Abstract

In the past few years, there has been a detonation of data in the field of biotechnology. Gene expression microarray experiments produce datasets with numerous missing expression values due to various reasons, e.g. insufficient resolution, image corruption, dust or scratches on the slides, or experimental error during the laboratory process. To improve these missing values, many algorithms for gene expression analysis oblige a complete matrix of gene array values as input, such as K nearest neighbor impute method, Bayesian principal components analysis impute method, etc. Accurate estimation of missing values is an important requirement for efficient data analysis. Main problem of existing methods for microarray data is that there is no external information but the estimation is based exclusively on the expression data. We conjectured that utilizing a priori information on functional similarities available from public databases facilitates the missing value estimation. Robust missing value estimation methods are required since many algorithms for gene expression analysis entail a complete matrix of gene array values. Either genes with missing values can be removed, or the missing
values can be replaced using prediction. Current methods for estimating the missing values include sample mean and K-nearest neighbors (KNN). Whether the accuracy of estimation methods depends on the actual gene expression has not been thoroughly investigated. Under this setting, we examine how the accuracy depends on the actual expression level and propose new method that provides improvements in accuracy relative to the current methods in certain ranges of gene expression.

Reference


Index Terms

Computer Science

Bioinformatics
Key words

Clustering

DNA Microarray
Fuzzy Logic