An Introduction to Classification and Regression Tree (CART) Analysis

Article · January 2000

CITATIONS
173

READS
3,862

1 author:

Roger J Lewis
Harbor-UCLA Medical Center
233 PUBLICATIONS **6,563** CITATIONS

SEE PROFILE

An Introduction to Classification and Regression Tree (CART) Analysis

Roger J. Lewis, M.D., Ph.D.
Department of Emergency Medicine
Harbor-UCLA Medical Center
Torrance, California

Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California.

Contact Information:

Roger J. Lewis, MD, PhD
Department of Emergency Medicine
Harbor-UCLA Medical Center, Box 21
1000 West Carson Street
Torrance, California 90509
Tel: (310) 222-6741
Fax: (310) 782-1763
Email: roger@emedharbor.edu

Introduction

A common goal of many clinical research studies is the development of a reliable clinical decision rule, which can be used to classify new patients into clinically-important categories. Examples of such clinical decision rules include triage rules, whether used in the out-of-hospital setting or in the emergency department, and rules used to classify patients into various risk categories so that appropriate decisions can be made regarding treatment or hospitalization.

Traditional statistical methods are cumbersome to use, or of limited utility, in addressing these types of classification problems. There are a number of reasons for these difficulties. First, there are generally many possible "predictor" variables which makes the task of variable selection difficult. Traditional statistical methods are poorly suited for this sort of multiple comparison. Second, the predictor variables are rarely nicely distributed. Many clinical variables are not normally distributed and different groups of patients may have markedly different degrees of variation or variance. Third, complex interactions or patterns may exist in the data. For example, the value of one variable (e.g., age) may substantially affect the importance of another variable (e.g., weight). These types of interactions are generally difficult to model, and virtually impossible to model when the number of interactions and variables becomes substantial. Fourth, the results of traditional methods may be difficult to use. For example, a multivariate logistic regression model yields a probability of disease, which can be calculated using the regression coefficients and the characteristics of the patient, yet such models are rarely utilized in clinical practice. Clinicians generally do not think in terms of probability but, rather in terms of categories, such as "low risk" versus "high risk."

Regardless of the statistical methodology being used, the creation of a clinical decision rule requires a relatively large dataset. For each patient in the dataset, one variable (the dependent variable), records whether or not that patient had the condition which we hope to predict accurately in future patients. Examples might include significant injury after trauma, myocardial infarction, or subarachnoid hemorrhage in the setting of headache. In addition, other variables record the values of patient characteristics which we believe might help us to predict the value of the dependent variable. For example, if one hopes to predict the presence of subarachnoid hemorrhage, a possible predictor variable might be whether or not the patient's headache was sudden in onset; another possible predictor would be whether or not the patient has a history of similar headaches in the past. In many clinically-important settings, the number of possible predictor variables is quite large.

Within the last 10 years, there has been increasing interest in the use of classification and regression tree (CART) analysis. CART analysis is a tree-building technique which is unlike traditional data analysis methods. It is ideally suited to the generation of clinical decision rules. Because CART analysis is unlike other analysis methods it has been accepted relatively slowly. Furthermore, the vast majority of statisticians have little or no experience with the technique. Other factors which limit CART's general acceptability are the complexity of the analysis and, until recently, the software required to perform CART analysis was difficult to use. Luckily, it is now possible to perform a CART analysis without a deep understanding of each of the multiple steps being completed by the software. In a number of studies, I have found CART to be quite effective for creating clinical decision rules which perform as well or better than rules developed using more traditional methods. In addition, CART is often able to uncover complex interactions between predictors which may be difficult or impossible to uncover using traditional multivariate techniques.

The purpose of this lecture is to provide an overview of CART methodology, emphasizing practical use rather than the underlying statistical theory.

Classification and Decision Problems

A classification problem consists of four main components. The first component is a categorical outcome or "dependent" variable. This variable is the characteristic which we hope to predict, based on the "predictor" or "independent" variables. Typical outcome variables are survival, need for surgery, and presence of myocardial infarction. The second component of a classification problem are the "predictor"

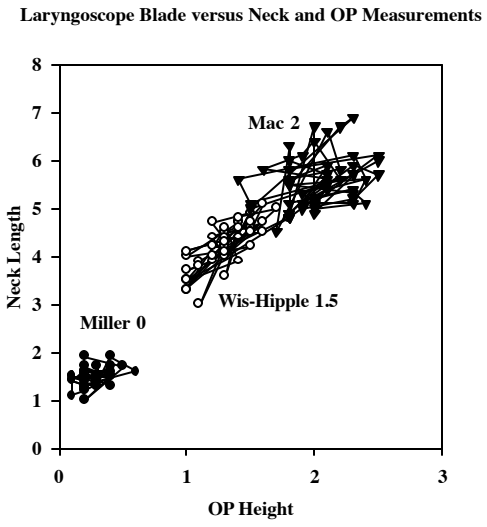
or "independent" variables. These are the characteristics which are potentially related to the outcome variable of interest. In general, there are many possible predictor variables. The third component of the classification problem is the learning dataset. This is a dataset which includes values for both the outcome and predictor variables, from a group of patients similar to those for whom we would like to be able to *predict* outcomes in the future. The fourth component of the classification problem is the test or future dataset, which consists of patients for whom we would like to be able to make accurate predictions. This test dataset may or may not exist in practice. While it is commonly believed that a test or validation dataset is required to validate a classification or decision rule, a separate test dataset is not always required to determine the performance of a decision rule.

A decision problem includes two components in addition to those found in a classification problem. These components are a "prior" probability for each outcome, which represents the probability that a randomly-selected future patient will have a particular outcome, and a decision loss or cost matrix. The decision cost matrix represents the inherent cost associated with misclassifying a future patient. For example, it is a much more serious error to classify a patient with an emergent medical condition as non-urgent, than to misclassify a patient with a non-urgent medical condition as urgent. A sample cost matrix is shown below, for a triage problem in which patients are classified as emergent, urgent, and non-urgent. The worst possible error, consisting of classifying a truly emergent patient as non-urgent (undertriage), is fifteen times as serious as misclassifying an urgent patient as emergent (overtriage).

