Genomics Data 8 (2016) 4–15

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data

Rabia Aziz *, C.K. Verma, Namita Srivastava

Department of Mathematics & Computer Application, Maulana Azad National Institute of Technology, Bhopal, 462003, MP, India

ARTICLE INFO

Article history:
Received 29 October 2015
Received in revised form 8 January 2016
Accepted 19 February 2016
Available online 23 February 2016

Keywords:
Fuzzy backward feature elimination (FBFE)
Independent component analysis (ICA)
Support vector machine (SVM)
Naïve Bayes (NB)
Classification

ABSTRACT

Feature (gene) selection and classification of microarray data are the two most interesting machine learning challenges. In the present work two existing feature selection/extraction algorithms, namely independent component analysis (ICA) and fuzzy backward feature elimination (FBFE) are used which is a new combination of selection/extraction. The main objective of this paper is to select the independent components of the DNA microarray data using FBFE to improve the performance of support vector machine (SVM) and Naïve Bayes (NB) classifier, while making the computational expenses affordable. To show the validity of the proposed method, it is applied to reduce the number of genes for five DNA microarray datasets namely; colon cancer, acute leukemia, prostate cancer, lung cancer II, and high-grade glioma. Now these datasets are then classified using SVM and NB classifiers. Experimental results on these five microarray datasets demonstrate that gene selected by proposed approach, effectively improve the performance of SVM and NB classifiers in terms of classification accuracy. We compare our proposed method with principal component analysis (PCA) as a standard extraction algorithm and find that the proposed method can obtain better classification accuracy, using SVM and NB classifiers with a smaller number of selected genes than the PCA. The curve between the average error rate and number of genes with each dataset represents the selection of required number of genes for the highest accuracy with our proposed method for both the classifiers. ROC shows best subset of genes for both the classifier of different datasets with propose method.

1. Introduction

Gene expression analysis using microarrays has become an important part of biomedical and clinical research. Recent advancements in DNA microarray technology have enabled us to monitor and evaluate the expression levels of thousands of genes simultaneously, which allows a great deal of microarray data to be generated [1]. Microarray techniques have been successfully employed virtually in every aspect of biomedical research because they exhibit the possibility to do massive tests on genome patterns [2]. Microarray gene expression data usually has a large number of dimensions and is permitted to evaluate each gene in a single environment in different types of tissues like various cancerous tissues [3]. Accordingly, microarray data analysis, which can supply useful data for cancer prediction and diagnosis, has also attracted many researchers from diverse areas. Progressively, the challenge is to translate such data to get a clear insight into biological processes and the mechanisms of human disease [4]. To aid such discoveries, mathematical and computational tools are required that are versatile enough to capture the underlying biology and simple enough to be applied efficiently on large datasets. Therefore, novel statistical methods must be introduced to analyze those large amounts of data generated from microarray experiments [5]. The process of microarray classification consists of two successive steps. The first step is to select a set of significant and relevant genes and the second step is to develop a classification model, which can produce accurate prediction for unseen data. One of the key goals of microarray data analysis is to distinguish the various categories of cancers. A true and accurate classification is essential for successful diagnosis and treatment of cancer. The enormous dimensionality of the DNA microarray data becomes a problem, when it is employed for cancer classification, as the sample size of DNA-microarray is far less than the gene size [6]. However, among the large number of genes, only a small fraction is effective for performing a classification task, so the choice of relevant genes is an important task in most microarray data studies that will give higher accuracy for sample classification (for example, to distinguish cancerous from normal tissues). This trouble can be alleviated by using machine learning with a gene selection problem. The goal of gene selection methods is to determine a small subset of informative genes that reduces processing time and provides higher classification accuracy [7]. There are a large number of methods, which have been developed and applied to do gene selection. A typical gene selection method has two constituents,

* Corresponding author.

http://dx.doi.org/10.1016/j.gdata.2016.02.012

an evaluation criterion and a searching scheme. As many evaluation criteria and searching schemes already exist, it is possible to develop many gene selection methods by just combining different evaluation criteria and searching schemes. Since, many of these combinations of evaluation criteria and searching schemes actually perform similarly, it is sufficient to compare the most commonly used combinations instead of all possible combinations [8]. The commonly used gene selection & extraction approaches are t-test, Relief-F, information gain, SNR-test and principal component analysis (PCA), linear discriminant analysis, independent component analysis (ICA). These methods are capable of selecting a smaller subset of genes for sample classification [9]. Recently, independent component analysis (ICA) method has received growing attention as effective data-mining tools for microarray gene expression data. As a technique of higher-order statistical analysis, ICA is capable of extracting biologically relevant gene expression features of microarray data [10]. The success of the ICA method depends upon the appropriate choice of best gene subset from given ICA feature vector and choice of an appropriate classifier [11].

In this study, fuzzy backward feature elimination (FBFE) scheme was introduced, in which features were eliminated successively from ICA feature vector according to their influence on a SVM and NB based evaluation criterion. FBFE is a backward feature elimination method based on fuzzy entropy measure. Several machine learning techniques, such as artificial neural networks (ANN), k-nearest neighbor (KNN), support vector machine (SVM), Naïve Bayes (NB), decision tree, random forest and kernel-based classifiers, have been successfully applied to microarray data and also for other biological data analyses in recent years [4,12]. From the study of Liwei Fan et al. and Chun-Hou Zheng, it was seen that NB and SVM were the best classifiers with ICA for microarray data, and feature subset selection from the ICA feature vector can significantly improve the performance of classifiers [3,13].

