

The Application of Rule-Based Methods to Class Prediction Problems in Genomics

GEORGE MICHAILIDIS and KERBY SHEDDEN

ABSTRACT

We propose a method for constructing classifiers using logical combinations of elementary rules. The method is a form of rule-based classification, which has been widely discussed in the literature. In this work we focus specifically on issues that arise in the context of classifying cell samples based on RNA or protein expression measurements. The basic idea is to specify elementary rules that exhibit a locally strong pattern in favor of a single class. Strict admissibility criteria are imposed to produce a manageable universe of elementary rules. Then the elementary rules are combined using a set covering algorithm to form a composite rule that achieves a perfect fit to the training data. The user has explicit control over a parameter that determines the composite rule's level of redundancy and parsimony. This built-in control, along with the simplicity of interpreting the rules, makes the method particularly useful for classification problems in genomics. We demonstrate the new method using several microarray datasets and examine its generalization performance. We also draw comparisons to other machine-learning strategies such as CART, ID3, and C4.5.

Key words: class prediction, genomics, gene expression, set covering.

1. INTRODUCTION

THE SIGNIFICANCE OF RNA AND PROTEIN EXPRESSION as fundamental processes in physiology, development, and pathology has been recognized for decades. Measurements of protein and RNA expression based on technologies such as hybridization to labeled probes, RT-PCR, RNase protection, Southern blots, and immunostaining have been commonplace in laboratories for a number of years. While the experimental analysis of gene and protein expression is not a new area, there have recently been a number of important technical advances, and the use of quantitative tools for analyzing the experimental data is undergoing a rapid expansion. In this paper, we describe a rule-based approach for addressing one of the key quantitative analysis problems in expression genomics: the class prediction problem.

The class prediction problem involves the prediction of a qualitative characteristic of a sample of cells based on expression assays covering large numbers of genes or proteins. For example, it may be of interest to predict the precursor cell type for a particular tumor, or whether a tumor is on the verge of becoming metastatic. The problem is difficult for several reasons. One reason is that the samples assigned to a single class may nevertheless exhibit a fair amount of heterogeneity. Part of this heterogeneity arises from measurement error. However, a substantial fraction is likely to be biological in origin, for instance, due to

tissue being acquired from patients with different disease stages, or exhibiting different levels of immune response. Another difficult feature is that, in general, there may be distinct mechanisms involving different genes that produce the same observable characteristic, and a single gene may play a role in determining many distinct characteristics. In particular, for the majority of problems, no single gene will serve as a universal marker for the characteristic of interest, so any useful decision procedure must be multivariate, in that it must rely on a number of genes.

Many of the popular methods for carrying out automated classification involve taking linear combinations of the features, which are gene or protein expression measurements in this case (see Ripley [1996] for an overview). While such methods can be powerful and may provide a great deal of insight in certain settings, they often will have a tendency to produce linear combinations of genes that are hard to interpret. Additionally, many of the popular “flexible” classification procedures, such as the nearest-neighbor and neural-network approaches, are widely considered to produce rules that are of a black-box nature and are also difficult to interpret. The rule-based classifiers that we describe in this paper are easy to interpret, as they resemble the kinds of decision procedures that trained experts, such as pathologists, have used for many years with assay systems such as immunostaining.

In this paper, we discuss a new mechanism for constructing a transparent class of rule-based classifiers, placing a special emphasis on issues that are specific to the problem of classifying cell samples based on gene expression. We demonstrate the procedure using two publicly available gene expression datasets, and we draw comparisons to other approaches for carrying out rule-based classification, both in terms of general properties and in terms of their suitability for the class prediction problem in genomics.

2. THE CLASSIFICATION PROBLEM

The expression data that we work with can be represented by an $n \times d$ matrix $X = [\mathbf{x}_1 \cdots \mathbf{x}_d]$, where $\mathbf{x}_j \in \mathcal{R}^n$ denotes the expression levels for the n genes in sample j and d is the number of training samples that are available. In addition, each sample is accompanied by a class label y_j that takes on a value in a set having K elements. In practice, the classes may correspond to the type or severity of disease, the type of treatment intervention, or the type of tissue that is affected, among other possibilities.