		Classified by Tree as		
		Emergent	**Urgent**	**Non-Urgent**
True Value of Outcome Variable	**Emergent**	0	5	15
	Urgent	1	0	5
	Non-Urgent	3	2	0

Example Decision Cost Matrix.

As the first example, consider the problem of selecting the best size and type of laryngoscope blade for pediatric patients undergoing intubation. The outcome variable, the best blade for each patient (as determined by a consulting pediatric airway specialist), has three possible values: Miller 0, Wis-Hipple 1.5, and Mac 2. The two predictor variables are measurements of neck length and oropharyngeal height. The learning dataset is shown below. As can be seen from the figure, the smallest patients are best intubated with the Miller 0, medium sized patients with the Wis-Hipple 1.5, and the largest patients with the Mac 2.



One possible approach to analyzing these data would be to use multivariate logistic regression, using neck length and oropharyngeal height as the two independent predictor variables. A multivariate logistic regression model yields regression coefficients which, when used in logit expressions, give the probability that each blade is the best for that patient. Logistic regression equations are very difficult to use in clinical practice, especially in situations such as this, in which the outcome variable has more than two levels. Furthermore, it is difficult to incorporate possible interactions in a multivariate logistic regression model, and the model makes parametric assumptions which may not be valid.

As shown on the next page, the logistic regression model yields regression coefficients for the two independent variables. Both of these regression coefficients are statistically significant, suggesting that both neck length and oropharyngeal height are important predictors of the best laryngoscope blade.

Although the interpretation of this output is beyond the scope of the current lecture, suffice it to say that this type of model is rarely clinically useful (especially when faced with a small child who needs emergent airway management).

Logistic Regression Results

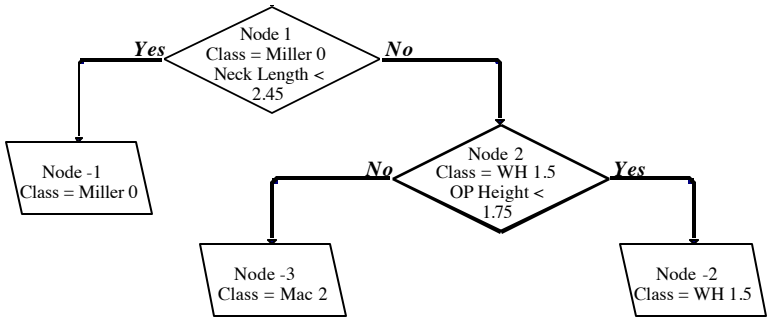
Variable	Parameter Estimate ± SE	P value	Odds Ratio
INTERCP1	22.4 ± 13.5	0.0974	.
INTERCP2	45.5 ± 13.5	0.0007	.
NECK_LEN	-5.8 ± 2.3	0.0114	0.003
OP_HEIGH	-10.5 ± 3.7	0.0051	0.000

Disclosure

The Classification and Regression Tree (CART) software to be illustrated in this lecture is a commercial product manufactured and sold by Salford Systems (http://www.salford-systems.com). Salford Systems has donated CDs which contain a trial version of their CART software, some additional modeling software not to be discussed in this lecture, and copies of the datasets used in this lecture (provided by the lecturer). *The lecturer does not intend this presentation to be an endorsement of this particular software package*. This is the only CART software with which the lecturer has significant personal experience, making it impossible for him to comment on the capabilities of other competitive products. The lecturer receives no financial support from Salford Systems, and SAEM has received no support from Salford Systems beyond the donation of the above-mentioned CDs for the use of meeting attendees.

Binary Recursive Partitioning

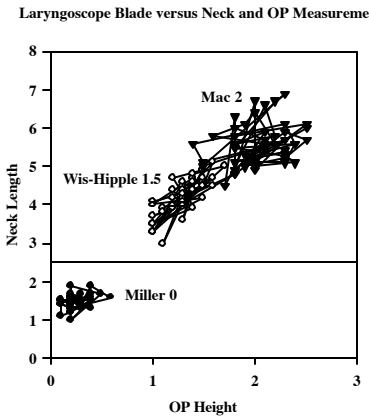
CART analysis is a form of binary recursive partitioning. The term "binary" implies that each group of patients, represented by a "node" in a decision tree, can only be split into two groups. Thus, each node can be split into two child nodes, in which case the original node is called a parent node. The term "recursive" refers to the fact that the binary partitioning process can be applied over and over again. Thus, each parent node can give rise to two child nodes and, in turn, each of these child nodes may themselves be split, forming additional children. The term "partitioning" refers to the fact that the dataset is split into sections or partitioned.

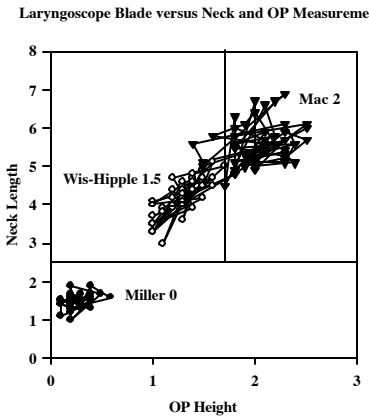


The figure to the left shows the classification and regression tree which results from analysis of the laryngoscope blade selection data shown above. This tree consists of a root node (Node 1), containing all patients. This node is split based on the value of the neck length variable. If the neck length is < 2.45 centimeters, then those patients are put in the first terminal node, denoted Node -1, and the best blade is predicted to be a Miller 0. All other patients are placed in Node 2. The group of patients in Node 2 is initially assigned a Wis-Hipple 1.5 blade but they are also split based on their oropharyngeal height. Those patients with an oropharyngeal height less than 1.75 are placed in terminal Node -2, and assigned a Wis-Hipple 1.5 blade, while those with an oropharyngeal height ≥ 1.75 are placed in terminal Node –3 and assigned a Mac 2 blade.

Several things should be pointed out regarding this CART tree. First, it is much simpler to interpret than the multivariate logistic regression model, making it more likely to be practical in a clinical setting. Secondly, the inherent "logic" in the tree is easily apparent, and it makes clinical sense. Interestingly, it has been shown that clinical decision rules which make sense to clinicians are more likely to be followed in clinical practice than rules in which the reasoning is not apparent.

