Abstract

Decision Tree which is one way in this classification is useful to obtain an appropriate rule set from a large number of instances. However, he had difficulty in obtaining relationship between the data points that have continuous values. Many of the Decision Tree algorithm can be developed to handle continuous attributes by applying the concept of pre-discretization. Attribute discretization become one important component of the data preparation stage for the induction of Decision Tree. Attribute discretization, a domain partition attribute to set the interval, involves two main steps: (1) determining the number of intervals of the attributes that should be didiskretisasi; and (2) determining the limits of each interval.

Some Decision Tree algorithm that applies the concept of pre-discretization is the C4.5 and the "state-of-art" algorithm NBTree because of its classification accuracy was very good even in a large data size. However, if further explored from the viewpoint of information theory, the loss of information due to the application of pre-discretization can reduce the classification accuracy.

In this final task,was made a software that implements the classification method of Self-adaptive NBTree, which induces a hybrid Decision Tree and Naive Bayes. Bayes measure, which is used to build a Decision Tree in a post-discretization, can directly handle continuous attributes and automatically find the number of interval-setabatas the most appropriate limits for diskretisasi. Naive Bayes node on the leaf which is helping to solve the problem of overgeneralization and overspecialization often seen in the Decision Tree.

Performance of Self-adaptive NBTree will compare to NBTree, and the two frames of it, the C4.5 (Decision Tree) and Naive Bayes. Test results showed self-adaptive NBTree performance have a better accuracy and smaller tree sizes than NBTree, C4.5, and Naive Bayes when dealing with a lot of continuous values. While NBTree and C4.5 works well on datasets with certain characteristics. Naive Bayes has the fastest model development and the classification time, but its accuracy is less kind to large size datasets.

Keyword : Decision tree, Naive Bayes,Discretization, Nbtree, Self-adaptive NBTree.