Genomics 109 (2017) 91–107

Contents lists available at ScienceDirect

Genomics

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

Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts

M. Dashtban *, Mohammadali Balafar

Department of Computer Engineering, Faculty of Electrical & Computer Engineering, University of Tabriz, Iran

ARTICLE INFO

Article history:
Received 5 August 2016
Received in revised form 9 January 2017
Accepted 24 January 2017
Available online 1 February 2017

Keywords:
Gene selection
Cancer classification
Microarray data analysis
Intelligent Dynamic Algorithm
Random-restart hill climbing
Reinforcement learning
Penalizing strategy
Cut and splice crossover
Self-refinement strategy
Feature selection

ABSTRACT

Gene selection is a demanding task for microarray data analysis. The diverse complexity of different cancers makes this issue still challenging. In this study, a novel evolutionary method based on genetic algorithms and artificial intelligence is proposed to identify predictive genes for cancer classification. A filter method was first applied to reduce the dimensionality of feature space followed by employing an integer-coded genetic algorithm with dynamic-length genotype, intelligent parameter settings, and modified operators. The algorithmic behaviors including convergence trends, mutation and crossover rate changes, and running time were studied, conceptually discussed, and shown to be coherent with literature findings. Two well-known filter methods, Laplacian and Fisher score, were examined considering similarities, the quality of selected genes, and their influences on the evolutionary approach. Several statistical tests concerning choice of classifier, choice of dataset, and choice of filter method were performed, and they revealed some significant differences between the performance of different classifiers and filter methods over datasets. The proposed method was benchmarked upon five popular high-dimensional cancer datasets; for each, top explored genes were reported. Comparing the experimental results with several state-of-the-art methods revealed that the proposed method outperforms previous methods in DLBCL dataset.

1. Introduction

Since the emergence of novel biotechnology, several methods have been proposed for microarray data analysis. Utilizing high-density oligonucleotide chips and cDNA arrays enable researchers to measure the expression levels of thousands of genes simultaneously in a single microarray experiment. The obtained genes could be used in various applications such as medical diagnosis and prognosis. One of the most important applications of microarray data is the classification of tissue samples into the normal or cancerous tissues. One of the most important applications of microarray data is the classification of tissue samples into the normal or cancerous tissues. However, a significant number of genes are irrelevant or insignificant to clinical applications [13,43,77]; consequently, they are unrelated to the classification tasks [31,84]. On the other hand, interpreting such huge number of genes is impossible. Therefore, selecting a proper number of most discriminating genes has been the most challenging task in microarray data analysis.

There are some major challenges associated with the analysis of microarray data, for example, they have a large number of genes (the curse of dimensionality) and a few number of experiments (the curse of data sparsity) usually <100 samples [12]. Moreover, they have high complexity because most of the genes are directly or indirectly correlated with each other, e.g., a gene with high expression level may simply be activated by a high-regulated gene. Several methods have been proposed in the literature to deal with such issues.

Several statistical methods have been developed to select the genes for disease diagnosis, prognosis, and therapeutic targets [30,73]. In addition to the statistical methods, recently, data mining and machine learning solutions have been widely used in genomic data analysis [41,56,88]. For example, Cho et al. [18] used a modified kernel Fisher discriminant analysis (KFDA) to analyze the hereditary breast cancer dataset [34]. The KFDA classifier used the mean-squared-error as the gene selection criterion. Besides, many hybrid evolutionary algorithms have been proposed to improve the accuracy of the classification methods [36,69,72]. Various evolutionary algorithms aim to find an optimal subset of features by using bio-inspired solutions (such as PSO, Honey Bee, Firefly algorithms). These kinds of algorithms have shown appropriate performances over various problems but are dependent on experts" intervention to obtain the desired performance. Gene selection is a subgroup of larger machine learning class of feature selection. Feature selection methods could be roughly classified into four distinct models: filter, wrapper, hybrid, and embedded models [41,65].

The filter model relies on the general statistical properties of training data without using any learning algorithm. In this model, genes are usually ranked individually using few criteria [61,62,70]. The genes with the highest rank can be selected for further analysis. Some of the successful

* Corresponding author.
E-mail addresses: dashtban@tabrizu.ac.ir, Dashtban.Edu@gmail.com (M. Dashtban).

http://dx.doi.org/10.1016/j.ygeno.2017.01.004
0888-7543/

filter approaches include Laplacian score [33], signal-to-noise ratio [30], mutual information [11], information gain [59], consensus independent component analysis that uses gene expression value for cancer classification [90], T-test feature ranking for gene selection [92], fuzzy logic for eliminating of redundant features, [37], maximum–minimum correntropy criterion [54], and receiver operating characteristics analysis [42]. Also, there is an excellent survey of filter techniques in [44] which focused on gene selection for microarray data analysis.

The wrapper model often utilizes evolutionary strategies to guide their searches. It often starts with a population of solutions; each contains a subset of features. Then each subset would be evaluated using a learner to assign fitness to each subset. Usually, an iterative process is used to improve the solutions (i.e., feature subsets). Some of the state-of-the-art wrapper approaches are particle swarm optimization [39], ant colony optimization [87], artificial bee colony (ABC) algorithm [29,32], ADSRPCL-SVM [81], genetic algorithm with SVM [78] and genetic programming (used for the prediction of alternative mRNA splice variants) [76].