Naïve Bayes (NB) classifier is a simple Bayesian network classifier, which is built upon the firm assumption that different attributes are independent of each other in the given course of instruction. There are two major challenges that may seriously affect the successful application of NB classifier to microarray data analysis. The first is the conditional independence assumption rooted in the classifier itself, which is hardly satisfied by the microarray data [14]. This limitation could be successfully resolved as the components extracted by the ICA are statistically independent therefore, gene extraction by ICA could effectively improve the performance of a NB classifier for microarray data. Second limitation is that, all the attributes have an influence on the classification; hence, the use of FBFE eliminates the inappropriate genes from ICA feature vector to improve the performance of a NB classifier during cross validation. It is therefore necessary to select genes to reduce the dimensionality of microarray data before applying a NB classifier [15].

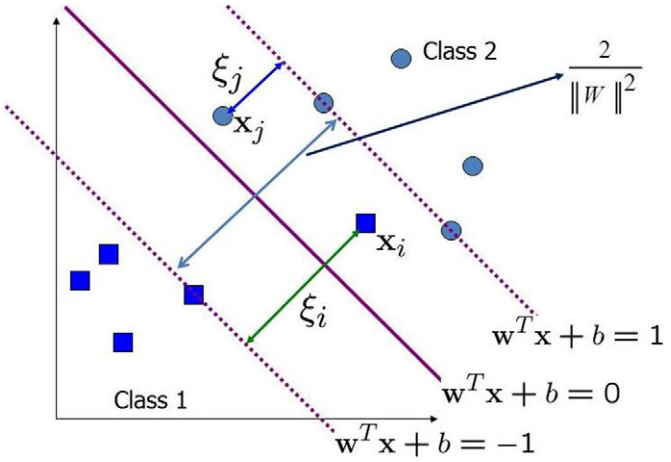

Fig. 2. Maximum margin hyperplanes for SVM divides the plane into two classes.

On the other hand the SVM-based classifier is superior, as it is less sensitive to the curse of dimensionality and more robust than other non-SVM classifiers [16]. The biggest drawback of an SVM is that it cannot directly obtain the genes of importance. Thus, during the fitting of an SVM model, a careful gene selection has to be done first and then the selected genes should be used to obtain improved classification results. If genes are not appropriately chosen, there may be a large number of redundant variables in the model, severely affecting its performance [17].

In this paper, a fuzzy backward feature elimination (FBFE) approach is used to eliminate the inappropriate genes from the independent components of the DNA microarray data for support vector machine (SVM) and Naïve Bayes (NB) classifiers. The proposed approach consists mainly of two steps. The original DNA microarray gene expression data are modeled by independent component analysis (ICA), and then the most discriminant features extracted by the ICA are selected by the fuzzy feature selection technique, which will be introduced and discussed in detail in Section 2. The next section explains the classification procedure of SVM and NB, followed by the details of used datasets and preprocessing step of datasets. In Section 4, the proposed method is compared and evaluated with PCA as a standard extraction method on several microarray datasets. The experimental results on five microarray datasets, show that the proposed approach can, not only improve the average classification accuracy rates, but also reduce the variance in classification performance of SVM and NB. Discussions and conclusions are presented in Section 5.

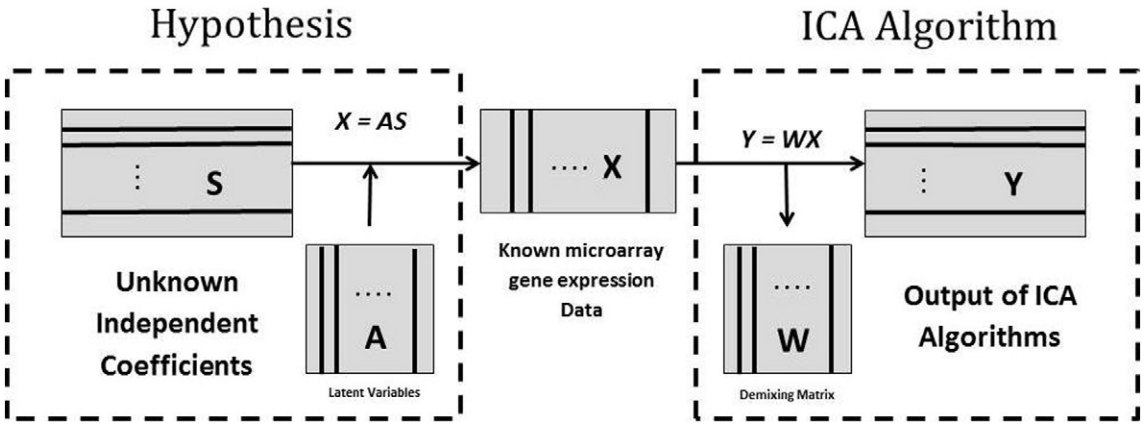

Fig. 1. Theoretical framework of ICA algorithms of microarray gene expression data.