Clustering and Classification In Large Biomedical Databases*

William D. Shannon
General Medical Sciences
Washington Univ. School of Medicine
St. Louis, MO 63110
shannon@osler.wustl.edu

1 Introduction

Analysis of large databases presents new challenges to statisticians, mathematicians, and computer scientists. Strategies developed for use with small data sets may be inappropriate or computationally difficult when applied to these larger data sets. In some cases, old approaches to data analysis may need to be abandoned altogether and replaced with computationally intensive data analytic techniques (e.g., bootstrapping, genetic algorithms). Several authors suggest that advanced computing provides an opportunity to develop an entirely new conception of mathematics[1,2].

In this paper I will describe several strategies of averaging, or combining, many fitted classification models to improve their accuracy and interpretability. I will then present an example of how multiple classification tree models can be 'averaged' to improve accuracy, while retaining the simple structure of a single binary tree for interpretation. In the final section I will describe several large biomedical databases where novel approaches for data analysis will be required to make full use of the information they contain.

2 Strategy: Averaging Multiple Models

A general classifier is unstable when small changes in the learning (training) data set cause large changes in the fitted model. To counteract this instability, research based on model averaging is showing great improvement in the accuracy of the final classification rules[3−6]. These strategies have so far come in one of two flavors: concatenation and parallelism[7].

Concatenation fits a sequence of prediction rules in a cascading manner with the output from the early models added to the learning data set for training subsequent models. Let (X, Y) be the data set, $T : X \mapsto Y$, be a classification model which uses the data X to predict the outcome Y. In concatenation there is a sequence of n classifiers, $T_i, i = 1, \ldots, n$, such that

$$\begin{aligned} T_1 &: X \mapsto \hat{Y}_1, \\ T_2 &: (X, \hat{Y}_1) \mapsto \hat{Y}_2, \\ &\vdots \\ T_n &: (X, \hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_{n-1}) \mapsto \hat{Y}_n. \end{aligned}$$

$\hat{Y}_n$ is the final prediction for an observation. (Note: the subscripts on the $\hat{Y}_i's$ indicate which classification rule in the concatenation sequence produced the prediction, and do not correspond to a case or record in the data set.)

Parallelism fits a sequence of classifiers independently to different data sets (e.g., bootstrapped samples), and combines the output predictions obtained from each independent classifier. For $i = 1, \ldots, n$, let X_i be the i^{th} (bootstrapped) data set, T_i the classifier fit to X_i, and $\hat{Y}_i$ the output from the i^{th} classifier. Then for the data set X,

$$\left.\begin{array}{l} T_1 : X_1 \mapsto \hat{Y}_1 \\ T_2 : X_2 \mapsto \hat{Y}_2 \\ \vdots \\ T_n : X_n \mapsto \hat{Y}_n \end{array}\right\} \rightarrow f(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n) = \hat{Y}.$$

The function f makes the final prediction, $\hat{Y}$, for an observation, and can be a vote-counting rule (i.e., classify as the most common class found among the $\hat{Y}_i's$), or average of the $\hat{Y}_i's$.

Concatenation and parallelism produces excellent predictive models with low error rates. However, in the process of building and using n distinct classifiers, the ability to learn the relationship between the predictors and outcome is lost. Sometimes, the increase in accuracy can not be justified if an understanding of this relationship

*Thanks to David Banks for guidance, discussion, and support.

is needed. In these cases, we would like to be able to average multiple models to remove the instability, and still retain a single model which lets us learn about this relationship. In the next section is an outline of a method we developed which averages classification trees and retains interpretability.