A classification rule is a function that predicts y_j based on \mathbf{x}_j . Such rules are constructed from a *training* dataset where an expert has assigned each sample to a particular class. The quality of the classification rule is assessed by applying it to an independent *test* dataset where the class labels are also known. One then compares the predicted and true class labels to estimate the error rate of the classifier.

Our classification rules are constructed in two stages. First, we identify a subset of genes such that either high or low expression of each gene in the subset implies strong evidence in favor of one class over the others. These genes define a set of *elementary rules*. At the second stage, these elementary rules are combined to generate *composite rules* that achieve a specified level of fit to the training set. Given that in the gene expression setting there are many more features (genes) than samples, one is virtually assured of obtaining a large number of composite rules that fit the training set to this specified level. Thus, for parsimony purposes, we require a composite rule to be minimal in size subject to meeting the fitness requirement to be specified below.

2.1. Identification of the elementary rules

We construct classification rules using logical combinations of the following elementary rules:

$$\begin{aligned} R_+(T, i) &= \mathcal{I}(\text{expression of gene } i \text{ is greater than } T) \\ R_-(T, i) &= \mathcal{I}(\text{expression of gene } i \text{ is less than } T). \end{aligned} \tag{1}$$

Letting $R_*(i, T)$ denote a rule of either of the two types, we consider an elementary rule for use in classification only if it satisfies the following admissibility criterion:

$$A(k) : R_*(T, i) = 1 \Rightarrow \text{class is } k.$$

We note that if there are K classes, then there are K distinct admissibility criteria.

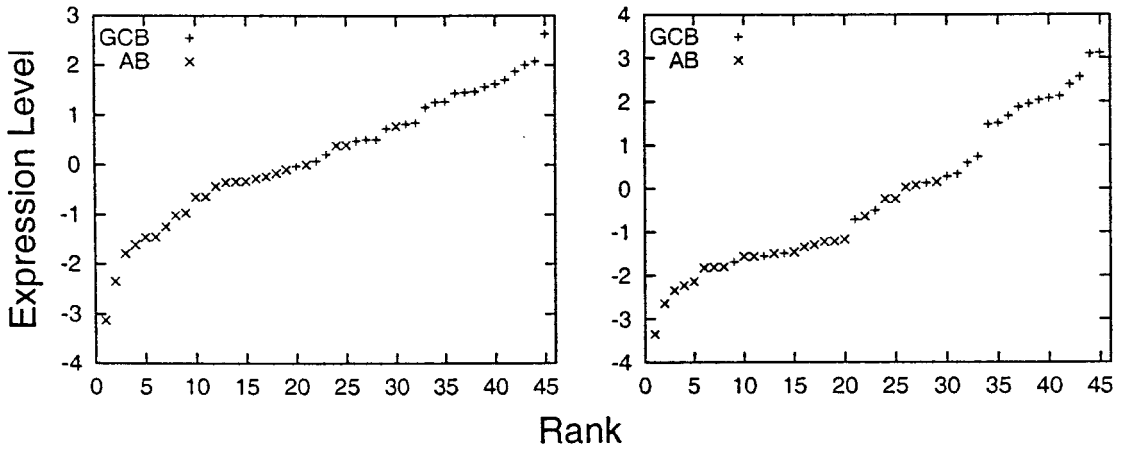


FIG. 1. Two examples of admissible rules using the lymphoma dataset. The **left** panel shows a rule $R_{-}(-.03, JAW1)$ that is admissible for the AB class based on the *JAW1* gene. The **right** panel shows a rule $R_{+}(.23, FMR2)$ that is admissible for the GCB class based on the *FMR2* gene.

Given a rule that satisfies admissibility criterion $A(k)$, we will say that a sample in class k is *covered* by the rule if the sample satisfies the rule. An alternative way of developing the elementary rules is to consider consecutive runs of samples that fall into a common class. An elementary rule is $A(k)$ -admissible if there is a run of samples in class k covered by the rule that includes one of the two samples taking on an extreme value. The threshold can be set to any value between the level of the innermost sample in the run and the level of the adjacent sample that does not belong to class k . Figure 1 shows two genes from the lymphoma dataset (introduced below in Section 3) that determine admissible elementary rules. The expression values are sorted from left to right. The abscissa for each point is the rank of the sample, and the ordinate is the expression level.

