Tutorial 3: Data Mining with Decision Trees

Johannes Gehrke
Cornell University

In this tutorial, we survey recent developments in learning tree-based models for classification and regression called predictor trees. The tutorial has three parts: (1) A general overview of tree-based classification and regression. (2) A survey of methods to construct predictor trees. (3) An overview of scalable data access methods to construct predictor trees from very large training databases.

In the first part, we motivate predictor trees and their use in a data mining environment. We show results from real-life studies that illustrate how predictor trees give understandable models where traditional models are hard or counter-intuitive to interpret, and compare related methods for classification and regression.

In the second part of the tutorial, we discuss choices involved in tree construction, including different split selection methods and tree pruning. Although we survey the most popular methods, including work from all KDD sub-communities, we emphasize recent work from the statistics literature. Wherever possible, we interleave results on real datasets. The methods presented in this part assume that the complete training database fits into main memory.

The third part of the tutorial covers scalable methods for predictor tree construction. We first motivate the concept of scalability and then survey recent work in the database literature on scalable data access methods for constructing predictor trees from very large training databases.

Bio

Johannes Gehrke is an assistant professor at Cornell University. He received his Ph.D. from the University of Wisconsin-Madison in 1999. His graduate research has been supported by a Fulbright scholarship and an IBM Graduate Fellowship. Johannes Gehrke’s research interests are in the areas of next-generation database architectures, knowledge discovery from databases, and mining and monitoring evolving datasets, and he is the co-author of two patents in this area.