3 Example: Classification Trees

3.1 Model

The following model is an extension of work begun by Banks and his colleagues[8,9] to classification trees[10,11]. Let G be a finite set of classification trees, $g \in G$ are the elements of G, and $d : G \times G \to \Re^{+}$ is an arbitrary metric on G. d will be defined later. Consider the following distribution on G,

$$P_{(g^*,\tau)}[g] = c(g^*,\tau) \times \exp[-\tau d(g,g^*)], \qquad \forall g \in G \quad (1)$$

where g^* is the modal or central tree, $\tau \geq 0$ is a concentration parameter, and $c(g^*,\tau)$ is the normalizing constant. Intuitively, interpret g^* as a 'mean' and τ as a measure of precision (the inverse of dispersion) of the distribution. As $\tau \to 0$, every $g \in G$ becomes equiprobable; thus all g are modal and there is no central tendency. But when τ is large, the observed trees strongly concentrate around g^*. In this case the data contain much useful information for the estimation of g^*.

For a random sample of n observed trees, $g_1, g_2, \ldots, g_n \in G$, the log-likelihood of (1) is

$$\begin{aligned} \ln L(g^*,\tau) &= n \ln c(g^*,\tau) - \tau \sum_{i=1}^{n} d(g_i,g^*) \\ &= -n \ln[\sum_{g\in G} \exp\left(-\tau d(g,g^*)\right)] \qquad (2) \\ &\quad -\tau \sum_{i=1}^{n} d(g_i,g^*) \end{aligned}$$

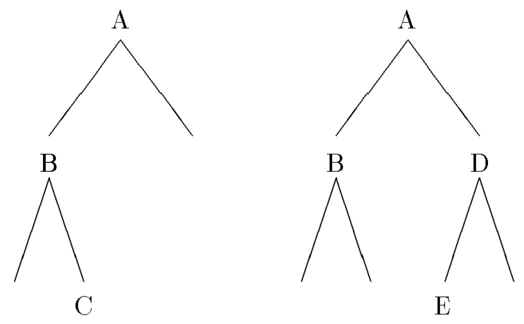
where $c(g^*,\tau) = \sum_{g\in G} \exp\left(-\tau \times d(g,g^*)\right)^{-1}$. The maximum likelihood estimates $(\hat{g}^*,\hat{\tau})$ must satisfy:

$$\frac{1}{n}\sum_{i=1}^{n} d(g_i,\hat{g}^*) = \frac{\sum_{g\in G} d(g,\hat{g}^*)\exp(-\hat{\tau} d(g,\hat{g}^*))}{\sum_{g\in G}\exp(-\hat{\tau} d(g,\hat{g}^*))} \qquad (3)$$

$$\begin{aligned} \hat{g}^* &= \min_{g\in G} \hat{\tau} \sum_{i=1}^{n} d(g_i,\hat{g}^*) \qquad (4) \\ &\quad + n \ln \sum_{g\in G} \exp[-\hat{\tau} d(g,\hat{g}^*)] \end{aligned}$$

3.2 Distance Metric

There is a 1-1 mapping from the tree icon to the complete list of decision rules the tree induces. A decision rule is any connected path from the root node to any descendant node. Let r be the number of variables (length) of a path. Order the variables from left to right starting with the root node and ending with a descendant node. To preserve the direction of splits in the tree we include a $'+'$ for a right child node and a $'-'$ for a left child node. In graph-theoretic terms the splitting variables label the vertices, and the $'+'$ and $'-'$ direction indicators are the edges of a graph indicating a right or left child node, respectively. To make this more concrete consider the two trees splitting on variables A, B, C, D, E:



Their path equivalent representations and path lengths are:

Paths	Lengths (r)
A	1
$A - B$	2
$\underline{A - B + C}$	$\underline{\mathbf{3}}$
A	1
$A - B$	2
$\underline{A + D}$	$\underline{\mathbf{2}}$
$\underline{A + D - E}$	$\underline{\mathbf{3}}$

The left tree has paths A, $A - B$ and $A - B + C$ with lengths 1, 2 and 3, respectively, and the right tree has paths A, $A - B$, $A + D$ and $A + D - E$ with lengths 1, 2, 2, and 3, respectively.