Every gene produces two admissible rules, one of the form R_{+} and one of the form R_{-} . The coverage for each rule must be at least 1, and it will be exactly 1 if the second-highest-expressing sample is of a different class than the highest-expressing sample (or analogously if the second-lowest-expressing sample is of a different class from the lowest-expressing sample in the case of R_{-} rules). The vast majority of the rules with coverage 1 or with very low coverage will be spurious. Therefore, we impose an additional admissibility constraint that requires each elementary rule to cover a minimal number of samples. Let N_k denote the number of samples in class k for $k = 1, \dots, K$. If there is no association between the expression of gene i and the characteristic, then the probability that a rule $R_{*}(T, i)$ is $A(k)$ -admissible and covers at least c samples in class k is given by

$$Q_c(k) = \frac{N_k(N_k - 1) \cdots (N_k - c + 1)}{d(d - 1) \cdots (d - c + 1)}.$$

For each class k , we require a minimal coverage c so the $Q_c(k)$ is smaller than a user-specified threshold. For the examples in this article, we use the familiar setting of .01.

2.2. Construction of Composite Rules

We make use of the following obvious fact to construct complex rules out of the elementary rules: disjunctions of rules satisfying $A(k)$ also satisfy $A(k)$. Therefore, if we consider a sufficiently long disjunction of rules satisfying $A(k)$, then we will eventually cover all samples in class k . This is not a mathematical fact, but it is virtually assured in our setting given the large number of elementary rules that are available. One could even proceed further to the point where each sample is covered $\kappa > 1$ times. In this case, we call the value κ the *covering multiplicity*. The use of disjunctions to represent rules for class prediction is reminiscent of the strategy followed by pathologists in the clinic, where a minimal number of features out of a specified universe of features is considered sufficient to make a prediction.

Each elementary rule covers a certain subset of samples, and the goal is to select a set of rules such that each sample is covered by a specified number κ of rules. More specifically, suppose we have m_k elementary rules that satisfy admissibility criterion $A(k)$ for $k = 1, \dots, K$, and let γ_{ijk} be the indicator that sample j is covered by the i^{th} $A(k)$ -admissible rule. The construction of a composite rule involving a minimal number of elementary rules corresponds to a set covering problem that can be formulated as the following integer program:

$$\begin{aligned} &\text{minimize } \sum_{k=1}^K \sum_{i=1}^{m_k} \delta_{ik} \\ &\text{s.t. } \sum_{k=1}^K \sum_{i=1}^{m_k} \gamma_{ijk} \delta_{ik} \geq \kappa, \quad j = 1, \dots, d \\ &\quad \delta_{ik} \in \{0, 1\}. \end{aligned} \tag{2}$$

Note that the first constraint ensures that each sample has the specified coverage multiplicity, while the second constraint imposes the integrality condition that indicator functions must satisfy. The objective function is formulated so that we minimize the number of rules that are involved, subject to achieving the specified coverage. It is also worth noting that one can introduce weights into the objective function that may give preference to certain of the elementary rules. For example, rules that cover many samples or that correspond to previously validated patterns may receive a higher weight.

From a computational point of view, Problem (2) is NP-complete (Papadimitriou, 1994), but a variety of good heuristics are available. For the problem at hand, we adopt a greedy approach based on forward selection to produce a sequence of elementary rules such that the composite rule has covering multiplicity κ . Let $C(j; r_1, \dots, r_\ell)$ denote the number of times that sample j has been covered using elementary rules r_1, \dots, r_ℓ . The next rule to be selected must be the solution to the following,

$$r_{\ell+1} = \operatorname{argmax}_r \sum_j \min\{C(j; r_1, \dots, r_\ell, r), \kappa\},$$

where r ranges over all admissible elementary rules, and j runs over the samples. That is, the next rule to be selected must make the greatest possible progress towards achieving the specified coverage multiplicity for all of the samples.

