Zhang, Heping

Working Paper

Recursive Partitioning and Tree-based Methods

Papers, No. 2004,30

Provided in Cooperation with:
CASE - Center for Applied Statistics and Economics, Humboldt University Berlin

Suggested Citation: Zhang, Heping (2004) : Recursive Partitioning and Tree-based Methods, Papers, No. 2004,30, Humboldt-Universität zu Berlin, Center for Applied Statistics and Economics (CASE), Berlin

This Version is available at:
https://hdl.handle.net/10419/22203

Recursive Partitioning and Tree-based Methods

Heping Zhang

Yale University School of Medicine, 60 College Street, New Haven, CT 06520-8034.
heping.zhang@yale.edu; http://peace.med.yale.edu

1 Introduction

Tree-based methods have become one of the most flexible, intuitive, and powerful data analytic tools for exploring complex data structures. The applications of these methods are far reaching. They include financial firms (credit cards: Altman, 2002; Frydman et al., 2002, and investments: Pace, 1995; Brennan et al., 2001), manufacturing and marketing companies (Levin et al., 1995), and pharmaceutical companies.

The best documented, and arguably most popular uses of tree-based methods are in biomedical research for which classification is a central issue. For example, a clinician or health scientist may be very interested in the following question (Goldman et al., 1996, 1982; Zhang et al., 2001): Is this patient with chest pain suffering a heart attack, or does he simply have a strained muscle? To answer this question, information on this patient must be collected, and a good diagnostic test utilizing such information must be in place. Tree-based methods provide one solution for constructing the diagnostic test.

Classification problems also frequently arise from engineering research. Bahl et al. (1989) introduced a tree-based language model for natural language speech recognition. Desilva and Hull (1994) used the idea of decision trees to detect proper nouns in document images. Geman and Jedynak (1996) used a related idea to form an active testing model for tracking roads in satellite images. In addition, decision trees have been used in scientific and social studies including astronomy (Owens et al., 1996), chemistry (Chen et al., 1998) and politics (http://www.dtreg.com/housevotes.htm). We will revisit some of these applications later in detail.

Most commercial applications of tree-based methods have not been well-documented through peer reviewed publications. In 1999 the author helped the CLARITAS, a marketing company, apply a tree-based method as described in Section 6 (Zhang, 1998) for marketing segmentation analysis. Tree-based methods have also been frequently used in the drug development pro-

cess. The author has personally provided consultations to Aventis, Inc. for drug approvals.

The purpose of this article is to provide an overview for the construction of the decision trees, and, particularly, the recursive partitioning technique, which is the thrust of this methodology. In their early applications, tree-based methods were developed primarily to facilitate the automation of classifications as an expert system (Breiman et al., 1984; Friedman, 1977; Wasson et al., 1985), although Morgan and Sonquist (1963) were motivated by the need to analyze survey data to identify interactions, particularly in the presence of non-numerical predictors. More recently, classification trees have not only been used for automated disease diagnosis, but also for selecting important variables that are associated with a disease or any response of interest (Zhang and Bracken, 1995, 1996; Zhang and Singer, 1999; Zhang et al., 2003, 2001).

There are different approaches to classification. First, it can be done intuitively. For example, a physician or a group of physicians may use their experience in caring for patients with chest pain to form a subjective opinion or an empirical decision as to whether a new patient with chest pain is likely to suffer a heart attack, and consequently, decide what treatment is most appropriate. Secondly, methods in both statistical and machine learning literature have been developed, such as Fisher linear discriminant analysis (Fisher, 1936) and support vector machine (Cristianini and Shawe-Taylor, 2000). These methods have the parametric flavor in the sense that the classification rule has an explicit form with only a few parameters to be determined from a given sample that is usually referred to as learning sample.

Classification trees belong to the third type of methods for which we allow a very general structure, e.g., the binary tree as displayed in Fig. 1, but the number of "parameters" also needs to be determined from the data, and this number varies. For this reason, classification trees are regarded as nonparametric methods. They are adaptive to the data and are flexible, although the large number of quantities (or parameters) to be estimated from the data makes the classification rule more vulnerable to noise in the data.

To be more precise about the statistical problem, let us define the data structure and introduce some notation. Suppose that we have observed p covariates, denoted by a p-vector $\mathbf{x}$, and a response y for n individuals. For the ith individual, the measurements are

$$\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})' \text{ and } y_i,\ i = 1, \cdots, n.$$

The objective is to model the probability distribution of $P(y\,|\mathbf{x})$ or a functional of this conditional distribution.

To appreciate how these variables are characterized in real applications, let us examine some of the published applications.

Example 1. Levin et al. (1995) described a probability-driven, customer-oriented decision support system for the marketing decisions of the Franklin Mint, a leading Philadelphia-based worldwide direct response marketer of

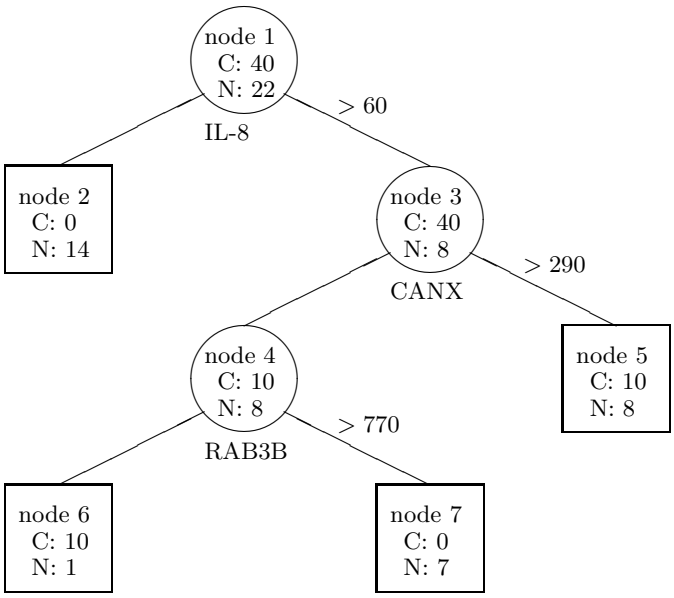

Fig. 1. Classification Tree for Colon Cancer Diagnosis Based on Gene Expression Data. Inside each node are the number of tumor (C) and normal (N) tissues. See Zhang et al. (2001) for more details.

quality collectibles and luxury home decor products. The purpose of the system is to target the"right" audience for each promotion from among a very large marketing database, based on the customers' attributes and characteristics. In this case, the customers' attributes and characteristics constitute the $\mathbf{x}$ variables. Whether the targeted client is desirable or not forms the basis for the response y.

Example 2. To screen large chemical databases in corporate collections and chemical libraries, Chen et al. (1998) used recursive partitioning to develop three-dimensional pharmacophores that can guide database screening, chemical library design, and lead optimization. Their idea was to encode the three-dimensional features of chemical compounds into bit strings, and those features are the $\mathbf{x}$ variables. Then, those features are selected in relation to the biological activities (i.e., y) of the compounds. Here, each compound contributes an observation. Using this idea, the authors successfully retrieved three-dimensional structure-activity relationships from a large heterogeneous dataset of 1644 monoamine oxidase inhibitors. We will revisit this example in detail in Section 4.

Like any multivariate regression model and as we can see from the above examples, the covariates or predictors in $\mathbf{x}$ may contain variables that can be categorical (nominal or ordinal) or continuous. For example, ethnicity is usually treated as categorical data and age as continuous. Some of the covariates

may have missing values, and we will discuss how missing values are handled in the tree framework. In a nutshell, unlike what is usually done in a simple linear regression, observations with missing information are not omitted from classification trees.

Not only can we have mixed types of predictors, but also the response variable can be discrete (binary or multiclass), continuous, and sometimes censored. The characteristics of the response, y, determines the method for estimating $P(y\ |\mathbf{x})$. We will review a variety of tree-based methods that are adaptable to the distribution of y. In Section 2, we will introduce the basic idea of classification trees using a dichotomous response. Section 2 is followed by some in-depth discussion of computational challenges and implementations in Section 3 and by examples in Section 4 to illustrate how we can interpret results from tree-based analyses. One of the most popular uses of tree-based methods is in the analysis of censored data in which y is the time to an event and is subject to censoring. As described in Section 5, such trees are referred to as survival trees (Bacchetti and Segal, 1995; Carmelli et al., 1991, 1997; Gordon and Olshen, 1985; Zhang, 1995). In Section 6, we will present an extension of the tree methodology to the classification of a response consisting of multiple components such as an array of respiratory symptoms (Zhang, 1998). Finally, we will conclude in Section 7 with some remarks on relatively recent developments such as forests and Bayesian trees. To illustrate the methods and their applications, some examples will be presented along with the methods.

2 Basic Classification Trees

We have highlighted some applications of decision trees. Here, we will explain how they are constructed. There has been a surge of interest lately in using decision trees to identify genes underlying complex diseases. For this reason, we will begin the explanation of the basic idea with a genomic example, and then will also discuss other examples.

