Available online at www.sciencedirect.com

ScienceDirect

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

Original Research Article

A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification

Maryam Mollaee [a,*], *Mohammad Hossein Moattar* [b]

[a] *Young Researchers and Elite Club, Mashhad Branch, Islamic Azad University, Mashhad, Iran*
[b] *Department of Software Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran*

ARTICLE INFO

Article history:
Received 22 November 2015
Received in revised form
13 May 2016
Accepted 20 May 2016
Available online 1 June 2016

Keywords:
Discriminant independent component analysis
Feature selection
Microarray classification
Particle swarm optimization
Bayesian logistic regression

ABSTRACT

Microarray data play critical role in cancer classification. However, with respect to the samples scarcity compared to intrinsic high dimensionality, most approaches fail to classify small subset of genes. Feature selection techniques can reduce the dimension of the problem, which can reduce computational cost of the microarray data classification. However, previous studies have shown that feature extraction methods can also be useful in improving the performance of data classification. In this paper, we propose an ensemble schema for cancer diagnosis and classification that has three stages. At first, a hybrid filter-based feature selection method using modified Bayesian logistic regression (BLogReg), Ttest and Fisher ratio is applied for selecting genes. In the second stage, selected genes are mapped via the proposed PSO-dICA method which is a modification of dICA. Finally, mapped features are classified using SVM classifier. To demonstrate the effectiveness of the proposed method, some traditional microarray data including Colon, Lung cancer, DLBCL, SRBCT, Leukemia-ALL and Prostate Tumor datasets are used. Experimental results show the efficiency and effectiveness of the proposed method.

1. Introduction

Every cell in body contains a combination of genes which dictate its unique characteristics. These genes makeup can be expressed by DNA microarray technology. This technology has the capability to express tens of thousands of genes at the same time and can be used to distinguish cancerous tissues from normal ones.

Recent progresses in microarray techniques allow scientists to classify and diagnose special cancers based on DNA microarray data [1]. Reliable classification methods for diagnosis and treatment of cancer are necessary using microarray gene expression. Samples of microarray datasets

* *Corresponding author at*: Young Researchers and Elite Club, Mashhad Branch, Islamic Azad University, Mashhad, Iran.
E-mail addresses: mollaee@mshdiau.ac.ir (M. Mollaee), moattar@mshdiau.ac.ir (M.H. Moattar).

http://dx.doi.org/10.1016/j.bbe.2016.05.001

often is fewer than 100 each having more than thousands of genes. In addition, these microarrays contain redundancy, data mortality and noise due to biological factors. Since the number of samples is low and the number of genes is very high, classification of microarray data is recognized as a challenge.

To confront with this problem, feature selection is a necessary preprocessing step. Generally, feature selection methods are divided into three classes. Filtering methods search for important features with a classifier independent test such as information entropy and statistical dependence test [2]. The other methods, which are known as wrapper, use a special machine learning algorithm such as neural networks or support vector machine (SVM) for feature selection instead of using independent test [3–5]. Wrapper methods are more accurate than filter method but are slower. In the embedded feature selection methods, feature selection and classifier learning are performed at the same time. These methods use the learning algorithm to select features. Compared with wrapper methods, these methods have lower computational complexities.

Many methods have been proposed in the field of dimensionality reduction for medical diagnosis. Such as Guyon and Elisseeff which proposed naive Sequential Forward Feature Selection (SFFS) for features subset selection from high-dimensional features [6]. Gan et al. proposed the Filter-Dominating Hybrid Sequential Forward Feature Selection (FDHSFFS) algorithm for this purpose which has high computational complexity [7].

Hall [8] introduced a feature selection algorithm based on correlation (called Correlation Feature Selection (CFS)), which is a simple multi-feature filter algorithm that determines the rank of subsets of the features according to a heuristic correlation function. A modification on the approach was proposed by Yu and Liu [9] which was named Fast Correlation-Based Filter (FCBF).