To define d, let $|\bullet|$ denote the cardinality of the argument set and $\alpha_r = \alpha(r)$ be a weight function that depends only on r. Then the distance $d(g_i,g_j)$ between trees g_i and g_j is

$$\sum_r \alpha_r \times |\{\text{discrepant } r\text{-paths between trees } g_i, g_j\}| \qquad (5)$$

where discrepant r-paths are paths of length r found in only one of the two trees. (Note the discrepant paths for the two trees above are underlined in the table). The distance between the two trees above is calculated as:

$$0\alpha_1 + 1\alpha_2 + 2\alpha_3$$

since there are no discrepant paths with length $r = 1$, one discrepant path with $r = 2$, and two discrepant paths with $r = 3$.

Selection of the weight function, $\alpha_r = \alpha(r)$, can be made to differentially penalize structural differences between trees based on where the differences occur. For example, a constant weight function $\alpha_r = 1$ does not distinguish between differences occurring early or late in the trees, but rather penalizes all discrepancies equally. An alternative weight function, $\alpha_r = \frac{1}{r}$, penalizes early discrepancies (i.e., those near the root with small r) more heavily than ones that occur later in the tree (i.e., those near terminal nodes with large r).

Other possible metrics for classification trees exist which might be useful. One reasonable metric penalizes discrepant paths with highly correlated variables less severely than discrepant paths with uncorrelated variables. Alternatively, instead of a metric on tree structure a metric over the partition and its values could be developed. This seems reasonable since it is entirely possible to have two trees that appear very different, but which impose essentially the same decision rule.

3.3 Numerical Approximation Search

Recall our goal is to obtain the maximum likelihood estimate of the central tree $\hat{g}^*$. This requires summing over all $g \in G$ in Equations (3) and (4) which is computationally expensive. (Note that even if we restrict the space of trees G to the set of trees with maximum path length $\leq$ some 'reasonable' constant, the number of elements in G is still often too large to enumerate.) One must use a numerical search algorithm to obtain $(\hat{g}^*, \hat{\tau})$ which maximizes the log-likelihood equation (2). Our algorithm employs an iterative process starting at each of the observed trees, $g_1, g_2, \ldots, g_n$, and is implemented as follows:

1. At initialization, choose a sample observation g_i as $\hat{g}_0^*$.
2. Conditional on $\hat{g}_i^*$ being the central tree, do a numerical search to estimate $\hat{\tau}$ according to equation (3). Calculate the log-likelihood $\ln L(\hat{g}_i^*, \hat{\tau})$.
3. Generate 200 neighboring trees $g_{n_1}, g_{n_2}, \ldots, g_{n_{200}}$ by randomly adding one split to a terminal node or deleting one terminal node from $\hat{g}_i^*$. Each of these trees will be 1 discrepant path away from $\hat{g}_i^*$.
4. For each of the 200 neighbor trees estimate $\hat{\tau}_{n_k}$ by solving equation (3) conditional on $\hat{g}_{n_k}$, and calculate the log-likelihood $\ln L(\hat{g}_{n_k}, \hat{\tau}_{n_k})$ for each pair.
5. If the maximum $\ln L(\hat{g}_{n_k}, \hat{\tau}_{n_k})$ is larger than the log-likelihood calculated in step 2, set $\hat{g}_{i+1}^* = \hat{g}_{n_k}$ and return to step 3. Otherwise, return to step 1 and select a new candidate tree from the set of observed trees and rerun the algorithm. When all trees have been tried, end the program and report the structure of the tree which produced the maximum log-likelihood over all n searches.

The right hand side of (3) is a function of the distance between $\hat{g}^*$ and every $g \in G$. Since the cardinality of G will typically be so large that enumerative computation is impossible, we consider only the trees $g \in G$ such that $d(g, \hat{g}^*) \leq k$, for k a conveniently chosen constant. Provided k is not too small, the approximation is justified by the discreteness of the space G and the fact that the contribution of distant trees decreases exponentially.

