Mobiles

Leland Wilkinson
SPSS, Inc.
Northwestern University
leland@spss.com

ABSTRACT: Mobiles are balanced trees for displaying the results of classification and regression trees. This paper describes these displays, implemented in the TREES module of SYSTAT, and discusses related research.

KEY WORDS: CART, Prediction Trees, Classification, AID, Recursive Partitioning

Leland Wilkinson is Senior VP, SPSS, Inc. and Adjunct Professor, Department of Statistics, Northwestern University. Email: leland@spss.com.

Prediction trees classify cases by successively splitting them into groups based on their values on a set of predictor variables. Morgan and Sonquist (1963) proposed a simple method for fitting these trees to data; they called it "automatic interaction detection" (AID). Kass (1980) extended the methodology to categorical data, calling it "chi-square AID" (CHAID). Breiman, Friedman, Olshen and Stone (1984) developed new algorithms and gave the previously ad-hoc methods more formal grounding in probability theory. Their computer program handled both categorical and continuous dependent variables, so they termed it "classification and regression trees" (CART). Quinlan (1986) developed independently a set of tree splitting algorithms based on information theory (ID3).

Graphical displays for these methods have not kept up with the pace of their numerical development. Most of the programs use standard tree displays derived from organizational chart software. One exception is the S system (Clark and Pregibon, 1992), which displays trees that can be modified dynamically. The S trees can also be vertically scaled to represent attributes such as split node importance. Another excepton is the classification tree display of Dirschedl (1991), also cited in Lausen et al. (1994) and Vach (1995), which represents subclass size by the width of a tree branch. The Dirchedl display is a variant of Hartigan and Kleiner's (1974) clustering tree icon.

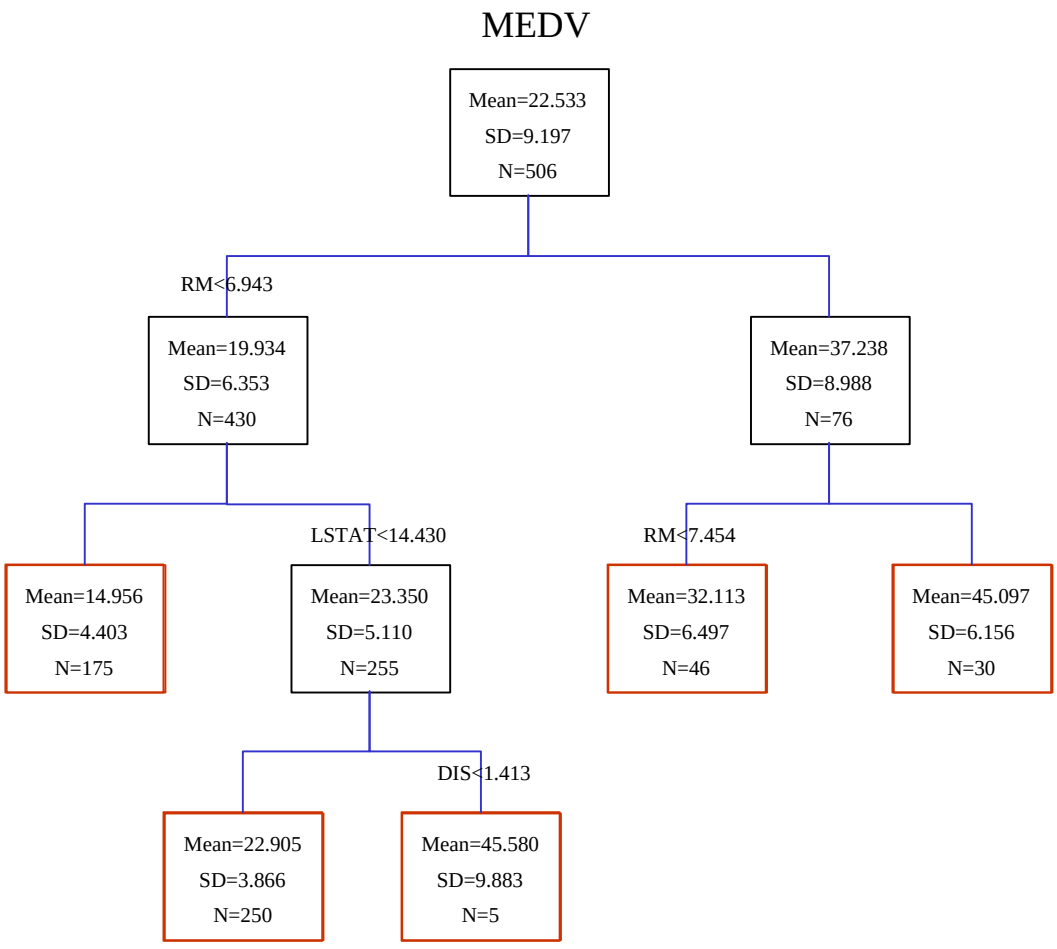
Figure 1 shows a typical classification tree for predicting median home prices from a set of environmental variables. This dataset was used in Breiman et al. (1984) to illustrate regression trees. It was originally collected by Harrison and Rubinfeld (1978) and used in Belsley, Kuh, and Welsch (1980) for demonstrating regression diagnostics. The variables in the dataset are:

- MEDV median value of owner-occupied homes in $1000's
- CRIM per capita crime rate by town
- ZN proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS proportion of non-retail business acres per town
- CHAS Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX nitric oxides concentration (parts per 10 million)
- RM average number of rooms per dwelling
- AGE proportion of owner-occupied units built prior to 1940
- DIS weighted distances to five Boston employment centers
- RAD index of accessibility to radial highways
- TAX full-value property-tax rate per $10,000
- PTRATIO pupil-teacher ratio by town
- B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT % lower status of the population

The loss function used for splitting nodes was based on ordinary least squares. Thus, the goodness of fit at each node is Fit=1-W/T, where W is within group sum-of-squares and T is total sum-of-squares on the dependent variable..

The first split is on RM (rooms). It separates 430 tracts with an average median value of 19.9 from 76 tracts with an average median value of 37.2. The left branch is then split by LSTAT (percent lower status of the population). The right branch is split by RM again. And so on.

FIGURE 1.
Classification tree for Boston Housing Data



This display has several drawbacks. It is difficult to discern the number of tracts in each node without reading text. Moreover, the effectiveness of the splits in classifying cases is not readily apparent.

Figure 2 remedies these two deficiencies. It is called a “mobile” because it balances nodes. If the terminal nodes were boxes and each case (tract) in a terminal node were a marble, then the tree would balance as shown. Several asymmetries are immediately apparent. The first split separates a relatively small proportion of high-valued tracts to the right of the tree. The most asymmetric split is at the lower left of the tree, where only 5 high-valued tracts are separated by DIS (distance to employment centers).