The forward selection procedure is carried out until the coverage multiplicity κ is achieved, giving as its final product a composite rule that is formed as a disjunction of a certain set r_1, \dots, r_h of elementary rules. Let $\delta(j)$ denote the class such that r_j satisfies admissibility criterion $A(\delta(j))$. In order to make a decision on a new sample Z , we allow the r_j to vote, with the votes weighted by the number of elementary rules corresponding to each class. That is, if there is a unique class k that maximizes

$$V(k) = \sum_j r_j(Z) \mathcal{I}(\delta(j) = k) / \sum_j \mathcal{I}(\delta(j) = k),$$

then the sample is predicted to belong to class k . If there is no such unique class, then the sample identity is considered to be indeterminate.

In general, there will be a large number of admissible elementary rules and an enormous number of composite rules that satisfy the coverage multiplicity κ . Therefore, it becomes computationally challenging to identify the composite rule of coverage multiplicity κ with minimal support (i.e., the rule involving the smallest possible number of elementary rules). The strategy that we adopt is to begin the forward selection procedure at each of the n genes, giving n composite rules with the required coverage. Among these rules, there is a rule with minimal support that is comprised of h elementary rules. We retain all of the composite rules produced by the forward selection procedure that have size equal to h .

As will be seen below, the rules that are generated will not even come close to using all of the information that is available (they will refer to only a tiny fraction of the genes). When only tens or a few hundred samples are available, the minimal support requirement alone will produce rules that can be considered to be too simple. Our second requirement, namely, that the rule must cover each sample several times,

is a way to mitigate this circumstance. The combination of the minimality and coverage multiplicity requirements is designed to achieve a balance between parsimony and the need to achieve a sufficient level of redundancy so that correct class predictions can be made in the presence of measurement error and expression heterogeneity.

3. EXAMPLE: PREDICTING SUBTYPES OF LYMPHOMA

Alizadeh *et al.* (2000) described a set of experiments using cDNA microarrays that identified two putative subtypes of diffuse large B-cell lymphoma (DLBCL). The subtypes were discovered by applying a hierarchical clustering technique to 97 samples that included 45 DLBCL samples, 20 samples from two other clinically distinct lymphoid malignancies, and 32 samples obtained from normal and transformed cell cultures of various types associated with the immune system (resting and activated B cells, T cells, tonsil and lymph node derived cells). Based on the hierarchical clustering, the 45 DLBCL cases were subdivided into 22 cases that were identified as having the “germinal center B” form of the disease (GCB-DLBCL) and 23 cases that were identified as having the “activated B” form of the disease (AB-DLBCL).

Treating the GCB-DLBCL/AB-DLBCL designations as fixed, we generated rule-based classifiers for the class prediction problem using the procedure described in Section 2. We used the 2,983 transcripts that are shown in Fig. 3b of Alizadeh (2000), for which the raw expression values can be obtained from *llmpp.nih.gov/lymphoma*. Using rules admissible under $A(1)$ (GCB) or $A(2)$ (AB) with coverage multiplicity $\kappa = 2$ and minimal coverage satisfying $Q_c(k) = .01$ for $k = 1, 2$, the smallest rules involved nine genes. There were 30 distinct rules having this minimal size. Although these 30 rules have “positions” for $30 \cdot 9 = 270$ genes, only 41 distinct genes occur. Six genes occurred in every one of the 30 composite rules. Eleven genes occurred in only a single composite rule.

It is easy to understand the structure of the composite rules through a graph such as is shown in Fig. 2. In the graph, each column corresponds to a sample, and each row corresponds to a gene. When an $R_+(\cdot, \cdot)$ rule covers a sample, then a “+” is placed at the point corresponding to the gene/sample pair. Similarly a “-” is used to indicate coverage by an $R_-(\cdot, \cdot)$ rule. It is easy to verify that every sample is covered twice and that, if any of the 9 genes were omitted, then at least one sample would be covered only once. Furthermore, we gain some insight into the way that the rule uses the expression information to produce the class prediction. For example, the genes Hs.186709 and JAW1 cover most of the AB cases twice; however, they miss samples 40 and 42 completely and cover samples 23, 29, 41, and 45 just once. The mRNA for T-cell tyrosine phosphatase and Hs.190288 fills in the missing coverage.

