Diagnosis of multiple cancer types by shrunken centroids of gene expression

Robert Tibshirani†‡, Trevor Hastie§, Balasubramanian Narasimhan§, and Gilbert Chu¶

Departments of †Health, Research and Policy, and Statistics, §Statistics and Health, Research and Policy, and ¶Medicine and Biochemistry, Stanford University, Stanford, CA 94305

Communicated by Bradley Efron, Stanford University, Stanford, CA, February 19, 2002 (received for review October 10, 2001)

We have devised an approach to cancer class prediction from gene expression profiling, based on an enhancement of the simple nearest prototype (centroid) classifier. We shrink the prototypes and hence obtain a classifier that is often more accurate than competing methods. Our method of “nearest shrunken centroids” identifies subsets of genes that best characterize each class. The technique is general and can be used in many other classification problems. To demonstrate its effectiveness, we show that the method was highly efficient in finding genes for classifying small round blue cell tumors and leukemias.

The problem of class prediction has recently received a great deal of attention in the context of DNA microarrays. Here the task is to classify and predict the diagnostic category of a sample on the basis of its gene expression profile. A problem of particular importance is the diagnosis of cancer type based on microarray data. Conventional diagnosis of cancer has been based on examination of the morphological appearance of stained tissue specimens in the light microscope. This method is subjective and depends on highly trained pathologists. Microarrays offer hope that cancer classification can be objective and rely on highly trained pathologists. Microarrays offer hope that cancer classification can be objective and can depend on highly trained pathologists.

Methods

Let \( x_i \) be the expression for genes \( i = 1, 2, \ldots, p \) and samples \( j = 1, 2, \ldots, n \). We have classes \( 1, 2, \ldots, K \), and let \( C_k \) be indices of the \( n_k \) samples in class \( k \). The \( k \)th component of the centroid for class \( k \) is \( \bar{x}_k = \frac{\sum_{i \in C_k} x_{ij}}{n_k} \), the mean expression value in class \( k \) for gene \( i \); the \( k \)th component of the overall centroid is \( \bar{x}_i = \frac{\sum_{i=1}^{n} x_{ij}}{n} \).

In words, we shrink the class centroids toward the overall centroid after standardizing by the within-class standard deviation for each gene. This standardization has the effect of giving higher weight to genes whose expression is stable within samples of the same class. Such standardization is inherent in other common statistical methods such as linear discriminant analysis.

Let

\[
d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_{ik}(s_i + s_k)},
\]

where \( s_i \) is the pooled within-class standard deviation for gene \( i \):

\[
s_i^2 = \frac{1}{n - K} \sum_{k} \sum_{j \in C_k} (x_{ij} - \bar{x}_j)^2,
\]

and \( m_{ik} = \sqrt{1/n_k + 1/n} \) makes the \( m_{ik}s_i \) equal to the estimated standard error of the numerator in \( d_{ik} \). In the denominator, the value \( s_i \) is a positive constant (with the same value for all genes), included to guard against the possibility of large \( d_{ik} \) values arising...
by chance from genes with low expression levels. We set \( s_0 \) equal to the median value of the \( s_i \) over the set of genes. A similar strategy was used in the SAM methodology of ref. 7.

Thus \( d_{ik} \) is a \( t \) statistic for gene \( i \), comparing class \( k \) to the overall centroid. We rewrite Eq. 1 as

\[
\frac{x}{H_{6126}}_{ik} = \frac{x}{H_{6126}}_{i} + \frac{m_k}{H_{20849}}(s_i + s_0)d_{ik}.
\]

Our method shrinks each \( d_{ik} \) toward zero, giving \( d_{ik} \) and yielding shrunken centroids or prototypes

\[
\tilde{x}_{ik} = \tilde{x}_i + m_k(s_i + s_0)d_{ik}.
\]

The shrinkage we use is called soft thresholding: each \( d_{ik} \) is reduced by an amount \( \Delta \) in absolute value and is set to zero if its absolute value is less than zero. Algebraically, soft thresholding is defined by

\[
\tilde{d}_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)^+,
\]

where \(+\) means positive part (\( t = t \) if \( t > 0 \) and zero otherwise). Because many of the \( \tilde{x}_{ik} \) values will be noisy and close to the overall mean \( \tilde{x}_i \), soft thresholding usually produces more reliable estimates of the true means (8, 9).

