A Generalized Machine Learning Framework for Fingerprinting Disease Based on Gene Expression Profile

Chakravarthy, Prathik (School: Beavercreek High School)
Chakravarthy, Sudarshan (School: Beavercreek High School)

Machine learning (ML) is a branch of artificial intelligence (AI) where systems learn from data, identify patterns, and make decisions with minimal human intervention. Classification algorithms use ML to analyze data samples and predict their class. Currently, large amount of data are available to the medical research community through microarrays - a high throughput technology that measures the expression of thousands of genes simultaneously on a genome-wide scale. Gene expression analysis reveals distinct patterns in the expression profile that can help differentiate disease from controls. Classification algorithms can thus leverage from the difference in expression profile of disease vs. controls and discriminate between them based on their genetic signatures, making it possible to accurately diagnose several diseases that are mainly based on behavioral models or inadequate pathological data, and also provide more information at the molecular level. Finally the classifier should extract a set of fingerprint genes for each disease to study the biological pathways affected. Currently, a classifier's performance is challenged by the high dimensionality of gene expression data. This research focuses on principal component analysis (PCA) as a dimensionality reduction and classification technique, through a multi-layer supervised learning approach. Gene expression data was downloaded from Gene Expression Omnibus (GEO), followed by statistical tests to identify differentially expressed genes (p-value < 10-5, FDR < 0.05), and implementation of the dual-layer PCA based classifier. The classifier was tested on data spanning several diseases such as autism, cancers, and neurodegenerative diseases, across multiple cell/tissue types, yielding over 90% sensitivity and specificity.