{K} \mathbf{I}\left(y_j = k\right)(\bar{x}_{ik} - \bar{x}_i)^2}{\sum_{j=1}^{n} \sum_{k=1}^{K} \mathbf{I}\left(y_j = k\right)(x_{ij} - \bar{x}_{ik})^2}, \quad (6)$$

$i = 1, 2, \ldots, m$. To form a classifier using only $\tilde{m}$ genes, class centroids are formed using only the genes corresponding to the largest $\tilde{m}$ *F*-statistics. All other genes are discarded.

Prediction Analysis of Microarrays

The current standard in LDA-based classification for microarrays is PAM (Tibshirani *et al*., 2002). Instead of *F*-statistics, PAM uses the statistics

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{w_k(s_i + s_0)} \quad (7)$$

to select genes, where $w_k = (1/n_k - 1/n)^{1/2}$ makes $w_k \times s_i$ equal to the standard error of the numerator, and s_0 is a fudge factor intended to guard against very large statistics for very small standard errors; by default, PAM chooses the median of the s_i for s_0. Without s_0, d_{ik} is just a *t*-statistic comparing the mean of gene i in class k with the overall mean of gene i. Hence, d_{ik} measures the difference between gene i in class k and gene i in all classes combined. A gene that discriminates one class from the rest will have a statistic of large absolute value. PAM then shrinks the d_{ik} toward zero, eliminating the genes that do not provide sufficient discriminatory information.

The shrinkage approach used by PAM is soft-thresholding. For a particular choice of shrinkage parameter Δ, the shrunken statistic is

$$\tilde{d}_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+, \quad (8)$$

where '+' means 'positive part' ($z_+ = z$ if $z > 0$, 0 otherwise). Thus, all d_{ik} less than Δ in absolute value are shrunken to zero, and the rest are shrunken to somewhere between zero and their original values. The shrinkage of the remaining statistics toward zero is intended as a 'de-noising' step. We can then rewrite Equation (7) with the shrunken statistics to produce corresponding shrunken centroids

$$\tilde{x}_{ik} = \tilde{d}_{ik} \times w_k \times (s_i + s_0) + \bar{x}_i, \quad (9)$$

where the shrinkage here is of the class centroids toward the overall centroid. Notice that the genes for which all shrunken class statistics $\tilde{d}_{i1}, \tilde{d}_{i2}, \ldots, \tilde{d}_{iK}$ equal zero have shrunken centroid components that equal the corresponding components of the overall centroid. When distances from a new sample to the shrunken class centroids are computed in Equation (5), the components for these inactivated genes are identical for each class. Hence, they do not contribute to the classification and do not need to be measured in a new sample. We call the genes with at least one shrunken centroid component different from the corresponding overall centroid component the *active* genes and the rest the *inactive* genes.

A new sample is classified by comparing its expression profile with each shrunken centroid, over the $\tilde{m}$ genes that remain active after shrinkage. Distances are assessed as in Equation (5), with the shrunken centroid $\tilde{x}_{ik}$ replacing μ_{ik} and $s_i + s_0$ replacing σ_i. Any prior information on class prevalences can be included in π_k. One simple choice is $\pi_k = n_k/n$, placing prior weights on each class in proportion to its sample prevalence; another is $\pi_k = 1/K$, placing equal prior weights on each class.

In a typical application of PAM, the shrinkage parameter Δ is allowed to vary over a wide range. For each Δ, a classifier is built and its error rate is estimated by cross-validation. The value of Δ to use in the final classifier is chosen from a plot of Δ against error rate, where the number of unique active genes corresponding to each Δ is also displayed. I argue that Δ does not mean as much as the number of active genes. The decision of how much error to tolerate will be made with cost and convenience in mind, and these are best gauged by the number of active genes.

In short, a PAM classifier selects $\tilde{m}$ genes by (1) soft thresholding the modified *t* statistics d_k, then (2) using the chosen, shrunken statistics to update the class centroids. The classifier can be represented by the shrunken centroid components and pooled standard deviations of the active genes, since these are the only components needed in the distance function (5). Each centroid is now of length $\tilde{m}$, with i-th component somewhere between that gene's class and overall means. Nothing is known about the distribution of active genes across classes. The selected genes are interpreted as simultaneously distinguishing all classes from each other.

