Chapter 20
Hybrid Wrapper/Filter Gene Selection Using an Ensemble of Classifiers and PSO Algorithm

Anouar Boucheham
Constantine 3 University, Algeria

Mohamed Batouche
Constantine 2 Abdelhamid Mehri University, Algeria

ABSTRACT

Bioinformatics has grown very quickly for the last 20 years, and it will grow even faster in the future. One of the long-standing open challenges in bioinformatics is biomarker identification and cancer diagnosis from gene expression. In this paper, the authors propose a novel hybrid wrapper/filter feature selection approach to identify the most informative genes for cancer diagnosis, named HWF-GS. It handles selection through two steps. The first one is an iterative filter-based mechanism to generate potential subsets of genes. The second step is the aggregation of the best-selected subsets by means of a wrapper-based consensus process that relies on a particle swarm optimization adapted to feature selection. An ensemble of classifiers (SVM and KNN) is employed to evaluate the selected genes. Experiments on nine publicly available cancer DNA microarray datasets have shown that HWF-GS selects robust signatures with high classification accuracy and competes with and even outperforms other methods in the literature.

1. INTRODUCTION

Several advanced genomic technologies developed last year's (DNA microarrays, NGS and RNAseq…), especially during the sequencing the human genome are being very helpful for molecular diagnostics, unveiling new insights into biology and have led to biomarker discovery (Mabert et al., 2014). Certainly, the use of molecular biomarkers will impact different areas of clinical practice and will give precious additional information for tumor diagnosis/prognosis and finally, contribute to personalized therapy of cancer. The ideal biomarker for cancer would have applications in (a) classification of tumors, (b)

DOI: 10.4018/978-1-5225-8903-7.ch020

prognosis of disease progression, (c) prediction of response to therapy, (d) monitoring of response to therapy and serve as a target for drug development (Stoss & Henkel, 2004).

Gene expression microarray is used to survey and measure genes activity in healthy and diseased tissues through various populations. It can measure and record the expression level of thousands of genes simultaneously in different samples types and specific experimental conditions (referred to as a sample) (Bolon-Canedo et al., 2014). In cancer examination these technologies have been broadly investigated for classification of different types of tumors and make the accurate prediction of cancer possible and easier using bioinformatics tools in machine learning and pattern recognition (Wu et al., 2012).

As a general observation, there are several problems studied in genes expression microarrays (GEM). All of them can be divided into three classes namely the class prediction which uses supervised machine learning approaches, the class discovery which uses unsupervised machine learning approaches (Banu & Andrews, 2015) and the class gene comparison that uses machine learning approaches in general (Golub et al., 1999). The direct application of these methods on high-dimensional data is usually ineffective (Wu et al., 2012). Since gene expression data consists of a high number of features (genes) and small sample sizes. However, there are a large number of irrelevant, redundant and noisy genes. Only a small set of genes contains useful biological interpretations and finally gives high accuracy for cancer diagnosis. In addition, the presence of many features affects not only the performance of prediction but also the computational time of learning algorithms (Bolon-Canedo et al., 2014).

To avoid the problem of the curse of dimensionality it becomes then necessary to select a small subset of features/genes that can separate healthy patients from cancer patients or in more general terms, genes which are relevant, non-redundant and discriminative for a particular genetic disease. These genes are called biomarkers, informative genes, parsimonious genes or differentially expressed genes.

Therefore, we require dimensionality reduction techniques, which identify a small set of genes that represent the most discriminant information of the original ensemble of genes to achieve better learning performance. This step plays a central role in the field of machine learning and more specifically in the classification task and allows many pros (Krishnapuram et al., 2004) (a) reduce the computational cost and storage space of the classification model, by constructing them using only a small subset of the original set of genes, (b) Improve significantly the intelligibility of the classifier, and maximize the prediction performance of a classification algorithm and (c) reduce the risk of ''overfitting'' when the number of samples is small. Subsequently, the prediction result of classifiers is more reliable, robust and can help doctors to take appropriate treatment solution which provide patients with better treatment or response to therapy, especially when the disease has been identified at its early time (Osl et al., 2012).

The identification of biomarkers and cancer classification are two closely related problems. From a machine learning viewpoint, biomarkers discovery is a feature selection problem and cancer diagnosis is a supervised classification problem. Features selection represents one of the recent growing areas of research in machine learning and it is defined as the process of identifying and removing unnecessary and redundant genes from the training data. Thus, it aims to choose a subset of available features which contribute most to the consistent classification of cancers, by eliminating features with little or no predictive information which contribute to the phenotype or symptom of disease (Bolon-Canedo et al., 2014). On the other hand, cancer classification or prediction refers to the procedure of constructing a model on the microarray dataset (using as inputs the resulting significant genes selected in the previous step), followed by affecting samples in suitable subtypes of disease by way of the constructed model (George & Raj, 2011).