Using Fig. 2, we can characterize the relationship between certain pairs of genes as being “complementary” or “redundant.” For classification purposes, Hs.186709 and JAW1 are redundant. Since they covary

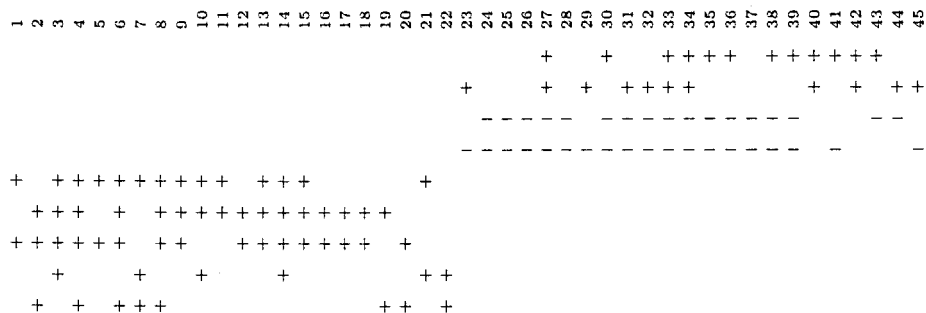


FIG. 2. A composite rule with coverage multiplicity $\kappa = 2$ for the lymphoma data. Each row corresponds to a gene, and each column corresponds to a sample. Columns 1–22 correspond to GCB-DLBCL cases, while columns 23–45 correspond to AB-DLBCL cases. The genes, listed from top to bottom, are T-cell protein-tyrosine phosphatase, Hs.190288, Hs.186709, JAW1, clone 1357367, FMR2, Hs.192708, diacylglycerol kinase delta, clone 1354788. A “+” indicates high expression for the corresponding gene/sample pair, while a “-” indicates low expression. A lack of a symbol indicates that there is no tendency for the gene/sample pair to give consistently high or low expression.

so closely, we would care to measure both genes only as a protection against measurement error. Both genes are being selected by our method since a coverage multiplicity of $\kappa = 2$ is imposed, reflecting a user-specified preference for a certain level of redundancy. On the other hand, Hs.190288 and Hs.186709 are fairly complementary. Although they cover a number of samples in common, each covers several samples that the other does not cover. The discrepancies between the expression levels of these two genes seem to be more than just measurement error. Rather, there may be a biological difference between AB cases expressing Hs.190288 and AB cases expressing Hs.186709 that requires both genes to be measured in order to achieve a good prediction.

4. EXAMPLE: BREAST TUMORS

Perou *et al.* (2000) described a study in which gene expression for 9,216 genes was measured on 55 breast tumor tissue samples using microarrays. The tumors were assigned to a mitotic grade of 1, 2, or 3, indicating the rate of tumor proliferation (higher numbers correspond to greater proliferation). We used the mitotic grade as the qualitative characteristic and built composite rules using our procedure. This demonstrates the natural way in which our procedure can handle problems with more than two classes. Using coverage multiplicity $\kappa = 1$ and minimal coverage satisfying $Q_c(k) = .01$ for $k = 1, 2$, the smallest rules used 12 genes. There were only four distinct rules of minimal size produced by the forward selection procedure. Figure 3 shows one of these rules. This rule uses four genes to cover each of the three mitotic grade levels.

5. GENERALIZATION ERROR RATE

A proposal for a new classifier is often evaluated primarily based on its generalization error rate. This quantity is obtained through a theoretical analysis when possible. However, for most classifiers such an approach is not tractable, in which case the issue is usually addressed empirically through cross-validation studies (Ripley, 1996). In order to evaluate the prediction error rate for our rule-based method, we carried out a cross-validation study using the lymphoma data that was discussed in Section 4. We note that for our rule-based procedure the *apparent* misclassification error rate (proportion of misclassified samples in the training data) is zero by construction.

Each of the 45 samples was omitted from the data set in turn, and a set of rules was obtained by applying the methods of section 2 to the remaining samples using coverage multiplicities $\kappa = 2$ and $\kappa = 10$. Denote by Q_{-j} the number of rules that are obtained when omitting sample j . Each of these Q_{-j} composite rules was applied to sample j using the voting procedure described above. Let ρ_j denote the proportion of the Q_{-j} rules that assigned sample j to the correct class. The average of the ρ_j over the samples estimates the misclassification error rate for a rule that is selected at random from the rules that are of minimal size subject to achieving the specified coverage multiplicity. The estimated error rates are reported in the first two rows of Table 1.