This method has the desirable property that many of the components (genes) are eliminated from the class prediction as the shrinkage parameter \( \Delta \) is increased. Specifically, if for a gene \( i \), \( d_{ik} \) is shrunken to zero for all classes \( k \), then the centroid for gene \( i \) is \( \tilde{x}_i \), the same for all classes. Thus gene \( i \) does not contribute to the nearest-centroid computation. We choose \( \Delta \) by cross-validation, as illustrated below.

**Results**

**Choosing the Amount of Shrinkage.** Fig. 2 shows the training, cross-validation, and test errors for different values of the shrinkage parameter \( \Delta \). We used 10-fold cross-validation, dividing the set of samples at random into 10 approximately equal-size parts. The 10 parts were roughly balanced, ensuring that the classes were distributed proportionally among each of the 10 parts. Ten-fold cross-validation works as follows: we fit the model on 90% of the samples and then predict the class labels of the remaining 10% (the test samples). This procedure is repeated 10 times, with each part playing the role of the test samples and the errors on all 10 parts added together to compute the overall error (see ref. 5 for details). Fig. 2 shows the results, from no shrinkage (Left) to complete shrinkage (Right). Both the cross-validated and test error were minimized near \( \Delta = 4.34 \), which is the value we used to produce the red bars in Fig. 1. The upper axis shows the number of active genes with at least one nonzero component, \( d_{ik} \), for each value of \( \Delta \). At \( \Delta = 4.34 \), there are 43 active genes.

**The Genes That Classify SRBCTs.** Fig. 3 shows the shrunken differences \( d_{ik} \) for the 43 genes having at least one nonzero difference. Comparing these genes to the 96 genes identified in ref. 4, the two lists have 27 genes in common. Fig. 4 shows the heat map of our 43 genes. The horizontal rows of the map represent genes,
whereas the columns represent samples. Each pixel represents the expression of one gene in one experiment: the colors depict intensity (log expression ratio) from blue (large negative) to yellow (large positive). We used hierarchical clustering to order the genes within each of the horizontal partitions and to order the samples within each vertical partition.

Consider the $2,308 \times 4 = 9,232$ unshrunken t statistics $d_{ik}$. The 43rd largest in absolute value is 4.34 (there are about 43 nonzero components in the model). Using the Bonferroni adjustment, in the context of 9,232 t statistics, this value would be barely significant at the 0.05 level. This example illustrates another way in which our procedure differs from the more common approach of screening genes by significance of their individual t statistics.

Our method uses soft thresholding rather than screening and focuses on misclassification error, which is the relevant measure for class prediction problems.

**Class Probabilities and Discriminant Functions.** We classify test samples to the nearest shrunken centroid, again standardizing by $s_i + s_0$. We also correct for the relative number of samples in each class.

Suppose we have a test sample (a vector) with expression levels $x^* = (x^*_1, x^*_2, \ldots, x^*_p)$. We define the discriminant score for class $k$ as

$$
\delta_k(x^*) = \sum_{i=1}^p \frac{(x^*_i - \bar{x}^*_i)^2}{(s_i + s_0)^2} - 2 \log \pi_k. \tag{6}
$$

The first term in Eq. 6 is simply the standardized squared distance of $x^*$ to the $k$th shrunken centroid. The second term is a correction based on the class prior probability $\pi_k$, where $\sum_{k=1}^K \pi_k = 1$. This prior probability gives the overall frequency of class $k$ in the population. The classification rule is then

$$
C(x^*) = \ell \text{ where } \delta_{\ell}(x^*) = \min_k \delta_k(x^*). \tag{7}
$$

If the smallest distances are close and hence ambiguous, the prior correction gives a preference for larger classes, because they potentially account for more errors. We usually estimate the $\pi_k$ by the sample priors $\hat{\pi}_k = n_k/n$. If the sample prior is not representative of the population, either more realistic priors or equal priors $\pi_k = 1/K$ can be used.
We can use the discriminant scores to construct estimates of the class probabilities, by analogy to Gaussian linear discriminant analysis:

\[
p_{k}^{\hat{}}(x^{*}) = \frac{e^{-\frac{1}{2} h_{k}(x^{*})}}{\sum_{\ell=1}^{K} e^{-\frac{1}{2} h_{\ell}(x^{*})}}.
\]

Fig. 5 displays these probabilities for the training data (Upper) and test data (Lower). Most samples have good separation between the highest and next highest probability, demonstrating that the sample is unambiguously classified by the method.

**Leukemia Classification**

Leukemia data from high-density Affymetrix oligonucleotide arrays were previously analyzed in Golub *et al.* (1). There were 7,129 genes and 34 samples: 20 in class ALL (acute lymphocytic leukemia) and 14 in class AML (acute mylogenous leukemia). The results for the methods of Golub *et al.* and nearest shrunken centroids are shown in Table 1.