FIGURE 2.
Regression Mobile for Boston Housing Dataset

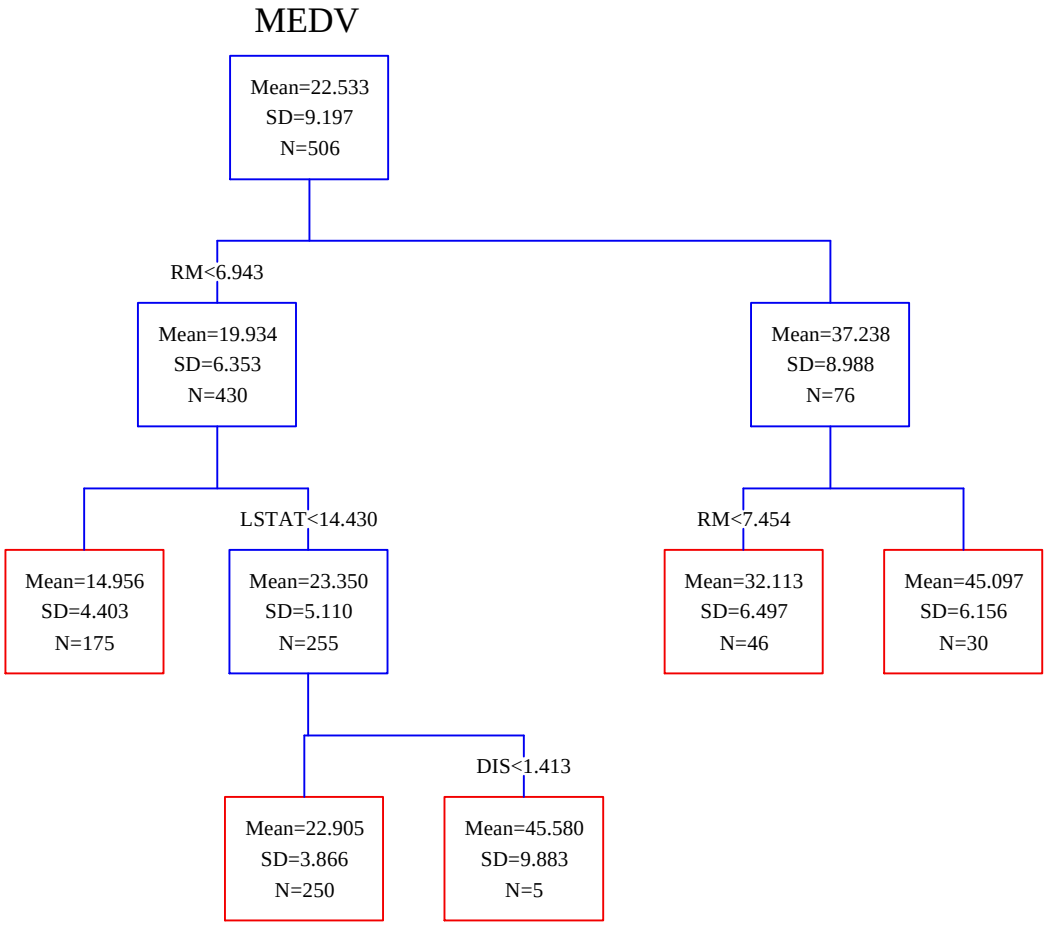
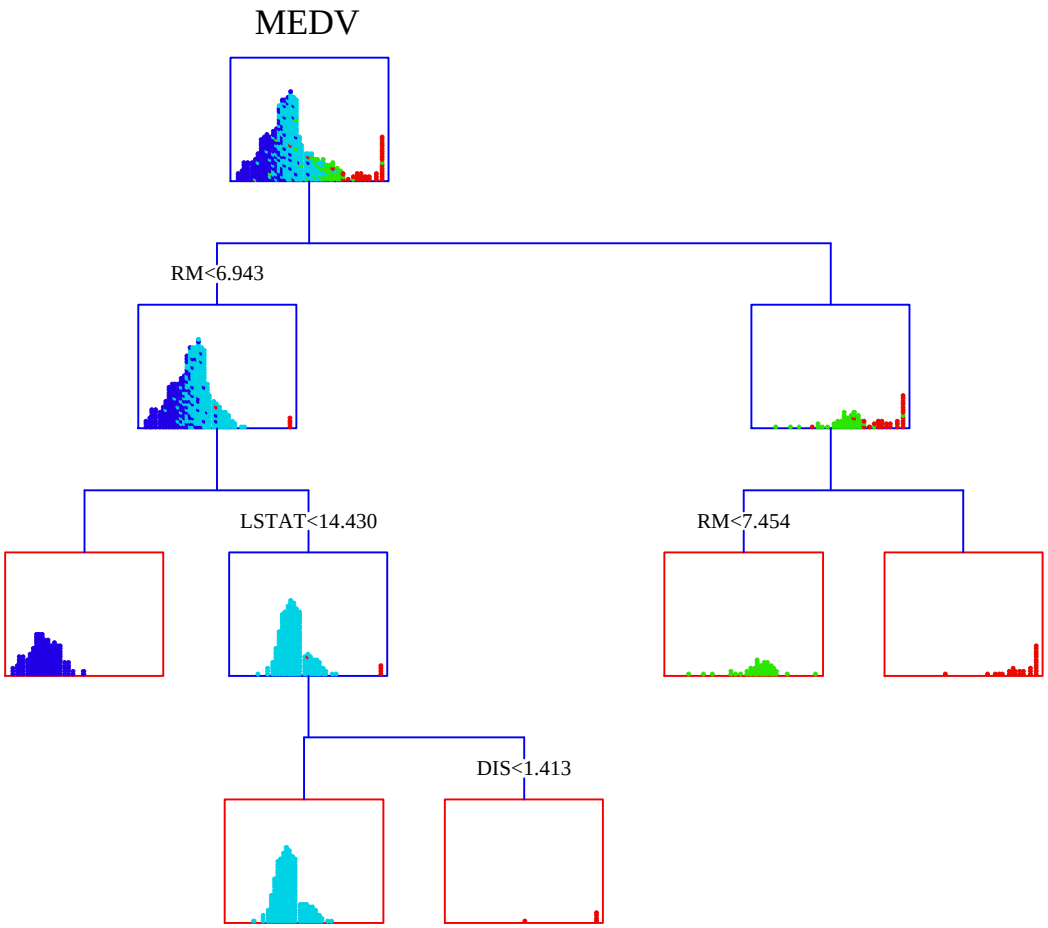


