Speeding up Incomplete Data Analysis using Matrix-Represented Approximations

Thin Thin Soe
thinthinsoe.cumdy@gmail.com
University of Computer Studies, Mandalay, UCSM

Myat Myat Min
myatiimin@gmail.com
University of Computer Studies, Mandalay, UCSM

The veracity related with data quality such as incomplete, imprecise and inconsistent data creates a major challenge to data mining and data analysis. Rough set theory provides a special tool for handling the imprecise and incomplete data in information systems. However, the existing rough set based incomplete data analysis methods may not be able to handle large amount of data within the acceptable time. This paper focuses on speeding up the incomplete data analysis. The computation of the lower and upper approximations is a vital step for improving the performance of rough set based data analysis process. In this paper, the lower and upper approximations are characterized as matrix-represented approximations. The resulting approximations are exploited as inputs for data analysis method LERS (Learning from Examples based on Rough Set) used with LEM2 (Learning from Examples Module, Version2) rule induction algorithm. Then, this paper provides a set of experiments on missing datasets with different missing percent. The experimental results on incomplete or missing datasets from UCI Machine Learning Repository show that the proposed system effectively reduces the computational time in comparison with the existing system.