Zhang et al. (2001) analyzed a data set from the expression profiles of 2,000 genes in 22 normal and 40 colon cancer tissues (Alon et al., 1999). In this data set, the response y equals 0 or 1 according to whether the tissue is normal or with cancer. Each element of $\mathbf{x}$ is the expression profile for one of the 2,000 genes. The objective is to identify genes and to use them to construct a tree so that we can classify the tumor type according to the selected gene expression profiles. Fig. 1 is a classification tree constructed from this data set. In what follows, we will explain how such a tree is constructed and how it can be interpreted.

2.1 Tree Growing and Recursive Partitioning

Tree construction usually comprises two steps: growing and pruning. The growing step begins with the root node, which is the entire learning sample.

In the present example, the root node contains the 62 tissues and it is labeled as node 1 on the top of Fig. 1. The most fundamental step in tree growing is to partition the root node into two subgroups, referred to as daughter nodes , such that one daughter node contains mostly cancer tissue and the other daughter node mostly normal tissue. Such a partition is chosen from all possible binary splits based on the 2,000 gene expression profiles via questions like "Is the expression level of gene 1 greater than 200?" A tissue is assigned to the right or left daughter according to whether the answer is yes or no. When all of the 62 tissues are assigned to either the left or right daughter nodes, the distribution in terms of the number of cancer tissues is assessed for both the left and right nodes using typically a node impurity. One of such criteria is the entropy function

$$i_t = -p_t \log(p_t) - (1 - p_t) \log(1 - p_t),$$

where p_t is the proportion of cancer tissue in a specified node t. This function is at its lowest level when $p_t = 0$ or 1. In other words, there is the least impurity when the node is perfect. On the other hand, it reaches the maximum when $p_t = \frac{1}{2}$, that is, the node is equally mixed with the cancer and normal tissues.

Let L and R denote the left and right nodes, respectively. The quality of the split s, resulting from the question "Is the expression level of gene 1 greater than 200?" is measured by weighing i_{L} and i_{R} as follows:

$$g_s = 1 - Pr(\mathrm{L}) i_{\mathrm{L}} - Pr(\mathrm{R}) i_{\mathrm{R}}, \tag{1}$$

where $Pr(\mathrm{L})$ and $Pr(\mathrm{R})$ are probabilities of tissues falling into the left and right nodes, respectively. The split with the lowest g_s is ultimately chosen to split the root node. This very same procedure can be applied to split the two daughter nodes, leading to the so-called recursive partitioning process. This process dies out as the sizes of the offspring nodes become smaller and smaller and the distribution of the tissue type becomes more and more homogeneous. The splitting stops when the node contains only one type of tissues.

The objective of the tree growing step is to produce a tree by executing the recursive partitioning process as far as possible. A natural concern is that such a saturated tree is generally too big and prone to noise. This calls for the second step to prune the saturated tree in order to obtain a reasonably sized tree that is still discriminative of the response whereas parsimonious for interpretation and robust with respect to the noise.

2.2 Tree Pruning and Cost Complexity

For the purpose of tree pruning, Breiman et al. (1984) introduced misclassification cost to penalize the errors of classification such as classifying a cancer tissue as a normal one, and vice versa. The unit of misclassification cost is chosen to reflect the seriousness of the errors because the consequence of classifying a cancer tissue as a normal one is usually more severe than classifying

a normal tissue as a cancer one. A common practice is to assign a unit cost for classifying a normal tissue as a cancer one and a cost, c, for classifying a cancer tissue as a normal one. Once c is chosen, the class membership for any node can be determined to minimize the misclassification cost. For example, the root node of Fig. 1 is classified as a cancer node for any c chosen to be greater than $\frac{22}{40}$. While c is usually chosen to be greater than 1, for the purpose of illustration here, if it is chosen to be 0.5, the root node is classified as a normal node because it gives rise to a lower misclassification cost.

When the class memberships and misclassification costs are determined for all nodes, the misclassification cost for a tree can be computed easily by summing all costs in the terminal nodes. A node is terminal when it is not further divided, and other nodes are referred to as internal nodes. Precisely, the quality of a tree, denoted by T, is reflected by the quality of its terminal nodes as follows:

$$R(T) = \sum_{t \in \tilde{T}} Pr(t)R(t), \tag{2}$$

where $\tilde{T}$ is the set of terminal nodes of tree T and $R(t)$ the within-node misclassification cost of node t.

The ultimate objective of tree pruning is to select a subtree of the saturated tree so that the misclassification cost of the selected subtree is the lowest on an independent, identically distributed sample, called a test sample. In practice, we rarely have a test sample. Breiman et al. (1984) proposed to use cross validation based on cost-complexity. They defined the number of the terminal nodes of T, denoted by $|\tilde{T}|$, as the complexity of T. A penalizing cost, the so-called complexity parameter, is assigned to one unit increase in complexity, i.e., one extra terminal node. The sum of all costs becomes the penalty for the tree complexity, and the cost-complexity of a tree is:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|, \tag{3}$$

where $\alpha(> 0)$ is the complexity parameter.

A useful and interesting result from Breiman et al. (1984) is that, for a given complexity parameter, there is a unique smallest subtree of the saturated tree that minimizes the cost-complexity measure (3). Furthermore, if $\alpha_1 > \alpha_2$ the optimally pruned subtree corresponding to α_1 is a subtree of the one corresponding to α_2. Therefore, increasing the complexity parameter produces a finite sequence of nested optimally pruned subtrees, which makes the selection of the desirably-sized subtree feasible.

Although the introduction of misclassification cost and cost complexity provides a solution to tree pruning, it is usually a subjective and difficult decision to choose the misclassification costs for different errors. Moreover, the final tree can be heavily dependent on such a subjective choice. From a methodological point of view, generalizing the concept of misclassification cost is difficult when we have to deal with more complicated responses, which

we will discuss in detail later. For these reasons, we prefer a simpler way for pruning as described by Segal (1988) and Zhang and Singer (1999).

Let us now return to the example. In Fig. 1, the 62 tissues are divided into four terminal nodes 2, 5, 6, and 7. Two of them (Nodes 2 and 7) contain 21 normal tissues and no cancer tissue. The other two nodes (Node 5 and 6) contain 40 cancer tissues and 1 normal tissue. Because this tree is relatively small and has nearly perfect classification, pruning is almost unnecessary. Interestingly, this is not accidental for analyses of many microarray data for which there are many genes and relatively few samples.

The construction of Fig. 1 follows the growing procedure as described above. First, node 1 is split into nodes 2 and 3 after examining all allowable splits from the 2000 gene expression profiles, and the expression level of gene IL-8 and its threshold at 60 are chosen because they result in the lowest weighted impurity of nodes 2 and 3. A tissue is sent to the left (node 2) or right (node 3) daughter node according to whether or not the IL-8 level is below 60. Because node 2 is pure, no further split is necessary and it becomes a terminal node. Node 3 is split into nodes 4 and 5 through recursive partitioning and according to whether or not the expression of gene CANX is greater than 290, while the partition is restricted to the 40 tissues in node 3 only. Furthermore, node 4 is subsequently partitioned into nodes 6 and 7 according to whether or not the expression of gene RAB3B exceeds 770.

There are also many interesting applications of simple classification trees. For example, Goldman et al. (1982) used classification trees to predict heart attack based on information from 482 patients. After a tree is constructed, the prediction is made from a series of questions such as "Is the pain in the neck only?" and/or "Is the pain in the neck and shoulder?" An appealing feature of tree-based classification is that the classification rule is based on the answers to simple and intuitive questions as posed here.

Although we present classification trees for a binary response, the method is similar for a mult-level response. The impurity function can be defined as

$$i_t = -\sum_{j=1}^{J} Pr(y = j) \log\{Pr(y = j)\},$$

for a J-level y. Everything else in the tree growing step as described above is applicable. For tree pruning, the only change to be made is to define the misclassification cost $c(j|k)$ from level k to level j, $j, k = 1, \ldots, J$.

3 Computational Issues

In Sections 2.1 and 2.2, we have explained the basic steps and concepts for tree construction. For most users of decision trees, the implementation aspect does not really affect the application. For methodological and software developments, however, it is imperative to understand the computational issues.

The most critical issue is to find the optimal split efficiently for any given node. The overall strategy is to identify the optimal split from each of the predictors and then choose the overall best one. Choosing the overall best one is straightforward, but identifying the optimal split from a predictor takes some efforts. The algorithm must take into account the nature of the predictor. Although we will use a dichotomous response to explain the ideas, the algorithm is also applicable for the other types of responses.

3.1 Splits Based on An Ordinal Predictor

Let us first consider a predictor with an ordinal scale such as gene expression in Fig. 1 or the ratio of cash flow to total debt in Fig. 2. Under the tree framework, as long as a predictor is ordinal, we will soon see that it does not matter whether the predictor is on a continuous or discrete scale.

