A CLASSIFICATION METHOD
BASED ON DATA GRAVITATION AND POSTERIOR PROBABILITY

Name : Muhamad Arief Hidayat
Registration Number : 5108.201.001
Advisor : Prof. Ir. Arif Djunaidy, M.Sc., Ph.D.

Abstract

The classification method based on data gravitation (DGC) is one of the new classification techniques that uses data gravitation as the criteria of the classification. In the case of DGC, an object is classified on the basis of the class that creates the largest gravitation in that object. However, the DGC method may cause inaccurate result when the training data being used suffer from the class imbalanced problem. This may be caused by the existence of the training data containing a class having excessively big mass that will in turn tend to classify an unknown object as a member of that class due to the high degree of the data gravitation produced, and vice versa.

In this research, a modification to the DGC method is performed by constructing a classification method that is based on both the data gravitation and posterior probability (DGCPP). In DGCPP, the mass concept defined in the DGC method as the prior probability is replaced by the posterior probability. By using this modification, data gravitation calculation process is expected to produce more accurate results compared to those produced by the DGC method. In addition, by improving the data gravitation calculation, it is expected that the DGCPP method will produce more accurate classification results compared to those produced by the DGC method for both normal datasets as well as datasets having class imbalanced problems.

A thorough test for evaluating the classification accuracy are performed using a ten-fold cross-validation method on several datasets containing both normal and imbalanced-class datasets. The results showed that DGCPP method produced positive average of accuracy differences compared to those produced by the DGC method. For the tests using the entire normal datasets showed that the average of accuracy differences are statistically significant with a 95% confidence level. In addition, results of the tests using the four imbalanced-class datasets also showed that the average accuracy differences are statistically significant with a 95% confidence level. Finally, results of the tests for evaluating the computing times required by the classification program showed that the additional computing time needed by DGCPP method to perform the classification process is insignificant and less than the human response time, in compared to that needed by DGC method for running all datasets being used.

Keywords: data gravitation-based classification, class imbalanced problem, posterior probability.