Golub *et al.* report a test error rate of 4/34 for their procedure (defined in the Appendix), using 50 genes to obtain their results. We found that if the number of genes was reduced to less than 47, the test error increased. The results for the nearest shrunken centroid method are shown in Fig. 6. The minimum cross-validation error occurs near \(\Delta = 1.4\) but leaves about 1,000 genes. In practice, this minimum error solution might be of interest. Here, to obtain a more manageable set of genes, we chose instead \(\Delta = 4.06\), the point at which the cross-validation error starts to rise quickly, yielding only 21 genes. Our method was superior to their procedure in both cross-validation error on the training data and test error.

**Discussion**

The method of nearest shrunken centroids was successful in finding genes that accurately predict classes. The method found a set of 43 genes that was able to assign SRBCTs to one of four classes, BL, EWS, NB, and RMS, with 100% accuracy. This result was superior to the neural network method of Khan *et al.* (4), which required 96 genes. Of our 43 genes, 27 were also found by the neural network. Thus, 69 of the genes from the neural network are not required for classification.

We also identified six genes not identified in ref. 4. Cold-shock domain protein A was underexpressed in NB. Farnesyl-diphosphate farnesyltransferase 1, neurofibromin 1, presenilin 2, homolog of mouse mesoderm specific transcript, and tissue plasminogen activator were overexpressed in RMS. It is interesting to note that cold-shock domain protein A is expressed in...
B cells and skeletal muscle but not in brain, consistent with our finding that it is expressed 6-fold less in NB compared to BL, EWS, and RMS. However, of the five genes with increased expression in RMS, neurofibromin 2 and tissue plasminogen activator are generally not expressed in muscle or heart tissue, and presenilin is expressed in brain as well as muscle and heart. Thus, genes that were predictive of a specific tumor did not always reflect the expression levels in the tissue from which the tumor was derived.

It is interesting to compare our list of 43 genes with genes that are presently considered to be diagnostic for SBRCTs. Several gene products are measured by immunostaining to distinguish the SBRCTs from each other: common leukocyte antigen (CD45) is specific for BL; MIC2 (CD99) is specific for EWS (10); chromogranin A and synaptophysin are specific for NB (11); and desmin, muscle-specific actin, myogenin, and MyoD1 are used for RMS (12, 13). MIC2 is on our list of genes (Fig. 3), but the other 7 genes are not. In addition, several genes play oncogenic roles in the SBRCTs. C-myc is activated as a result of chromosomal translocation to one of the immunoglobulin loci in BL. N-myc is commonly amplified in NB (14). One of the fusion proteins EWS/FLI-1 or EWS/ERG is overexpressed from chromosomal translocation in EWS (15). One of the fusion proteins PAX3/FKHR or PAX7/FKHR is expressed in RMS (16). We could find only three of the aforementioned genes in the cDNA microarray of ref. 4; their data are displayed in Fig. 4 to demonstrate their shortcomings in comparison to the genes identified in ref. 1. Our list of 43 genes included myeloperoxidase and barely missed terminal deoxynucleotidyl transferase, which were not identified in ref. 1 but are known to be excellent markers for AML and ALL, respectively.

One goal of our method is to find the smallest set of genes that can accurately classify samples. The efficiency of our method in finding a relatively small number of predictive genes will facilitate the search for new diagnostic tools. There is often significant correlation in the expression patterns of groups of genes in an experiment. Hence, after a minimal gene list is found, it may be useful to search for other genes that are highly correlated with the genes in this list. Gene interaction is another important issue. Because genes often work in biological pathways, there could be strong interactions between genes in the same pathway. Careful post hoc analysis will be required to discover such interactions.

The success of our methodology has implications for improving the diagnosis of cancer. The method efficiently finds and ranks genes that can distinguish one type of tumor from another. Ultimately, it may be used to search for genes that are predictive for response to chemotherapy. For SBRCTs and leukemias analyzed here, the predictive genes are attractive candidates for raising antibodies suitable for immunostaining. Immunohistochemistry has an advantage for analyzing difficult specimens, because it allows the pathologist to localize the stain to tumor cells. In addition, our results suggest that RNA-based diagnostic tests may soon become feasible, based on either small-scale microarrays or quantitative PCR.