Table 1. Expression Level of Gene IL-8 in 22 Normal and 40 Colon Cancer Tissues Used in Fig. 1

Expression Level	Colon Cancer	Expression Level	Colon Cancer	Expression Level	Colon Cancer	Expression Level	Colon Cancer
23.74	N	35.95875	N	33.9725	N	45.1	N
56.91875	N	28.7675	N	28.00875	N	39.7575	N
11.37625	N	31.6975	N	30.57875	N	171.4525	N
36.8675	N	40.33875	N	76.9875	N	97.92	N
55.2	N	238.58625	N	645.99375	N	117.6025	N
113.91375	N	567.13125	N	1528.4062	Y	306.30875	Y
76.125	Y	169.1375	Y	213.6275	Y	326.42625	Y
370.04	Y	114.92375	Y	311.4375	Y	186.2775	Y
131.65875	Y	412.135	Y	284.14625	Y	1178.9188	Y
75.81375	Y	1007.5262	Y	120.72	Y	227.70625	Y
80.73875	Y	2076.9025	Y	93.3575	Y	1813.4562	Y
170.11875	Y	737.695	Y	270.19625	Y	75.95	Y
62.7375	Y	148.04125	Y	599.6975	Y	247.52625	Y
390.31125	Y	222.55875	Y	391.355	Y	249.15125	Y
117.185	Y	104.78125	Y	124.91875	Y	210.90125	Y
519.08125	Y	175.55125	Y				

Table 1 displays the expression levels of gene IL-8 in 22 normal and 40 colon cancer tissues. Our objective for the time being is to split these 62 tissues into two subsamples according to whether the expression level of gene IL-8 is greater than a given threshold. In theory, this threshold can be anything, but practically, there is only a finite number of them that make a difference. In other words, it takes a finite number of steps to find an optimal threshold, although the solution is not unique.

The first step in finding an optimal threshold is to sort all expression levels, say, in an ascending order as displayed in Table 2. If the threshold is

Table 2. Sorted Expression Level of Gene IL-8 in 22 Normal and 40 Colon Cancer Tissues Used in Fig. 1

Expression Level	Colon Cancer	Expression Level	Colon Cancer	Expression Level	Colon Cancer	Expression Level	Colon Cancer
11.37625	N	23.74	N	28.00875	N	28.7675	N
30.57875	N	31.6975	N	33.9725	N	35.95875	N
36.8675	N	39.7575	N	40.33875	N	45.1	N
55.2	N	56.91875	N	62.7375	Y	75.81375	Y
75.95	Y	76.125	Y	76.9875	N	80.73875	Y
93.3575	Y	97.92	N	104.78125	Y	113.91375	N
114.92375	Y	117.185	Y	117.6025	N	120.72	Y
124.91875	Y	131.65875	Y	148.04125	Y	169.1375	Y
170.11875	Y	171.4525	N	175.55125	Y	186.2775	Y
210.90125	Y	213.6275	Y	222.55875	Y	227.70625	Y
238.58625	N	247.52625	Y	249.15125	Y	270.19625	Y
284.14625	Y	645.99375	N	306.30875	Y	311.4375	Y
326.42625	Y	370.04	Y	390.31125	Y	391.355	Y
412.135	Y	519.08125	Y	567.13125	N	599.6975	Y
737.695	Y	1007.5262	Y	1178.9188	Y	1528.4062	Y
1813.4562	Y	2076.9025	Y				

below the minimum (11.37625) or above the maximum (2076.9025), it produces an empty subsample. Thus, the threshold should be between 11.37625 and 2076.9025. If we take a look at the two lowest levels, 11.37625 and 23.74, it is clear that any threshold between these two levels produces the same two subsamples (or daughter nodes). In this example, there are 62 distinct levels of expression. Thus, we have $62 - 1 = 61$ distinct ways to split the 62 samples into two daughter nodes. It is noteworthy that, unlike this example, the number of unique levels of a predictor is usually lower than the number of samples.

The second step in finding an optimal threshold is to move along the intervals defined by two adjacent, distinct levels of the sorted predictor values. In Table 2, we move along as follows:

$$[11.37625, 23.74),\ [23.74, 28.00875), \ldots, [56.91875, 62.7375),$$
$$\ldots, [1528.4062, 1813.4562), [1813.4562, 2076.9025).$$

For computation, the threshold can be chosen as the middle point of the above intervals. For interpretation, the threshold can be rounded-off as is done to the first split in Fig. 1.

We have determined the pool of the potential thresholds, which is sometimes referred to as the allowable splits. Obviously, we can examine each threshold one at a time and assess its quality according to (1).

For a large data set, this means a lot of wasted computing time. To reduce the computation to a minimal level, let us take a careful look as to what

Table 3. Search for the Optimal Split

	Left Node			Right Node			Split
	No. of	No. of	Node	No. of	No. of	Node	Quality
Interval	Sample	Cancer	Impurity	Sample	Cancer	Impurity	g_s
[11.37625, 23.74)	1	0	0	61	40	0.6438	0.3666
[23.74, 28.00875)	2	0	0	60	40	0.6365	0.3849
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
[56.91875, 62.7375)	14	0	0	48	40	0.4506	0.6512
[62.7375, 75.81375)	15	1	0.1030	47	39	0.4562	0.6292
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
[1528.4062, 1813.4562)	60	38	0.6572	2	2	0	0.3640
[1813.4562, 2076.9025)	61	39	0.6538	1	1	0	0.3568

happens when we move the threshold from one interval to the next. In Table 3, as the threshold is moved up to the next interval, the samples that were already assigned to the left daughter stay on the left side because their expression levels are still below the new threshold. Most of the samples that were assigned to the right daughter stay on the right side, except those samples whose expression levels are equal to the lower limit of the new interval. In this particular case, there is only one sample that we need to move from the right side to the left every time we move the threshold by one interval. This observation implies that the node impurities and the split quality can be computed by updating the information slightly for the small set of the samples that are affected. Every of such a small set of samples is affected only once in the entire search of the predictor. In summary, after the values of a predictor are sorted, we can find an optimal threshold to split a node in the number of steps proportional to the number of distinct values of the predictor, which is at most the number of samples in the node. For the present example, any value in $[56.91875, 62.7375)$ is an optimal split. Intuitively from Table 3, we push the threshold as high as possible to maintain the perfect purity of the left daughter node. In the meantime, if we look bottom-up from the table, we also push the threshold as low as possible to maximize the purity of the right daughter node. The interval $[56.91875, 62.7375)$ offers the best balance. In Fig. 1, the split is chosen at 60, although any number in this interval is a legitimate choice.

Overall, if we have n samples in a node and p predictors, excluding the sorting time, the final threshold for the node can be identified in at most $O(np)$ steps.

3.2 Splits Based on A Nominal Predictor

For a nominal variable, we cannot sort the values of the variable as we did in Table 2. For a predictor of k levels, there are a total of $2^{k-1}-1$ ways to split a

node. To explain the algorithm, let us use an artificial example as summarized in Table 4.

Table 4. An Artificial Data Set

Predictor Value	No. of Normal	No. of Cancer	Rate of Cancer
A	5	10	0.67
B	10	5	0.33
C	20	30	0.60
D	35	25	0.42

In Table 4, the predictor has 4 levels, giving rise to 7 possible ways to split a node. A naive way is to assess every allowable split on an individual basis. This could be an extensive computation when the number of levels is 10 or higher. Thus, it is important to find a way to compute the quality of all splits in a gradual manner as in Section 3.1. If we focus on the levels of the predictor for the left daughter node, we can travel all 7 possible splits as follows: $\{A\}$, $\{AB\}$, $\{B\}$, $\{BC\}$, $\{C\}$, $\{AC\}$, and $\{ABC\}$. The key is that every move requires either the deletion or addition of a single level, which keeps the computation at the minimal level. Such a path of traveling through all $2^{k-1} - 1$ splits can be defined for any k.

There is actually a simple and quick solution for a dichotomous response. As shown in Table 4, we can compute the cancer rate for every level of the nominal predictor. During the splitting, the rates can substitute for the corresponding nominal levels. Because the rates are ordinal, the method described in Section 3.1 can be applied. After the optimal split is determined, we can map the rate back to the original nominal level. For example, for the data in Table 4, the optimal threshold based on the rate is in the interval $[0.42, 0.6)$, which means that the left daughter node contains samples with levels B and D, and the right daughter node with levels A and C. For a multiclass response, there is no apparent way to form an ordinal surrogate for a nominal predictor.

3.3 Missing Values

An important feature of decision trees is their ability to deal with missing predictor values. There are several solutions. Although there have been limited attempts (Quinlan, 1989) to compare some of them, the performance of the various solutions is largely unexplored. The choice mostly depends on the objective of the study.

The easiest approach is to treat the missing attribute as a distinct value and to assign all samples with missing values to the same node (Zhang et al., 1996). This approach is not only simple, but also provides clear paths as to where the samples with missing attributes end up in the tree structure.

