Hearne
Scientific Software

Australia
Phone: (03) 9602 5088
Fax: (03) 9602 5050
E-mail: info@hearne.com.au
Web: www.hearne.com.au

New Zealand
Phone: (09) 358 0350
Fax: (09) 359 0325
E-mail: info@hearne.co.nz
Web: www.hearne.co.nz

5th March 2004

CART's proven methodology is characterized by:

A reliable pruning strategy,

CART's developers determined definitively that no stopping rule could be relied on to discover the optimal tree, so they introduced the notion of over-growing trees and then pruning back; this idea, fundamental to CART, ensures that important structure is not overlooked by stopping too soon. Other decision tree techniques use problematic stopping rules.

A powerful binary split search approach, and

CART's binary decision trees are more sparing with data and detect more structure before too little data is left for learning. Other decision tree approaches use multi-way splits that fragment the data rapidly, making it difficult to detect rules that require broad ranges of data to discover.

Automatic self validation procedures.

In the search for patterns in databases it is essential to avoid the trap of "overfitting," or finding patterns that apply only to the training data. CART's embedded test disciplines ensure that the patterns found will hold up when applied to new data. Further, the testing and selection of the optimal tree are an integral part of the CART algorithm. Testing in other decision tree techniques is conducted after-the-fact and tree selection is left up to the user.

In addition, CART accommodates many different types of real world modeling problems by providing a unique combination of automated solutions:

Surrogate splitters intelligently handle missing values,

CART handles missing values in the database by substituting "surrogate splitters," which are back-up rules that closely mimic the action of primary splitting rules. The surrogate splitter contains information that is typically similar to what would be found in the primary splitter. Other products' approaches treat all records with missing values as if the records all had the same unknown value; with that approach all such "missings" are assigned to the same bin. In CART, each record is processed using data specific to that record; this allows records with different data patterns to be handled differently, which results in a better characterization of the data.

Adjustable misclassification penalties help avoid the most costly errors, and

CART can accommodate situations in which some misclassifications, or cases that have been incorrectly classified, are more serious than others. CART users can specify a higher penalty for misclassifying certain data, and the software will steer the tree away from that type of error. Further, when CART cannot guarantee a correct classification, it will try to ensure that the error it does make is less costly. If credit risk is classified as low, moderate, or high, for example, it would be much more costly to classify a high risk person as low risk than as moderate risk. Traditional data mining tools cannot distinguish between these errors.

Alternative splitting criteria make progress when other criteria fail.

CART includes seven single variable splitting criteria - Gini, symmetric Gini, twoing, ordered twoing and class probability for classification trees, and least squares and least absolute deviation for regression trees - and one multi-variable splitting criteria, the linear combinations method. The default Gini method typically performs best, but, given specific circumstances, other methods can generate more accurate models. CART's unique "twoing" procedure, for example, is tuned for classification problems with many classes, such as modeling which of 170 products would be chosen by a given consumer. To deal more effectively with select data patterns, CART also offers splits on linear combinations of continuous predictor variables.