THESIS ABSTRACT:

Scalable and Efficient Algorithms and Techniques for Learning Large Decision Trees

Introduction

Decision tree learning is one of the most widely used and practical methods for inductive learning. Among early and basic works on decision tree learning are Hunt's Concept Learning System (Hunt et al. 1966), Friedman and Breiman's CART system (Friedman 1977; Breiman et al. 1984), and Quinlan's ID3 system (Quinlan 1986). Following them, numerous researches continue searching for alternative techniques to improve the effectiveness (predictive accuracy) and the efficiency of decision tree learning. The development of decision tree learning leads to and is encouraged by a growing number of commercial systems such that C5.0/See5 (RuleQuest Research), MineSet (SGI), and Intelligent Miner (IBM).

Classical research issues in decision tree learning include splitting criteria, pruning, missing values, discretization, and nonstandard forms of decision trees (e.g., oblique decision trees). When decision tree learning becomes one of the most applicable methods in data mining, a rapidly growing area of research and application, a new challenge is how the method deals with very large databases. For that, several research issues arise such as data reduction, scalable algorithms, and visualization of large data and decision trees.

My research in decision tree learning has three objectives: a splitting criterion based on rough sets, a visualization technique for large decision trees, and a scalable algorithm for rule post-pruning. For the first objective, as there is no dominated splitting criterion, my research aims to provide an alternative criterion that may outperforms other criteria in several circumstances. My second objective is developing a visualization technique for large decision trees. This objective arises from the fact that though visualizing decision trees is very helpful for the user to understand and use them, it is extremely difficult to view and navigate the large ones with available techniques. My last objective concerns the problem of rule post-pruning in decision tree learning. As a rule set has advantages over a decision tree in many cases, some decision tree learning systems such as C4.5 (Quinlan 1993) uses a rule post-pruning algorithm to generate rules from a decision tree. However, due to the algorithm complexity, it fails to deal with large databases. My research aims at developing a more scalable algorithm for generating rules from decision trees.

Research Content

My research focuses on three topics in decision tree learning: attribute selection, tree visualization, and rule post-pruning.

A Measure for Attribute Selection based on Rough Sets

The performance of a decision tree learning system depends mainly on techniques to solve two problems: attribute selection and pruning. For attribute selection, there is no dominated criterion to select attributes, a criterion may fit for some applications but does not for others. Most criteria (measures) for attribute selection are either information theory-based such as information gain and gain-ratio (Quinlan 1993), or statistics-based such as χ^2, gini-index (Breiman et al. 1984), etc. There are still many researches searching for alternative criteria to increase the predictive accuracy or to reduce the size of generated decision trees. Beginning with an extended model of rough sets, my research aims at developing an attribute selection criterion that may outperform other criteria in several circumstances.

In developing the new criterion, I propose a variant of the probabilistic model of rough sets (Pawlak 1998) in order (1) to overcome the limitations of the original model in case of noisy data, (2) to make the model more coherent, and (3) to preserve the convenience of non-parameter. Based on this model, R-measure is developed to measure how much the class attribute depends on a predictive attribute. Using R-measure as an attribute selection criterion, an experimental comparative evaluation on 18 datasets from UCI repository of machine learning shows that it can be considered as a good alternative criterion for attribute selection.

Visualizing Large Decision Trees

Though a decision tree is a simple notion, we can understand its content and hierarchical structure easily if it is small but cannot understand or understand difficultly if it is large. Research on visualization of decision trees has recently received a great attention from the data mining community because of its practical importance. Many works have been done, e.g., the 3D Tree Visualizer in the system MineSet, CAT scan (classification aggregation tablet) for inducing bagged decision trees (Rao 1997), the interactive visualization in decision tree construction (Ankerst 1999), the tree visualizer with a tree map in the system CART, etc. However, it is difficult to view and navigate large trees with these systems. On the other hand, there are also several approaches in the field of information visualization for visualizing large trees, e.g., cone tree (Robertson 1991), Tree-Map (Johnson 1991), and hyperbolic trees (Lamping 1997), but none of them seems directly appropriate for decision trees learning.

To fulfill the need of appropriate visualizers for large decision trees, I am developing a new technique called T2.5D (Tree 2.5 Dimensions) that combines the clearness of 2D displaying techniques and the compactness of 3D ones. At a moment, T2.5D displays the active path of a tree in a 2D form to give a clear view and other nodes in a 3D form to save space. T2.5D has several advantages comparing to other techniques: (1) it easily handles decision trees with more than 20000 nodes, and more than 1000 nodes can be displayed together on the screen, (2) it gives the user a clear view of an active path and an image of the overall structure of the tree at the same time, (3) it facilitates the tree navigation as only a minimum number of operations (e.g., click, scroll, etc.) is needed.

Generating Rules from Decision Trees

If-then rules are the basis for some of the most popular concept description languages used in machine learning. A variety of approaches to learning rules have been investigated. One is to begin by generating a decision tree, then to transform it into a rule set, and finally to simplify the rules such as that in C4.5 (Quinlan 1993). Another is to use the ‘‘separate-and-conquer” strategy such as that in RIPPER (Cohen 1995). The first approach usually requires a complex process of global optimization and therefore extremely slow on large databases, while the second approach can lead to a particularly problematic form of over-pruning. A combined approach is proposed in PART (Frank et al. 1998) to overcome those weak points. However, to make a single rule, PART has to build a pruned decision tree then choose the most ‘‘effective” path. That makes the algorithm quite complex (the time complexity of the algorithm is approximately $O(n^2)$.

I propose a new algorithm that also is a combination of two initial approaches but to generate a rule set just only one pruned tree needs to be built (therefore the complexity will be approximately O(nlogn)). In essence, to generate a rule set, the algorithm first uses the ‘‘separate-and-conquer” strategy to build a special decision tree called an attribute-value decision tree. After that a pruning technique will be applied to that tree, and the last step is to transform directly the pruned tree to a rule set. The efficiency (time complexity) of the algorithm is theoretically proved. I am doing an experiment comparative evaluation to confirm its effectiveness (predictive accuracy).