We compared our procedure to a nearest neighbor classifier in order to assess whether comparable results could be obtained using a simpler method. In order to produce a meaningful comparison, we took into account the design constraint that our method must produce rules that rely on a small number of genes. The intent of this constraint is to enable the production of simple and transparent rules, even at a moderate cost in predictive power. For comparison, we considered nearest-neighbor classifiers using 9 genes and 21 genes, which are the numbers of genes used by the rule-based procedure with coverage multiplicities $\kappa = 2$ and $\kappa = 10$. It is relevant to note that at present routine clinical immunohistochemistry work-ups generally use fewer than five expression assays to make a prediction.

The nearest-neighbor classifier is defined by three steps: (1) standard two-sample t-tests were computed for each gene, and the $N = 9$ or $N = 21$ genes with smallest p-value were retained, (2) the correlation coefficient between the held-out sample and each of the remaining samples was computed using the N selected genes, (3) the classifications of the five samples with the greatest correlation coefficient were considered, with the majority vote determining the class prediction for the held-out sample. The results are shown in rows three and four of Table 1.

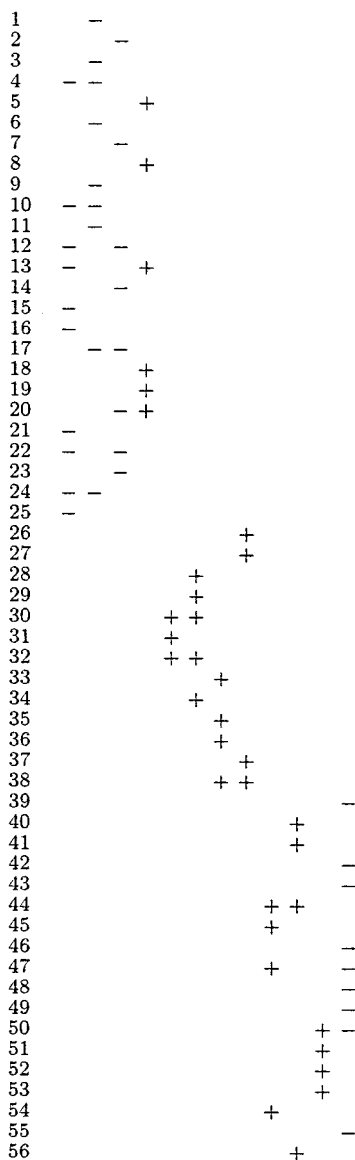


FIG. 3. A coverage rule with multiplicity $\kappa = 1$ for the breast tumor data. Samples 1–25 have mitotic grade 1, samples 26–38 have mitotic grade 2, and samples 39–56 have mitotic grade 3. The genes, listed from left to right, are PIR121, (EST), (EST), (EST), TLH6, YWHAH, HNRPC, (EST), SGCA, AA609920, SMARCC2, (EST).

TABLE 1. SUMMARY OF THE GENERALIZED MISCLASSIFICATION ERROR RATES FOR THE RULE-BASED APPROACH (RB) AND THE NEAREST-NEIGHBOR APPROACH (NN)^a

<i>Method</i>	<i># Genes</i>	<i>Error rate</i>
RB	9	.19
RB	21	.06
NN	9	.35
NN	21	.04

^aThe rule-based approach was applied using coverage multiplicities of $\kappa = 2$ and $\kappa = 10$, which lead to rules involving 9 or 21 genes. The nearest-neighbor approach was applied as described in the text using either 9 or 21 genes.