In the two figures below, a visual illustration of the CART approach is given. The root node, which contains all patients, is split in two, analogous to a horizontal line being drawn at neck length = 2.45. All patients below the line, which are those found in the first terminal node, are assigned a predicted class of Miller 0. The group of patients above the original line are then split by a second line drawn at oropharyngeal height = 1.75. Those to the left of this line are assigned a class of Wis-Hipple 1.5, while those to the right of the line are assigned the class Mac 2. It is important to note that this second line applies only to one of the regions, which corresponds to the second parent node (Node 2). This process of partitioning is easy to visualize in two dimensions (i.e., when there are only two possible predictor variables) but is difficult or impossible to picture when there are five, ten, or dozens of possible predictors.





The text box to the right shows the set of commands, contained within a "command" file, which were used to produce the analysis represented by the CART tree. While there are 16 lines of commands, half are generic formatting commands or the definition of the misclassification costs. This command file is included in the CD which has been distributed, as is the dataset "airway.sys."

```
USE 'c:\cart\data\airway.sys'
LOPTIONS MEANS = NO, PREDICTION = NO,
     PLOTS = YES, TIMING = YES, PRINT
FORMAT =  3
MODEL best_bla
KEEP neck_len op_heigh
CATEGORY best_bla = 3 [min = 1]
PRIORS EQUAL
Misclass UNIT
Misclasify Cost = 1 Classify 1 as 2
Misclasify Cost = 1 Classify 1 as 3
Misclasify Cost = 1 Classify 2 as 3
Misclasify Cost = 1 Classify 3 as 1
Misclasify Cost = 1 Classify 3 as 2
Misclasify Cost = 1 Classify 2 as 1
LIMIT DEPTH=16
ERROR CROSS = 20
```

Advantages and Disadvantages of CART

CART analysis has a number of advantages over other classification methods, including multivariate logistic regression. First, it is inherently non-parametric. In other words, no assumptions are made regarding the underlying distribution of values of the predictor variables. Thus, CART can handle numerical data that are highly skewed or multi-modal, as well as categorical predictors with either ordinal or non-ordinal structure. This is an important feature, as it eliminates analyst time which would otherwise be spent determining whether variables are normally distributed, and making transformation if they are not.

As discussed below, CART identifies "splitting" variables based on an exhaustive search of all possibilities. Since efficient algorithms are used, CART is able to search all possible variables as

splitters, even in problems with many hundreds of possible predictors. [While some listeners may shudder at possible problems with overfitting and data dredging, these issues are dealt with in depth later].

CART also has sophisticated methods for dealing with missing variables. Thus, useful CART trees can be generated even when important predictor variables are not known for all patients. Patients with missing predictor variables are not dropped from the analysis but, instead, "surrogate" variables containing information similar to that contained in the primary splitter are used. When predictions are made using a CART tree, predictions for patients with missing predictor variables are based on the values of surrogate variables as well.

Another advantage of CART analysis is that it is a *relatively* automatic "machine learning" method. In other words, compared to the complexity of the analysis, relatively little input is required from the analyst. This is in marked contrast to other multivariate modeling methods, in which extensive input from the analyst, analysis of interim results, and subsequent modification of the method are required.

Finally, CART trees are relatively simple for nonstatisticians to interpret. As mentioned above, clinical decision rules based on trees are more likely to be feasible and practical, since the structure of the rule and its inherent logic are apparent to the clinician.

Despite its many advantages, there are a number of disadvantages of CART which should be kept in mind. First, CART analysis is relatively new and somewhat unknown. Thus, there may be some resistance to accept CART analysis by traditional statisticians (some of whom consult for prestigious medical journals). In addition, there is some well-founded skepticism regarding tree methodologies in general, based on unrealistic claims and poor performance of earlier techniques. Thus, some statisticians have a generalized distrust of this approach. Because of its relative novelty, it is difficult to find statisticians with significant expertise in CART. Thus, it may be difficult to find someone to help you use CART analysis at your own institution. Because CART is not a standard analysis technique, it is not included in many major statistical software packages (e.g., SAS).

Steps in Cart

CART analysis consists of four basic steps. The first step consists of tree building, during which a tree is built using recursive splitting of nodes. Each resulting node is assigned a predicted class, based on the distribution of classes in the learning dataset which would occur in that node and the decision cost matrix. The assignment of a predicted class to each node occurs whether or not that node is subsequently split into child nodes. The second step consists of stopping the tree building process. At this point a "maximal" tree has been produced, which probably greatly overfits the information contained within the learning dataset. The third step consists of tree "pruning," which results in the creation of a sequence of simpler and simpler trees, through the cutting off of increasingly important nodes. The fourth step consists of optimal tree selection, during which the tree which fits the information in the learning dataset, but does not overfit the information, is selected from among the sequence of pruned trees. Each of these steps will be discussed in more detail below.

<u>Tree Building</u>

Tree building begins at the root node, which includes all patients in the learning dataset. Beginning with this node, the CART software finds the best possible variable to split the node into two child nodes. In order to find the best variable, the software checks all possible splitting variables (called splitters), as well as all possible values of the variable to be used to split the node. A number of clever programming tricks are used to reduce the time required to search through all possible splits. In the case of a categorical variable, the number of possible splits increases quickly with the number of levels of the categorical variable. Thus, it is useful to tell the software the maximum number of levels for each categorical variable.

In choosing the best splitter, the program seeks to maximize the average "purity" of the two child nodes. A number of different measures of purity can be selected, loosely called "splitting criteria" or

"splitting functions." The most common splitting function is the "Gini", followed by "Twoing." Although the CART software manual recommends experimenting with different splitting criteria, these two methods will give identical results if the outcome variable is a binary categorical variable.

As discussed below, each node (even the root node) is assigned a predicted outcome class. The process of node splitting, followed by the assignment of a predicted class to each node, is repeated for each child node and continued recursively until it is impossible to continue.