3.4 An Application

This work was motivated by a secondary analysis of observational cancer data. Immunological measurements were available on 460 people (283 with advanced cancer patients from 13 different clinical trials, 177 normal controls used for laboratory quality control), and we wanted to used the immunological measurements to predict health status. Our goal was to gain insight into how the immune system differs between cancer patients and controls. We generated 13 different classification trees by sequentially removing data for one of the clinical trials and fitting the tree to the remaining data (e.g., jackknifed the clinical trials). These were our set of observed trees which we wanted to combine into a single tree. In the numerical approximation we used the constant weight function $\alpha_r = 1$. Starting at each of the 13 observed trees we consistently converged on the same structure for the MLE of the central tree.

The standard and MLE tree structures were nearly identical in structure[10,11]. Compared with the standard tree, the MLE tree has a larger log-likelihood (-481.6 vs. -599.3), was more centrally located, measured as the mean distance to to the 13 jackknife trees (4.23 vs. 5.13), and had a lower resubstitution error rate (23.7% vs. 24.1%). It also has a lower expected loss (0.236 vs. 0.25) and lower cost-complexity (0.284 vs. 0.29). These measures of performance suggest this strategy can produce improved classification trees which retain interpretability.

4 Health Services Research

Health services research measures the utilization and effectiveness of different treatments and health delivery services in hospitals, clinics and physician offices. The goal of this research is to reduce health care costs and optimize patient outcomes and satisfaction. We are conducting secondary analyses of observational databases which contain information on the real-world utilization of health care resources. The secondary analysis makes use of existing data to try to answer questions the data were not initially intended to answer.

Specifically, we are investigating statistical clustering and classification methods for use in *case mix*[12] determination and *propensity score adjustment*[13,14]. *Case mix* is defined as the patient population characteristics in a health delivery service. For example, academic research medical centers generally have sicker patients than would be found in a small, rural hospital. A direct comparison of mortality, without adjusting for case mix (patient disease severity) differences, could be misleading. *Propensity score adjustment* is used to match patients receiving different treatments in a nonrandomized setting so that we can make causal inference of treatment effectiveness. In a nonrandomized setting treatment assignment is based on physician decision making based on the health status of the patient. For example, more severe forms of cardiac disease may be treated surgically, while less severe forms may be treated with medication. Since surgical patients are sicker than meedically treated patients, a direct comparison of mortality may again be misleading. Propensity score adjustment attempts to match patients with similar disease severity (reduce bias) and compare outcomes in matched samples.

Examples of the types of databases we are analyzing include administrative claims data, survey data, and electronic medical records (e.g., patient data). These databases are often tens or hundreds of megabytes in size. Administrative claims databases contain a patients' billing history for all diagnoses, treatments and procedures. A complete record can be used to reconstruct the medical history, and often can be used to study outcomes. A recent study examined the claims pattern for 55,000 children enrolled in a large medical insurance company to develop a predictive model to identify children who will suffer a severe head trauma in one year given there claims pattern in the previous year. The 55,000 children had a total of about one million distinct claims, and provided an extremely large database for this type of analysis.

Survey data is gathered to measure rates of an event within the population being studied. Because these data are observational in nature, it is difficult to make inferences about cause and effect relationships among the variables. However, recent work on propensity score adjustment is changing this. We are currently working with the *National Longitudinal Alcohol Epidemiological Survey* database, a publicly available database from the *National Institute of Alcohol Abuse and Alcoholism*[15]. This survey questioned 42,000 adults on depression, drug (e.g. alcohol, tobacco, narcotics) use, and treatment history. In addition, standard demographic and family history data were gathered. We are currently using this database and propensity score adjustment to answer two questions. The first is whether treatment for alcohol use reduces subsequent alcohol dependence. We are adjusting for the propensity to have received treatment based on their prior alcohol abuse, education, age, gender, etc., matching subjects with the same propensity for having received the treatment, and comparing post-treatment alcohol dependence rates between the treated and untreated groups. The second question is whether smoking at a young age increases the risk of developing alcohol dependency later in life. This is known as the *gateway hypothesis*. In this analysis we are matching on the propensity to smoke.