For the smaller rule ($\kappa = 2$ or $N = 9$), the nearest-neighbor classifier performs substantially worse than the rule-based method, while for the larger rule ($\kappa = 10$ or $N = 21$), the error rates are comparable. We believe that this reflects a general point of strength for our method when working with very small rules. This point of strength derives from the issue of “complementarity” versus “redundancy,” which was discussed above. By requiring every sample to be covered, we implicitly enforce a minimum degree of independence across the genes that are included in the composite rule. Specifically, given a set of genes that are very highly differentially expressed, but that provide essentially the same information in terms of sample coverage, at most κ members of the set will be included. The remainder of the rule must be built from complementary genes. These complementary genes help to produce correct predictions for some of the less common variants of a class of samples.

In contrast, in order to build a parsimonious rule using a classifier that does not possess a built-in method for variable selection (such as the nearest-neighbor approach), the variable selection must be done in a separate stage. The most commonly used methods for this second stage are marginal in nature, such as the t-test procedure described above. These first-stage variable selection procedures will tend to select genes that are highly correlated and that are strong predictors for the majority of the samples in a class, but that do not provide additional information for the more rare variants. This is consistent with the finding that as the rules grow larger, the rule-based method and nearest neighbor method become comparable. We can speculate that certain genes that fall in the bottom half of the top 20 genes based on the t-test are decisive in correctly predicting the classifications for about 20% of the samples that might be somewhat less prototypical.

We have claimed that it is gene selection rather than the details of the rule construction method that controls the error rate in the case of the smaller set of genes. Further evidence for this conclusion follows by using the set of genes selected by the rule-based procedure to drive the nearest neighbor algorithm. In this case, we achieved an error rate of .17, which is comparable to the error rate of .19 achieved by the pure rule-based procedure.

Certainly, an alternative to the t-test could be used to select genes for use with the nearest neighbor approach that would provide for greater independence among the genes that determine the neighborhoods. Some alternative methods for gene selection are discussed by Ramaswamy *et al.* (2001) and Dudoit *et al.* (2002). It is quite possible that error rates that are even better than those obtained for the rule-based procedure could be obtained using a carefully constructed method. Our assertion is simply that the proposed rule-based procedure produces simple and transparent rules, with error rates that are comparable to or lower than those of other simple methods. As is pointed out by Dudoit *et al.* (2000), the identification of “marker” genes to be used in classification procedures remains a very important issue. In our approach, the derivation of elementary rules and the construction of the minimal composite rule automates and integrates the processes of gene selection and rule construction in a natural way.

6. RELATIONSHIP TO TREE-STRUCTURED METHODS

The tree-structured methods such as CART (Breiman *et al.*, 1984), ID3, and C4.5 (Quinlan, 1993) are three of many possible alternatives to our rule-based method. In this section, we briefly describe the relationship between our proposal and these widely used classification procedures.

The tree-structured methods produce rules that are described by binary trees. At each node of the tree, a rule of the form (1) is applied. A number of algorithms have been devised for fitting trees to data. Finding the optimal fit is known to be NP-hard (Grigni *et al.*, 2000), just like the set covering problem that arises in our approach. Therefore, in practice, heuristic fitting procedures are used. Like the set covering strategy described above, the usual methods for fitting trees are sequential and greedy, adding a new split at each step to maximize the homogeneity at each node.

One obvious difference between our method and the tree-structured methods is that the latter have no analogue to our admissibility criteria $A(\cdot)$. These admissibility criteria require that each elementary rule in our method must make a definitive decision about a certain subset of the training samples. Specifically, the samples that are covered by an elementary rule that is $A(k)$ admissible are known to belong to class k without reference to the other rules. The tree-structured methods, on the other hand, use an entropy criterion to define the splits. This criterion refines the classification at each step, but until the final step a definitive call cannot be made for any of the training samples.

A more important difference between our procedure and the tree-structured methods is that the latter methods apply the constituent rules sequentially in defining the composite rule. The specific rule that is applied at the k^{th} step (i.e., at the k^{th} level of the tree) depends on the result of the rule that was applied at step $k - 1$. The rules for our method can be applied in any order, or in parallel. There is no interaction between the rules, as the result depends on a simple majority vote.

The aspects of the tree-structured methods that differ from our proposal allow the former methods to be more flexible. In many contexts, this will work to the advantage of the tree-structured methods, in particular if the feature space is small and the classes are hard to separate. In the genomics context, however, the feature space is enormous, and the classes are always linearly separable. Thus, the ability to obtain simpler rules using a more constrained approach may be valuable.

