Research Progress Report

Advanced Computational Intelligence and Deep Machine Learning for Early Detection and Diagnosis of Diseases

Overview

Machine Learning is a powerful tool for extracting meaningful trends and patterns from large amounts of data. It has received wide interest in biomedical areas due to the high potential of being used in the detection and diagnosis of diseases. The increasing complexity of molecular biology data has made it difficult if not impossible to use traditional patient diagnosis and screening techniques which would allow medical specialists to detect possible early signs and symptoms of diseases.

Traditional techniques for disease diagnosis have mainly relied on the understanding of biological processes and the experience of medical experts in the analysis of medical patient data. The diagnosis would result in decisions for various therapeutic interventions in the form of medications and the quality of the diagnosis is evaluated by the analysis of the resulting patient clinical trials.

The advances in DNA microarray technology has provided the capabilities to measure the expression levels of thousands of genes during various biological processes, collected from different experimental samples and conditions. The collective data samples are commonly referred to as gene expression data, where the underlying samples could represent different environmental conditions or taken at different time intervals. Different samples could represent different organs collected from either healthy or infected tissues and could also be from different patients.

Analysis of biomedical gene expression data of this type for disease diagnosis is extremely challenging given the complexity and high dimensionality of the data. Machine Learning offers a potentially powerful range of solutions for the analysis and research of biomedical data. In particular there has been active research for the use of classification and cluster analysis algorithms in biomedical research especially in the areas of disease diagnosis, genomic sequence analysis and biomedical image and MRI analysis.

Among the great challenges in dealing with data is to be able classify or group data sets into a set of categories or clusters to better comprehend the nature of the data.

For unsupervised classification techniques, no training data is made available which is why it is commonly referred to as clustering or exploratory data analysis. The objective of a clustering technique is to distinguish or separate a finite unlabeled data set into a finite and discrete set of data structures.
which were previously hidden. The output of clustering algorithms is to partition the input data set into a certain number of groups or subsets referred to as clusters.

The objective of the research is the Early detection of Diseases by leveraging the latest Deep Machine Learning techniques for exploration of biomedical gene expression data.

OVERVIEW OF BIOLOGICAL GENE EXPRESSION DATA

A. DNA Microarray technology

DNA microarray technology provides the capabilities to measure the expression levels of thousands of genes during various biological processes, collected from different experimental samples and conditions. A DNA microarray is a small chip typically manufactured using chemically coated glass or silicon. Thousands of probes representing DNA molecules are attached on the chip in a grid layout where each grid cell relates to a specific DNA sequence. Two samples representing a test sample and a control sample are typically reverse transcribed and labeled using fluorescent dyes and then hybridized with the probes on the surface of the chip.

In a microarray experiment, DNA sequences under multiple conditions are captured for analysis and are represented as gene expression data. The variations in conditions of the data samples could represent different time intervals in a specific biological process or they could represent different samples from different organs or tissues. For example the gene expressions could represent the DNA sequence progression of infected cancer cells at different stages, or they could represent samples from different tissues of healthy and infected patients.

B. Gene Expression Data Representation

The Gene expression data that is produced from microarray experiments can be represented in a matrix representation where each row represents the expression patterns of genes and each column represents the expression profiles of samples such that the value in cell wij represents the expression level measured for a typical gene i in a specific sample j. Among the challenges faced in such a data representation is that it is highly likely that only a small subset of the genes could be influencing the disease infection being monitored and also it is possible that interesting features of the disease are only present in a subset of the .

C. Gene Expression Data Quality

It is very common that the data in the gene expression matrix resulting from the microarray process contains many data anomalies such as noise and missing values which are always expected in any biological experimental procedure. Accordingly, preprocessing the gene data is a crucial step before attempting any data analysis procedures for disease diagnosis to ensure the quality and accuracy of the results. Some of the preprocessing tasks involved are data normalizing, estimating missing values and also filtering gene expression data which are not relevant or significant to the biological process being analyzed.
RESEARCH SURVEY FOR GENE EXPRESSION DATA CLUSTERING

UPGMA Clustering

This algorithm called Unweighted Pair Group Method with Arithmetic Mean (UPGMA) has produced a technique for graphically representing the clustered data set. Each gene row of the gene data matrix is reordered based on the hierarchical dendrogram and a predefined ordering rule in addition to that, each data cell in the gene matrix is colored based on the fluorescence ratio measurement. The clustering process transforms the gene data matrix into a colored cluster image table where any large colored regions will indicate genes that share similar expression patterns such as common disease characteristics.

Sample Based Clustering Algorithms

Sample-based clustering attempts to determine the phenotype structures present in the sample gene data. Research has shown that phenotypes of samples can be identified through only a small subset of genes whose expression levels correlate strongly with the class being identified. These genes are referred to as informative genes, where the rest of the genes in the data matrix are irrelevant to the division of samples and can be considered as noise in the data set.

Although the previous clustering methods such as K-means, hierarchical clustering and Self organizing maps (SOM) can be used to cluster samples using all the genes as features, the signal to noise ratio is usually smaller than 1:10, which may result in poor clustering results.

In order to solve this problem, it is common to combine these algorithms with supervised clustering approaches by including phenotype information integrated with the sample gene data which represent class labels for the data samples that can determine which class represents disease infected genes from normal health samples.

BiClustering Algorithms

Biclustering refers to a distinct class of clustering algorithms that perform simultaneous row and column clustering of data objects. Biclustering techniques are very important in the analysis of biological gene expression data since the matrix data representation of microarray experiments is constructed such that each row represents the expression patterns of genes and each column represents the expression profiles of samples. Traditional clustering algorithms are applied on either the rows or columns of the data matrix separately, where each gene in a cluster is defined using all the experimental sample conditions. Biclustering techniques perform clustering in both dimensions simultaneously by allowing both genes and sample conditions in a bicluster to be selected using only a subset of the gene data.

The objective of biclustering is to identify sub clusters of genes and samples by performing simultaneous clustering of both rows and columns of the gene matrix. The advantage of biclustering over traditional clustering algorithms is that it allows the identification of groups of genes that show similar activity
patterns under a specific subset of the experimental conditions. This will provide biological researchers with more in-depth analysis and diagnosis of disease patterns in correlation with the underlying experimental conditions they were captured from. For example different time intervals of progression stages of a disease or different conditions from different tissues representing both healthy and infected samples.

Deep Machine Learning

Neural Networks trained using traditional Back propagation have shown limitations in processing high dimensional data. Recent neuroscience findings have provided additional insight into the principles governing information representation in the brain. The discovery motivated the emergence of deep machine learning in 2006 which has since revolutionized the capabilities of processing high dimensional data.

Convolutional Neural Networks

Advanced Multi-layer network architectures initially designed for deep image and video classification. Minimal data preprocessing requirements and capable of processing huge amounts of high dimensional data. Leverages spatial relationships among data to reduce number of dimensions to be learned which dramatically improves learning compared to traditional feed-forward back propagation.

Research Challenges

- Deep Learning involves non-convex optimization models, which make it very difficult to prove built models.
- Designing the Architecture and optimal number of deep learning layers.
- Formulate learning algorithm and mathematical proof of convergence.
- Demonstrating successful model on the latest benchmark datasets.