Abstract

Statistical Imputation Techniques have been proposed mainly with the aim of predicting the missing values in the incomplete sets as an essential step in any data analysis framework. K-means-based Imputation, as a representative statistical imputation method, has been producing satisfied results in terms of effectiveness and efficiency in handling popular and freely available data set (e. g., Bupa, Breast Cancer, Pima, etc.). The main idea of K-means based methods is to impute the missing value relying on the prototypes of the representative class and the similarity of the data. However, such kinds of methods share the same limitations of the K-means as data mining technique. In this paper and motivated by such drawbacks, we introduce simple and efficient imputation methods based on K-means to deal with the missing data from various classes of data sets. Our proposed methods give higher accuracy than the one given by the standard K-means.
Scalable Algorithms for Missing Value Imputation


Index Terms

Computer Science
Data Mining
Keywords

Statistical Imputation  Clustering  K-mean