We note that a compromise between our method and the tree-structured methods arises out of the various proposals for growing many small trees and then using a majority vote of the trees to determine the class prediction (see, for example, Breiman [1996]). If this idea is applied to stumps (trees with a single split), then there is a very close connection with our method, with the stumps playing the role of the elementary rules.

7. DISCUSSION

In this paper, we have introduced a rule-based method for carrying out class prediction that produces compact, transparent rules with competitive error rates. We evaluated the procedure by applying it to two gene expression datasets that have been widely discussed in the literature. An earlier work (Michailidis and Shedden, 2000) presented more examples, including a widely discussed leukemia study (Golub *et al.*, 1999).

Recently, a number of different classification methods have been applied to classification problems involving gene expression measurements. For a comparative study, see Dudoit *et al.* (2002). Many of the initial results are quite promising, although it is still unclear how much information the data actually carry about different characteristics, and which (if any) methods are achieving near-optimal results. Since many of the methods are applied in their generic form, it is expected that there is a great deal of room for improvement by considering the special nature of the problem. This improvement may take the form of better error rates, or it may arise through the construction of rules that are simpler or more interpretable, or that address certain preferences that are specific to the domain of genomics. Our proposal emphasizes the latter goal.

While microarrays permit large numbers of genes to be considered during the discovery phase, in practical applications such as clinical diagnosis, a cheaper assay that covers far fewer genes will be used. Therefore, there is substantial interest in achieving reliable class prediction using only a few expression measurements. On the other hand, a certain amount of redundancy must be built into the procedure in order to provide some robustness to measurement error and expression heterogeneity. The procedure that we have described allows the user to balance in a natural way the opposing goals of obtaining a compact rule and having a desired level of redundancy.

It is generally accepted that from the point of view of prediction error rates, it is preferable to retain all of the features and subject the model to a regularization or shrinkage, rather than to drop many of the features completely (Copas, 1983). Therefore, it is quite possible that our method, which ignores nearly all of the features, will not be able to achieve as low an error rate as would be achieved by a method that can retain hundreds of genes. However, there continues to be an emphasis on developing transparent and small classifiers for use in class prediction with gene expression data. Our method performs well from the prediction point of view when compared to other methods that use very few genes.

Note: A C-language program that implements our procedure is available from the second author.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the University of Michigan Bioinformatics Program, Pfizer Global Research and Development, Ann Arbor, and the Howard Hughes Medical Institute. GM is supported by NSF grant IIS-9988095.

REFERENCES

- Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson Jr, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., and Staudt, L.M. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511.
- Breiman, L. 1996. Bagging predictors. *Machine Learning* 24, 2, 123–140.
- Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and Regression Trees*, Wadsworth, Belmont, CA.
- Copas, J.B. 1983. Regression, prediction, and shrinkage. *J. R. Statist. Soc., Series B.* 45, 3, 311–354.
- Dudoit, S., Fridlyand, J., and Speed, T.P. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA* 97(457), 77–87.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., 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.
- Grigni, M., Mirelli, V., and Papadimitriou, C.H. 2000. On the difficulty of designing good classifiers. *SIAM J. Computing* 30, 1, 318–323.
- Michailidis, G., and Shedden, K. 2000. A minimal support set approach to structure discovery and classification in large scale genomics and proteomics databases. In Technical Report #537, University of Michigan Department of Statistics.
- Papadimitriou, C.H. 1994. *Computational Complexity*, Addison-Wesley, Reading, MA.
- Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lonning, P.E., Borresen-Dale, A.L., Brown, P.O., and Botstein, D. 2000. Molecular portraits of human breast tumours. *Nature* 406, 747–752.
- Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufman, San Mateo, CA.
- Ramaswamy, S., Tamayo, P., Rifkin, R., Mukerjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., and Golub, T.R. 2001. Multiclass cancer diagnosis using tumor gene expression signatures, *PNAS* 98(26), 15149–15154.
- Ripley, B.D. 1996. *Pattern Recognition and Neural Networks*, Cambridge University Press, London.

Address correspondence to:
Kerby Shedden
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1027

E-mail: kshedden@umich.edu