THE PERFORMANCE OF LEARNING ALGORITHMS ON REDUCED DATA SETS

Kittisak Kerdprasop and Nittaya Kerdprasop

School of Computer Engineering Suranaree University of Technology
Nakorn Ratchasima 30000, THAILAND

Abstract

Knowledge discovery is the process of extracting useful and previously unknown information from the very large data set. But extracting knowledge from a large data set is computationally inefficient. Using a sample from the original data can speed up the data mining process, but this is only acceptable if it does not reduce the quality of the induced information. We thus investigate the behavior of learning algorithms on different sampling sizes to decide which sample is sufficiently similar to the original data. We observe the accuracy of the induced rules extracted from training samples of decreasing sizes and use these results to determine when a sample is sufficiently small, yet maintain the acceptable accuracy rate. We evaluate random and stratified sampling methods on data from the UCI repository with three learning algorithms.

Key Words: Data Mining, Data Reduction, Sampling, Accuracy

Proceedings of International Conference on Artificial and Computational Intelligence (ACI 2002), September 26-28, Tokyo, Japan