Use of Distribution Algorithms, for the Construction of a Classification and Regression Tree

MSc. Adem Meta
Cuyahoga Community College, USA

Abstract

One of the most important processes in the construction of the classification and regression trees is the distribution of a given data. There are numerous algorithms for predicting continuous variables or categorical variables from a set of continuous predictors and/or categorical factor effects. In this paper I address the problem of learning various types of algorithms to be used to get a optimal decision trees from data base. In particular, we study online machine learning algorithms for learning classification and regression trees, linear model trees, option trees for regression, multi-target model trees, and ensembles of model trees from a given data. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. The core algorithm for building decision trees called, greedy search through the space of possible branches uses *Entropy* and *Information Gain* to construct a decision tree. In this paper, through a concrete example, I will explicitly look at the use of four algorithms such as the Gini Index, Chi-Square, Entropy and the Variance Reduction on which node will be the distribution of a database. Once the data base is small, no doubt that the calculation is much more simple than in the case of a large database.

Keywords: distribution algorithms, regression tree, classification.

Full Text: PDF