FUZZY INFERENCE SYSTEM FOR CLASSIFICATION OF CANCER CELLS FROM MICROARRAY GENE EXPRESSION DATASET

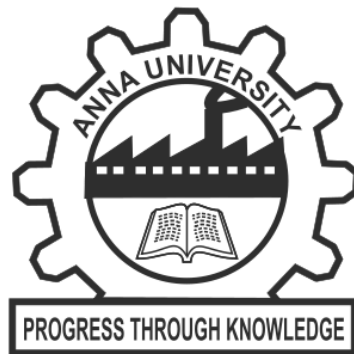A THESIS

Submitted by

TAMILSELVI MADESWARAN

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

FACULTY OF SCIENCE AND HUMANITIES

ANNA UNIVERSITY

CHENNAI 600 025

SEPTEMBER 2014

ABSTRACT

Data mining has emerged as a successful solution for the identification of data hidden in databases. It is the non-trivial extraction of implicit, formerly unknown and beneficial information from data in databases. Apart from the raw analysis process, data mining involves data pre-processing, complexity considerations, post-processing of identified structures etc. Microarray technology has evolved as a significant tool in the tracking of genome-wide expression levels of genes. The microarray technology is also utilized to automate the diagnostic task and to improve the accuracy of conventional diagnostic approaches.

The simultaneous analysis of gene expressions is facilitated by the microarray technology. By utilizing the gene expression data from microarrays, prediction, classification and clustering methods are used to analyze and interpret the data. A particular application of microarray data is the determination of molecular variation among cancers. The samples of gene expression data are classified into categories based on the type of tumor. The fundamental problems relating to the cancer diagnosis and drug discovery can be addressed by the microarray data. It is also used to determine previously undiscovered subtypes of cancer.

sequence and iv) Development of a fuzzy inference system that classifies the MAGED into AML and ALL.

In order to reduce the dimension of the outsized microarray gene expression data set, this research applied three statistical methods namely Mean, Median and Standard deviation. This phase extracts the data relating with significant genes representing AML and ALL from MAGED. The binary session phase finds out the threshold value based on which the equivalent binary sequence of MAGED is generated. The pattern generation phase locate all possible patterns in the form of a two dimensional matrix.

A fuzzy inference system (FIS) is proposed to classify the gene expression dataset. The fuzzy rules are derived from the generated pattern. These fuzzy rules are used to train the FIS classifier. The dimensionality reduction has a direct impact on the robustness of the generated fuzzy rules.

The performance of the proposed dimensionality reduction approach is compared with the existing dimensionality reduction techniques like principal component analysis (PCA) and multi-linear principal component analysis (MPCA). The work also provides comparison of classification performance of the proposed FIS approach with Naive Bayesian Classifier, k-nearest neighbor (KNN) and Support Vector Machine (SVM). The classification performance of FIS is simpler and easier to understand. The experimental results of the proposed FIS approach assist physicians in diagnosing the cancer disease.

sequence and iv) Development of a fuzzy inference system that classifies the MAGED into AML and ALL.

In order to reduce the dimension of the outsized microarray gene expression data set, this research applied three statistical methods namely Mean, Median and Standard deviation. This phase extracts the data relating with significant genes representing AML and ALL from MAGED. The binary session phase finds out the threshold value based on which the equivalent binary sequence of MAGED is generated. The pattern generation phase locate all possible patterns in the form of a two dimensional matrix.

A fuzzy inference system (FIS) is proposed to classify the gene expression dataset. The fuzzy rules are derived from the generated pattern. These fuzzy rules are used to train the FIS classifier. The dimensionality reduction has a direct impact on the robustness of the generated fuzzy rules.

The performance of the proposed dimensionality reduction approach is compared with the existing dimensionality reduction techniques like principal component analysis (PCA) and multi-linear principal component analysis (MPCA). The work also provides comparison of classification performance of the proposed FIS approach with Naive Bayesian Classifier, k-nearest neighbor (KNN) and Support Vector Machine (SVM). The classification performance of FIS is simpler and easier to understand. The experimental results of the proposed FIS approach assist physicians in diagnosing the cancer disease.