Development of an electronic medical record is a major priority of the *National Library of Medicine*[16]. This will replace the usual paper chart by having patient data entered directly into the computer, and viewed on a workstation at the patient bedside or examination room. This allows the health care worker to access all patient information (e.g., lab results, x-rays, histories) instantaneously. The electronic medical record also produces large volumes of patient data for secondary analyses. We are currently using k-means clustering to determine patient case mix within an academic medical center intensive care units. We have accessed data on 1,414 ICU patients admitted during a three month period. Data values for 18 important clinical measurements were obtained for a total of 363,698 total measurements. 9.8% of these patients died. We are using k-means clustering to identify homogeneous subgroups within the 1,414 patients based on clinical measurements. Once these subgroups are identified and validated we will compare them for differences in resource utilization, primary diagnoses responsible for the ICU admission, and mortality.

5 Conclusion

Large amounts of observational biomedical data are becoming available and viewed as an important resource for health services research. Clustering and classification of patients based on this data is important for making causal inferences and identifying subgroups to measure and compare treatment effectiveness. However, the increasing size of the databases are making standard clustering and classification methods unusable. We therefore need to develop new model averaging strategies to uncover the information in these databases. There will be many opportunities to develop new statistical, mathematical and computational methods in health services research to answer important public health questions

References

[1] Bailey, J. *After Thought.* New York, Basic Books, (1996).
[2] Vivaldi, F. 'An experiment with mathematics', In: *Exploring Chaos*, ed. Nina Hall, New York, Norton and Company, Inc, (1994).
[3] Wolpert, D. 'Stacked generalization', *Neural Networks*, 5: 241-259, (1992).
[4] Breiman, L. 'Stacked regressions', *Machine Learning*, 24: 49-64 (1995).
[5] Breiman, L. 'Bagging predictors',*Machine Learning*, 26: 123-140. (1996).
[6] LeBlanc, M., Tibshirani, R. 'Combining estimates in regression and classification', *Journal of the American Statistical Association*, 91(436): 1641-1650, (1997).
[7] Schurman, J. *Pattern classification: A unified view of statistical and neural approaches*, New York, John Wiley and Sons, Inc., (1996).
[8] Banks, D. and Carley, K., 'Metric inference for social networks', *Journal of Classification*, 11: 121-149, (1994).
[9] Banks, D. and Constantine, G., 'Maximum entropy models for graph-valued random objects', *Journal of Classification*, Submitted, (1997).
[10] Shannon, W. and Banks, D. 'An MLE strategy for combining CART Models', *Proceedings of the 1997 Interface*, In Press, (1997).
[11] Shannon, W. and Banks, D. 'Combining classification trees using maximum likelihood estimation', *Statistics in Medicine*, Under Review, (1997).
[12] Iezzoni, L. 'Risk adjustment for medical outcome studies', In: *Summary Report: Medical Effectiveness Resaerch Data methods*, eds. Grady, M., and Schwartz, H., U.S. Dept. of Health and human Services, AHCPR Pub. No. 92-0056, (1992).
[13] Rosenbaum, P., and Rubin, D. 'Reducing bias in observational studies using subclassification on the propensity score,' *Journal of the American Statistical Association*, 79(387): 516-524, (1984).
[14] Rosenbaum, P. *Observational Studies*, New York, Springer-Verlag, (1995).
[15] Grant, B., and Harford, T. 'The alcohol use disorder and associated disabilities interview schedule: Reliability of alcohol and drug modules in a general population sample,' *Drug and Alcohol Dependence*, 39: 37-44, (1995).
[16] Institute of Medicine, *The Computer-Based Patient Record: An Essential technology for Health Care*,' Washington, D.C., National Academy Press, (1991).