Minimizing Structural Risk on Decision Tree Classification

DaeEun Kim

Cognitive Robotics
Max Planck Institute for Human Cognitive & Brain Sciences
Munich, 80799, Germany
daeeun@cbs.mpg.de

Summary. Tree induction algorithms use heuristic information to obtain decision tree classification. However, there has been little research on how many rules are appropriate for a given set of data, that is, how we can find the best structure leading to desirable generalization performance. In this chapter, an evolutionary multi-objective optimization approach with genetic programming will be applied to the data classification problem in order to find the minimum error rate or the best pattern classifier for each size of decision trees. As a result, we can evaluate the classification performance under various structural complexity of decision trees. Following structural risk minimization suggested by Vapnik, we can determine a desirable number of rules with the best generalization performance. The suggested method is compared with C4.5 application for machine learning data.

11.1 Introduction

The recognition of patterns and the discovery of decision rules from data examples is one of the challenging problems in machine learning. When data points with numerical attributes are involved, the continuous-valued attributes should be discretized with threshold values. Decision tree induction algorithms such as C4.5 build decision trees by recursively partitioning the input attribute space [24]. Thus, a conjunctive rule is obtained by following the tree traversal from the root node to each leaf node. Each internal node in the decision tree has a splitting criterion or threshold for continuous-valued attributes to partition a part of the input space, and each leaf represents a class depending on the conditions of its parent nodes.

The creation of decision trees often relies on heuristic information such as information gain measurement. Yet how many nodes are appropriate for a given set of data has been an open question. Mitchell [22] showed the curve of the accuracy rate of decision trees with respect to the number of nodes over the independent test examples. There exists a peak point of the accuracy rate in a certain size of decision trees; a larger size of decision trees can
increase its classification performance on the training samples but reduces the accuracy over the test samples which have not been seen before. This problem is related to the overfitting problem to increase the generalization error\(^1\). Many techniques such as tree growing with stopping criterion, tree pruning or bagging [25, 21, 24, 5] have been studied to reduce the generalization error. However, the methods are dependent upon a heuristic information or measure to estimate the generalization error, and they do not explore every size of trees.

An evolutionary approach to decision trees has been studied to obtain optimal classification performance [16, 15, 4], since the decision tree based on heuristics is not optimal in structure and performance. Freitas et al. [15] have shown evolutionary multi-objective optimization to obtain both the minimum error rate and minimum size of trees. Their method was based on the information gain measurement; it followed the C4.5 splitting method and selected the attributes with genetic algorithms. They were able to reduce the size of decision trees, but had higher test error rates than C4.5 in some data sets. Recently a genetic programming approach with evolutionary multi-objective optimization (EMO) was applied to decision trees [4, 3]. A new representation of decision trees for genetic programming was introduced [3], where the structure of decision trees is similar to linear regression trees [6]. Two objectives, tree size and accuracy rate in data classification, were considered in the method. The method succeeded in reducing both error rates and size of decision trees in some data sets. However, searching for the best structure of decision trees has not been considered in their works.

It has been shown that EMO is very effective for optimization of multi-objectives or constraints in continuous range [27, 7]. Also EMO is a useful tool even when the best performance for each discrete genotype or structure should be determined [19, 17]. The EMO approach was used to minimize the training error and the tree size for decision tree classification [3, 8]. Also other works using fitness and size or complexity as objectives have been reported [2, 20]. Yet there has been no effort so far to find what is the best structure of decision trees to have the minimal generalization error. Vapnik [26] showed an analytic study to reduce the generalization error, and he suggested the structural risk minimization to find the best structure. It can be achieved by exploring the empirical error (training error) and generalization error (test error) for various structure complexity. We will follow the approach and the tree size will be the parameter to control the structure.

In this work, the EMO with genetic programming for two objectives, the tree size and the training error, is first used to obtain the Pareto-optimal solutions, that is, the minimum training error rate for each size of trees. Then the best tree for each size will be examined to see the generalization performance for a given set of test data. By observing the distribution of the test error rates over the size of trees, we can pinpoint the best structure to minimize the generalization error. In our EMO approach, a special elitism strategy

\(^{1}\) This is also called test error in this chapter.