Assignment of Node Classes

Each node, even the root node, is assigned a predicted class. This is necessary, as there is no way to know during the tree-building process which nodes will end up being terminal nodes after pruning. The predicted class assigned to each node depends on three factors: (1) the assumed prior probability of each class within future datasets; (2) the decision loss or cost matrix; and (3) the fraction of subjects with each outcome in the learning dataset that end up in each node. The function used to assign predicted classes to each node is shown at right. This method of node class assignment ensures that the tree has a minimal expected average decision cost for future datasets similar to the learning dataset in which the probability of each outcome is equal to the assumed prior probabilities.

Criteria for Assigning Classes to Nodes:

$C(j|i)$ is cost of classifying i as j.

$\pi(i)$ is prior probability of i.

N_i is number of class i in dataset.

$N_i(t)$ is number of class i in node.

Node is class i, if

$$\frac{C(j|i)\pi(i)N_i(t)}{C(i|j)\pi(j)N_j(t)} > \frac{N_i}{N_j}$$

for all values of j.

Missing Variables

For each node, the "primary splitter" is the variable that best splits the node, maximizing the purity of the resulting child nodes. When the primary splitting variable is missing for an individual observation, that observation is not discarded but, instead, a surrogate splitting variable is sought. A surrogate splitter is a variable whose pattern within the dataset, relative to the outcome variable, is similar to the primary splitter. Thus, the program uses the best *available* information in the face of missing values. In datasets of reasonable quality this allows all observations to be used. This is a significant advantage of this methodology over more traditional multivariate regression modeling, in which observations which are missing *any* of the predictor variables usually are often discarded.

Stopping Tree Building

As mentioned above, the tree building process goes on until it is impossible to continue. The process is stopped when: (1) there is only one observation in each of the child nodes; (2) all observations within each child node have the identical distribution of predictor variables, making splitting impossible; or (3) an external limit on the number of levels in the maximal tree has been set by the user ("depth" option).

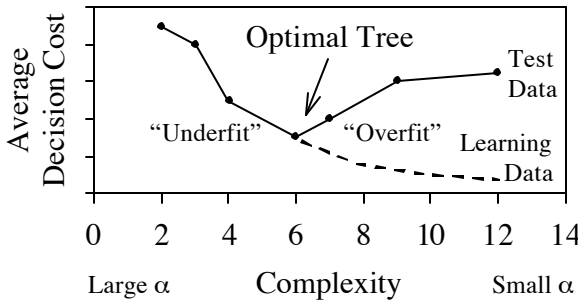
The "maximal" tree which is created is generally very overfit. In other words, the maximal tree follows every idiosyncrasy in the learning dataset, many of which are unlikely to occur in a future independent group of patients. The later splits in the tree are more likely to represent over fitting than the earlier splits, although one part of the tree may need only one or two levels, while a different branch of the tree may need many levels in order to fit the true information in the dataset. A major breakthrough of the CART methodology was the realization that there is no way during the tree-building process to know when to stop, and that different parts of the tree may require markedly different depths.

Tree Pruning

In order to generate a sequence of simpler and simpler trees, each of which is a candidate for the appropriately-fit final tree, the method of "cost-complexity" pruning is used. This method relies on a complexity parameter, denoted α, which is gradually increased during the pruning process. Beginning at the last level (i.e., the terminal nodes) the child nodes are pruned away if the resulting change in the predicted misclassification cost is less than α times the change in tree complexity. Thus, α is a measure of how much additional accuracy a split must add to the entire tree to warrant the additional complexity. As α is increased, more and more nodes (of increasing importance) are pruned away, resulting in simpler and simpler trees.

Optimal Tree Selection

The maximal tree will always fit the learning dataset with higher accuracy than any other tree. The performance of the maximal tree on the original learning dataset, termed the "resubstitution cost," generally greatly overestimates the performance of the tree on an independent set of data obtained from a similar patient population. This occurs because the maximal tree fits idiosyncrasies and noise in the learning dataset, which are unlikely to occur with the same pattern in a different set of data. The goal in selecting the optimal tree, defined with respect to expected performance on an independent set of data, is to find the correct complexity parameter α so that the information in the learning dataset is fit but not overfit. In general, finding this value for α would require an independent set of data, but this requirement can be avoided using the technique of cross validation (see below).

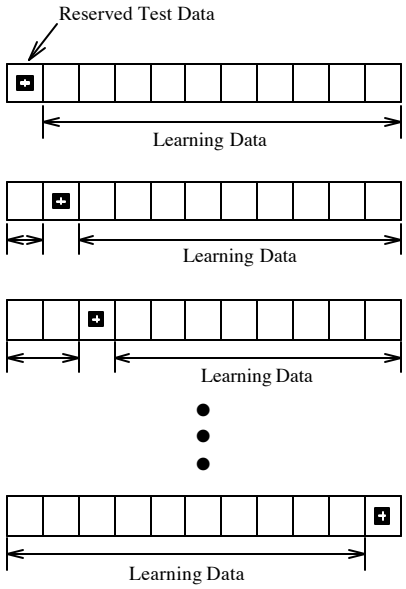


The figure to the right shows the relationship between tree complexity, reflected by the number of terminal nodes, and the decision cost for an independent test dataset and the original learning dataset. As the number of nodes increases, the decision cost decreases monotonically for the learning data. This corresponds to the fact that the maximal tree will always give the best fit to the learning dataset. In contrast, the expected cost for an independent dataset reaches a minimum, and then increases as the complexity increases. This reflects the fact that an overfitted and overly complex tree will not perform well on a new set of data.

Cross Validation

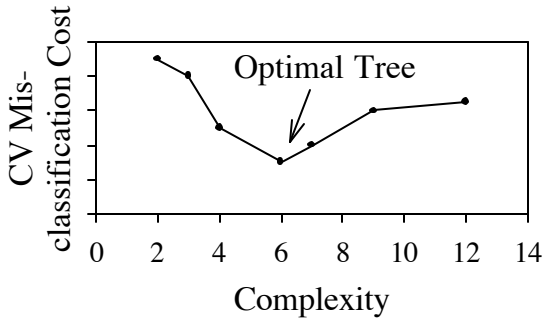
Cross validation is a computationally-intensive method for validating a procedure for model building, which avoids the requirement for a new or independent validation dataset. In cross validation, the learning dataset is randomly split into *N* sections, stratified by the outcome variable of interest. This assures that a similar distribution of outcomes is present in each of the *N* subsets of data. One of these subsets of data is reserved for use as an independent test dataset, while the other *N*-1 subsets are combined for use as the learning dataset in the model-building procedure (see the figure on the next page). The entire model-building procedure is repeated *N* times, with a different subset of the data reserved for use as the test dataset each time. Thus, *N* different models are produced, each one of which can be tested against an independent subset of the data. The amazing fact on which cross validation is based is that the *average* performance of these *N* models is an excellent estimate of the performance of the original model (produced using the entire learning dataset) on a future independent set of patients.

