Computer-Intensive Methods in Classification

William Shannon
Dept. of Biostatistics, Washington University
Campus Box 8005, 660 South Euclid Ave.
St. Louis, MO 63110, U.S.A.
wshannon@im.wustl.edu

David L. Banks
Bureau of Transportation Statistics, U.S. DOT
400 7th Street, SW
Washington, D.C. 20590, U.S.A.
david.banks@bts.gov

Cezary Janikow, Tomasz Mozolewski
Department of Mathematics, University of Missouri at St. Louis
8001 Natural Bridge Road
St. Louis, MO 63121, U.S.A.

1. Introduction

Modern classification researchers make extensive use of computer-intensive methods. These often arise from large data sets, and can involve unusual random objects such as graphs, sequences, partitions, and permutations. Recent innovations in classification include the boosting algorithm, various model-averaging techniques, and analogues of nonparametric regression that fit flexible discriminant surfaces.

This paper focuses upon the application of computer-intensive methods to classification trees, and examines the situation in which there are multiple classification trees from which a consensus tree must be obtained. This would arise, for example, in multicenter clinical trials in which each site's data are used to develop a decision tree about the efficacy of a cancer therapy (cf. Shannon and Banks, 1999). Similar problems applications arise in cluster analysis and phylogenetic analysis (cf. Banks and Constantine, 1998).

The second section of the paper lays out general issues in extending a method pioneered by Mallows (1957) to general random objects, such as classification trees. The third section specializes this approach to classification and regression trees, including a description of competing methods of estimation. The last section describes work that compares these competing estimates.

2. The General Model

Let $\mathcal{S}$ be a set of objects with elements s; depending on the application, these objects might be trees, directed graphs, partitions, sequences, or even stranger things. Let $\mathbb{R}^+$ be the nonnegative reals, and denote by $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^+$ an arbitrary metric on $\mathcal{S}$. Given d, one can mimic Mallows's method (1957) of setting probabilities on the set of permutations. This approach yields the probability measure $F(s^*, \tau)$, defined by

$$P_{(s^*,\tau)}[\, s\,] = c(s^*,\tau)e^{-\tau\, d(s,s^*)} \qquad \forall\, s \in \mathcal{S}, \tag{1}$$

where $s^* \in \mathcal{S}$ is the modal or central object, τ is a scale parameter, and $c(s^*,\tau)$ is the normalizing constant. Thus s^* and τ are analogous to the mean and precision of a normal distribution, respectively, and index the family of probability distributions $\{F(s^*,\tau)\}$. If $\tau = 0$, one has uniform measure on $\mathcal{S}$, but $\tau >> 0$ implies concentration about s^*. These parameters index the family of probability distributions $\{F(s^*,\tau)\}$.

The key practical advantage of this formulation is that it provides a location-scale family of models which apply to observations that are not the usual vector-space quantities. The key theoretical advantage of this formulation is that it is a maximum entropy model. And the key computational advantage is that probability mass falls off exponentially with distance from the central object, enabling practical computer-intensive solutions for otherwise intractable problems.

Model (1) assumes a unimodal probability distribution, which is not always valid. However, unimodal models provide a starting point from which to evaluate and build more complex models for inference on tree structures. For example, mixtures of this model can flexibly describe truly multimodal situations, and the combination of the EM algorithm with the numerical search procedure we describe in section 3 would support inference in such situations.

In our context, we suppose that $\mathcal{S}$ is a space of decision trees. This space will have special restrictions that depend upon the application. For example, it will vary according to the number of explanatory variables that are used in the classification, and according to how the number of cases in the training sample interact with the tree-building algorithm to determine the maximum possible complexity.

3. Example: Classification and Regression Trees

Recursive partitioning is an unstable classifier; small changes in the learning set can cause major structural changes in fitted trees. To address this problem, research into the combination of classification rules has significantly increased during this decade. Methods for combining classifiers generally use of one of two strategies:

> *Concatenation* produces a sequence of classifiers with output from early classifiers added to the learning set and used as input to subsequently built classifiers. Thus, early prediction results are used to strengthen subsequent classifier fitting.
>
> *Parallel classifiers* produce multiple classifiers using independent learning sets (e.g., bootstrapped learning sets sampled randomly from the original data). A new observation is classified by each of the classifiers, and the different results are combined (e.g., majority-rule vote counting) to assign the final classification to the observation.

Both of these strategies can greatly improve the predictive accuracy of unstable classifiers.

However, when the goal of the statistical analysis is to learn about the relationship between outcome and predictors these strategies for combining classifiers are unacceptable since they produce a large number of trees, making interpretation difficult. By using the model described

in (1), we have a novel method for combining a sample of classification trees that uses maximum likelihood estimation to produce a single, interpretable tree.

The key ingredient in applying the model in (1) is an appropriate distance metric. This should incorporate a sense of discrepancy that is scientifically meaningful within the application, as well as the restrictions that determine the space of allowable trees. In Shannon and Banks (1999), we used a metric that measures the amount of rearrangement needed to change one of the trees so that it has structure identical to the other. Using this distance metric we were able to perform a numerical search to find the maximum likelihood estimate of the central tree parameter. This estimate of the central tree was our proposed representation of the true tree structure.

