# Abstract


## Khaled Mahar


## Gene Function Prediction from Expression Data Using Machine Learning Techniques


Gene function prediction is one of the primary goals of bioinformatics. Using machine learning techniques for this task faces many difficulties including high dimensionality nature of the input data and redundancy of the target annotation data. Machine learning techniques such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), Naïve Bayesian (NB) and Decision Trees (DT) have been extensively applied to the field of gene expression analysis. This paper focuses on analyzing high throughput gene expression data coming from microarray and gene ontology domains. In this paper a survey of existing machine learning techniques used in this area is presented. A new approach for gene function prediction is then proposed to overcome the main difficulties of this problem. It is based on quantitative analysis of gene expression data. The proposed approach is based on dimensionality reduction and a super classifier followed by subcategory classifiers. The proposed approach applies Independent Component. Analysis (ICA) proceeded by Principal Component Analysis (PCA) to reduce the dimension of the data before feeding it to an ANN classifier. This is important to speed up the learning and classification processes and improve the accuracy. To assess the performance of the proposed technique, a comparative study is conducted by using the most used machine learning techniques in bioinformatics, which are ANN, SVM, NB, and DT. Also, another comparative study regarding the dimension reduction methods and their effect on classification accuracies and performances is reported. The proposed approach is applied to the data set of a mouse gene expression microarray source, where the experimental results show the superiority of ANN over other techniques. It has also been shown that performing ICA after PCA gives the best results in both performance and accuracy.