PREFACE

Recent technological advances have facilitated the mass collection and gathering of data. Consequently, it is of growing importance to develop tools and models that enable us to better understand this data and to design data-driven algorithms that leverage this information. Decision trees are potentially powerful predictors and explicitly represent the structure of a dataset. A complete description of decision tree and related study is given in chapter 1 and 2. This dissertation contains a set of contributions that deal with classification in data mining, in particular. Each contribution can be considered as a solution to a specific problem, in an attempt to map out a common ground. It provides several fundamental theoretical and experimental results.

In this thesis, we have developed a new composite splitting criterion, Randomized Gain Ratio (RGR) and a novel pruning technique, Expert Knowledge Based Pruning (EKBP) for generating efficient and simpler decision trees. The use of this composite splitting criterion is motivated by the fact that gain ratio (the heuristic function in C4.5) selects improper attributes when the attributes are irrelevant and independent with each other. Using composite splitting criterion, results in much more simpler trees while retaining or improving the classification accuracy. A detail analysis about RGR is given in Chapter 3. The use of the Expert Knowledge Based Pruning (EKBP) is motivated by the idea of using expert knowledge in decision tree pruning for giving users more flexibility in incorporating domain specific knowledge in pruning to obtain optimal results for increased performance in prediction. A complete description of EKBP is in Chapter 4. Finally, in Chapter 5, we implemented our new composite splitting criterion, RGR and new pruning technique, EKBP to form a decision tree model. We call our model as CURE – ‘Classifier Using Randomization and Expert knowledge’.

We then discuss the application of our approach as a case study in Chapter 6 regarding a heart disease dataset collected from local hospital and where we exhibit the benefit of our approach for early detection of heart disease by identification of the bio-markers by using our improved decision tree.