One of the important stages of data mining is preprocessing, where the data is prepared for different mining tasks. Often, the real-world data tends to be incomplete, noisy, and inconsistent. It is very common that the data are not obtainable for every observation of every variable. So the presence of missing variables is obvious in the data set. A most important task
when preprocessing the data is, to fill in missing values, smooth out noise and correct inconsistencies.

This paper presents the missing value problem in data mining and evaluates some of the methods generally used for missing value imputation. In this work, three simple missing value imputation methods are implemented namely (1) Constant substitution, (2) Mean attribute value substitution and (3) Random attribute value substitution.

The performance of the three missing value imputation algorithms were measured with respect to different rate or different percentage of missing values in the data set by using some known clustering methods. To evaluate the performance, the standard WDBC data set has been used.

Reference

- Thomas Lumley, "Missing data", A Lecture Note, BOST 570, 2005-11-9

**Index Terms**

Computer Science  
Knowledge Discovery

**Key words**

Datamining  
Preprocessing  
Imputation methods  
Missing data