Classification and Regression Trees

Johannes Gehrke
Cornell University, USA

INTRODUCTION

It is the goal of classification and regression to build a data-mining model that can be used for prediction. To construct such a model, we are given a set of training records, each having several attributes. These attributes either can be numerical (e.g., age or salary) or categorical (e.g., profession or gender). There is one distinguished attribute—the dependent attribute; the other attributes are called predictor attributes. If the dependent attribute is categorical, the problem is a classification problem. If the dependent attribute is numerical, the problem is a regression problem. It is the goal of classification and regression to construct a data-mining model that predicts the (unknown) value for a record, where the value of the dependent attribute is unknown. (We call such a record an unlabeled record.) Classification and regression have a wide range of applications, including scientific experiments, medical diagnosis, fraud detection, credit approval, and target marketing (Hand, 1997).

Many classification and regression models have been proposed in the literature; among the more popular models are neural networks, genetic algorithms, Bayesian methods, linear and log-linear models and other statistical methods, decision tables, and tree-structured models, which is the focus of this article (Breiman, Friedman, Olshen & Stone, 1984). Tree-structured models, so-called decision trees, are easy to understand; they are non-parametric and, thus, do not rely on assumptions about the data distribution; and they have fast construction methods even for large training datasets (Lim, Loh & Shih, 2000). Most data-mining suites include tools for classification and regression tree construction (Goebel & Gruenwald, 1999).

BACKGROUND

Let us start by introducing decision trees. For the ease of explanation, we are going to focus on binary decision trees. In binary decision trees, each internal node has two children nodes. Each internal node is associated with a predicate, called the splitting predicate, which involves only the predictor attributes. Each leaf node is associated with a unique value for the dependent attribute. A decision encodes a data-mining model as follows. For an unlabeled record, we start at the root node. If the record satisfies the predicate associated with the root node, we follow the tree to the left child of the root, and we go to the right child otherwise. We continue this pattern through a unique path from the root of the tree to a leaf node, where we predict the value of the dependent attribute associated with this leaf node. An example decision tree for a classification problem, a classification tree, is shown in Figure 1. Note that a decision tree automatically captures interactions between variables, but it only includes interactions that help in the prediction of the dependent attribute. For example, the rightmost leaf node in the example shown in Figure 1 is associated with the classification rule: “If (Age >= 40) and (Gender=male), then YES”; as classification rule that involves an interaction between the two predictor attributes age and salary.

Decision trees can be mined automatically from a training database of records, where the value of the dependent attribute is known: A decision tree construction algorithm selects which attribute(s) to involve in the splitting predicates, and the algorithm decides also on the shape and depth of the tree (Murthy, 1998).

MAIN THRUST

Let us discuss how decision trees are mined from a training database. A decision tree usually is constructed in two phases. In the first phase, the growth phase, an overly large and deep tree is constructed from the training data. In the second phase, the pruning phase, the final size of the tree is determined with the goal to minimize the expected misprediction error (Quinlan, 1993).

Figure 1. An example classification tree