Appendix

Relationship to Other Approaches. The discriminant scores in Eq. 6 are similar to that used in linear discriminant analysis (LDA), which arises from using the Mahalanobis metric in computing distance to centroids:

\[ d_k^{\text{LDA}}(x^*) = (x^* - \bar{x}_k)^TW^{-1}(x^* - \bar{x}_k) - 2 \log \pi_k. \]  

Here we are using a vector notation, and \( W \) is the pooled within-class covariance matrix. With thousands of genes and tens of samples (\( p \gg n \)), \( W \) is huge, and any sample estimate will be singular (hence its inverse is undefined). Our scores can be seen to be a heavily restricted form of LDA, necessary to cope with the large number of variables (genes). The differences are that:

- we assume a diagonal within-class covariance matrix for \( W \), without which this LDA would be ill-conditioned and would fail;
- we use shrunken centroids rather than simple centroids as a prototype for each class;
- as the shrinkage parameter \( \Delta \) increases, an increasing number of genes will have all their \( d_k^* = 0, k = 1, \ldots, K \), because of the soft thresholding in Eq. 5. Such genes contribute no discriminatory information in Eq. 6 and in fact cancel in Eq. 8.

Refs. 1 and 2 present linear scoring procedures for the two-class problems, which are very similar to each other and also to our procedure and hence LDA in the two-class case. Both methods work essentially by:

- selecting a set of genes based on a two-sample \( t \) statistic, and then
- forming a linear classifier in the selected \( x_i^* \) using these \( t \) statistics as weights.

We describe the procedures in turn, using our own re-expression and terminology but remaining faithful to the original in each case. Ref. 2 uses a standard \( t \) test at level \( \alpha \) to select differentially expressed genes in two breast cancer types. They then construct a “compound covariate” class predictor, which they threshold to classify a sample into one of the two classes. Their compound covariate is

\[ h(x^*) = \sum_{i \in S(\alpha)} t_i x_i^*. \]  

where \( x_i^* \) is the gene expression vector for the sample to be classified, \( t_i \) is the \( t \) statistic for comparing the two classes in the training data, and \( S(\alpha) \) is the subset of genes with a significant \( t \) statistic at level \( \alpha \). They use as a threshold the same compound formula applied to \( (\bar{x}_1 + \bar{x}_2)/2 \). Here
Here it is clear that if both genes are shrunk, differences, after a little manipulation, can be seen to be a linear function, we can form the difference and threshold at zero. These weighted vote, our selection is based on soft thresholding of t statistics. Soft thresholding is known to perform better in other settings (see, e.g., ref. 17).

We use cross-validation to select α, and Golub et al. (1) use it to select their m. We could not determine whether Hedenfalk et al. (2) use cross-validation to select α, although it would be reasonable to do so.

Our procedure can adjust the classification according to the class prior probabilities.

Our procedure provides estimates of probabilities in Eq. 8 for each of the classes.

Our nearest prototype procedure works for more than two classes.

With more than two classes, the method of nearest shrunken centroids uses soft thresholds for all of the differences between the class centroids and the overall centroid. In the process, it chooses different sets of genes for characterizing each class, as illustrated in Fig. 3.

The neural network approach (4) can be interpreted as a form of damped discriminant analysis. In that paper, Khan et al. actually use a linear network, using the first 10 principal components (eigengenes). In statistical parlance, this is known as principal components regression and requires no iterative learning procedure or learning curves. Principal component regression is a hard-threshold version of ridge regression. Khan et al. (4) do use a model-averaging procedure, similar to “bagging” (18) but based on 3-fold cross-validation, to regularize the procedure further. Although their procedure also produced zero training and test errors on the SRBCT data, it is far more complex than nearest shrunken centroids. With so many genes and so few samples, it is very likely that restricted versions of simpler statistical methods will do as well as or better than neural networks, as is the case here.

Software and Computational Details. The computations involved in the shrunken nearest centroid method are straightforward. We have developed a program similar to the popular sam package to implement nearest shrunken centroid classification. This program incorporates methods for automatic threshold choice and graphical methods for application of the procedure to the results of an unsupervised clustering procedure. Details may be found at http://www-stat.stanford.edu/~tibs/PAM.

We thank two referees for their helpful comments. R.T. was supported by National Institutes of Health (NIH) Grant 2 R01 CA72028 and National Science Foundation (NSF) Grant DMS-9971405. T.H. was supported by NSF grant DMS-9803645 and NIH grant ROI-CA-72028-01. G.C. was supported by National Cancer Institute Grant 5R01 CA77302 and a Burroughs-Wellcome Clinical Scientist Award.