R-TREE: Implementation of Decision Trees using R

Margaret Miró-Juliá[1,⋆], Arnau Mir[1] and Monica J. Ruiz-Miró[2]

1. Departamento de Ciencias Matemáticas e Informática, Universidad de las Islas Baleares, SPAIN
2. Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, SPAIN

⋆Contact author: margaret.miro@uib.es

Keywords: R-software, classification, attribute selection, information gain.

The work presented here deals with the application of intelligent approaches to data analysis using R. The use of intelligent strategies allow for the construction of efficient patterns or structures for data classification. Knowledge discovery in Databases can be viewed as a three phase process: the data processing phase, the data mining phase and the evaluation phase. The main elements are the original data base, the transformed or processed data, the patterns or models of classified data and the knowledge extracted from the data. The data processing phase selects from the original data base a data set that focuses on a subset of attributes or variables on which knowledge discovery has to be performed; it also removes outliers and redundant information. The data mining phase converts this target data into useful patterns. Among all available classification methods, decision trees were selected for their simplicity and intuitiveness. The evaluation phase proves the consistency of the decision tree by means of a testing set.

Decision tree structures provide a common and easy way to organize data. Every decision tree begins with a root node. Each node in the tree evaluates an attribute from the data and determines which path it should follow. Classification using a decision tree is performed by routing from the root node until arriving at a leaf node. Decision trees can represent different types of data. The simplest and most familiar is numerical data. It is often desirable to organize nominal data as well. Nominal quantities are formally described by a discrete set of symbols. The type of data organized by a tree is important for understanding how the tree works at the node level. Recalling that each node is effectively a test, numeric data is often evaluated in terms of simple mathematical inequality, whereas nominal data is tested in Boolean fashion.

Decision tree induction algorithms function recursively. First, an attribute must be selected as the root node. In order to create the most efficient tree, the root node must effectively split the data. Each split attempts to pare down a set of instances (the actual data) until they all have the same classification. The best split is the one that provides what is termed as the most information gain. Information in this context comes from the concept of entropy from information theory developed by Claude Shannon. Although information has many contexts, it has a very specific mathematical meaning relating to certainty in decision making. Information can be expressed as a mathematical quantity as follows: $I = -\sum_{i=1}^{m} p_i \log_2 p_i$. Information gain is defined as information before splitting minus information after splitting. Ideally, each split in the decision tree should bring us closer to a classification.

The aim of this paper is to present an algorithm written with R: the `R-tree` algorithm. This algorithm will implement a decision tree by calculation of the information gain of the different attributes available in the data frame. This will allow for the selection of the attribute that offers the best split at each level of the tree. An important feature of this algorithm is that it handles both nominal and numerical data.

References

Fayyed U., Piatetsky-Shapiro G., Smyth P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence, AI Magazine Fall 96*, 37–54.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, vol. 1, 81–106.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 4, 379–423.