Breiman et al. (1984) introduced and advocated surrogate splits to deal with missing attributes. The idea is very intuitive. For example, in Table 2, we considered using expression levels from gene IL-8 to split the 62 samples. What happens if the expression level from one of the samples, say, the first one, was not recorded? This happens in microarray experiments. Because IL-8 level is missing for the first sample, we cannot determine whether the level is below or above 60 and hence cannot decide whether the first sample should be assigned to the left or right daughter node. To resolve this ambiguity, Breiman et al. (1984) proposed to seek help from other genes that act "similarly" to IL-8. Since there are many other genes, we can use the one that is most similar to IL-8, which leads to a surrogate for IL-8.

What we need to clarify is the meaning of similarity. To illustrate this concept, let us consider gene CANX. Using the method described in Section 3.1, we can find an optimal split from gene CANX. The similarity between CANX and IL-8 is the probability that the optimal splits from these two genes assign a sample with complete information in these two genes into the same node. This strategy is similar to replacing a missing value in one variable in linear regression by regressing on the non-missing value most highly correlated with it. Then, why can't we use the same strategy as in the linear regression? According to Breiman et al. (1984), their strategy is more robust. The main reason is that their strategy is more specific to the particular sample with missing attributes, and does not result in a potential catastrophic impact for other samples with missing attributes.

The surrogate splits have some advantages over the simpler approach as described earlier. It makes use of other potentially useful information. Breiman et al. (1984) also proposed to rank the importance of variables through surrogate splits. The surrogate splits also have some limitations. First, it is uncommon, if at all, that surrogate splits are provided in published applications. Thus, it is unrealistic to know what the surrogate splits are and how we assign a sample with a missing attribute. Second, there is no guarantee in a data set that we can find a satisfactory surrogate split. Lastly, while it is a sensible idea to rank the variable importance based on surrogate splits, there is no assurance that a predictor ranked relatively high is necessarily predictive of the outcome, which can create a dilemma for interpretation. More recently, the importance of a variable tends to be evaluated on the basis of its performance in forests (Breiman, 1994; Zhang et al., 2003) rather than on a single tree.

In the construction of random forests, Breiman proposed another way of replacing missing values through an iterative process. A similar idea can be applied for tree construction. To initialize the process, we can fill in the missing values by the median of an ordered variable or by the category of a nominal variable with the highest frequency. An initial tree can be constructed once all missing data are imputed. In the next step, suppose again that in Table 2, the expression of gene IL-8 is missing for the first sample. The unobserved level is estimated by a weighted average over the samples with observed expressions for gene IL-8. Here, the weight is the so-called proximity, which is a similarity

measure between a pair of samples. Intuitively, if the second sample is more similar to the first sample than to the third one, we give more weight to the second sample than to the third one if the first sample is not observed. How is the proximity defined for a pair of samples? We can set it to zero before the initial tree is grown. Then, whenever a tree is grown, if two samples end up in the same terminal nodes, its promixity is increased by one unit. After the missing data are updated, a new tree is grown. Breiman recommends to continue this process at most five times in the random forest construction. For tree construction, it may take longer for the process to "converge," especially when the number of predictors is large. Nonetheless, it may still be worthwhile to repeat a few iterations. In addition to this convergence issue, it is also difficult to track where the samples with missing values are assigned as with the use of surrogate splits.

4 Interpretation

Interpretation of results from trees is usually straightforward. In Figure 1, we identified 3 genes IL-8, CANX, and RAB3B whose expression levels are highly predictive of colon cancer. However, this does not necessarily mean that these genes cause colon cancer. Such a conclusion requires a thorough search of the literature and further experiments. For example, after reviewing the literature, Zhang et al. (2001) found evidence that associates IL-8 with the stage of colon cancer (Fox et al., 1998), the migration of human clonic epithelial cell lines (Toshina et al., 2000), and metastasis of bladder cancer (Inoue et al., 2000). In addition, the expression of the molecular chaperone CANX was found to be down-regulated in HT-29 human colon adenocarcinoma cells (Yeates and Powis, 1997) and to be involved in apoptosis in human prostate epithelial tumor cells (Nagata et al., 1997). Lastly, RAB3B is a member of the RAS oncogene family. Therefore, these existing studies provide independent support that the three genes identified in Fig. 1 may be in the pathways of colon cancer. If this hypothesis could be confirmed from further experiments, Fig. 1 would have another important implication. Pathologically speaking, the 40 colon cancer samples are indistinguishable. Fig. 1 indicates that those 40 samples are not homogeneous in terms of gene expression levels. If confirmed, such a finding could be useful in cancer diagnosis and treatment.

As we stated earlier, there are numerous applications of decision trees in biomedical research, including the example above. To have a glimpse of the diverse applications of decision trees, let us review two different examples.

Example 3. Frydman and colleagues introduced recursive partitioning for financial classification (Frydman et al., 2002). They considered a financial dataset of 58 bankrupt ($y = 1$) industrial companies that failed during 1971-81, and 142 non-bankrupt ($y = 0$) manufacturing and retailing companies randomly selected from COMPUSTAT universe. Each company forms an observational unit or the so-called sample. Twenty financial variables with prior

evidence of predicting business failure are considered. They include the ratio of cash to total assets, the ratio of cash to total sales, the ratio of cash flow to total debt, the ratio of current assets to current liabilities, the ratio of current assets to total assets, the ratio of current assets to total sales, the ratio of earnings before interest and taxes to total assets, interest coverage, the ratio of market value of equity to total capitalization, the ratio of net income to total assets, the ratio of quick assets to current liabilities, the ratio of quick assets to total assets, the ratio of quick assets to total sales, the ratio of retained earnings to total assets, the ratio of total debt to total assets, the ratio of total sales to total assets, and the ratio of working capital to total sales.

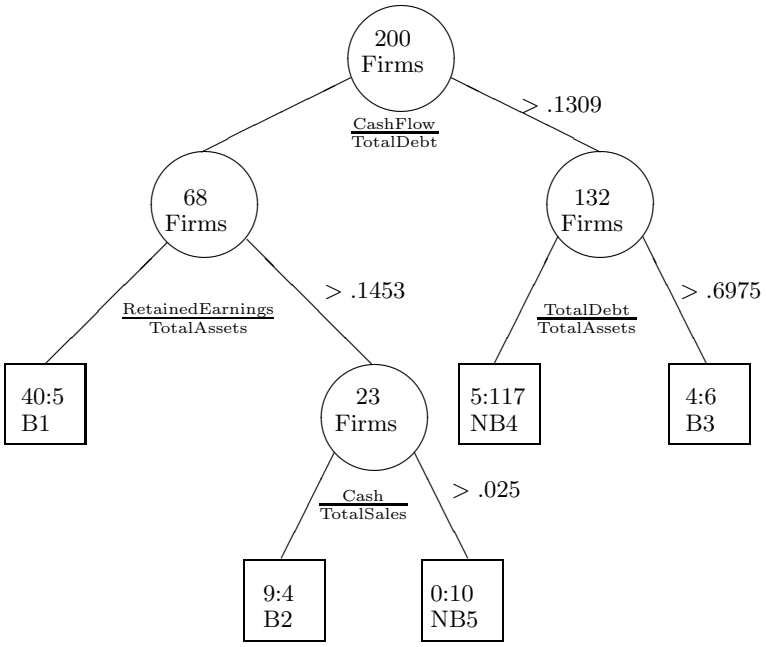

Fig. 2. Classification Tree for Bankruptcy. B1, B2, and B3 are three groups of relatively high risk of bankruptcy, and NB1 and NB2 are two groups of likely non-bankrupt companies. Inside the terminal nodes (boxes) are the numbers of bankrupt and non-bankrupt companies. See Frydman et al. (2002) for more details.

We can see from Fig. 2 that the risk of bankruptcy is relatively high if the ratio of cash flow to total debt is below 0.1309, unless both the ratio of retained earnings to total assets and the ratio of cash to total sales are above certain levels, i.e., 0.1453 and 0.025, respectively. Even if the ratio of cash flow to total debt is above 0.1309, there can be elevated risk of bankruptcy if the ratio of total debt to total assets is high (above 0.6975). A tree diagram as

in Fig. 2 offers a very clear and simple assessment of the financial state of a company.

Example 2 (continued) We indicated earlier what the predictors and response are for Example 2. Let us revisit this example. Unlike the other examples that we have introduced so far, this example uses a continuous response y – the compound potency. Because of this difference, the resulting tree is called a regression tree. To utilize the information from the 3-dimensional structures of compounds, Chen et al. (1998) used atom pair descriptors that are composed of the atom types of the two component atoms and the "binned" Euclidean distance between these two atoms. The width of each distance bin was chosen as 1.0 Å. To define predictors $\mathbf{x}$ from the atom pair descriptors, the authors characterized the atom pair descriptors in 17 types including negative charge center (e.g., sulfinic group), positive charge center (e.g., the nitrogen in primary, secondary, and tertiary amines), hydrogen bond acceptor (e.g., oxygen with at least one available lone pair electron), triple bond center, aromatic ring center, and H-bond donor hydrogen.