Figure 3 adds one more useful feature to the tree: a dot plot. This reveals the distribution of the dependent variable at each node. Dot plots are especially useful for this purpose because they work for continuous or categorical variables. Continuous variable dot plots resemble histograms or stem-and-leaf diagrams. Categorical dot plots look like bar charts. Furthermore, dot plots behave like a collection of marbles, which is consistent with the mobile metaphor. Coloring the dots according to the terminal node they end up in also helps reveal the stratification in the pooled dot plot at the top.

FIGURE 3.
Graphical Regression Mobile for Boston Housing Dataset



Conclusion

As Hartigan (1975) points out, the AID algorithm with a least squares loss function is prone to splitting off singeltons or outliers. Whether this is a defect or advantage in a given context, mobiles will reveal this behavior and assist in pruning. Interestingly, the Boston data mobile becomes fairly symmetric when 20 percent trimmed means are used for the loss function. Finally, graphical mobiles are especially suitable for large trees because text is not needed to reveal the conditional distributions.

References

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

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

Clark, L.A. and Pregibon, D. (1992). Tree-Based Models. In J.M. Chambers and T.J. Hastie (eds.), *Statistical Models in S.* Pacific Grove, CA: Wadsworth.

Dirschedl, P. (1991). Klassifikationsbaume - Grundlagen und Neuerungen. In W. Flischer, M. Nagel, and R. Ostermann (eds.), *Interaktive Datenalyse mit ISP.* Essen: Westart Verlag, 15-30.

Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management, 5,* 275-286.

Hartigan, J.A. (1975). *Clustering Algorithms.* New York: John Wiley & Sons.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29,* 119-127.

Kleiner, B., and Hartigan, J.A. (1981). Representing points in many dimensions by trees and castles. *Journal of the American Statistical Association, 76,* 260-269.

Lausen, B.,Sauerbrei, W., and Schumacher, M. (1994). Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In P. Dirschedl and R. Ostermann (eds). *Computational Statistics.* Heidelberg: Physica-Verlag, 483-496.

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.

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

Vach, W. (1995). Classification trees. *Computational Statistics, 10,* 9-14.