Classification to Nearest Centroids

Shrinkage and fudge factors are intended to denoise the data, stabilizing the statistics used. Although shrinkage has been shown to produce more accurate mean estimates in noisy data (Donoho and Johnstone, 1994), it is not clear that this should translate to increased prediction accuracy. Furthermore, PAM's shrinkage procedure makes the class centroids look more similar to each other. It is unclear why this should be expected to make it easier to distinguish the classes from each other; this issue is discussed further below. For consistency with the preceeding discussion, we can formulate the gene selection process without shrinkage as

$$\tilde{d}_{ik} = d_{ik}\mathbf{I}(|d_{ik}| > \Delta), \quad (10)$$

where $\mathbf{I}(\cdot)$ is the indicator function.

The stated justification for fudge factors is that they 'guard against the possibility of large (test statistics) arising by chance from genes with low expression values' (Tibshirani *et al*., 2002). However, an extreme *t*-statistic indicates a 'significant' mean difference, regardless of whether its numerator is large or small. Mean differences that arise by chance will tend to have small *t*-statistics. In practice, then, fudge factors may remove from consideration informative genes with small mean differences and thus actually increase error.

Typically, genes are selected in a *class-nonspecific* manner by applying a common threshold to all test statistics. Suppose that class k_0 is more heterogeneous than the others and hence more difficult to characterize; alternatively, k_0 may simply have less training samples than the rest of the classes. Then the components of the vector d_{k_0} will tend to be closer to zero than their counterparts in the other classes, and class-nonspecific selection may not choose any class k_0 genes. With this in mind, I consider *class-specific* selection

$$\tilde{d}_{ik} = d_{ik}\mathbf{I}(|d_{ik}| > \Delta_k), \quad (11)$$

where different thresholds Δ_k are chosen for each class. In practice, there is still a single tuning parameter: the number of active genes per class.

I note that a subsequent PAM publication proposed something similar to this (Tibshirani *et al*., 2003). Class-specific thresholds were chosen in an adaptive manner by including another scale parameter in the denominator of Equation (7), as in

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{w_k\theta_k(s_i + s_0)}, \quad (12)$$

where the θ_k are restricted to $\min(\theta_k) = 1$. In classes for which prediction error is highest, the scale parameter is decreased to allow more active genes for those classes, and the θ_k are rescaled so that $\min(\theta_k) = 1$ (see the comments in the Discussion section).

It is possible that a single gene will have more than one extreme *t*-statistic. In this case, if we eliminate all but the top $\tilde{m}$ statistics, the number of (unique) active genes is less than $\tilde{m}$. It may be that the multiple extreme statistics for a

class are redundant, and hence it would be inefficient to include more than one. Furthermore, it may be intuitively appealing for each selected gene to characterize a single class. With this in mind, I consider allowing each gene to be active in at most one class. If a gene has more than one extreme t-statistic, only the largest in absolute value is chosen. The others are set to zero. Note that this does not exclude genes with more than one extreme t-statistic. Such genes will be included in the classifier, but they will only be active in one class.

Combining all of the above, I propose an alternative LDA-based classifier called ClaNC, for Classification to Nearest Centroids. ClaNC (1) does not shrink centroids, (2) does not use fudge factors, (3) carries out class-specific gene selection, and (4) allows each gene to be active in at most one class. Thus, each active gene characterizes exactly one class. Its centroid component for that class equals its class-specific mean. Its centroid components in all other classes equal its overall mean. A ClaNC classifier can be represented by the centroid components and pooled standard deviations of the active genes. Out of $\tilde{m}$ active genes, there are $\tilde{m}/K$ genes that characterize each class by default. Note, however, that ClaNC allows one to easily choose a different distribution of genes across classes. For example, it is straightforward to form a ClaNC classifier with more than $\tilde{m}/K$ characteristic genes in one class and less in another. The selected genes are interpreted as uniquely characterizing each class.