Fig. 3 presents part of the regression tree that is constructed by Chen et al. (1998). We trimmed the left hand side to fit into the space here; however, we can get the idea from the right hand side of tree. Generally speaking, a node of size 3 or 6 such as nodes 6 and 8 is too small to be reliable. Since we do not have the data to re-grow the tree, let us pretend that the node sizes are adequate, and concentrate on the interpretation instead. Since the main objective of Chen et al. appears to identify active nodes (i.e., those with high potencies), a small, inactive node is not of great concern.

First, there is one highly active node (node 7 with potency greater than 2) in Fig. 3. There are also two highly active nodes on the left hand side which are not shown in Fig. 3. Supported by the literature, Chen et al. (1998) postulated that there might be different mechanisms of action because the active nodes contain compounds of very different characteristics. This is similar to the hypothesis suggested by Fig. 1 that the 40 colon cancer tissues might be biologically heterogeneous. Chen et al. concluded further that their tree demonstrates the ability to detect multiple mechanisms of action coexisting in a large three-dimensional chemical data set. In addition, the selected atom pair descriptors also reveal interesting features of the monoamine oxidase (MAO) inhibitors. For instance, the "aromatic ring center–triple bond center" pair in the first split is the structural characteristic of pargyline, a well known MAO inhibitor.

We can see from these examples that tree-based methods tend to unravel integrated, intuitive results whose pieces are consistent with prior findings. Not only can we use trees for prediction, but also we may use them to identify potentially important mechanisms or pathways for further investigation.

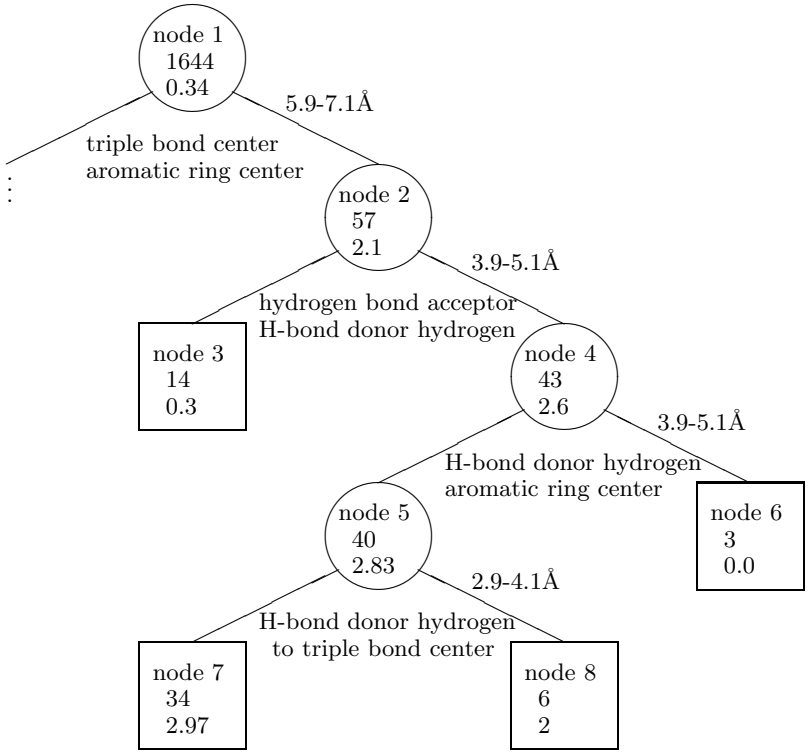

Fig. 3. Regression Tree for Predicting Potencies of Compounds. Inside each node are the number of compounds (middle) and the average potency of all compounds within the node (bottom). Underneath each node is the selected atom pair descriptor. Above the arm is the interval for the distance between the selected atom pair descriptor that assigns the compounds to the right daughter node. See Chen et al. (1998) for more details.

5 Survival Trees

The most popular use of tree-based methods is arguably in survival analysis for censored time, particularly in biomedical applications. The general goal of such applications is to identify prognostic factors that are predictive of survival outcome and time to an event of interest. For example, Banerjee et al. (2000) reported a tree-based analysis that enables the natural identification of prognostic groups among patients in the perioperative phase, using information available regarding several clinicopathologic variables. Such groupings are important because patients treated with radical prostatectomy for clinically localized prostate carcinoma present considerable heterogeneity in terms of disease-free survival outcome, and the groupings allow physicians to make early yet prudent decisions regarding adjuvant combination therapies. See,

e.g., Bacchetti and Segal (1995), Carmelli et al. (1991), Carmelli et al. (1997) and Kwak et al. (1990) for additional examples.

Before pointing out the methodological challenge in extending the basic classification trees to survival trees, let us quickly introduce the censored data. Let z denote the time to an event, which can be death or the occurrence of a disease. For a variety of reasons including losts to follow-up and the limited period of a study, we may not be able to observe z until the event occurs for everyone in the study. Thus, what we actually observe is a censored time y which is smaller than or equal to z. When z is observed, $y = z$. Otherwise, z is censored and $y < z$. Let $\delta = 1$ or 0 denote whether z is censored or observed.

The question is how to facilitate the censored time y in the tree-based methods. As in Section 2, we need to define a splitting criterion to divide a node into two, and also to find a way to choose a "right-sized" tree. Many authors have proposed different methods to address these needs. Here, we describe some of the methods. See Crowley et al. (1995), Intrator and Kooperberg (1995), LeBlanc and Crowley (1995), Segal (1988), Segal (1995), Zhang et al. (2001) and Zhang and Singer (1999) for more details.

5.1 Maximizing Difference between Nodes

Gordon and Olshen (1985) are among the earliest to have developed survival trees. Earlier, we focused on reducing the impurity within a node by splitting. When two daughter nodes have low impurities, the distributions of the response tend to differ between the two nodes. In other words, we could have achieved the same goal by maximizing the difference between the distributions of the response in the two daughter nodes. There are well established statistics that measure the difference in distribution. In survival analysis, we can compute the Kaplan-Meier curves (see, e.g., Miller, 1981) separately for each node. Gordon and Olshen used the so-called L^p Wasserstein metrics, $d_p(\cdot, \cdot)$, as the measure of discrepancy between the two survival functions. Specifically, for $p = 1$, the Wasserstein distance, $d_1(S_L, S_R)$, between two Kaplan-Meier curves, S_L and S_R, is illustrated in Fig. 4.

A desirable split maximizes the distance, $d_1(S_L, S_R)$, where S_L and S_R are the Kaplan-Meier curves for the left and right daughter nodes, respectively. Replacing g_s in (1) with $-d_1(S_L, S_R)$ we can split the root node into two daughter nodes and use the same recursive partitioning process as before to produce a saturated tree.

To prune a saturated survival tree, T, Gordon and Olshen (1985) generalized the tree cost-complexity for censored data. The complexity remains the same as before, but we need to redefine the cost $R(t)$, which now is measured by how far node t deviates from a desirable node in lieu of a pure node in the binary response case. In the present situation, a replacement for a pure node is a node τ in which all observed times are the same, and hence its Kaplan-Meier curve, δ_τ, is a piecewise constant survival function that has at most one point of discontinuity. Then, the within-node cost, $R(t)$, is defined as

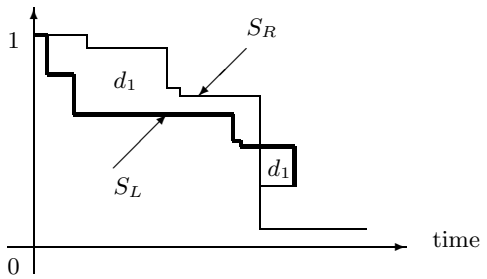

Fig. 4. The L^1 Wasserstein Distance between Two Kaplan-Meier Curves as measured by the area marked with d_1. Note that one curve (S_L) is thicker than the other (S_R).

$d_1(S_t, \delta_\tau)$. Combining this newly defined cost-complexity with the previously described pruning step serves as a method for pruning survival trees.

Another, perhaps more commonly used way to measure the difference in survival distributions is to make use of the log-rank statistic. Indeed, the procedures proposed by Ciampi et al. Ciampi et al. (1986) and Segal (1988) maximize the log-rank statistic by comparing the survival distributions between the two daughter nodes. The authors did not define the cost-complexity using the log-rank statistic. However, LeBlanc and Crowley (1993) introduced the notion of "goodness-of-split" complexity as a substitute for cost-complexity in pruning survival trees. Let $G(t)$ be the value of the log-rank test at node t. Then the split-complexity measure is

$$G(T) = \sum_{t \notin \tilde{T}} G(t) - \alpha(|\tilde{T}| - 1).$$

Therneau et al. (1990) proposed another way to define $R(t)$ that makes use of the so-called martingale residuals by assuming within-node proportional hazard models and then the least squares are computed as the cost.