When cross validation is used in CART, the entire tree building and pruning sequence is conducted *N* times. Thus, there are *N sequences* of trees produced. Trees within the sequences are matched up, based on their number of terminal nodes, to produce an estimate of the performance of the tree in predicting outcomes for a new independent dataset, *as a function of* the number of terminal nodes or complexity. This allows a data-based estimate of the tree complexity which results in the best performance with respect to an independent dataset. Using this method, a minimum cost occurs when the tree is complex enough to fit the information in the learning dataset, but not so complex that "noise" in the data is fit. The figure below right shows a typical minimum which should occur in the cross-validation estimate of the misclassification cost, as a function of the number of terminal nodes or complexity.



Example: HIV-Triage

As our next example, consider a dataset involving the triage of self-identified HIV-infected patients who present to the emergency department (ED) for care. The outcome variable is the "urgency" of the visit, which has three levels: emergent, urgent, and non-urgent. These urgency levels are based on a retrospective evaluation of the final diagnosis and clinical course. The patient's historical and presenting features, and the results of an abbreviated and focussed review of systems are the predictor variables. The dataset includes 389 observations and is included on the CD distributed during the lecture.



Initial Screen

The initial screen which appears when the CART program is run includes the standard "File" and "View" menus. The "File" menu includes an option which allows one to submit an entire command file in order to conduct an analysis. In addition, one can use the "View" to open a "Notebook" window, through which a command file may be viewed, edited, and submitted for processing.

The command files included on the CD can be opened in the Notepad window or submitted directly from the File option. The use of the Notepad window allows editing of the command file (e.g., to give the correct location for the data on your computer) followed by submission for processing. Alternatively, the sample dataset may be opened, and a graphical user interface for model building may be used in lieu of a command file. When the graphical model-building interface is used, the associated command file may be saved, so that it can be viewed, edited, and reused later.

The text box on the next page shows the command file ("hivtree.cmd") which can be used to analyze the HIV triage dataset included on the CD. The MODEL statement shows the outcome variable, in this case a numerical representation of the urgency of the ED visit. The KEEP command lists the possible predictor variables to be considered in building the decision tree. [The version on the CD does not use the KEEP command, as only the correct variables have been included in the dataset.] The PRIORS statement shows that the actual distribution of outcomes in the learning dataset is to be used as the prior probabilities for outcomes when assigning outcome classes to nodes. The MISCLASSIFY statement defines the costs of various misclassification errors which may occur. In this case, the most serious error is to misclassify a patient who is truly in class 0 (emergent) as class 2 (non-urgent). The

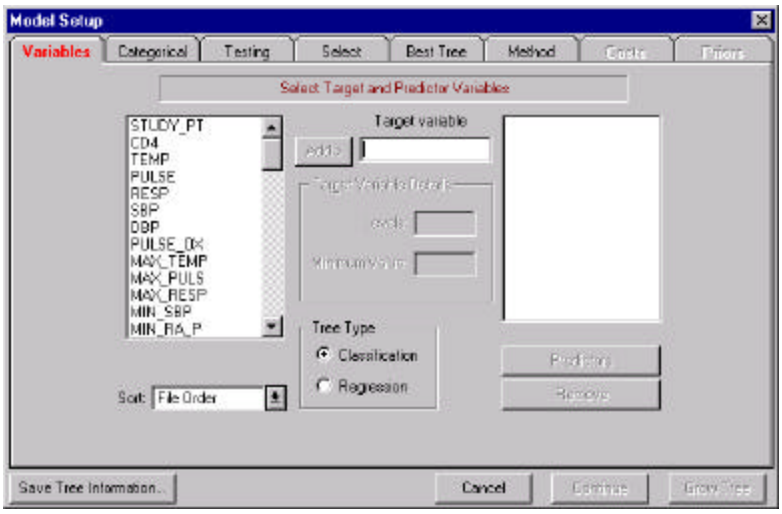
model is limited to a depth of 16, which limits the size of the maximal tree which may be constructed. In addition, cross validation is used, with the sample size divided into 10 subsets. Each of the commands used in the command file is described in detail in the online documentation included with the software.

The graphical user interface used for model setup, as well as the screen which appears during processing of the command file are shown (below right and top of next page). In the latter case, 20-fold cross validation is being used, which requires the creation of 21 tree sequences (one sequence using all learning data, and 20 tree sequences in which part of the data has been withheld for testing).

```
USE 'C:\CART\data\HIVNEW.SYS'
LOPTIONS MEANS = NO, PREDICTION = NO,
PLOTS = YES, TIMING = YES, PRINT
FORMAT =  3
MODEL NVISITCA
KEEP NSOBE, NDIARR, NVOMIT, NDYSPH,
NATAXI, NBEHAV, NWORSEHA, NHEADA,
NSWEAT, NCHILL, NSOBA, NDIZZY, NCOUGH,
SBP, RESP, PULSE, TEMP, NCCSOB,
NCCVOMDI, NCCHEADA, NCCFEVCH, NCCWEAKN,
NCCOTHER, CD4
PRIORS DATA
CATEGORY NVISITCA =   3 [min=   0]
Misclass UNIT
Misclasify Cost = 4 Classify 0 as 1
Misclasify Cost = 8 Classify 0 as 2
Misclasify Cost = 4 Classify 1 as 2
Misclasify Cost = 6 Classify 2 as 0
Misclasify Cost = 4 Classify 2 as 1
Misclasify Cost = 1 Classify 1 as 0
LIMIT DEPTH=16
ERROR CROSS = 10
BOPTIONS SURROGATES = 5 COMPETITORS = 3,
TREELIST = 10, BRIEF
BOPTIONS SERULE  = 1 ,IMPORTANCE = 1
BOPTIONS  COMPLEXITY = 0.0, NCLASSES = 8
TREE 'C:\CART\data\HIVTREE.TR1'
```

Results and Output

The original CART software was written for non-graphical computers, and all output had to be printable on old style line printers. The current CART software still produces the old text-based reports, although it provides a Window interface for navigating the text reports. The software also provides a fully graphical navigator, which presents the same information in a more user-friendly graphical and tabular format.



