Applied Soft Computing 38 (2016) 548–560

Contents lists available at ScienceDirect

Applied Soft Computing

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

Classification of DNA microarrays using artificial neural networks and ABC algorithm

Beatriz A. Garro [a,*], Katya Rodríguez [a], Roberto A. Vázquez [b]

[a] Instituto en Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Ciudad Universitaria, México, D.F., Mexico

[b] Intelligent Systems Group, Facultad de Ingeniería – Universidad La Salle, Benjamin Franklin 47, Col. Condesa, CP 06140 México, D.F., Mexico

ARTICLE INFO

Article history:
Received 27 April 2015
Received in revised form
13 September 2015
Accepted 3 October 2015
Available online 17 October 2015

Keywords:
DNA microarrays
Artificial neural networks
Pattern recognition
Cancer classification
Artificial Bee Colony algorithm

ABSTRACT

DNA microarray is an efficient new technology that allows to analyze, at the same time, the expression level of millions of genes. The gene expression level indicates the synthesis of different messenger ribonucleic acid (mRNA) molecule in a cell. Using this gene expression level, it is possible to diagnose diseases, identify tumors, select the best treatment to resist illness, detect mutations among other processes. In order to achieve that purpose, several computational techniques such as pattern classification approaches can be applied. The classification problem consists in identifying different classes or groups associated with a particular disease (e.g., various types of cancer, in terms of the gene expression level). However, the enormous quantity of genes and the few samples available, make difficult the processes of learning and recognition of any classification technique. Artificial neural networks (ANN) are computational models in artificial intelligence used for classifying, predicting and approximating functions. Among the most popular ones, we could mention the multilayer perceptron (MLP), the radial basis function neural network (RBF) and support vector machine (SVM). The aim of this research is to propose a methodology for classifying DNA microarray. The proposed method performs a feature selection process based on a swarm intelligence algorithm to find a subset of genes that best describe a disease. After that, different ANN are trained using the subset of genes. Finally, four different datasets were used to validate the accuracy of the proposal and test the relevance of genes to correctly classify the samples of the disease.

1. Introduction

DNA microarray is an essential technique in molecular biology that allows, at the same time, to know the expression level of millions of genes. The DNA microarray consists in immobilizing a known deoxyribonucleic acid (DNA) molecule layout in a glass container and then this information with other genetic information are hybridized. This process is the base to identify, classify or predict diseases such as different kind of cancer [1–4].

The process to obtain a DNA microarray is based on the combination of a healthy DNA reference with a testing DNA. Using fluorophores and a laser it is possible to generate a color spot matrix and obtain quantitative values that represent the expression level of each gene [5]. This expression level is like a signature useful to diagnose different diseases. Furthermore, it can be used to identify genes that modify their genetic expression when a medical treatment is applied, identify tumors and genes that make regulation genetic networks, detect mutations among other applications [6].

Computational techniques combined with DNA microarrays can generate efficient results. The classification of DNA microarrays can be divided into three stages: gene finding, class discovery, and class prediction [7,8]. The DNA microarray samples have millions of genes and selecting the best genes set in such a way that get a trustworthy classification is a difficult task. Nonetheless, the evolutionary and bio-inspired algorithms, such as genetic algorithm (GA) [9], particle swarm optimization (PSO) [10], bacterial foraging algorithm (BFA) [11] and fish school search (FSS) [12], are excellent options to solve this problem. However, the performance of these algorithms depends of the fitness function, the parameters of the algorithm, the search space complexity, convergence, etc. In general, the performance of these algorithms is very similar among them, but depends of adjusting carefully their parameters. Based on that, the criterion that we used to select the algorithm for finding the set of most relevant genes was in term of the number of

* Corresponding author. Tel.: +52 55 5622 3899x44304.
E-mail addresses: beatriz.garro@iimas.unam.mx (B.A. Garro), katya.rodriguez@iimas.unam.mx (K. Rodríguez), ravem@lasallistas.org.mx (R.A. Vázquez).

http://dx.doi.org/10.1016/j.asoc.2015.10.002
1568-4946/

parameters of each algorithm. In that sense, the ABC algorithm was chosen because it has fewer parameters to adjust compared with other evolutionary algorithms. Moreover, literature reports that the ABC algorithm presents faster convergence than other techniques. According to [13], results to solve multi-modal and multi-variate problems are better or similar to other evolutionary algorithms. Additionally, ABC presents a higher population diversity avoiding premature convergence. However, other bio-inspired techniques, such as differential evolution, particle swarm optimization, etc., will be evaluated and compared in future works.