In our experience, we found that Segal's bottom-up procedure (Segal, 1988) is practical and easy to use. That is, for each internal node (including the root node) of a saturated tree, we assign it a value that equals the maximum of the log-rank statistics over all splits starting from the internal node of interest. Then, we plot the values for all internal nodes in an increasing order and decide a threshold from the graph. If an internal node corresponds to a smaller value than the threshold, we prune all of its offspring. Zhang and Singer (1999) pointed out that this practical procedure can be modified in a broad context by replacing the log-rank statistic with a test statistic that is appropriate for comparing two samples with a defined outcome.

5.2 Use of Likelihood Functions

Although the concept of node impurity is very useful in the development of tree-based methodology, that concept is closely related to the concept of

likelihood as pointed out by Zhang et al. (2001). In fact, the adoption of likelihood makes it much easier to extend the tree-based methodology to analysis of complex dependent variables including censored time. For example, Davis and Anderson (1989) assume that the survival function within any given node is an exponential function with a constant hazard. LeBlanc and Crowley (1992) and Ciampi et al. (1988) assume different within-node hazard functions. Specifically, the hazard functions in two daughter nodes are assumed proportional, but are unknown. In terms of estimation, LeBlanc and Crowley (1992) use the full or partial likelihood function in the Cox proportional hazard model whereas Ciampi et al. (1988) use a partial likelihood function.

The most critical idea in using the likelihood is that within-node survival functions are temporarily assumed to serve as a vehicle of finding a split instead of believing them to be the true ones. For example, we cannot have a constant hazard function in the left daughter node, and then another constant hazard function in the right daughter node while assuming that the parent node also has a constant hazard function. Here, the constant hazard function plays the role of the "sample average." However, after a tree is constructed, it is both reasonable and possible that the hazard functions within the terminal nodes may become approximately constant.

5.3 A Straightforward Extension

Zhang (1995) examined a straightforward tree-based approach to censored survival data by observing the fact that the response variable involves two dimensions: a binary censoring indicator and the observed time. If we can split a node so that the node impurity is "minimized" in both dimensions, the within-node survival distribution is expected to be homogeneous. Based on this intuitive idea, Zhang (1995) proposed to compute the within-node impurity in terms of both the censoring indicator and the observed time first separately, and then together through weighting. Empirically, this simple approach tends to produce trees similar to those produced from using the log-rank test. More interestingly, empirical evidence also suggests that this simple approach outperforms its more sophisticated counterparts in discovering the underlying structures of data. Unfortunately, there need to be more comparative studies to scrutinize these different methods, even though limited simulations comparing some of the methods have been reported in the literature (Crowley et al., 1995, 1997; Zhang, 1995).

5.4 Other Developments

The methods that we described above are not designed to deal with time-dependent covariates. Bacchetti and Segal (1995) and Huang et al. (1998) proposed similar approaches to accommodate the time-dependent covariates in survival trees. The main concern with these existing approaches is that the same subject can be assigned to both the left and right daughter nodes, which

is distinct from any other tree-based methods and is potentially confusing in interpretation.

It is common in survival tree analysis that we want to stratify our sample into at a few groups that define the grades for the survival. To this end, it is useful to combine some terminal nodes into one group, which is loosely called "amalgamation." Ciampi et al. (1986) used the log-rank statistic for this purpose. LeBlanc and Crowley (1993) proposed constructing an ordinal variable that describes the terminal nodes. Often, we can simply examine the Kaplan-Meier curves for all terminal nodes to determine the group membership (Carmelli et al., 1997).

6 Tree-based Methods for Multiple Correlated Outcomes

As pointed out by Zhang (1998), multiple binary responses arise from many applications for which an array of health-related symptoms are of primary interest. Most of the existing methods are parametric; see, e.g., Diggle et al. (1994) for an excellent overview. In this section, we will describe a tree-based alternative to the parametric methods.

Motivated by both the broad application as well as by the need to analyze building-related occupant complaint syndrome (BROCS), Zhang (1998) proposed a tree-based method to model and classify multiple binary responses. Let us use the BROCS study to explain the method.

To understand the nature of BROCS, data were collected in 1989 from 6,800 employees of the Library of Congress (LOC) and the headquarters of the Environmental Protection Agency (EPA) in the United States. The data contain many explanatory variables, but Zhang (1998) extracted a subset of 22 putative risk factors, most of which are answers to "yes or no" or frequency (never, rarely, sometimes, etc.) questions. For example, is working space an enclosed office with door, a cubicle without door, stacks, etc? See Table 1 of Zhang (1998) for a detailed list. In this data set, BROCS is represented by six binary responses that cover respiratory symptoms in the central nervous system, upper airway, pain, flu-like, eyes, and lower airway. The primary purpose with this extracted data set is to evaluate the effect of the important risk factors on the six responses by constructing trees.

In terms of notation, the primary distinction is that the response y for each subject is a 6-vector. Consequently, we need to generalize the node-splitting criterion and cost-complexity to this vector-response. As we indicated earlier, one solution is to assume a certain type of within-node distribution for the vector-response and then maximize the within-node likelihood for splitting. One such distribution is

$$f(y;\Psi,\theta) = \exp(\Psi' y + \theta' w - A(\Psi,\theta)), \tag{4}$$

where Ψ and θ are node-dependent parameters, $A(\Psi,\theta)$ is the normalization function depending on Ψ and θ, and $w = \sum_{i<j} y_i y_j$. Zhang (1998) chose

this distribution because it is commonly used in the parametric models for multiple binary responses. See, e.g., Cox (1972), Fitzmaurice and Laird (1993) and Zhao and Prentice (1990). The negative of the likelihood based on (4) now serves as the impurity function, and the rest of the recursive partitioning as described before applies.

A naive approach is to treat y as a numerical vector and use a function such as the determinant of the within-node covariance matrix of y as a measure of impurity. If y were continuous, this approach is what Segal (1992) proposed to construct regression trees for repeatedly measured continuous y. For binary outcomes, however, this approach appears to suffer the well-known end-cut preference problem in the sense that it gives preference to the splits that result in two unbalanced daughter nodes in terms of their sizes.

One advantage of the likelihood based method is that the negative of the within-node likelihood can also be used as the within-node cost $R(t)$ for tree pruning. The main difficulty with this method is the computational burden, because every allowable split calls for a maximization of the likelihood derived from (4). Some strategies for reducing the computational time are discussed in Zhang (1998).

The criterion based on (4) ultimately leads to a 9 terminal nodes tree as displayed in Fig. 5, which suggests that respondents belonging to terminal nodes 7 and 17 have high incidence of respiratory symptoms. This is because the working area air quality of the people within these terminal nodes was poor, namely, often too stuffy or sometimes dusty. On the other hand, for example, subjects in terminal node 14 experienced the least discomfort because they had the best air quality. The basic message from this example is that tree-based analyses often reveal findings that are readily interpretable.

7 Remarks

In Breiman et al. (1984), tree-based methods are presented primarily as an automated machine learning technique. There is now growing interest in applying tree-based methods in biomedical applications, partly due to the rising challenges in analyzing genomic data in which we have a large number of predictors and a far smaller number of observations (Zhang et al., 2001). In biomedical applications, scientific understanding and interpretation of a fitted model are an integral part of the learning process. In most situations, an automated tree as a whole has characteristics that are difficult or awkward to interpret. Thus, the most effective and productive way of conducting tree-based analyses is to transform this machine learning technique into a human learning technology. This requires the users to review the computer-generated trees carefully and revise the trees using their knowledge, which not only often simplifies the trees, but also may improve the predictive precision of the trees, because recursive partitioning is not a forward looking process and does not

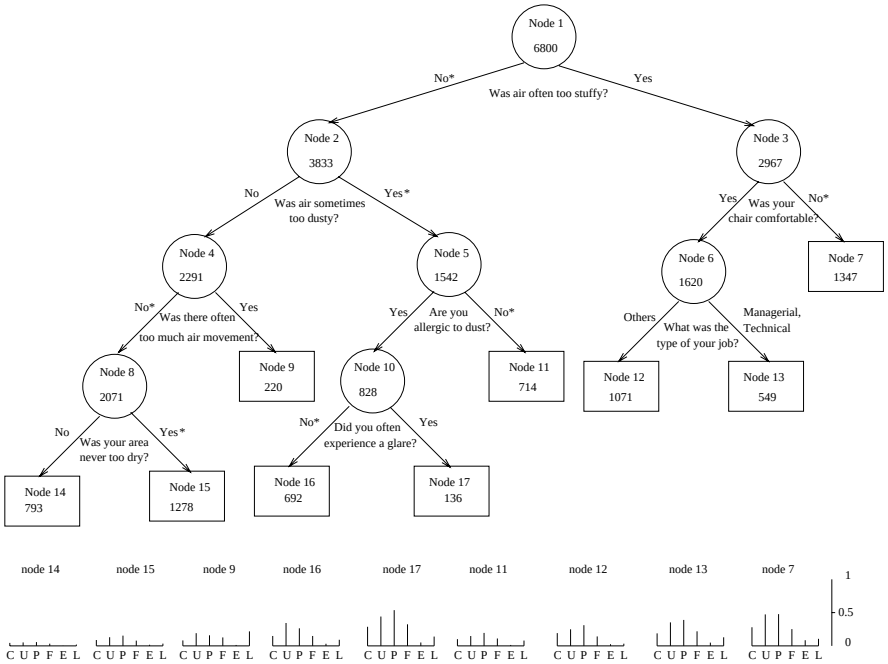

Fig. 5. Tree Structure for the Risk Factors of BROCS based on (4). Inside each node (a circle or a box) are the node number and the number of subjects. The splitting question is given under the node. The asterisks indicate where the subjects with missing information are assigned. The pin diagrams under the tree show the incidence rates of the six clusters (C: CNS; U: upper airway; P: pain; F: flu-like; E: eyes; and L: lower airway) in the nine terminal nodes. The side bar on the right end indicates the range of 0 and 1 for the rates of all symptoms.

guarantee any optimality of the overall tree. Zhang et al. (1996) called this step tree repairing.

