Chapter 2
Genetic Algorithm Based Pre-Processing Strategy for High Dimensional Micro-Array Gene Classification:
Application of Nature Inspired Intelligence

Deepak Singh
National Institute of Technology Raipur, India

Dilip Singh Sisodia
National Institute of Technology Raipur, India

Pradeep Singh
National Institute of Technology Raipur, India

ABSTRACT

Discretization is one of the popular pre-processing techniques that helps a learner overcome the difficulty in handling the wide range of continuous-valued attributes. The objective of this chapter is to explore the possibilities of performance improvement in large dimensional biomedical data with the alliance of machine learning and evolutionary algorithms to design effective healthcare systems. To accomplish the goal, the model targets the preprocessing phase and developed framework based on a Fisher Markov feature selection and evolutionary based binary discretization (EBD) for a microarray gene expression classification. Several experiments were conducted on publicly available microarray gene expression datasets, including colon tumors, and lung and prostate cancer. The performance is evaluated for accuracy and standard deviations, and is also compared with the other state-of-the-art techniques. The experimental results show that the EBD algorithm performs better when compared to other contemporary discretization techniques.

DOI: 10.4018/978-1-5225-5852-1.ch002

INTRODUCTION

Advancement in healthcare technology enabled a revolution in health research that could expedite endurance of the leaving being (Acharya & Dua, 2014). Early diagnosis with higher accuracy that provides the faster treatment conditions is feasible due to the enormous findings in technology. The approach to the study of the biological data had a significant contribution in discovering medical illness. However, the challenges associated with biomedical problem solving is handling and management of complex data sets. Here we consider one such example is DNA microarray gene dataset (Statnikov, Tsamardinos, Dosbayev, & Aliferis, 2005) where thousands of gene expressions measured for each biological sample using microarray and used for the diagnosis of cancer and its classification. The heterogeneity, ambiguity and inconsistencies persist with microarray data sets are biggest hurdle to tackle. Mostly these data are inconsistent because of noise, missing values, outliers, redundant values and data imbalance. Computational approaches can provide the means to resolve these challenges (Le, Paul, & Ong, 2010). Decision making, knowledge extraction, data management and data transmissions are the complex tasks that were effectively performed by the computational models. The recent trend in the computational methods evolves opportunities to disseminate the newly efficient techniques that can be helpful in designing prominent health care systems (Tsai, Chiang, Ksentini, & Chen, 2016). With the advent of nature inspired intelligence technique and the current paradigm of machine learning together can accelerate the current biomedical computational models.

The foundation of Machine learning is laid around the Knowledge discovery in database principle, consists of three basic steps namely the data preprocessing, learning phase, and validation phase (García, Luengo, & Herrera, 2015). The pre-processing phase has the objective of transforming data and discovering patterns by removing redundant features. Moreover, the pre-processing data phase is helpful in identifying the influential factors that contribute towards the classification. These techniques play a vital role in machine learning for improving the system performance. Various pre-processing (García et al., 2015) techniques are used for handling data inconsistencies which in turn help learner for efficient classification of the data. The performance of a learner is heavily relied on the class-attribute dependency of the training data. Dealing with a large number of instances or attributes in heterogeneous data could agitate the dependency (class-attribute) and hence requires preprocessing strategies (Liu, Motoda, Setiono, & Zhao, 2010; Molano, Cobos, Mendoza, & Herrera-viedma, 2014; Houari, Bounceur, Kechadi, Tari, & Euler, 2016) which is an essential step for eliminating lesser informative features and instances. Reduction of inconsistencies varies according to the preprocessing strategies considered. Feature selection (Diao & Shen, 2015) and Discretization (García et al., 2013) are the most popular measures for reduction of unnecessary information in data mining. Feature selection is the process of selecting the subset of the most relevant features from the set of features whereas discretization achieves the reduction by (Maimon, Oded, and Rokach, 2002) converting continuous values into discrete values with fixed interval span.

Feature selection has numerous methods which can be grouped into two categories: filters and wrappers (Liu et al., 2010). Filter method searches and evaluates either each gene individually (univariate filters) or the subset of genes (multivariate filters) by measuring their intrinsic properties related to class discrimination, independent of a learning method. Wrapper method encapsulates a global search method and the classifier in a single approach. The search method explores the gene space for all possible gene subsets, and the goodness of each subsets evaluated by a specific classifier for sample classification. In discretization, data partition is done through cut-point selection by applying different types of heuristics including Equal-interval-width (Chan, 1991), Equal-frequency-per-interval, minimal-class-entropy (Ar-