Analysis of Missing Data Using Matrix-Characterized 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

Nowadays, the veracity related to data quality such as incomplete, inconsistent, vague or noisy data creates a major challenge to data mining and data analysis. Rough set theory presents a special tool for handling the incomplete and imprecise data in information systems. In this paper, rough set based matrix represented approximations are presented to compute lower and upper approximations. The induced approximations are conducted 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. Analyzes are performed on missing datasets with "do not care" conditions and missing datasets with lost values. In addition, experiments on missing datasets with different missing percent by using different thresholds are also provided. The experimental results show that the system outperforms when missing data are characterized as "do not care" conditions than represented as lost values.