While the full potential of tree-based applications remains to be seen and exploited, it must be made crystally clear that parametric methods such as logistic regression and Cox models will continue to be useful statistical tools. We will see more applications that use tree-based methods together with parametric methods to take advantages of various types of methods. The main advantage of tree-based methods is their flexibility and intuitive structures. However, because of their adaptive nature, statistical inference based on tree-based methodology is generally difficult. Despite the difficulty, some progress has been made to understand the asymptotic behavior of tree-based inference (Breiman, 1994; Buhlmann and Yu, 2003; Donoho, 1997; Gordon and Olshen, 1978, 1980, 1984; Lugosi and Nobel, 1996; Nobel, 1996; Nobel and Olshen, 1996).

Some attempts have been made to compare the tree-structured methods with other methods (Long et al., 1993; Segal and Bloch, 1989; Selker et al., 1995). More comparisons are still warranted, particularly in the context of genomic applications where data reduction is necessary and statistical inference is also desirable.

One exciting development in recent years is the expansion of trees into forests. In a typical application such as Banerjee et al. (2000) and Carmelli et al. (1997), constructing one or several trees is usually sufficient to unravel relationships between predictors and a response. Nowadays, many studies produce massive information such as recognizing spam mail from numerous characteristics and identifying disease genes. One or even several trees are no longer adequate to convey all of the critical information in the data. Construction of forests enables us to discover data structures further and in the meantime improves classification and predictive precision (Breiman, 1994; Zhang et al., 2003). So far, most forests are formed through some random perturbations and are hence referred to as random forests (Breiman, 1994). For example, we can draw bootstrap samples (Efron and Tibshirani, 1993) from the original sample and construct a tree as described above. Every time we draw a bootstrap sample, we produce a tree. Repetition of this process yields a forest. This is commonly called bagging (Breiman, 1994). The emergence of genomic and proteomic data afford us the opportunity to construct deterministic forest (Zhang et al., 2003) by collecting a series of trees that have a similarly high predictive quality. Not only do forests reveal more information from large data sets, but they also outperform single trees (Breiman, 1994; Buhlmann and Yu, 2003, 2002; Zhang et al., 2003).

A by-product of forests is a collection of variables that are frequently used in the forests, and the frequent uses are indicative of the importance of those variables. Zhang et al. (2003) examined the frequencies of the variables in a forest and used them to rank the variables. It would be even more helpful and informative if a certain probability measure could be assigned to the ranked the variables.

Bayesian approaches may offer another way to construct forests by including trees with a certain level of posterior probability. These approaches may also help us understand the theoretical properties of tree-based methods. However, the existing Bayesian tree framework focuses on providing an alternative method to those that exist. We would make an important progress if we could take full advantage of the Bayesian approach to improve our tree-based inference.

Classification and regression trees assign a subject to a particular node following a series of boolean statements. Ciampi et al. (2002) considered a "soft" splitting algorithm that at each node an individual goes to the right daughter node with a certain probability, which is a function of a predictor. This approach has the spirit of random forests. In fact, we can construct a random forest by repeating this classification scheme.

Several companies including DTREG.com, Insightful, Palisade Corporation, Salford Systems, and SAS market different variants of decision trees. In addition, there are many versions of free-ware including my own version, which is distributed from my website.

Acknowledgment

This work is supported in part by grant R01DA12468 from the National Institutes of Health. The author wishes to thank Norman Silliker, Elizabeth Triche, Yuanqing Ye and Yinghua Wu for their helpful assistance.

References

Altman, E.I. (2002). Bankruptcy, Credit Risk and High Yield Junk Bonds. *Blackwell Publishers*, Malden, MA.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96: pp.6745–6750.

Bacchetti, P. and Segal, M.R. (1995). Survival trees with time-dependent covariates: application to estimating changes in the incubation period of AIDS. *Lifetime Data Analysis*, 1: 35–47.

Bahl, L.R., Brown, P.F., de Sousa, P.V., Mercer R.L. (1989). A tree-based language model for natural language speech recognition. *IEEE Trans. on AS and SP*, 37: 1001–1008.

Banerjee, M., Biswas, D., Sakr, W., Wood, D.P. Jr. (2000). Recursive partitioning for prognostic grouping of patients with clinically localized prostate carcinoma. *Cancer*, 89: 404–411.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont, California.

Breiman, L. (1994). Bagging predictors. *Machine Learning*, 26: 123–140.

Brennan, N., Parameswaran, P. et al. (2001). *A Method for Selecting Stocks within Sectors*, Schroder Salomon Smith Barney.

Buhlmann, P. and Yu, B. (2003). Boosting with the L-2 loss: Regression and classification. *Journal of the American Statistical Association*, 98: 324-339.

Buhlmann, P. and Yu, B. (2002). Analyzing bagging. *Annals of Statistics*, 30: 927-961.

Buntine, W. and Niblett, T. (1992). A further comparison of splitting rules for decision-tree induction. *Machine Learning*, 8: 75–85.

Carmelli, D., Halpern, J., Swan, G.E., Dame, A., McElroy, M., Gelb, A.B., Rosenman, R.H. (1991). 27-year mortality in the western collaborative group study: construction of risk groups by recursive partitioning. *Journal of Clinical Epidemiology*, 44: 1341–1351.

Carmelli, D., Zhang, H.P. and Swan, G.E. (1997). Obesity and 33 years of coronary heart disease and cancer mortality in the western collaborative group study. *Epidemiology*, 8: 378–383.

Carter, C. and Catlett, J. (1987). Assessing credit card applications using machine learning. *IEEE Expert*, 2: 71–79.

Chen, X., Rusinko, A. and Young, S.S. (1998). Recursive partitioning analysis of a large structure-activity data set using three-dimensional descriptors. *Journal of Chemical Information and Computer Sciences*, 38: 1054–1062.

Ciampi, A., Couturier, A. and Li, S.L. (2002). Prediction trees with soft nodes for binary outcomes. *Statistics in Medicine*, 21: 1145–1165.

Ciampi, A., Hogg, S., McKinney, S. and Thiffault, J.(1988). A computer program for recursive partition and amalgamation for censored survival data. *Computer Methods and Programs in Biomedicine*, 26: 239–256.

Ciampi, A., Thiffault, J., Nakache J.-P. and Asselain, B. (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates. *Computational Statistics and Data Analysis*, 4: 185-204.

Cox, D.R. (1972). The analysis of multivariate binary data. *Applied Statistics*, 21: 113–120.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, Cambridge.

Crowley, J., LeBlanc, M., Gentleman, R. and Salmon S. (1995). Exploratory methods in survival analysis. In Koul, H.L. and Deshpande, J.V. (eds), *IMS Lecture Notes - Monograph Series 27*, pp.55-77, IMS, Hayward, CA.

Crowley, J., LeBlanc, M., Jacobson, J. and Salmon S. (1997). Some exploratory methods for survival data. In Lin, D.Y. and Fleming, T.R. (eds), *Proceedings of the First Seattle Symposium in Biostatistics*, Springer, New York.

Davis, R. and Anderson, J. (1989). Exponential survival trees. *Statistics in Medicine*, 8: 947–962.

Desilva, G. L. and Hull, J. J. (1994). Proper noun detection in document images. *Pattern Recognition*, 27: 311-320.

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*, Oxford Science Publications, New York.

Donoho, D.L. (1997). CART and best-ortho-basis: A connection. *Annals of Statistics*, 25: 1870–1911.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7: 179–188.

Fitzmaurice, G. and Laird, N.M. (1993). A likelihood-based method for analyzing longitudinal binary responses. *Biometrika*, 80: 141–151.

