Abstract

Data cleaning is one of the important step of KDD (Knowledge discovery in database) process. One critical problem in data cleaning is the presence of missing values. Various approaches have proposed to find & replace such missing data including use of mean value, use of global constant, replace by more probable value etc. Imputation is one of the important procedures in statistics that is used to replace the missing values in a data set. One advantage of this approach is that the missing data treatment is independent of the learning algorithms that are used. This allows the user to select the most suitable and appropriate imputation method for each situation. This paper analyze the six different imputation methods proposed in the field of statistics and implement them in Data mining environment. An artificial data set of 1000 records is used to analyze the performance of these methods. For testing the significance of these methods Z-test approach were used. Exhaustive experiments show the effectiveness of the proposed methods. It is assumed that all the attributes of input data are of numeric data type.

References

Comparative Analysis of Different Imputation Methods to Treat Missing Values in Data Mining Environment


Index Terms

Computer Science Data Mining

Keywords

KDD Data mining Imputation methods Data pre-processing sampling attribute
missing values.