The CART report consists of 7 sections. The first section consists of the tree sequence, which includes the primary sequence of trees based on the entire learning dataset, as well as cross validation estimates of tree misclassification costs for an independent set of data. The second section of the report gives detailed node information, including the splitting criteria of each node, surrogate variables to be used if the primary splitter is missing, and the distribution of outcomes for the learning dataset at each node. The third section of the report gives information on the terminal nodes, including the distribution of outcome classes.

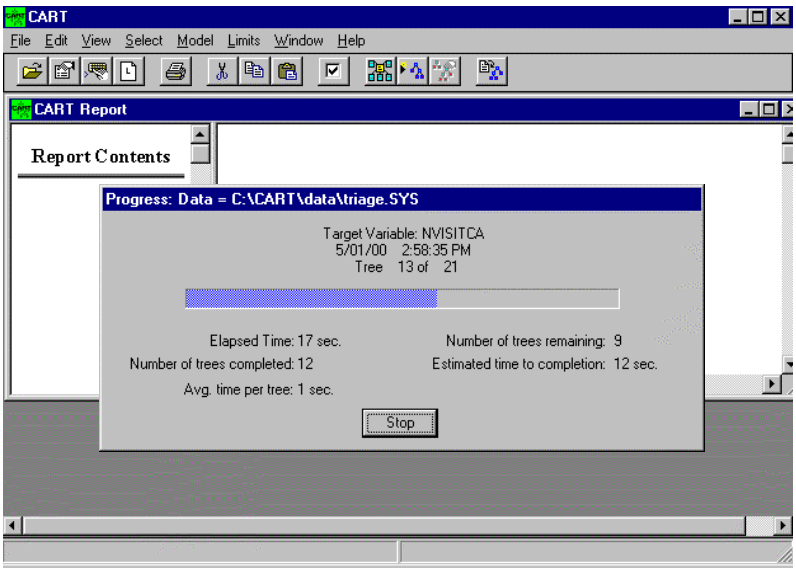
The next three sections of the report all involve misclassification rates. The fourth section of the report gives general misclassification information for the learning dataset (which overestimates the performance of the tree), while the fifth section give cross validation estimates of misclassification rates for an independent dataset. The sixth section give additional information on misclassif ication rates for the learning dataset. The last section of the report gives information on the relative importance of different

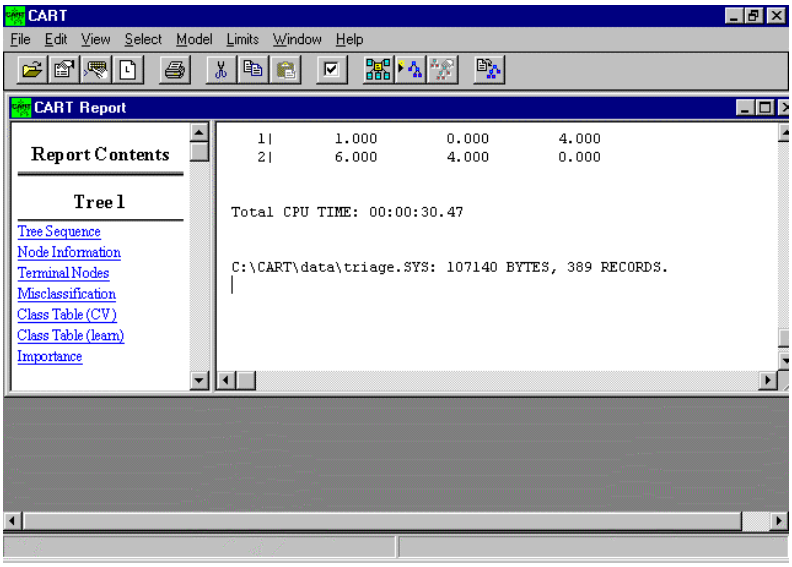
variables in the dataset. These importance measures incorporate information both on the use of variables as primary splitters, and also their relative worth as surrogate variables when primary splitters are missing. A discussion of variable importance is beyond the scope of this lecture.

The accompanying figures on pages 11-13 show the primary Report Screen, the tree sequence from a typical report (the HIV Triage example), as well as the graphical navigator which can be used to examine any tree in the tree sequence in detail. In addition, the navigator can be used to view detailed reports on misclassification rates, both for the learning dataset and for an independent dataset (based on cross validation).

The CART software can also produce a script for the program allCLEAR to be used to generate graphical representations of trees for exporting to other programs (e.g., Microsoft Word or Microsoft Powerpoint). The decision tree resulting from an analysis of the HIV triage data, represented using the allCLEAR software program is illustrated on page 13. The program allCLEAR is no longer automatically included with the software, and the CART program can also print nice looking trees by itself, as well as export graphical formats.

In the HIV Triage example, the root node is split on the initial heart rate, with patients with a heart rate of 104 or less going to the left, and those with a

