Decision Tree Induction

Roberta Siciliano
University of Naples Federico II, Italy

Claudio Conversano
University of Cassino, Italy

INTRODUCTION

Decision Tree Induction (DTI) is an important step of the segmentation methodology. It can be viewed as a tool for the analysis of large datasets characterized by high dimensionality and nonstandard structure. Segmentation follows a nonparametric approach, since no hypotheses are made on the variable distribution. The resulting model has the structure of a tree graph. It is considered a supervised method, since a response criterion variable is explained by a set of predictors.

In particular, segmentation consists of partitioning the objects (also called cases, individuals, observations, etc.) into a number of subgroups (on the basis of suitable partitioning of the modalities of the explanatory variables, the so-called predictors) in a recursive way, so that a tree-structure is produced. Typically, partitioning is in two subgroups yielding to binary trees, although ternary trees as well as r-way trees also can be built up.

Two main targets can be achieved with tree-structures—classification and regression trees—on the basis of the type of response variable, which can be categorical or numerical.

Tree-based methods are characterized by two main tasks: exploratory and decision. The first is to describe with the tree structure the dependence between the response and the predictors. The decision task is properly of DTI, aiming to define a decision rule for unseen objects for estimating unknown response class/values as well as validating the accuracy of the final results.

For example, trees often are considered in credit-scoring problems in order to describe and classify good and bad clients of a bank on the basis of socioeconomic indicators (e.g., age, working conditions, family status, etc.) and financial conditions (e.g., income, savings, payment methods, etc.). Conditional interactions describing the client profile can be detected looking at the paths along the tree, when going from the top to the terminal nodes. Each internal node of the tree is assigned a partition (or a split for binary tree) of the predictor space, and each terminal node is assigned a label class/value of the response. As a result, each tree path, characterized by a sequence of predictor interactions, can be viewed as a production rule yielding to a specific label class/value. The set of production rules constitutes the predictive learning of the response class/value of new objects, where only measurements of the predictors are known. As an example, a new client of a bank is classified as a good client or a bad one by dropping it down the tree according to the set of splits (binary questions) of a tree path, until a terminal node labeled by a specific response-class is reached.

BACKGROUND

The appealing aspect for the segmentation user is that the final tree provides a comprehensive description of the phenomenon in different contexts of application, such as marketing, credit scoring, finance, medical diagnosis, and so forth.

Segmentation can be considered as an exploratory tool but also as a confirmatory nonparametric model. Exploration can be obtained by performing a recursive partitioning of the objects until a stopping rule defines the final structure to interpret. Confirmation is a different problem, requiring definition of decision rules, usually obtained by performing a pruning procedure soon after a partitioning one. Important questions arise when using segmentation for predictive learning goals (Hastie et al., 2001; Zhang, 1999). The tree structure that fits the data and can be used for unseen objects cannot be the simple result of any partitioning algorithm. Two aspects should be jointly considered: the tree size (i.e., the number of terminal nodes) and the accuracy of the final decision rule evaluated by an error measure. In fact, a weak point of decision trees is the sensitivity of the classification/prediction rules measured by the size of the tree and its accuracy to the type of dataset as well as to the pruning procedure. In other words, the ability of a decision tree to detect cases and take right decisions can be evaluated by a simple measure, but it also requires a specific induction procedure. Likewise, in statistical inference, where the power of a testing procedure is judged with respect to changes of the alternative hypotheses, decision tree induction strongly

depends on both the hypotheses to verify and their alternatives. For instance, in classification trees, the number of response classes and the prior distribution of cases among the classes influence the quality of the final decision rule. In the credit-scoring example, an induction procedure using a sample of 80% of good clients and 20% of bad clients likely will provide reliable rules to identify good clients and unreliable rules to identify bad ones.

MAIN THRUST

Exploratory trees can be fruitfully used to investigate the data structure, but they cannot be used straightforwardly for induction purposes. The main reason is that exploratory trees are accurate and effective with respect to the training data used for growing the tree, but they might perform poorly when applied to classifying/predicting fresh cases that have not been used in the growing phase.

DTI Main Tasks

DTI definitely has an important purpose represented by understandability: the tree structure for induction needs to be simple and not large; this is a difficult task since a predictor may reappear (even though in a restricted form) many times down a branch. At the same time, a further requirement is given by the identification issue: on one hand, terminal branches of the expanded tree reflect particular features of the training set, causing over-fitting; on the other hand, over-pruned trees necessarily do not allow identification of all the response classes/values (under-fitting).

Tree Model Building

Simplification method performance in terms of accuracy depends on the partitioning criterion used in the tree-growing procedure (Buntine & Niblett, 1992). Thus, exploratory trees become an important preliminary step for DTI. In tree model building, it is worth distinguishing between the optimality criterion for tree pruning (simplification method) and the criterion for selecting the best decision rule (decision rule selection). These criteria often use independent datasets (training set and test set). In addition, a validation set can be required to assess the quality of the final decision rule (Hand, 1997). In this respect, segmentation with pruning and assessment can be viewed as stages of any computational model-building process based on a supervised learning algorithm. Furthermore, growing the tree structure using a Fast Algorithm for Splitting Trees (FAST) (Mola & Siciliano, 1997) becomes a fundamental step to speed up the overall DTI procedure.

Tree Simplification: Pruning Algorithms

A further step is required for DTI relying on the hypothesis of uncertainty in the data due to noise and residual variation. Simplifying trees is necessary to remove the most unreliable branches and improve understandability. Thus, the goal of simplification is inferential (i.e., to define the structural part of the tree and reduce its size while retaining its accuracy). Pruning methods consist in simplifying trees in order to remove the most unreliable branches and improve the accuracy of the rule for classifying fresh cases.

The pioneer approach of simplification was presented in the Automatic Interaction Detection (AID) of Morgan and Sonquist (1963). It was based on arresting the recursive partitioning procedure according to some stopping rule (pre-pruning).

Alternative procedures consist in pruning algorithms working either from the bottom to the top of the tree (post-pruning) or vice versa (pre-pruning). CART (Breiman et al., 1984) introduced the idea to grow the totally expanded tree for removing retrospectively some of the branches (post-pruning). This results in a set of optimally pruned trees for the selection of the final decision rule.

The main issue of pruning algorithms is the definition of a complexity measure that takes account of both the tree size and accuracy through a penalty parameter expressing the gain/cost of pruning tree branches. The training set is often used for pruning, whereas the test set for selecting the final decision rule. This is the case of both the error-complexity pruning of CART and the critical value pruning (Mingers, 1989). Nevertheless, some methods require only the training set. This is the case of the pessimistic error pruning and the error-based pruning (Quinlan, 1987, 1993) as well as the minimum error pruning (Cestnik & Bratko, 1991) and the CART cross-validation method. Instead, other methods use only the test set, such as the reduced error pruning (Quinlan, 1987). These latter pruning algorithms yield to just one best pruned tree, which represents in this way the final rule.

In DTI, accuracy refers to the predictive ability of the decision tree to classify/predict an independent set of test data. In classification trees, the error rate, measured by the number of incorrect classifications of the tree on test data, does not reflect accuracy of predictions for classes that are not equally likely, and those with few cases are usually badly predicted. As an alternative to the CART pruning, Cappelli, et al. (1998) provided a pruning algorithm based on the impurity-complexity measure to take account of the distribution of the cases over the classes.