Gene selection is a subgroup of larger machine learning class of feature selection. Feature selection methods could be roughly classified into four distinct models: filter, wrapper, hybrid, and embedded models [41,65]. The filter model relies on the general statistical properties of training data without using any learning algorithm. In this model, genes are usually ranked individually using few criteria [61,62,70]. The genes with the highest rank can be selected for further analysis. Some of the successful filter approaches include Laplacian score [33], signal-to-noise ratio [30], mutual information [11], information gain [59], consensus independent component analysis that uses gene expression value for cancer classification [90], T-test feature ranking for gene selection [92], fuzzy logic for eliminating of redundant features, [37], maximum–minimum correntropy criterion [54], and receiver operating characteristics analysis [42]. Also, there is a good survey of filter techniques in [44] which focused on gene selection for microarray data analysis.

The wrapper model often employs evolutionary strategies to guide their searches. It often starts with a population of solutions; each contains subset of features. Then each subset would be evaluated using a learner to assign fitness to each subset. Usually, an iterative process is used to improve the solutions (i.e., feature subsets). Some of the state-of-the-art wrapper approaches are particle swarm optimization [39], ant colony optimization [87], artificial bee colony (ABC) algorithm [29,32], ADSRPCL-SVM [81], genetic algorithm with SVM [78] and genetic programming (used for the prediction of alternative mRNA splice variants) [76].

The performance of wrapper approaches is typically better than filter models because they employ the interactions between the solutions and predictors. However, the high time complexity of this model, particularly for high-dimensional data makes the need for using hybrid approaches that have lower time complexity, while a hybrid approach uses a filter model to reduce the dimensionality, it is aimed to achieve a trade-off between the time complexity and feature space size. Some state-of-the-art hybrid approaches are information gain with a novel mimetic algorithm [94], chi-square statistics with GA [45], mRMR with GA [2], a novel similarity scheme with ABC [32], and hybrid between genetic algorithm and SVM [49].

There are other feature selection approaches in which the process of learning a classifier is concurrent with feature selection. It does not use a filter model and accordingly does not shrink the feature space, instead, it tries to select high discriminant features and remove poor ones by analyzing and measuring their effects upon constructing a classifier. Furthermore, their time complexities are relatively high, particularly for high-dimensional data such as microarray data. Some of the latest embedded methods are the random forest for genomic data analysis [17], convergent random forest for predicting drug response [8], and artificial neural network approach for improving classification of precursor microRNA [63].

In the present study, a novel hybrid evolutionary algorithm called intelligent dynamic genetic algorithm (IDGA) based on genetic algorithm and some artificial intelligence concepts and techniques is described. The proposed method mainly consists of two major steps. In the first step, a score-based method is used to reduce the dimensionality and more importantly to provide statistically significant genes to the next step. In the second step, quite different scoring methods are used, the Fisher score and the Laplacian score. The performance of Fisher score and its robustness to noise has already been proven in the literature for various applications [51,58,82]. Furthermore, the high performance of Fisher score for gene selection against other widely used methods such as T-test [16], information gain, and *Z*-score was shown by [83,85]. Nonetheless, each method has its own characteristics that affect the stability of final results [23,91]. Further, Laplacian discriminant analysis [50,57] shows its competitive performance for identifying predictive genes in cancer datasets. The Laplacian score is an unsupervised method that relies on the underlying structure of a dataset. This characteristic motivated us to utilize and investigate it as a preprocessing step despite that is an unsupervised method. The proposed IDGA is benchmarked in combination with both Laplacian and Fisher score. Beforehand, a comparison based on dissimilarity of selected top M genes are performed. To the best of our knowledge, it is the first time that the Laplacian score is used directly as a gene ranking method for reducing dimensionality in a hybrid method for cancer classification.

After reducing dimensionality and selecting statistically significant genes, the IDGA method is applied. The presented evolutionary strategy is, in fact, an integer-coded genetic algorithm with dynamic-length genotype, intelligent adaptive parameters, and modified genetic operators, followed by some random initializations (populations including chromosomes with randomly generated length). To the best of our knowledge, it is the first genetic algorithm-based strategy that uses dynamic length with integer-encoding scheme for feature selection at all. The fast convergence of this method motivates us to exploit it on the high-dimensional microarray data. In fact, the variable length chromosomes with integer-encoding scheme followed by adapted genetic operators capable of dealing with dynamic chromosomes with straightforward and random initialization made this algorithm quite effective.

A few genetic algorithms with adaptive mutation and crossover probability for feature selection have proposed in the literature [45,68], which have several parameters to adjust. In the present study, an adaptive crossover and mutation rate based on the social concept of 'encouraging and penalizing strategy,' is proposed. The IDGA algorithm obtains its parameters simply with regards to the quality of explored solution compared with its pair and with total solutions. One significant advantage of this method is that the probability of promoting poor solutions would become higher by adopting more forces on the poor solutions to mutate and to cross over. Furthermore, this algorithm practically uses the well-known artificial intelligence as a concept, the random restart hill climbing for avoiding biased or improper initializations. It performs through initializing new population with randomly generated length a few times and running evolutionary process on each.

Five high-dimensional microarray cancer datasets are used to demonstrate the performance of the proposed evolutionary algorithm. The general convergence trend of IDGA is also studied, and its convergence over five different datasets regarding the number of selected genes and error loss are exhibited. The experimental results demonstrate the comparable performance of simple IDGA regarding prediction error and number of selected genes.

2. Materials and methods

Microarray dataset is usually represented as N by M matrix, where N is the number of experimental samples and N is the corresponding gene expressions. In the present study, five duplicate microarray datasets were utilized to evaluate the proposed method for gene selection and