Available online at www.sciencedirect.com
ScienceDirect
journal homepage: www.elsevier.com/locate/bbe

Original Research Article

Gene selection from large-scale gene expression data based on fuzzy interactive multi-objective binary optimization for medical diagnosis

Q1 Saleh Shahbeig, Akbar Rahideh [*], Mohammad Sadegh Helfroush, Kamran Kazemi

Department of Electrical and Electronics Engineering, Shiraz University of Technology, Shiraz, Iran

ARTICLE INFO

Article history:
Received 16 September 2017
Received in revised form
4 February 2018
Accepted 8 February 2018
Available online xxx

Keywords:
Fuzzy interactive method
Sub-optimal subset
Gene selection
Large-scale data
Multi-objective
Adaptive binary particle swarm optimization

ABSTRACT

An efficient fuzzy interactive multi-objective optimization method is proposed to select the sub-optimal subset of genes from large-scale gene expression data. It is based on the *binary particle swarm optimization* (BPSO) algorithm tuned by a chaotic method. The proposed method is able to select the sub-optimal subset of genes with the least number of features that can accurately distinguish between the two classes, e.g. the normal and cancerous samples. The proposed method is evaluated on several publicly available microarray and RNA-sequencing gene expression datasets such as leukemia, colon cancer, central nervous system, lung cancer, ovarian cancer, prostate cancer and RNA-seq lung disease. The results indicate that the proposed method can identify the minimum number of genes to achieve the most *accuracy*, *sensitivity* and *specificity* in the classification process. Achieving 100% accuracy in six out of the seven datasets investigated in this study, demonstrates the high capacity of the proposed algorithm to find the sub-optimal subset of genes. This approach is useful in clinical applications to extract the most influential genes on a disease and to find the treatment procedure for the disease.

1. Introduction

Microarray technology provides the opportunities to simultaneously monitor the expression of thousands genes in a single hybridization experiment. In addition to cell biological understanding, the analysis of microarray data can also be useful to diagnose and treat diseases. In microarray technology, by studying the gene expression in normal and abnormal cells, it is possible to identify the most influential genes involved in the formation of a disease. The "most influential genes" are those genes which divide the groups in any given datasets (disease vs. control or different disease sub-types) with the largest margin, i.e. highest classification accuracy in

[*] *Corresponding author at*: Department of Electrical and Electronics Engineering, Shiraz University of Technology, Shiraz, Iran.
E-mail addresses: s.shahbeig@sutech.ac.ir (S. Shahbeig), rahide@sutech.ac.ir (A. Rahideh).

https://doi.org/10.1016/j.bbe.2018.02.002
0208-5216/© 2018 Published by Elsevier B.V. on behalf of Nalecz Institute of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences.

the test stage. As the publicly available gene expression datasets are expanded, several algorithms have been proposed to analyze the datasets. There are two major challenges facing the algorithms of gene expression data analysis: excessive number of genes compared to the number of samples, which is known as curse of dimensionality; and difficulty to interpret the obtained results. So, the feature selection method should be able to overcome the mentioned challenge and select a sufficient number of informative genes to achieve a good *accuracy* for the classification task. It is highly regarded by geneticists and physicians as a valuable research.

The difficulty to interpret the results is more pronounced when the purpose is to identify the genes by which the classes can be distinguished. For example, in the *support vector machine* (SVM)-based classification [1–4], a separator function is found and placed between samples belonging to different classes, while in KNN-based classification [5] the Euclidean distance between samples is used to distinguish between the classes. As a result, in these algorithms, the identification of the genes with the most capability to separate the samples of different classes is difficult. Statistical methods, such as Bayesian-based model [6,7], have also been used in this realm with the drawback of high computational burden.

The feature selection techniques fall into three following categories, as shown in Fig. 1 [8], depending on the interaction of feature selector and the classifier: (a) filtering methods; (b) wrapper methods; and (c) embedded methods. Filtering techniques are independent from the classifier and select the features based on the intrinsic (e.g. statistical) properties of the data. In contrast, the wrapper and embedded techniques are classifier dependent in which not only the classifier tunable parameters need to be sub-optimally obtained but also a sub-optimal subset of features should be selected. Since the proposed method in this manuscript fall into the category of wrapper methods, the choice of classifier in order to achieve the highest classification accuracy with the lowest number of genes is important. In this case, classifier is a main component in the structure of the feature selection algorithm. The difference between the wrapper and embedded techniques is the order of the classifier tunable parameters space (also known as the hypothesis space) and the feature selection space as depicted in Fig. 1.

1.1. *Literature review*

An overview of some of the feature selection algorithms reported in recent years has been presented in this section.

A feature selection method has been proposed in [4]. It uses a backward elimination procedure similar to that of implemented in support vector machine recursive feature elimination (SVM-RFE). In this method, at each step, the ranking score of each feature is computed by using a statistical analysis of weight vectors of multiple linear SVMs trained on sub-samples of the original training data. A feature selection method has been reported for the classification of autism data in which the genes with similar information are excluded in the pre-processing stage [9]. Then 8 methods of statistical feature selection have been applied to the remaining genes. A number of genes with the highest rank have been selected using each of the statistical methods and the obtained genes from all the methods have been merged. K-means clustering technique has been employed to identify the sub-optimal number of genes and finally a SVM classifier has been used to separate the categories.

Wang and Han [10] have presented a hybrid feature selection algorithm for the sake of gene expression data analysis. The algorithm creates a connection between the purely statistical methods of feature selection and those based on optimization techniques. In this way, the number of genes has been reduced to a desired level.

An optimization method based on binary biogeography has been proposed to select a small subset of informative genes [11]. To this end, the Fisher–Markova selector algorithm has initially been used to select 60 genes with the highest statistical scores. Then *binary biogeography-based optimization* (BBBO) algorithm based on a model of binary migration and mutation has been employed to optimize the discrete problem. Finally, *multi-objective* BBBO (MOBBBO) has been proposed by integrating the non-dominated sorting method and the crowding distance method into the BBBO framework. Detection of cancer by feature selection and classification through the use of adjustable SVM has been presented in [12]. *Consistency-based feature selection* (CBFS) method and *signal-to-noise ratio* (STNR) technique has been performed and the transductive SVM has been used for a semi-supervised classification.

The feature selection and clustering of gene expression data have been reported by using the so-called *clustering large applications based on RAN-domized search* (CLARANS) method [13]. The gene ontology based CLARANS algorithm has been used to select a number of effective genes for the clustering purposes. A combination of the simulated annealing and discretized multivariate joint entropy has been used for the feature selection of microarray gene expression data [14]. Wang et al. [15] have proposed a sub-optimal feature selection approach for sparse linear discriminant analysis to be used for gene expression data. Cui et al. [16] have presented an algorithm for dimensionality reduction of gene expression data based on the so-called sparse maximum margin discriminant analysis to extract influential features and select

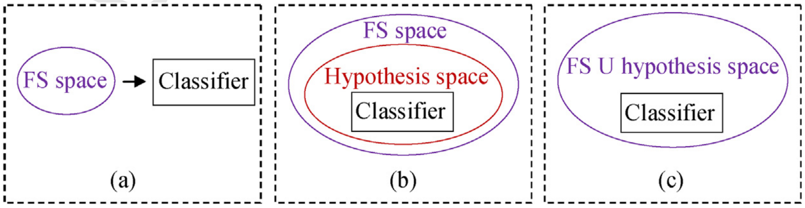

Fig. 1 – (a) Filtering method, (b) wrapper method, and (c) embedded method [8].