Abstract

A DNA microarray commonly known as DNA chip or biochip is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously. Gene expression using DNA microarrays provides high throughput investigation of gene expressions by simultaneously measuring the expression of thousands of genes under a certain experimental condition.

Gene expression microarray experiments sometimes generate data sets with multiple missing expression values. Effective missing value estimation methods are needed since many algorithms for gene expression data analysis requires a complete matrix of gene array values as inputs. Most of the gene expression data analysis algorithms, such as clustering, classification and network design, require complete information, i.e. without any missing values. It is therefore very important to accurately impute the missing values before applying the data analysis algorithms. Therefore methods for imputing missing data are needed to minimize the effect of incomplete data sets on analysis, and to increase the range of data sets to which these algorithms can be applied.

Here we have discussed the most famous algorithms for missing value estimations and have implemented them using C++ programming language. We have analysed them with different datasets and with different percentage of missing values and later on we have plotted graphs and presented them here. The different algorithms implemented are row average, k-nearest neighbor and the Single value distribution. The plotted graph shows that K-nearest neighbor algorithm remains reliable with increase percentage of missing values. Among the implemented algorithms average method gives the highest NRMSE and K-nearest neighbor algorithms give the least NRMSE error.

Keywords:pseudo-inverse, K-nearest neighbor, Singular Value Decomposition, Multiple linear regressions.