Artificial neural networks (ANN) are excellent computational models that have been implemented to solve different kind of problems. The pattern classification, forecasting and regression problems are areas where the ANN have demonstrated to be an efficient technique [14]. ANN have been widely applied in DNA microarrays. For example, in [15], the authors used a multilayer perceptron (MLP) with back-propagation learning and a dimensional reduction method based on k-means and principal component analysis (PCA) techniques. In [16], the authors described an application based on ANN aimed to cancer studies. In [17], the authors diagnosed disease categories using small round blue cell tumors (SRBCT) by means of reducing the dimensionality data using PCA and training ANN models with no hidden layers. In other works, like [18], the authors selected a set of genes using a filter and k-means technique to train a support vector machine (SVM) and a multilayer perceptron (MLP). In [19], the author used mutual information techniques for selecting the most relevant genes before performing the classification task. In [20], an ANN with a sample filtering algorithm is designed for separating the wrongly labeled samples from the training set, and used to construct one more ANN just for the wrong samples classified. In [21], the authors described the singular value decomposition (SVD) technique for training a single layer feed-forward neural network. The authors in [22] performed a selection of genes based on k-means and PCA; finally, an ANN was training during a recursive feature elimination to classify BRCA1 and BRCA2 mutations and childhood SRBCT. In [23], the authors performed a feature selection in DNA microarrays using an ensambling learning technique. Also, they used an algorithm that converts a multiclass problem into multiple binary classes to reduce the complexity of the problem. In [24], the authors analyzed a generalized radial basis function (GRBF), where the coefficients of the neural network were tuned by a hybrid evolutionary algorithm. In [25], the authors used a neurofuzzy model (NFM) for identify distinct prognostic genes with a carcinogenic pathways.

On the other hand, bioinspired and evolutionary algorithms have been widely applied to select the set of genes that best describe a disease. For example, in [26], the authors used the ant colony optimization algorithm (ACO) for selecting the most representative genes from a DNA microarray. A nonparallel plane proximal classifier (NPPC) is described in [27], where the authors used genetic algorithms for selecting genes for a cancer diagnosis and the results are compared against a support vector machine (SVM). In [28], the authors described a genetic bee colony algorithm in order to select the most predictive and informative genes for cancer classification. In [29], the authors presented an improved genetic algorithm that selects the gene subset from the high dimensional gene data for breast cancer diagnosis. In [30], the authors proposed a novel feature selection approach for the classification of high dimensional cancer microarray data, which uses filtering technique such as signal-to-noise ratio (SNR) score and optimization techniques as particle swarm optimization (PSO).

In this research, we introduce a new approach for classifying DNA microarray data based on artificial neural networks and dimensional reduction technique, previously described in [31]. The proposed methodology uses the Artificial Bee Colony (ABC) algorithm as an optimization technique for selecting the set of genes, from a DNA microarray, that best described a particular disease. After that, this information is used to train three types of ANN (multilayer perceptron (MLP), radial basis function (RBF) and support vector machines (SVM)) for classifying the DNA microarrays associated to a disease. In order to test the accuracy of the proposed methodology, four different datasets were used.

It is important to remark that other strategies, applied to DNA microarrays classification, implement the ANN in the fitness function and at the same time perform a dimensional reduction, provoking that the individual evaluation be more expensive in time and computational resources. The main contribution of this paper is firstly reduced the number of genes by means of the ABC algorithm. The proposed fitness function was computed in terms of the classification error using an Euclidean distance. Then, the reduced genes set is used to train an ANN in order to classify the DNA microarray data.

The rest of this paper is organized as follows: Section 2 presents an introduction to DNA microarrays. A brief explanation of ABC algorithm is presented in Section 3. Section 4 presents the basic concepts related to artificial neural networks. In addition, the propose methodology is outlined in Section 5 followed by the experimental results in Section 6. Finally, conclusions of this research are given in Section 7.

2. DNA microarrays

The human genome sequencing was completed in 2001 [32,33]. This discovery impacted the world because has allowed better diagnostics, to know the genes that participate in an illness for doing a better treatment and even more, to know about the human evolution and other advantages in sciences like biomedics, genetics, biology and so forth. In [34], the authors described the use of DNA microarray technologies, presented an overview of their frequent biomedical applications and described the steps of a typical laboratory procedure to obtain information with this powerful technique.

DNA microarray is a container that immobilize DNA molecule, complementary DNA or oligonucleotides for hybridizing with DNA molecule marked to be analyzed. The container is made of glass, nylon or silicone. There are two types of DNA microarrays: the ergonomic and the transcriptomics. The first one is divided into two kinds: that can detect lost or profit genes, and that can detect mutations. The second one measures the mRNA levels [35]. DNA microarray allows to use the genome sequencing information to measure quantitatively the expression level of millions of genes at the same time. This expression level is like a signature useful to diagnose diseases, identify tumors, select the best treatment to resist illness and detect mutations.

To obtain the expression level of a DNA microarray sample is necessary to compare the healthy DNA reference, called “data control”, against a testing DNA (the sample to be studied), see Fig. 1 [34]. First, the messenger ribonucleic acid (mRNA) of both tissues is isolated. Then, it is necessary to obtain the corresponding complementary DNA (cDNA). Additionally, these molecules should be marked with a different fluorophore: Cy3 for the experimental sample (red color) and Cy5 for the control sample (green color). Furthermore, the marked molecules are mixed for the hybridizing process that consists of the union of the cDNA of each sample [36]. The result is a matrix with many colored spots. The red color indicates that a particular gene (spot) is more expressed in the diseased sample. The green color means that a particular gene is more expressed in the healthy sample. The yellows spots indicate that the gene is equally expressed in healthy and diseased samples.

DNA microarray is an efficient technology that presents many advantages. It allows the analysis of thousand of genes at the same time, decreasing the spend time to its study. Also it increases