Chapter 38
Hybrid Ensemble Learning Methods for Classification of Microarray Data:
RotBagg Ensemble Based Classification

Sujata Dash
North Orissa University, India

ABSTRACT

Efficient classification and feature extraction techniques pave an effective way for diagnosing cancers from microarray datasets. It has been observed that the conventional classification techniques have major limitations in discriminating the genes accurately. However, such kind of problems can be addressed by an ensemble technique to a great extent. In this paper, a hybrid RotBagg ensemble framework has been proposed to address the problem specified above. This technique is an integration of Rotation Forest and Bagging ensemble which in turn preserves the basic characteristics of ensemble architecture i.e., diversity and accuracy. Three different feature selection techniques are employed to select subsets of genes to improve the effectiveness and generalization of the RotBagg ensemble. The efficiency is validated through five microarray datasets and also compared with the results of base learners. The experimental results show that the correlation based FRFR with PCA-based RotBagg ensemble form a highly efficient classification model.

INTRODUCTION

Cancer is caused due to the changes or mutation in the expression profiles of certain genes which elevates the importance of feature selection techniques to find relevant genes for classification of the disease. The most significant genes selected from the process are useful in clinical diagnosis for identifying disease profiles (Yang et al., 2006). The discriminative genes are selected through feature selection techniques that aim to select an optimal subset of genes. But, high dimension and small sample size characteristics of microarray dataset creates lot of computational challenges for selecting optimal subsets of genes

DOI: 10.4018/978-1-7998-1204-3.ch038

such as the problem of "curse of dimensionality" and over-fitting of training dataset. Feature selection is often used as a preprocessing step in machine learning. Only non-redundant and relevant features are sufficient enough to provide effective and efficient learning. However, selecting an optimal subset is very difficult (Kohavi & John, 1997) as the possible number of subsets grows exponentially when the dimension of the dataset increases.

The feature selection techniques can be broadly classified into filter (Hall, 2000; Liu, Motoda & Yu, 2002; Yu & Liu, 2003) and wrapper model (Hsu et al., 2011; Dash, Patra & Tripathy, 2012). The filter model uses specific evaluation criterion which is independent of learning algorithm to select feature subset from the dataset. It depends on various evaluation measures which are employed on the general characteristics of the training data such as information, distance, consistency and dependency. The wrapper method measures the goodness of the selected subsets using the predictive accuracy of the learning algorithm. But these methods require intensive computation for high dimensional dataset. Apart from this another key factor in feature selection is search strategy. The trade-off between optimal solution and computational efficiency is attained by adopting an appropriate search strategy such as random, exhaustive and heuristic search (Dash & Liu, 2003).

There are feature selection methods available for supervised (Yu & Liu, 2003; Dash & Liu, 1997) and unsupervised (Dash, Choi, Scheuermann & Liu., 2002) learning methods and it has been applied in several applications like genomic microarray data analysis, image retrieval, text categorization, intrusion detection etc. But, the theoretical and empirical analysis has demonstrated that the presence of irrelevant and redundant features (Kohavi & John, 1997; Hall, 2000) in the dataset reduces the speed and accuracy of the learning algorithms, thus need to be removed from the dataset. Most of the feature selection techniques employed so far has considered individual feature evaluation and feature subset evaluation (Guyon & Elisseeff, 2003; Abraham, 2004). Individual feature evaluation method ranks the features with respect to their capability of differentiating instances of different classes and eliminates the irrelevant and redundant features likely to have the same rankings. The feature subset evaluation method finds a subset of minimum features satisfying measure of goodness removes irrelevant and redundant features. It is observed that the advance search strategies like heuristic search and greedy search used for subset evaluation even after reducing the search space from $O(2^N)$ to $O(N^2)$ prove to be inefficient for high-dimensional dataset. This shortcoming encourages exploring different techniques for feature selection which will address both feature relevance and redundancy for high-dimensional microarray dataset.

There are various uncertainties associated with fabrication of microarray data such as the gathering of data, hybridization and image processing. They introduce lot of noises which need to be addressed by a robust and reliable classification model (Piao, Piao, Park & Ryu, 2012). On the other hand, conventional machine learning algorithms encounter many challenges to develop an effective and reliable classification model. The generalization capability of such kind of classifier algorithm based on a few significant genes and small number of training samples cannot be dependable. Therefore, it is essential to devise generalized robust classification methods which could overcome the constraints of small sample size and uncertainties associated with the datasets. This study motivates to develop ensemble classifier which is not very sensitive to the above specified constraints.

Ensemble learning technique is an advanced mechanism to combine multiple number of learning techniques to achieve better prediction accuracy (Liu et al., 2010; Dietterich, 2000; Yang, Yang, Zhou & Zomaya, 2010). It has the advantage of ignoring the constraint of sample size and the potential threat of overfitting by averaging and incorporating over multiple learning models (Hansen & Salamon, 1990). That is how the dataset is being used in an effective way, which was highly difficult for many bioin-