Chapter 20
TreeWorks:
Advances in Scalable Decision Trees

Paul Harper
Cardiff University, UK

Evandro Leite Jr.
University of Southampton, UK

ABSTRACT

Decision trees are hierarchical, sequential classification structures that recursively partition the set of observations (data) and are used to represent rules underlying the observations. This article describes the development of TreeWorks, a tool that enhances existing decision tree theory and overcomes some of the common limitations such as scalability and the ability to handle large databases. We present a heuristic that allows TreeWorks to cope with observation sets that contain several distinct values of categorical data, as well as the ability to handle very large datasets by overcoming issues with computer main memory. Furthermore, our tool incorporates a number of useful features such as the ability to move data across terminal nodes, allowing for the construction of trees combining statistical accuracy with expert opinion. Finally, we discuss ways that decision trees can be combined with Operational Research health care models, for more effective and efficient planning and management of health care processes.

INTRODUCTION

Since the second half of the 20th Century, an upsurge of electronic data has been taking place worldwide. Studies such as Frawley, Piatestsky-Shapiro, and Matheus (1991) show that the amount of data is doubling each year. Contributing factors for the data explosion include the widespread use of computer systems for nearly any commercial, financial, governmental, or research activity; easier access to large storage capacity media; and advances in data collection tools. The Internet and all of its related services like the World Wide Web, e-mail, and online databases as a global information system have flooded humanity with a tremendous amount of data and information.

With the continuous growth in size and complexity of information, there is an urgent need for a new generation of computational theories and tools to assist us in extracting useful information

(knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the fields of data mining and knowledge discovery in databases (KDD). The need to understand data is extremely important for business, government, and research. Examples are numerous and varied and include applications in optimising market shares by increasing competitive advantage using knowledge extracted from sales transactions (Rygielski, Wang, & Yen, 2002), maximising customer retention (Edelstein, 1998), predicting the size of television audiences (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro, & Simoudis, 1996), and improving ovarian cancer detection (Li et al., 2004).

Decision trees are one such data mining technique that learn from data and generate models containing explicit rule-like relationships among the variables. Decision tree algorithms begin with the entire training set of data, split into two or more subsets until the split size reaches an appropriate level. The entire modelling process can be visualised in a tree structure. This structure maps observations between dependent and independent variables. An arc between two nodes in the tree represents a partition of the parent node into child nodes. All observations follow a path from the root (initial node) and are assigned to a leaf (terminal node) based on splitting criteria (values of the independent variables). The two best-known and most widely used decision tree algorithms are Classification and Regression Trees (CART) and C4.5 (a successor of ID3). CART was developed by the statisticians Breiman, Friedman, Olshen, and Stone (1984), and C4.5 was developed by Quinlan, a computer scientist in the field of machine learning (Quinlan, 1993).

The original methods to grow decision trees are not ideal for handling some of the features that are present in modern-day data sets such as categorical variables with many distinct values and the ability to handle extremely large datasets. In order to help overcome such issues, Catlett (1991) proposed sampling at each node of the classification tree, but considers in his studies only datasets that could fit in main memory which were rather limited in size. Methods for partitioning the dataset such that each subset fits in main memory are considered by Chan and Stolfo (1993). Even though this method makes classification of large datasets possible, their studies show that the quality of the resulting decision tree is not as good as if the classifier had used all the available data. Most existing methods for automatic construction of classification trees utilise greedy heuristics, choosing locally optimal splits to divide the data at each level. Unfortunately these locally optimal values cannot be obtained quickly when a large dataset is analysed.

In this article, we present TreeWorks, a CART tool that enhances existing decision tree theory and overcomes some of these common decision tree limitations. Whilst the primary focus of our research, as presented here, is on scalable decision trees, the resulting TreeWorks tool is highly practical and user-friendly and has already found considerable application by the UK National Health Service (NHS). The NHS handles millions of patient records each year and TreeWorks has assisted the NHS Information Centre with the redesign of Healthcare Resource Groups (HRGs). HRGs, which are similar to DRGs as used in the U.S., are standard groupings of clinically similar treatments which use common levels of healthcare resource, and are fundamental for standardising healthcare commissioning across the country as part of the UK Government's policy of Payment by Results (PbR) (Department of Health, 2008).

In this article, we also highlight ways in which decision trees can support Operational Researchers building health care models. A particular feature of health care processes is the inherent variation and uncertainty in treating individuals. Homogeneity leads to increased certainty in individual patient predictions (resource consumption, outcomes, pathways, etc.), which in turn results in the potential for more effective and efficient planning and management of health care processes.