Microarray Gene Expression Data Dimensionality Reduction Using PCA

Authors : Fuad M. Alkoot

Abstract : Different experimental technologies such as microarray sequencing have been proposed to generate high-resolution genetic data, in order to understand the complex dynamic interactions between complex diseases and the biological system components of genes and gene products. However, the generated samples have a very large dimension reaching thousands. Therefore, hindering all attempts to design a classifier system that can identify diseases based on such data. Additionally, the high overlap in the class distributions makes the task more difficult. The data we experiment with is generated for the identification of autism. It includes 142 samples, which is small compared to the large dimension of the data. The classifier systems trained on this data yield very low classification rates that are almost equivalent to a guess. We aim at reducing the data dimension and improve it for classification. Here, we experiment with applying a multistage PCA on the genetic data to reduce its dimensionality. Results show a significant improvement in the classification rates which increases the possibility of building an automated system for autism detection.