The role of shrinkage in classification

As mentioned earlier, the LDA classification rule minimizes misclassification error. This is because the LDA rule equals the Bayes' rule (Mardia *et al*., 1979). For simplicity, assume all genes are independent with variance one, and that each class has equal prior probability. Then the Bayes' rule is to classify a new sample to the class for which $\sum_{i=1}^{m}\left(x_i^* - \mu_{ik}\right)^2$ is smallest. However, we must estimate the centroids μ_k in practice, using $\hat{\boldsymbol{\mu}}_k$ in their place. Suppose x^* comes from class k_0. Then, expanding the squared distance between x^* to class k_0 and taking expectations, we have

$$E\sum_{i=1}^{m}\left(x_i^* - \hat{\boldsymbol{\mu}}_{ik_0}\right)^2 = E\sum_{i=1}^{m}\left(x_i^* - \mu_{ik_0} + \mu_{ik_0} - \hat{\boldsymbol{\mu}}_{ik_0}\right)^2$$
$$= E\sum_{i=1}^{m}\left(x_i^* - \mu_{ik_0}\right)^2 + E\sum_{i=1}^{m}\left(\mu_{ik_0} - \hat{\boldsymbol{\mu}}_{ik_0}\right)^2, \qquad (13)$$

or the Bayes' rule plus the mean squared error (MSE) of the centroid estimate.

Reducing the MSE of $\hat{\boldsymbol{\mu}}_{k_0}$ will bring us closer to the Bayes' rule. According to the Stein Paradox of statistics (Stein, 1956), we can reduce the MSE of $\hat{\boldsymbol{\mu}}_{k_0}$ by shrinking toward $1/m\sum_{i=1}^{m}\hat{\boldsymbol{\mu}}_{ik_0}$ (or any other constant). In our setting, this suggests shrinking each centroid across its m components. Note, however, that PAM shrinks each gene toward its overall mean. In other words, PAM shrinks each gene across its K classes. Thus, although shrinkage could in principle improve prediction accuracy, PAM apparently shrinks in the wrong direction. Furthermore, by shrinking across classes, all class centroids are made more similar to each other. It is unclear why this would be expected to improve classification accuracy. It is true, as one reviewer maintained, that shrinkage *can* be carried out across classes. However, owing to the much higher number of genes than classes, one would expect the greatest gain from shrinking across genes.

Even when shrinking each centroid across genes, it is not necessarily clear that prediction accuracy will be improved. If, for example, all class centroid estimates have their MSEs reduced equally, then there may be no change in the relative relationships of the centroid estimates. Preliminary investigation suggests that this is the case. However, I intend to perform a more thorough investigation of shrinkage for classification in future work.

RESULTS

I first consider three simulated set-ups, with 35 simulations per set-up. There are $K = 4$ classes, with $n_k = 30$ samples in each class and $m = 5000$ genes under consideration. Half of the genes provide no discriminatory information. In the other half, the class means differ randomly from each other. The first simulation allows for 'noisy' data, where there is significant heterogeneity between genes. Overall means μ_{i0}, $i = 1, 2, \ldots, m$, were generated from the $N(0, 1)$ distribution. Each class mean μ_{ik}, $i = 1, 2, \ldots, m$, $k = 1, 2, \ldots, K$, was the overall mean plus a draw from the $U(-1, 1)$ distribution. Observations were then generated from the $N(\mu_{ik}, \sigma_i)$ distributions, where the squares of the σ_i were drawn from the χ^2_3 distibution. The second simulation has one class being easier to distinguish than the rest. Overall means were generated as above. The class means for class one were the overall means plus draws from the $U(-2, 2)$ distribution. Those for the other three classes were the overall means plus draws from the $U(-1, 1)$ distribution. Observations were then generated from the $N(\mu_{ik}, 1)$ distributions. The third simulation is similar, now with one class being more difficult to distinguish than the rest. Class one means were overall means plus draws from the $U(-1, 1)$ distributions, while those for the other three classes were the overall means plus draws from the $U(-2, 2)$ distribution.

I compared six classifiers: PAM, PAM without shrinkage (I), PAM without fudge factors (II), PAM with class-specific gene selection (III), PAM with unique features (IV) and ClaNC. The first 15 samples in each class were used for training the classifiers, and the remaining 15 samples were used as test data. For each of 35 simulations within each simulation setup, misclassification error was computed using the 15 test samples. Table 1 shows the average test error over the 35 simulations, together with standard error estimates. In all three simulations, all classifiers considered attain zero error when using all 2500 relevant genes. The most apparent differences occur at low numbers of active genes, and so we only report these. Note that, although ClaNC is formulated in terms of the number of active genes per class, all results are presented in terms of the total number of active genes.

