

Jordan University of Science and Technology

A New Sampling Technique for Association Rule Mining

Authors: Basel Mahafzah, A. Al-Badarneh, and Mohammed Zakaria

Abstract: Association Rule Mining (ARM) is one of the data mining techniques used to extract hidden knowledge from datasets, that can be used by organizations' decision makers to improve overall profit. However, performing ARM requires repeated passes over the entire database. Obviously, for large database, the role of I/O overhead in scanning the database is very significant. A popular solution to improve the speed of ARM is to apply the mining algorithm on a sample instead of the entire database. In this paper, a parameterized sampling algorithm for ARM is presented. This algorithm extracts sample datasets based on three parameters: transaction frequency, transaction length, and transaction frequency-length. To evaluate its performance and accuracy, a comparison against Two-Phase sampling algorithm is performed using real and synthetic datasets. The experimental results show that the proposed sampling algorithm in some cases outperforms Two-Phase sampling algorithm, and achieves up to 98% accuracy.