A comparative study of missing value estimation methods: Which method performs better?

Abstract

Missing data is a problem that permeates much of the research bring done today. Some data frequently contain missing values such as gene expression data, which most of its down stream analyses for microarray experiments require complete data. In the literature many methods have been proposed to estimate missing values via information of the correlation patterns within the data matrix. In this report we describe an evaluation of top three current methods including a neural network method and two imputation methods on multiple types of data including microarray data, time series data such as air pollutant data and phytoplankton data. Based on the overall performance of the method, we then determine the most appropriate method that can be applied to various data sets. We found that the optimal method (Local Least Square Imputation (LLS) and Bayesian Principle Component Analyses (BPCA)) are all highly competitive to each other in overall results. We tested with Radial Basis Function (RBF) network method which is one of the neural network methods and found that, the overall performance of RBF network is lower than BPCA method and LLS method. According to the overall NRMSE of the three methods, the BPCA method provides the most accurate estimation for missing values.