Abstract

Good data preparation is a key prerequisite to successful data mining. Conventional wisdom suggests that data preparation takes about 60 to 80% of the time involved in a data mining exercise. There have been good reviews of the problems associated with data preparation. However the data preprocessing is a crucial step used for variety of data warehousing and mining. Real world data is noisy and can often suffer from corruptions or incomplete values that may impact the models created from the data. Accuracy of any mining algorithm greatly depends on the input datasets. In this paper we describe a novel idea of predicting the missing values in the dataset by a well known principle of Maximum likelihood EM (Expectation
An Efficient Approach for Filling Incomplete Data

Maximization). After doing implementing and applying the EM filter, the dataset is completed with the estimated values, based on the well known principle of expected maximization of attribute instance. We demonstrate the efficacy of the approach on real data sets as a preprocessing step.

References

- Sameer S. Prabhune, Dr. S. R. Sathe "Reconstruction of a Complete Dataset from an Incomplete Dataset by Expectation Maximization Technique", International Journal of Computer Science and Network Security, VOL. 10 No. 11, November 2010
- UCI Machine Learning Repository,
- Jiawei Han and Micheline Kamber "Data Mining Concepts and techniques;"
- M. Richardson and P. Domingos. Mining Knowledge –sharing sites for viral marketing.
- Data Mining Leading Edge: Insurance & Banking, D Romano in Proceedings of Knowledge Discovery and Data Mining, Unicom, Brunel University, 1997.

Index Terms

Computer Science
Emerging Trends in Technology

Keywords

Data Mining  Data Preprocessing  Missing Data