Weighted Majority Decision among Several Region Rules for Scientific Discovery

Akihiro Nakaya, Hideharu Furukawa*, and Shinichi Morishita

University of Tokyo, 4-6-1 Shirokanedai, Minato Ward, Tokyo 108-8639, JAPAN
{nakaya, moris}@ims.u-tokyo.ac.jp, furukawa@is.s.u-tokyo.ac.jp

Abstract. We consider the classification problem of how to predict the values of a categorical attribute of interest using the other numerical attributes in a given set of tuples. Decision by voting such as bagging and boosting attempts to enhance the existing classification techniques like decision trees by using a majority decision among them. However, a high accuracy ratio of prediction sometimes requires complicated predictors, and makes it hard to understand the simple laws affecting the values of the attribute of interest. We instead consider another approach of using at most several fairly simple voters that can compete with complex prediction tools. We pursue this idea to handle numeric datasets and employ region splitting rules as relatively simple voters. The results of empirical tests show that the accuracy of decision by several voters is comparable to that of decision trees, and the computational cost is inexpensive.

1 Introduction

1.1 Motivating Example

Large Decision Tree. Fig. 1A shows a sample decision tree generated by the See5/C5.0 program [13] with the default parameters. The 900-tuple training dataset is randomly selected from the german-credit (24 numerical) dataset of 1000 tuples, which was obtained from the UCI Machine Learning Repository [9]. This tree contains 87 leaf nodes, and its prediction error ratio for the remaining 100 tuples of the dataset is 27.0%. However, as shown by the number of the leaf nodes and the nested structure of the tree, this decision tree is relatively difficult to interpret and thus not always suitable for representation of scientific knowledge.

Region Rules for Readability. To address the present classification problem, we used a weighted majority decision among region rules in which the component rules were relatively powerful and visually understandable. When we consider a pair of numeric attributes as a two-dimensional attribute, we can compute a line partition of the corresponding two-dimensional space so that it properly
(A) An example of a large decision tree generated by the See5/C5.0 program. The training dataset is 900 tuples of the german-credit dataset. This tree contains 87 leaf nodes. The prediction error ratio for the remaining 100 tuples of the dataset was 27.0%. $A_i$ denotes the $i$th attribute.

(B) Region rules for the german-credit dataset. Used training dataset (900 tuples) and test data (100 tuples) were the same as for the above decision tree. The prediction error ratio of weighted majority decision among these five rules was 27.0%. $A_i$ denotes the $i$th attribute.

Fig. 1. Decision tree and region rules.