Abstract

Modeling Gene Expression with Artificial Neural Networks

While the prime focus of this work was on the application of LVQ for modeling the gene Expression data which gave highly consistent results, different variants of SOM and LVQ algorithms were applied to datasets related to breast cancer, mouse (Musmusculus), Arabidopsis thaliana, Homo sapiens, sugarcane, etc. the LVQ1 algorithm provided the best results compared to other variants of LVQ and its unsupervised counterpart SOM.

The application of LVQ was followed by enhancement or fine tuning the map generated by SOM using LVQ. It was established that when SOM was used as a pattern classifier, the accuracies produced by various experiments of SOM improved considerably after application of LVQ.

Extraction of differentially expressed genes from the given dataset was also discussed in detail, as a case study the dataset of breast cancer was used for the said extraction. This work also brings forth some visualization techniques under one platform that are easy to learn, and can help the researchers can easily tracking out desired clusters / classes of genes. The popular visualization of Eisen et al (1998) named Tree View was reviewed and modified to accommodate additional features of generating clusters with more than two genes under the same parent node, using the gene expression data matrix as input instead of the distance matrix, etc.

The above datasets were also subjected to other algorithms such as k-means, HC and PCA. Combination of PCA with SOM and LVQ was also compared.

The enormous growth of bio-molecular databases makes it increasingly important to have fastest methods to process, analyze and understand such massive amounts of data. In the domain of genomics, the microarray gene expression data processing is the new area of interest for many researchers. Though lot of progress has been made in the microarray manufacturing process and to some extent the data mining from these microarrays, there are very few software suits available, either commercially or publicly, to make available different tools under one environment.

In pursuit to attain the above objective, well established works implemented through the GEDA suit of University of Pennsylvania, GEPAS suite of CNIO, Spain, Cluster 3.0 software by Hoon et al.(2004) and expression Profiler: Next Generation by European Bioinformatics institute (EBI), UK were extensively studied and improved. Software called gene expression data analysis suite (GEDAS) was designed and developed. The motivation behind development of this software was primarily based on the following ideas.

- Application of the popular ANN model for classification called the learning vector quantization (LVQ) through the three form / variants viz., LVQ1, LVQ2 and LVQ3.
- Bringing a number of data mining algorithms under one umbrella software environment. The algorithms include SOM, LVQ, k-means, hierarchical clustering, SVM and PCA.
- Support of a number of visualization techniques and gene expression data preprocessing algorithms has been provided in this software. It also contains a host of 19 distance measures.

This work also proposes obtaining ideal number of clusters for any given dataset. Debate of this aspect was on since decades. It was attempted to narrow down the search scope by suitably defining the upper and lower bounds.