In simulation one, fudge factors increase error. The heterogeneity across genes in variance creates many large t-statistics corresponding to relatively small mean differences. However, an extreme t-statistic indicates a 'significant' mean difference, regardless of whether its numerator is large or small. It is not surprising then that fudge factors increase error here, since many informative genes have been excluded. In simulation two, class-nonspecific gene selection increases error. This is because the t-statistics for class one tend to be more extreme than those for the other classes. As a result, more genes are selected that characterize class one than are selected for the other classes, leading to poor classification in the other classes. In simulation three, class-nonspecific selection and shrinkage increase error. This is because the t-statistics for class one tend to be less extreme than those for the other classes. As a result, fewer genes are selected that characterize class one than are selected for the other classes, leading to poor classification in class one. Note finally that the PAM classifiers on these simulations tend to have greater variability than their ClaNC analogs, probably as a result of the increased simplicity of ClaNC.

I now compare PAM and ClaNC on four previously published cDNA microarray experiments. In each analysis, any missing values were imputed using k-nearest neighbors (Troyanskaya *et al*., 2001) with $k = 10$. I compare the methods on the basis of error rates from 5-fold cross-validation. I avoid gene-selection bias by completely rebuilding classifiers to identical specifications in each

cross-validation iteration (Ambroise and McLachlan, 2002). Cross-validated error rates are nearly unbiased, being slightly conservative (Ambroise and McLachlan, 2002; Hastie *et al*., 2001), and they are thus sufficient for comparing classifiers. Figure 1 and Tables 2–5 compare the performance of PAM and ClaNC on the examples; nominal error rates will change with example, being the expected error for non-informative data. I have listed the top few genes selected by ClaNC in each example in Supplementary Tables 1–4. There is very good agreement between the genes chosen by ClaNC and those chosen by previously published classifiers.

Small round blue cell tumors

The first example involves small round blue cell tumors (SRBCT) of childhood (Khan *et al*., 2001). Expression measurements were made on 2307 genes in 83 SRBCT samples. The tumors were classified as Burkitt lymphoma, Ewing sarcoma, neuroblastoma or rhabdomyosarcoma. There are 11, 29, 18 and 25 samples in each respective class. As seen in Figure 1 and Table 2, PAM requires 20 genes to drop the cross-validation error below 10%, whereas ClaNC needs only 8. Note that the PAM misclassification error estimate for 40 active genes (0.05) is slightly lower than that for ClaNC (0.06). However, the differences are not statistically significant. The standard errors are 0.05 and 0.025, with a two-sample *t*-test giving a *P*-value of 0.58. Shrinkage and fudge factors increase error in this example.

Lymphoma

In the second example, expression measurements were made on 4026 genes in 58 lymphoma patients (Alizadeh *et al*., 2000). The tumors were classified as diffuse large B-cell lymphoma and leukemia, follicular lymphoma, and chronic lymphocytic leukemia. There are 42, 6 and 10 samples in each respective class. As seen in Figure 1 and Table 3, PAM error rates are >10% with even 45 genes. Meanwhile, ClaNC requires only *three* genes for 10% error. Each of shrinkage, fudge factors, class-nonspecific selection and repeat representations of a gene increase error in this example.

NCI cancer cell lines

The third example involves the cell lines used in the National Cancer Institute's screen for anti-cancer drugs (Ross *et al*., 2000; Scherf *et al*., 2000). Expression measurements were made on 6830 genes in 60 cell tumors. There are representative cell lines for each of lung cancer, prostate cancer, CNS, colon cancer, leukemia, melanoma, NSCLC, ovarian cancer, renal cancer and one unknown sample. I filtered out 988 genes for which 20% or more of the tumors had missing values. I also excluded samples from prostate cancer (there being only two samples) and the one unknown sample. There are 9, 6, 7, 6, 8, 7, 6 and 8 samples in each remaining respective class. Classification is more difficult in this example, at least partly owing to there being so many classes and few samples per class. PAM requires 80 genes for <45% error, whereas ClaNC needs only 16. The PAM misclassification error estimate for 102 genes (0.30) is less than that for ClaNC (0.35). Again, the differences are not statistically significant. The standard errors are 0.07 and 0.045, with a two-sample *t*-test giving a *P*-value of 0.67. Class-nonspecific selection and shrinkage increase error in this example. There is weak evidence that fudge factors decrease error.