Hall and Smith [10] used information gain which is one of the most common methods of assessment of features. Peng et al. [2] offered minimum Redundancy Maximum Relevance method (mRMR) that selects the features that have the highest correlation with the target class and the minimum redundancy; i.e. it selects the features that have the maximum difference with each other. Both optimization criteria (i.e. minimum redundancy and maximum relevance) are based on mutual information. Relief is proposed by Kira and Rendell [11] in which feature weights are estimated according to their ability to discriminate between nearby instances.

Inza et al. [12] used classic wrapper search algorithms (i.e. Sequential Forward Floating Selection (SFFS) feature selection and Best-First Search) on microarray data sets. Ruiz et al. offered an increasing wrapper method called BIRS for selecting genes. Successive Feature Selection (SFS) algorithm was proposed by Sharma et al. [13]. The proposed method had high classification accuracy on several DNA microarray data sets.

Recently, gene selection methods have been applied based on evolutionary processing methods such as Particle swarm optimization (PSO) for identification of important genes. PSO is a meta-heuristic optimization which has very simple concepts and only needs basic mathematical operators and it is suitable in terms of speed and cost. PSO also can be combined with a machine learning method for feature selection [3]. Chuang et al. [14] applied a combination of improved binary PSO with SVM classifier to select informative genes. Li applied combined PSO-GA for gene selection [4].

Wrapping model has high computational cost and becomes aggravated by high dimensions of the microarray data. Moderate solutions use embedded approaches using classifier core criteria by creating a criterion for ranking the attributes. A new feature selection method based on Recursive Feature Elimination (RFE) has been proposed by Guyon [5] which is based on SVM and is named SVM-RFE. Wang et al. [15] presented First-Order Inductive Learner (FOIL) feature selection algorithm for a subset of features. Canul-Reich et al. [16] presented a new algorithm based on Iterative Feature Perturbation (IFP) called FRFS.

Today, not only classic feature selection methods are used, but new hybrid algorithms or group methods are also used [17]. Mondra and Rejapeks [18] combined two of the most famous feature selection methods (i.e. SVM-RFE and mRMR) for microarray data. Lee and Leu [19] proposed a hybrid approach that used genetic algorithm with dynamic parameter setting to generate subsets of genes and with respect to the frequency of genes occurrence, ranked them in the subset of genes. Then used χ^2-test to select the top-ranked genes. Finally, SVM was used to evaluate the efficiency of selected genes. Also, Leung and Hung [20] presented Multiple Filter Multiple Wrapper (MFMW) method.

Recently, Zhao proposed a hybrid method which combines information gain (IG), F-score, Genetic Algorithm (GA), PSO, and SVM for gene selection and classification of microarray data [21]. Chen et al. and Kar et al. [22,23] used PSO + kNN for feature selection. Chen et al. [24] proposed a combination of PSO and C4.5 decision tree for cancer classification on gene expression data. Shen et al. [25] proposed a modified particle swarm optimization algorithm for joint gene and sample selection. Alshamlan et al. [26] proposed a Genetic Bee Colony (GBC) algorithm that combines GA with Artificial Bee Colony (ABC) algorithm to select the most informative genes for cancer classification. Wang et al. [27] proposed an approach based on an ensemble of probabilistic neural networks (PNN) and neighborhood rough set models for gene selection for tumor classification.

One of the important tools for analysis and interpretation of microarray data is unsupervised machine learning method of which the most important ones are Principal Component Analysis (PCA) and Independent Component Analysis (ICA). PCA has been widely applied in high dimension data analysis and feature extraction. In bioinformatics, its usual application is for analysis of high-dimension microarray data and is able to find useful genes [28]. ICA [29–31] is regarded as a development of PCA method and has more capability than PCA which has been successfully applied for analysis of microarray data. Lotfi and Keshavarz [32] proposed a hybrid method based on PCA and Brain Emotional Learning (BEL) network for gene-expression microarray data classification. Liu et al. [33] proposed a genetic algorithm based ensemble independent component selection (EICS) system that GA is applied to select a set of optimal IC subsets. Fan et al. [34] proposed a sequential feature extraction approach for classification of microarray data that consists of feature selection by stepwise regression and feature extraction using class-conditional independent component analysis component analysis. Li et al. [35] proposed an algorithm based on