Fox, S. H., Whalen, G. F., Sanders, M. M., Burleson, J. A., Jennings, K., Kurtzman, S. and Kreutzer, D. (1998). Angiogenesis in normal tissue adjacent to colon cancer. *Journal of Surgical Oncology*, 69: 230–234.

Friedman, J.H. (1977). A recursive partitioning decision rule for nonparametric classification. *IEEE Trans. Computers.*, C-26: 404–407.

Frydman, H, Altman, E.I. and Kao, D.-I. (2002). Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress. In *Altman ed. Bankruptcy*, Credit Risk and High Yield Junk Bonds, pp.37–59.

Geman, D. and Jedynak, B. (1996). An active testing model for tracking roads in satellite images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18: 1–14.

Goldman, L., Cook, F., Johnson, P., Brand , D., Rouan, G. and Lee, T. (1996). Prediction of the need for intensive care in patients who come to emergency departments with acute chest pain. *The New England Journal of Medicine*, 334: 1498–504.

Goldman, L., Weinberg, M., Olshen, R.A., Cook, F., Sargent, R., et al. (1982). A computer protocol to predict myocardial infarction in emergency department patients with chest pain. *The New England Journal of Medicine*, 307: 588–597.

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.

Gordon, L. and Olshen, R.A. (1978). Asymptotically efficient solutions to the classification problem. *Annals of Statistics*, 6: 515–533.

Gordon, L. and Olshen, R.A. (1980). Consistent nonparametric regression from recursive partitioning schemes. *Journal Multivariate Analysis*, 10: 611–627.

Gordon, L. and Olshen, R.A. (1984). Almost surely consistent nonparametric regression from recursive partitioning schemes. *Journal Multivariate Analysis*, 15: 147–163.

Gordon, L. and Olshen, R.A. (1985). Tree-structured survival analysis. *Cancer Treatment Reports*, 69: 1065–1069.

Huang, X., Chen, S.D. and Soong, S.J. (1998). Piecewise exponential survival trees with time-dependent covariates. *Biometrics*, 54: 1420-14333.

Inoue, K., Slaton, J. W., Karashima, T., Shuin, T., Sweeney, P., Millikan, R. and Dinney, C. P. (2000). The prognostic value of angiogenesis factor expression for predicting recurrence and metastasis of bladder cancer after neoadjuvant chemotherapy and radical cystectomy. *Clinical Cancer Research*, 6: 4866–4873.

Intrator, O. and Kooperberg, C. (1995). Trees and splines in survival analysis. *Statistical Methods in Medical Research*, 4: 237–262.

Kwak, L.W., Halpern, J., Olshen, R.A. and Horning, S.J. (1990). Prognostic significance of actual dose intensity in diffuse large-cell lymphoma: results of a tree-structured survival analysis. *Journal of Clinical Oncology*, 8: 963–977.

LeBlanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*, 48: 411–425.

LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness-of-split. *Journal of the American Statistical Association*, 88: 457–467.

LeBlanc, M. and Crowley, J. (1995). A review of tree-based prognostic models. In Thall, P.F. (ed), *Recent Advances in Clinical Trial Design and Analysis*, pp.113-124, Kluwer, New York.

Levin, N., Zahavi, J. and Olitsky, M. (1995). Amos - A probability-driven, customer-oriented decision support system for target marketing of solo mailings. *European Journal of Operational Research*, 87: 708–721.

Loh, W.Y. and Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association*, 83: 715–725.

Long, W.L., Griffith, J.L., Selker, H.P. and D'Agostino, R.B. (1993). A comparison of logistic regression to decision tree induction in a medical domain. *Computers and Biomedical Research*, 26: 74–97.

Lugosi, G. and Nobel, A.B. (1996). Consistency of data-driven histogram methods for density estimation and classification. *Annals of Statistsics*, 24: 687–706.

Miller, R.G. (1981). *Survival Analysis*, Wiley, New York.

Mingers, J. (1989). An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, 3: 319–342.

Mingers, J. (1989). An empirical comparison of pruning methods for decision-tree induction. *Machine Learning*, 4: 227–243.

Morgan, J.N. and Sonquist, J.A. (1963). Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association*, 58: 415–434.

Nagata, K., Okano, Y. and Nozawa, Y. (1997). Differential expression of low Mr GTP-binding proteins in human megakaryoblastic leukemia cell line, MEG-01 and their possible involvement in the differentiation process. *Thrombosis and Haemostasis*, 77: 368–375.

Nobel, A.B. (1996). Histogram regression estimation using data-dependent partitions. *Annals of Statistics*, 24: 1084–1105.

Nobel, A.B. and Olshen, R.A. (1996). Termination and continuity of greedy growing for tree structured vector quantizers. *IEEE Transactions on Information Theory*, 42: 191–206.

Owens, E. A., Griffiths, R. E. and Ratnatunga, K. U. (1996). Using oblique decision trees for the morphological classification of galaxies. *Monthly Notices of the Royal Astronomical Society*, 281: 153–157.

Pace, R. K. (1995). Parametric, semiparametric and nonparametric estimation of characteristic values within mass assessment and hedonic pricing models. *Journal of Real Estate, Finance and Economics*, 11: 195–217.

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning*, 1: 81–106.

Quinlan, J.R. (1989). Unknown attribute values in induction. In *Proceedings of the Sixth International Machine Learning Workshop*, Cornell, New York, Morgan Kaufmann.

Segal, M.R. (1988). Regression trees for censored data. *Biometrics*, 44: 35–48.

Segal, M.R. (1992). Tree-structured methods for longitudinal data. *Journal of American Statistical Association*, 87: 407–418.

Segal, M.R. (1995). Extending the elements of tree-structured regression. *Statistical Methods in Medical Research*, 4: 219–236.

Segal, M.R. and Bloch, D.A. (1989). A comparison of estimated proportional hazards models and regression trees. *Statistics in Medicine*, 8: 539–550.

Selker, H.P., Griffith, J.L, Patil, S., Long, W.L. and D'Agostino, R.B. (1995). A comparison of performance of mathematical predictive methods for medical diagnosis: Identifying acute cardiac ischemia among emergency department patients. *Journal of Investigative Medicine*, 43: 468–476.

Sitaram, V. S., Huang, C. M. and Israelsen, P. D. (1994). Efficient codebooks for vector quantization image compression with an adaptive tree-search algorithm. *IEEE Transactions on Communications*, 42: 3027–3033.

Therneau, T.M., Grambsch, P.M. and Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77: 147–160.

Toshina, K., Hirata, I., Maemura, K., Sasaki, S., Murano, M., Nitta, M., Yamauchi, H., Nishikawa, T., Hamamoto, N. and Katsu, K. (2000). Enprostil, a prostaglandin-E-2 analogue, inhibits interleukin-8 production of human colonic epithelial cell lines. *Scandinavian Journal of Immunology*, 52: 570–575.

Wasson, J.H., Sox, H.C., Neff, R.K. and Goldman, L. (1985). Clinical prediction rules: Applications and methodologic standards. *The New England Journal of Medicine*, 313: 793–799.

Yeates, L. C. and Powis, G. (1997). The expression of the molecular chaperone calnexin is decreased in cancer cells grown as colonies compared to monolayer. *Biochemical and Biophysical Research Communications*, 238: 66–70.

Zhang, H.P. (1995). Splitting criteria in survival trees. In *Statistical Modelling: Proceedings of the 10th International Workshop on Statistical Modeling*, pp.305–314, Springer.

Zhang, H.P. (1998). Classification trees for multiple binary responses. *Journal of the American Statistical Association*, 93: 180–193.

Zhang, H.P. and Bracken, M.B. (1995). Tree-based risk factor analysis of preterm delivery and small-for-gestational-age birth. *American Journal of Epidemiology*, 141: 70–78.

Zhang, H.P. and Bracken, M.B. (1996). Tree-based, two-stage risk factor analysis for spontaneous abortion. *American Journal of Epidemiology*, 144: 989–996.

Zhang, H.P., Crowley, J., Sox, H. and Olshen, R.A. (2001). Tree structural statistical methods. *Encyclopedia of Biostatistics*, 6: pp.4561–4573, Wiley, Chichester, England.

Zhang, H.P., Holford, T. and Bracken, M.B. (1996). A tree-based methods of analysis for prospective studies. *Statistics in Medicine*, 15: 37–49.

Zhang, H.P. and Singer, B. (1999). *Recursive Partitioning in the Health Sciences*, Springer, New York.

Zhang, H.P., Yu, C.Y. and Singer, B. (2003). Cell and tumor classification using gene expression data: Construction of forests. *Proceedings of the National Academy of Sciences*, 100: 4168-4172.

Zhang, H.P., Yu, C.Y., Singer, B. and Xiong, M.M. (2001). Recursive partitioning for tumor classification with gene expression microarray data. *Proceedings of the National Academy of Sciences*, 98: 6730–6735.

Zhao, L.P. and Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, 77: 642–648.

Index