Classification and Regression Trees as Alternatives to Regression

Mandy C. Phelps & Edgar C. Merkle
Department of Psychology, Fairmount College of Liberal Arts and Sciences

Abstract. Traditional statistical analyses, such as ANOVA and Regression, require that many assumptions be fulfilled in order to obtain accurate results. For example, there are assumptions of normally-distributed data, linear relationships between the dependent variable(s) and independent variables, and homogeneity of variance. In this presentation, we will describe the use of Classification and Regression Trees (CART) to sidestep the assumptions required by traditional analyses. CART has the added benefit of not requiring large sample sizes in order to obtain accurate results, although larger sample sizes are preferred. There is also a difference in the goal of CART compared to traditional analyses. CART is geared toward prediction, whereas traditional analyses are geared toward developing a specific model for your data. The poster will contain specific information about the procedures underlying CART, as well as an example involving data from a legibility of fonts study.

1. Introduction

Classification and Regression Trees are nonparametric therefore, they "can handle numerical data that are highly skewed or multi-modal, as well as categorical predictors with either ordinal or non-ordinal structure" [1]. This will save researchers time because they will not have to check the normality of their distribution, as well as other assumptions, which will allow them to focus more on the interpretation of their results. Researchers also have more leniencies in choosing which form of independent variables to use. Classification trees are capable of handling independent variables that are continuous, categorical, or a mixture of both, and unlike traditional statistics, actually perform better when there are numerous independent variables. The final tree will only include the independent variables that were predictive of the dependent variable; the other non-predictive independent variables will have no affect on the final results, as they do with more traditional statistics, such as regression.

Missing data also poses less of a problem with classification trees because classification trees have a built-in algorithm that readily handles missing data by replacing it with a surrogate variable that is most similar to the primary splitter [1]. The graphical presentation of the results also allows the researcher to see the complex interactions among the independent variables which is not as easily accomplished with traditional statistics. Lastly, the results of classification trees generalize well to new data when the necessary steps are taken in the growing and pruning process, which will be explained shortly. Now that some of the background has been covered it is time to move on to how classification trees actually work.

The structure of a classification tree includes parent nodes, child nodes, and terminal nodes. The parent node contains all observations that are to be divided into two groups. Child nodes are the nodes resulting from the splitting of the parent node. Each child node becomes a parent node itself and then splits into two additional child nodes. Nodes that satisfy the split criterion are labeled with a one and those that do not are labeled with a zero. This process continues until the tree has finished growing. Terminal nodes are the final nodes of the tree which contain the most heterogeneous subsets of the data and give the predictions for the observations that were ran down the tree.

Some important steps to follow when growing a tree according to Breiman [2] include a set of binary questions, a goodness of split criterion, a stopping rule, and determining how to assign every terminal node to a class. The binary questions determine if the observations goes to the right child node or left child node. They are set up as "yes" or "no" responses that if the observation meets the criterion then the observation goes to the left and if the observation does not meet the criterion it goes to the right. This process continues until all the observations have been have been divided into child nodes that are as different as possible. There is no set number of observations that are required to be in each node so they do not have to be divided equally, it is based on the properties of the observations and any criteria set by the researcher, such as a splitting criterion.

The next step is determining the goodness of split criterion. The objective here is to determine the split that results in the minimum amount of node impurity, which is the amount of difference between the observations within each node. There are several different techniques to determine when to split a node, the most commonly used technique is the Gini improvement measure. With this measure, values are determined to "indicate the greater difference with respect to the prevalence of the dependent measure in the two child nodes" [3]. Next, the "independent variable whose split provides the largest value of the improvement measure is selected for splitting at each step" [3].

Once the splitting criterion is set, the next step is to determine the stopping rules. This is an optional step but is sometimes used in order to reduce the likelihood of overfit, which reduces the probability that your results will generalize well to new data. The researcher can allow the tree to grow as large as possible but by setting a stopping rule the researcher determines how large the tree can become. This can be done in several ways. The researcher can set a minimum number of cases to be in each node, the maximum number of splits, or by setting a minimum value of the splitting criterion [3]. Researchers can use all three of these rules or however many they feel is necessary.

However, there is a potential problem with relying only on stopping rules to determine the size of your tree, which is the researcher, could potentially stop the tree too soon. Another strategy is to allow the tree to grow large and then use a procedure called pruning to get rid of unnecessary nodes. The concept of pruning has been compared to trimming the branches of a tree. Pruning begins with the terminal nodes and proceeds up the tree. Nodes that do not improve the predictive accuracy of the tree by a predetermined amount are removed. The one standard error rule is a procedure used where "the smallest tree whose cost is within on standard error of the tree with minimum cost is selected" [3]. It can also be used along with stopping rules however, the pruning method works best when used alone. Using the pruning method results in a tree that is the most generalizeable to future data sets.

The final step is to determine how to assign every terminal node to a class. There are cases where a terminal node is not assigned to a class or every terminal node belongs to a different class. The researcher should set class assignment rules, such as assumed prior probabilities or cost matrix, so that the terminal node falls in a class for which the probability of the node being in that class is the highest [4]. A class assignment rule reduces the likelihood of misclassification resulting in a more accurate tree, which will be able to better generalize to future data sets [2].

2. Experiment, Results, Discussion, and Significance

The font legibility experiment was conducted by Doug Fox & Dr. Barbara Chaparro for Microsoft, Inc. Microsoft was interested in which attributes determine font legibility. There were 20 different fonts being analyzed. Ten participants were tested over 3 trials resulting in 600 responses for each of the 15 commonly misclassified characters. Using Classification Trees we were able to determine the attributes that were most predictive of whether a font was legible.

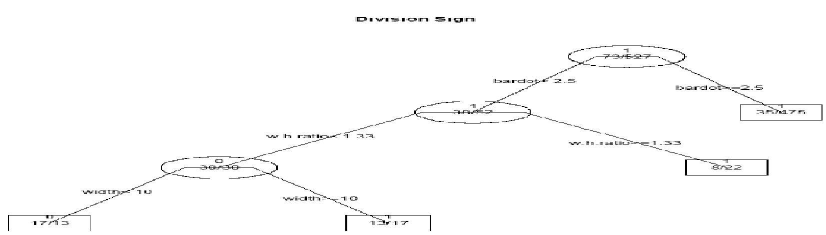

Fig. 1. Classification tree of predictive attributes for the division sign

3. Conclusions

Classification trees are a useful tool for psychologists. They require very few assumptions and can be carried out with relatively little effort. Results are more intuitive and generalize better to future data.

[1] Lewis, R. J. (2000). An introduction to classification and regression tree (CART) analysis. Paper presented at the annual meeting of the Society for Academic Emergency Medicine, San Francisco, CA.

[2] Breiman, L. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

[3] Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *Annals of Behavioral Medicine, 26(3),* 172-181.

[4]Dana, J., & Dawes, R. M. (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Education and Behavioral Statistics, 29(3),* 317-331.