Leukemia

The fourth example involves acute myeloid leukemia (AML) and acute lymphblastic leukemia (ALL) (Golub *et al*., 1999). The public version of the training data used in the original analysis include expression measurements on 3857 genes in 38 leukemia patients. There are 11 and 27 samples in each respective class. PAM requires 20 genes for 5% error, whereas ClaNC needs only 10. Shrinkage increases error in this example. Fudge factors, class-nonspecific gene selection and repeat representations of a gene increase error for very small numbers of active genes and decrease error somewhat for larger numbers of active genes.

DISCUSSION

LDA-based methods have been successful in classifying microarrays. Surprisingly, I have found that shrinkage and fudge factors tend to actually increase misclassification error. Also, selecting genes by class appears to offer improvements in performance. Finally, I have suggested the selection of genes that uniquely characterize each class. Based on these observations, I have proposed a new LDA-based classifier called ClaNC. The classifier ClaNC does not use shrinkage or fudge factors and hence is very simple. Also, selected genes in a ClaNC classifier are naturally interpreted as uniquely characterizing a single class. I have demonstrated that ClaNC error rates tend to be substantially lower than their PAM counterparts in each of several examples, both simulated and real. Finally, I have provided freely available point-and-click software for ClaNC.

Tibshirani *et al*. (2003) demonstrated that adaptive thresholds sometimes improve PAM performance. Applying the adaptive thresholds to the examples listed above, I found that ClaNC errors are still lower than those for PAM. Furthermore, the adaptive thresholds are not generally available to users, as they are not implemented in the PAM point-and-click software.

Other aspects of LDA-based classification for microarrays could be further investigated. For example, all LDA-based classifiers use univariate statistics to select genes. The top genes from this list will be those that *individually* best distinguish the classes. However, we would ideally like to select the best *collection* of genes. It may be inefficient to select genes from the list of *t*-statistics if some of the top genes act in similar ways. I am currently working on an LDA-based classifier that would choose optimal collections of genes.

ACKNOWLEDGEMENTS

I greatly appreciate the helpful comments of John D. Storey on Stein's Paradox in the context of classification. I am also thankful to the reviewers for many helpful comments and suggestions. This research was supported in part by the Cancer-Epidemiology and Biostatistics Training Grant 5T32CA009168-29, NIH grant 1 U54 GM2119-03, and a traineeship at the Los Alamos National Laboratory.

Conflict of Interest: none declared.

REFERENCES

Alizadeh,A. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.
Breiman,L. (2001) Random forests. *Machine Learning*, **45**, 5–32.
Breiman,L., Friedman,J.H., Olshen,R. and Stone,C.J. (1984) *Classification and Regression Trees*. Wadsworth, Belmont, CA.
Donoho,D. and Johnstone,I. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.
Dudoit,S. *et al.* (2000) Comparison of discriminant methods for the classification of tumors using gene expression data. *University of California, Berkeley, Technical Report #576*.
Dudoit,S. *et al.* (2002) Comparison of discriminant methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.
Golub,T. *et al.* (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–536.
Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.
Khan,J. *et al.* (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.
Lee,J.W. *et al.* (2005) An extensive comparison of recent classification tools applied to microarray data. *Comput. Stat. Data Anal.*, **48**, 869–885.
Mardia,K., Kent,J. and Bibby,J. (1979) *Multivariate Analysis*. Academic Press, London.
Nguyen,D.V. and Rocke,D.M. (2004) On partial least squares dimension reduction for microarray-based classification: a simulation study. *Comput. Stati. Data Anal.*, **46**, 407–425.
Ramaswamy,S. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signature. *Proc. Nati. Acad. Sci. USA*, **98**, 15149–15154.
Ross,D. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.
Schena,M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
Scherf,U. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, **24**, 236–244.
Stein,C. (1956) Inadmissability of the usual estimator for the mean of a multivariate distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.*, **1**, 197–206.
Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci.*, **99**, 6567–6572.
Tibshirani,R. *et al.* (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.*, **18**, 104–117.
Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.
Zhu,J. and Hastie,T. (2004) Classification of gene microarrays by penalized logistic regression. *Biostatistics*, **5**, 427–443.