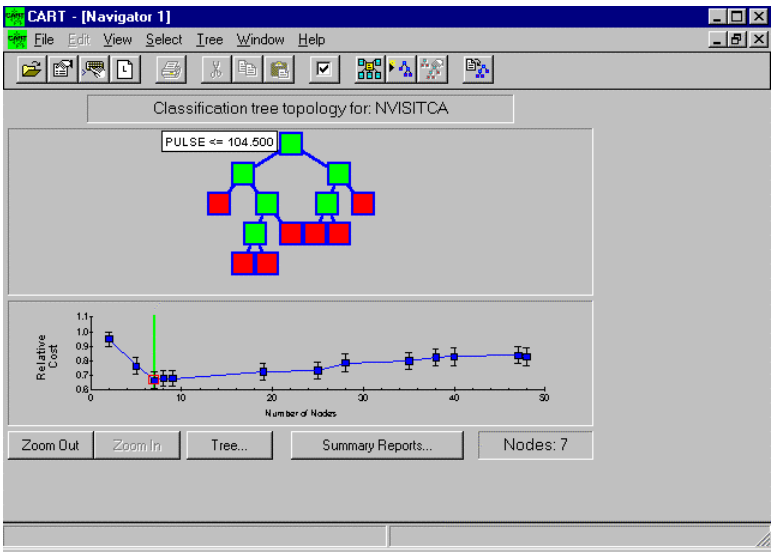


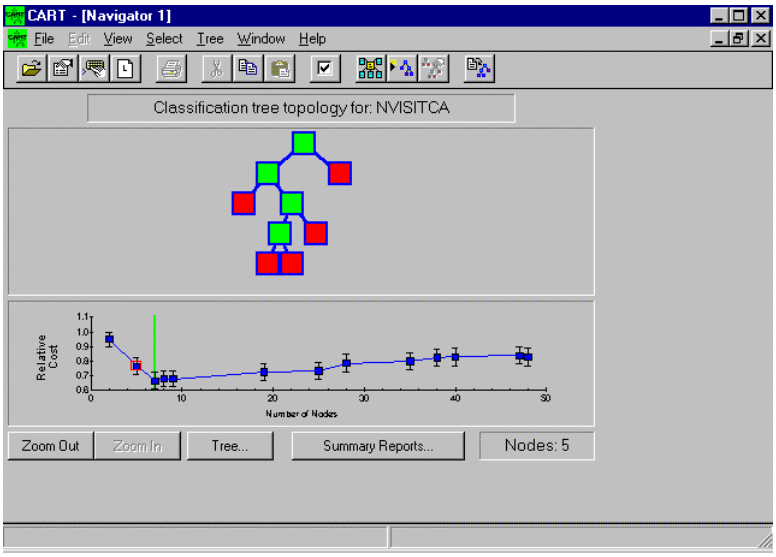


Dependent variable: NVISITCA

Terminal Tree	Nodes	Cross-Validated Relative Cost	Resubstitution Relative Cost	Complexity Parameter
1	48	0.828 +/- 0.059	0.256	0.000
5	35	0.798 +/- 0.059	0.299	0.008
6	28	0.784 +/- 0.059	0.341	0.010
7	25	0.731 +/- 0.059	0.361	0.011
8	19	0.722 +/- 0.058	0.433	0.021
9	9	0.678 +/- 0.057	0.558	0.022
10	8	0.678 +/- 0.057	0.578	0.033
11**	7	0.663 +/- 0.057	0.603	0.044
12	5	0.762 +/- 0.056	0.692	0.076
13	2	0.951 +/- 0.050	0.901	0.120
14	1	1.000 +/- 0.000	1.000	0.170

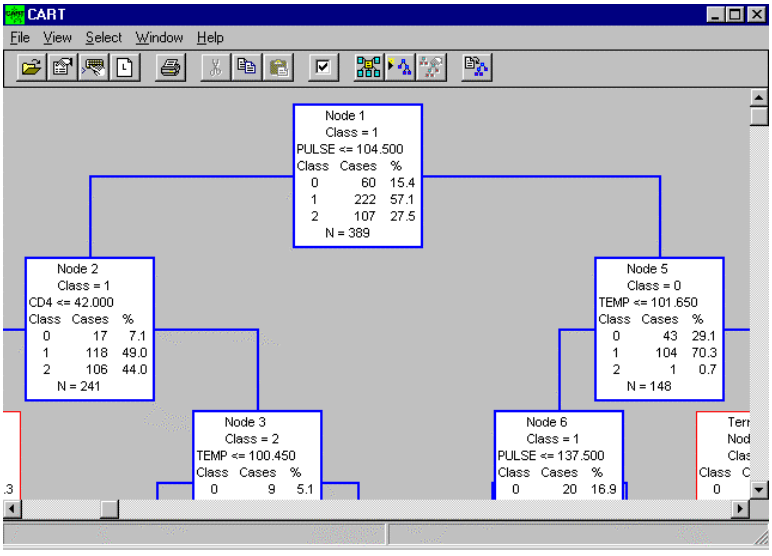
higher heart rate going to the right. As can be seen from the tree, the CD4 count is only important for patients who are not tachycardic, as the CD4 count appears only in the left-hand side of the tree. This is typical of insights which can be obtained through this type of analysis, namely, the importance of a variable (e.g., CD4 count) can vary tremendously based on the value of other variables. This makes clinical sense--a patient with markedly abnormal vital signs probably has an emergent medical condition regardless of the degree of immunosuppression. In contrast, a patient who is markedly immunocompromised may have an urgent or emergent medical condition even with relatively normal vital signs at triage.



