Effect of missing value methods on Bayesian network classification of hepatitis data

ABSTRACT

Missing value imputation methods are widely used in solving missing value problems during statistical analysis. For classification tasks, these imputation methods can affect the accuracy of the Bayesian network classifiers. This paper study's the effect of missing value treatment on the prediction accuracy of four Bayesian network classifiers used to predict death in acute chronic Hepatitis patients. Missing data was imputed using nine methods which include, replacing with most common attribute,support vector machine imputation (SVMI), K-nearest neighbor (KNNI), Fuzzy K-means Clustering (FKMI), K-means Clustering Imputation (KMI), Weighted imputation with K-Nearest Neighbor (WKNNI), regularized expectation maximization (EM), singular value decomposition (SVDI), and local least squares imputation (LLSI). The classification accuracy of the naive Bayes (NB), tree augmented naive Bayes (TAN), boosted augmented naive Bayes (BAN) and general Bayes network classifiers (GBN)were recorded. The SVMI and LLSI methods improved the classification accuracy of the classifiers. The method of ignoring missing values was better than seven of the imputation methods. Among the classifiers, the TAN achieved the best average classification accuracy of 86.3% followed by BAN with 85.1%.

Keyword: Bayesian network classifiers; Missing data; Imputation; Hepatitis dataset; Classification and data mining.