THE EFFECT OF SAMPLING TECHNIQUES TO ACCURACY ESTIMATION

Kittisak Kerdprasop, Nittaya Kerdprasop, Pongden Punpakdeewong, and Petchpirin Doungsuwan
School of Computer Engineering Suranaree University of Technology 111 Muang District Nakorn Ratchasima 30000

Abstract

Knowledge discovery is the process of extracting useful and previously unknown information from the very large data set. Among many discovering methods, decision rules extracting is one of the most extensively studied techniques. But extracting rules from a large database is computationally inefficient. Using a sample from the database can speed up the data mining process, but this is only acceptable if it does not reduce the quality of the induced rules. We thus investigate the criteria to decide whether a sample is sufficiently similar to the original database. 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 systematic sampling methods on data from the UCI repository.