Chapter 36
Genetic Diagnosis of Cancer by Evolutionary Fuzzy-Rough based Neural-Network Ensemble

Sujata Dash

North Orissa University, India

Bichitrananda Patra

KMBB College of Engineering and Technology, India

ABSTRACT

High dimension and small sample size is an inherent problem of gene expression datasets which makes the analysis process more complex. The present study has developed a novel learning scheme that encapsulates a hybrid evolutionary fuzzy-rough feature selection model with an adaptive neural net ensemble. Fuzzy-rough method deals with uncertainty and impreciseness of real valued gene expression dataset and evolutionary search concept optimizes the subset selection process. The efficiency of the hybrid-FRGSNN model is evaluated by the proposed neural net ensemble learning algorithm. Again to prove the learning capability of ensemble algorithm, performance of the component classifiers pairing with FR, GSNN and FRGSNN are compared with proposed hybrid-FRGSNN based ensemble model. In addition to this, efficiency of neural net ensemble is compared with two classical and one advanced ensemble learning algorithms.

1. INTRODUCTION

Microarray technology produces high dimensional datasets by measuring the expression levels of tens of thousands of genes in a single experiment under varying conditions. It has become an indispensable tool for biological, medical and pharmaceutical researchers to get a better understanding of the diseases at genomic level. On the other hand, the inherent problem i.e., large number of features and small sample

size of microarray dataset makes the analysis process difficult for the problem. Typically, a relatively small number of features are found to be strongly correlated with the phenotypes in question. Therefore, to identify these discriminative features from gene expression dataset, a data mining tool (Witten, Frank & Hall, 2011) known as feature selection technique plays an important role.

The methods specifically used for feature selection can be categorized into two major groups namely, filter and wrapper methods (Saeys, Inza & Larranaga, 2007; Guyon, Nikravesh, Zadeh, 2006). Filter methods select features considering individual characteristic of each feature without taking into account the mutual dependencies among features. Then the features are sorted by their assigned ranks. The top ranked features are kept for further analysis by removing low ranked features. Actually, these selected features are used to develop the diagnostic model to efficiently predict the diseases. On the contrary, in wrapper methods a search algorithm is wrapped around a learning algorithm: so that an estimated learning accuracy for all subsets can be calculated to derive an optimal one. This method is computationally intensive in comparison to filter methods because to obtain an optimal subset of features all possible subsets need to be examined which is practically a difficult task. Thus, to alleviate this difficulty in wrapper methods an intelligent search strategy (Ghosh & Jain (Eds), 2005; Akadi, Amine, Ouardighi & Aboutajdine, 2009; Dash, 2016) may be adopted from a set of heuristic or stochastic algorithms which can be able to generate an optimal subset effectively.

Rough set theory (RST) was introduced by Pawlak (1982) and has been used widely by researchers as a classifier and selection technique. Primarily, it removes redundant data (Dash & Patra, 2012) while keeping intact the characteristics of the problem for classification. The main drawback of rough set (Dash & Patra, 2013) is it cannot handle real valued data whereas microarray dataset contains all real valued gene expression data. In this case, the expression values of genes are discredited before being used and the dataset loses some information during this process which makes the dataset inadequate for analysis purpose. The best alternative to get rid of this problem is either using real valued rough set or fuzzy-rough set theory (Zahi, 2011; Jensen & Shen, 2004; Dash, 2015). Both of them handle the concept of vagueness and indiscernibility which is prevalent in most real life problems (Xu, Miao & Wei, 2009) by complementing each other.

The ensemble learning approach constructs several classifier models for the original dataset and then combines the predictive outputs to identify an unknown sample. The motivation of combining several classifiers is to improve the classification efficiency which in turn depends on the accuracy and diversity (Yang P., Yang H., Bing, Zomaya, 2010) of the base classifiers. The ensemble technique is very popular in the field of classification and pattern recognition as it increases the generalization and percentage of classification by aggregating (Chen, Hong, Deng, Yang, Wei & Cui, 2015) the outcome of finite number of neural network classifiers (Lee, Hong & Kim, 2009a). However, neural network ensemble learning has been used in many problems, such as, face recognition (Lee, Hong & Kim, 2009 b), digital image processing (Liu, Cui, Jiang & Ma, 2004) and medical diagnosis (Huang, Zhou, Zhang & Chen, 2000) and has given outstanding performance in terms of classification accuracy.

In this study, a hybrid gene selection model (Dash & Patra, 2012) is introduced by hybridizing fuzzy-rough filter model (FR) with an evolutionary genetic search based wrapper model (GSNN) (Ghosh & Jain (Eds), 2005). The efficiency of the hybrid model (FRGSNN) (Yang, Chuang & Yang, 2009) is evaluated by the proposed adaptive neural net ensemble learning algorithm. Again to prove the learning capability of ensemble algorithm, performance of the component classifiers pairing with FR, GSNN and FRGSNN are compared with proposed hybrid FRGSNN based ensemble model. In addition to this, efficiency of neural net ensemble is compared with two ensemble learning algorithms.

Related Content

Data Collection and Analyses Applying Unmanned Helicopter (UAV) Remote Sensing to Survey Water Chestnut Invasive Species
Tao Tang, Chenliuli Jiangand Mary Perrelli (2020). *International Journal of Data Analytics (pp. 38-51).*
www.irma-international.org/article/data-collection-and-analyses-applying-unmanned-helicopter-uav-remote-sensing-to-survey-water-chestnut-invasive-species/244168

Data Collection and Analyses Applying Unmanned Helicopter (UAV) Remote Sensing to Survey Water Chestnut Invasive Species
Tao Tang, Chenliuli Jiangand Mary Perrelli (2020). *International Journal of Data Analytics (pp. 38-51).*
www.irma-international.org/article/data-collection-and-analyses-applying-unmanned-helicopter-uav-remote-sensing-to-survey-water-chestnut-invasive-species/244168

Characterization and Predictive Analysis of Volatile Financial Markets Using Detrended Fluctuation Analysis, Wavelet Decomposition, and Machine Learning
Manas K. Sanyal, Indranil Ghoshand R. K. Jana (2021). *International Journal of Data Analytics (pp. 1-31).*
www.irma-international.org/article/characterization-and-predictive-analysis-of-volatile-financial-markets-using-detrended-fluctuation-analysis-wavelet-decomposition-and-machine-learning/272107

A Comprehensive Review on Leveraging Business Intelligence for Enhanced Marketing Analytics
P. S. Venkateswaran, Divya Marupaka, Sachin Parate, Amit Bhanushali, Latha Thammareddiand P. Paramasivan (2024). *Data-Driven Decision Making for Long-Term Business Success (pp. 34-48).*
www.irma-international.org/chapter/a-comprehensive-review-on-leveraging-business-intelligence-for-enhanced-marketing-analytics/335562

A Multi-Objective Ensemble Method for Class Imbalance Learning: Application in Prediction of Life Expectancy Post Thoracic Surgery
Sajad Emamipour, Rasoul Saliand Zahra Yousefi (2017). *International Journal of Big Data and Analytics in Healthcare (pp. 16-34).*
www.irma-international.org/article/a-multi-objective-ensemble-method-for-class-imbalance-learning/197439