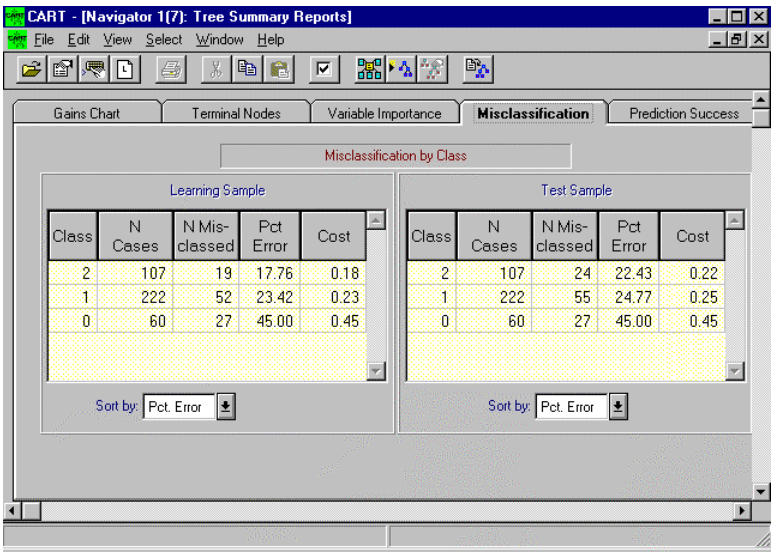


Future Datasets

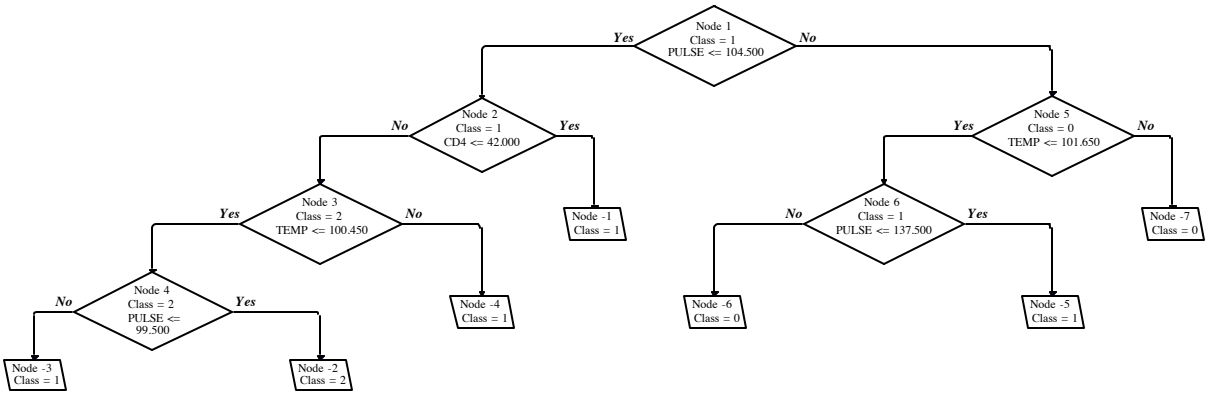
The purpose of a decision tree is usually to allow the accurate prediction of outcome for future patients, based on the values of their predictor variables. Similarly, the best way to test a tree using an independent dataset is to "drop" cases from a new dataset through the tree in order to determine the observed misclassification rates and costs. The CART software provides a command (the "tree" command) which allows the decision tree to be saved, so that it can be used with a new set of data in the future to *predict* outcome. This allows the testing of the tree on a new independent dataset.



Additional Topics



There are a number of additional topics that, although of practical importance to those using CART analysis, are beyond the scope of this lecture. These include the choice and use of different splitting rules and purity measures, the choice of alternative prior probability distributions, the interpretation of information on surrogate and competitive variables, methods used to rate the importance of different variables, and details of CART's handling of missing variables. Insights into these topics can best be obtained by actual use of the software, reference to the manual and online documentation, and comparison of results while varying user options. The dataset and command files included in the distributed CD should allow the listener to begin using the CART software and gain experience with its use.



Conclusions

Classification and Regression Tree (CART) analysis is a powerful technique with significant potential and clinical utility. Nonetheless, a substantial investment in time and effort is required to use the software, select the correct options, and interpret the results. Nonetheless, the use of CART has been increasing and is likely to increase in the future, largely because of the substantial number of important problems for which it is the best available solution.

References

Primary Reference

1. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Chapman & Hall (Wadsworth, Inc.): New York, 1984.

Examples

2. Steadman HJ, Silver E, Monahan J, Apelbaum PS, Robbins PC, Mulvey EP, Grisso T, Roth LH, Banks S. A classification tree approach to the development of actuarial violence risk assessment tools. Law and Human Behavior 2000;24:83-100.
3. Hess KR, Abbruzzese MC, Lenzi R, Raber MN, Abbruzzese JL. Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma. Clinical Cancer Research 1999;5:3403-3410.
4. Rainer TH, Lam PK, Wong EM, Cocks RA. Derivation of a prediction rule for post-traumatic acute lung injury. Resuscitation 1999;42:187-196.
5. Dart RG, Kaplan B, Varaklis K. Predictive value of history and physical examination in patients with suspected ectopic pregnancy. Annals of Emergency Medicine 1999;33:283-290.
6. Tsien CL, Fraser HS, Long WJ, Kennedy RL. Using classification tree and logistic regression methods to diagnose myocardial infarction. Medinfo 1998;9:493-497.
7. Nelson LM, Bloch DA, Longstreth WT Jr., Shi H. Recursive partitioning for the identification of disease risk subgroups: a case-control study of subarachnoid hemorrhage. Journal of Clinical Epidemiology 1998;51:199-209.
8. Germanson TP, Lanzino G, Kongable GL, Torner JC,Kassell NJ. Risk classification after aneurysmal subarachnoid hemorrhage. Surgical Neurology 1998;49:155-163.
9. Kastrati A, Schomig A, Elezi S, Schuhlen H, Dirschinger J, Hadamitzky M, Wehinger A, Hausleiter J, Walter H, Neuman FJ. Predictive factors of restenosis after coronary stent placement. Journal of the American College of Cardiology 1997;30:1428-1436.
10. Crichton NJ, Hinde JP, Marchini J. Models for diagnosing chest pain: Is CART helpful? Statistics in Medicine 1997;16:717-727.
11. Hadzikadic M, Hakenewerth A, Bohren B, Norton J, Mehta B, Andrews C. Concept formation vs. logistic regression: Predicting death in trauma patients. Artificial Intelligence in Medicine 1996;8:493-504.
12. Mair J, Smidt J, Lechleitner P, Dienstl F, Puschendorf B. A decision tree for the early diagnosis of acute myocardial infarction in nontraumatic chest pain patients at hospital admission. Chest 1995;108:1502-1509.
13. Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB. A comparison of performance of mathematical predictive methods for medical diagnosis: Identifying acute cardiac ischemia among emergency department patients. Journal of Investigative Medicine 1995;43:468-476.
14. Li D, German D, Lulla S, Thomas RG, Wilson SR. Prospective study of hospitalization for asthma. A preliminary risk factor model. American Journal of Respiratory and Critical Care Medicine 1995;151:647-655.
15. Falconer JA, Naughton BJ, Dunlop DD, Roth EJ, Strasser DC, Sinacore JM. Predicting stroke inpatient rehabilitation outcome using a classification tree approach. Archives of Physical Medic ine and Rehabilitation 1994;75:619-625.
16. Hasford J, Ansari H, Lehmann K. CART and logistic regression analyses of risk factors for first dose hypotension by an ACE-inhibitor. Therapie 1993;48:479-482.