Formally, the log-likelihood of (1) is

$$\begin{aligned} \ln L(s^*, \tau) &= n \ln c(s^*, \tau) - \tau \sum_{i=1}^{n} d(s_i, s^*) \\ &= -n \ln[\sum_{s \in S} \exp\left(-\tau d(s, s^*)\right)] - \tau \sum_{i=1}^{n} d(s_i, s^*) \end{aligned} \tag{2}$$

where $c(s^*, \tau) = \sum_{s \in S} \exp\left(-\tau \times d(s, s^*)\right)^{-1}$. It can be shown[11] that the maximum likelihood estimates $(\hat{s}^*, \hat{\tau})$ must satisfy:

$$\frac{1}{n} \sum_{i=1}^{n} d(s_i, \hat{s}^*) = \frac{\sum_{s \in S} d(s, \hat{s}^*) \exp(-\hat{\tau} d(s, \hat{s}^*))}{\sum_{s \in S} \exp(-\hat{\tau} d(s, \hat{s}^*))} \tag{3}$$

$$\hat{s}^* = \min_{s \in S} \hat{\tau} \sum_{i=1}^{n} d(s_i, \hat{s}^*) + n \ln \sum_{s \in S} \exp[-\hat{\tau} d(s, \hat{s}^*)] \tag{4}$$

In our application, we assume s^* and τ are unknown parameters to be estimated from the data. Solving (3) and (4) is hard, and thus we resort to a steepest ascent numerical approximation. Given the sample, we start with a candidate central tree $\hat{s}^*$ and estimate $\hat{\tau}$ by solving (4). For the current estimates $(\hat{s}^*, \hat{\tau})$, calculate the log-likelihood by solving (2). Next, generate all neighbor trees around $\hat{s}^*$ by adding and deleting splits that result in a tree that is one change away from $\hat{s}^*$. For each neighbor tree calculate τ and the log-likelihood. Select a tree with increased log-likelihood and repeat the process until no improvement is obtained. Restart this process at each of the sample trees. From all these iterations select the tree that produced the maximum log-likelihood and call this the MLE tree.

The left hand side of (3) is the mean distance of the candidate $\hat{s}^*$ to the observed trees $s_1, s_2, \ldots, s_n$; this can be calculated exactly using the distance metric defined below. The right hand side is a function of the distance between $\hat{s}^*$ and every $s \in S$. The cardinality of S is typically so large that enumerative computation is impossible. To avoid this problem we consider only the trees $s \in S$ such that $d(s, \hat{s}^*) \leq k$, for k a conveniently chosen constant. Provided k is not too small, the approximation is justified by the discreteness of the space S and the fact that the contribution of distant trees decreases exponentially.

An alternative to estimating s^* using MLE is to find the central tree which has the minimum mean distance to each of the observed trees in the sample. For a random sample of n observed trees, $s_1, s_2, \ldots, s_n \in S$, we wish to obtain the minimum mean distance (MMD) tree, s^*_{MMD} that satisfies

$$\frac{\sum_{i=1}^{n} d(s^*_{MMD}, s_i)}{n} \leq \frac{\sum_{i=1}^{n} d(s^*, s_i)}{n} \tag{5}$$

for all possible central trees s^*.

The search for the MMD tree is an iterative steepest ascent like that used for the MLE tree, but can be expected to be more rapid since neither τ or the log-likelihood need to be estimated. Instead, each candidate MMD tree is scored by (3), and that tree which has the smallest mean distance is the final MMD tree, s^*_{MMD}.

In developing the MMD approach we have examined several reasonable metrics on trees (cf. Shannon and Banks, 1999, and Shannon, 1998). A *topology* metric based on structural similarity defined by nodes and variables splitting the nodes has proven most robust. This allows differences in trees at various heights to be weighted differently to accentuate differences in the early splits or near the terminal nodes. A *partitioning* metric is based on integrated mean square error between partitions induced by two trees. However, this has proven difficult to compute, and convergence onto the MMD tree is difficult. Finally, a *matching* metric is defined based on pairs appearing in the same or different terminal nodes in the two trees.

4. Results

We have previously reported results which suggest the MLE and MMD trees perform better than a single tree obtained using the standard recursive partitioning algorithm. In Shannon and Banks (1999) we modeled the relationship of cancer to immunologic parameters using a standard recursive partitioning model and the MLE tree. The MLE tree was more accurate in predicting cancer than the standard tree. In Shannon (1999), a mixture model approach produced three distinct trees for simulated waveform data, each of which performed better than the standard tree. Finally, Cappelli and Shannon (COMPSTAT, 2000) showed that using the MLE tree from a set of trees pruned using different pruning algorithms performed better than any of the pruned trees.

We are currently conducting a large scale simulation in S-Plus testing the MLE and MMD tree in various situations.

REFERENCES

Banks, D.L. and Constantine, G.M. (1998). Metric models for random graphs. *Journal of Classification*, **15**, 199-224.

Capelli, M. and Shannon, W. (2000). An MLE strategy for identifying the best decision tree. *COMPSTAT*, in press.

Mallows, C. (1957). Non-null ranking models I. *Biometrika*, *44*, 114-130.

Shannon, W. (1999). Averaging classification tree models. *Computing Science and Statistics* **31**, 67-72.

Shannon, W. and Banks, D.L. (1999). Combining classification trees using MLE. *Statistics in Medicine*, **18**, 727-740.

RESUME

This paper describes a probability model for random samples of classification trees, and a computer-intensive algorithm to apply it.