Signal Processing 131 (2017) 58–65

Contents lists available at ScienceDirect

Signal Processing

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

A fuzzy multi-objective hybrid TLBO–PSO approach to select the associated genes with breast cancer

Saleh Shahbeig, Mohammad Sadegh Helfroush *, Akbar Rahideh

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

ARTICLE INFO

Article history:
Received 28 May 2016
Received in revised form
1 July 2016
Accepted 30 July 2016
Available online 1 August 2016

Keywords:
Multi-objective binary optimization
Hybrid method
Mutated method
Fuzzy adaptive adjusting
Gene selection
Breast cancer

ABSTRACT

When the genes associated with breast cancer are mutated, they may not function normally and breast cancer risk increases. Therefore the method that among huge number of unrelated genes identifies the genes associated with breast cancer is an efficient method for diagnosis of breast cancer before the progression of the disease. In this paper, a new hybrid algorithm is proposed to identify the most relevant genes involved in breast cancer development. A combination of the teaching learning-based optimization (TLBO) algorithm and the proposed mutated fuzzy adaptive particle swarm optimization (PSO) algorithm is employed to find the smallest subset of genes involved in breast cancer with the highest amount of classification *accuracy*, *sensitivity* and *specificity*. Due to the presence of the two conflicting goals, i.e. minimization of the number of selected genes and maximization of the classification performance, the optimization problem is represented in a multi-objective form and solved using the Pareto technique. The obtained results show that the proposed technique is able to achieve the *accuracy* of 91.88%, the *sensitivity* of 90.55% and the *specificity* of 93.33% in the breast cancer microarray data by selecting 195 genes.

1. Introduction

Breast cancer is a disease that affects breast tissue and can occur in both men and women but, it is very rare in men. In the event of early diagnosis, breast cancer can be prevented from propagating to other tissues of the body. Microarray gene expression technology has been widely used for gene expression analysis and is able to measure the expression levels of thousands of genes simultaneously. Microarray data analysis can also be beneficial to diagnose diseases by identifying the most influential genes involved in the disease creation. The fact is that a very large number of genes have no information and the large number of genes to the small number of samples challenges the classification of the microarray gene expression data. It is important that the diagnosis research performs the diagnosis of disease based on microarray gene expression data which contains a sufficient number of genes to diagnose intended disease in order to get a good *accuracy* for the classification task.

There are a number of researches have been reported on the classification or clustering of microarray data.

Bosio et al. [1] have proposed a framework for microarray data classification. In the first step, the original feature set has been enhanced by a new set of features which are obtained through a hierarchical clustering process on the original microarray data. Then in the second step, the Improved Sequential Floating Forward Selection algorithm (IFFS) has been applied to choose a subset of features for classification of microarray data. Sasikala et al. [2] have proposed shapely value embedded genetic algorithm (SVEGA) to gene selection for breast cancer diagnosis. The SVEGA has included two operators namely “remove” and “include” genes to discover the genetic algorithm solution. The proposed method is ranking the genes according to its ability to differentiate the classes. In other words, the proposed algorithm selects the genes that can maximize the ability to discriminate between the two classes. An ant colony optimization-based dimension reduction method has been presented for microarray gene expression data analysis in [3]. The proposed algorithm consists of two steps, namely ant system and ant colony system, which have been applied to search for genes, respectively. In the first step, an ant system has been used to filter the non-significant genes, and a number of genes have been stored for the next step. In the next step, a reclaimed ant colony system has been applied for the gene selection process. Gonzalez-Navarro et al. [4] have used a combination of the simulated annealing and discretized multivariate joint entropy for gene selection from microarray data. The discretized multivariate joint entropy has been used for gene correlation evaluation in a microarray gene expression context. In the proposed algorithm, a simulated annealing

* 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), ms_helfroush@sutech.ac.ir (M.S. Helfroush), rahide@sutech.ac.ir (A. Rahideh).

http://dx.doi.org/10.1016/j.sigpro.2016.07.035
0165-1684/

algorithm has been combined with the incrementally computed joint entropy. A gene selection algorithm has been proposed in [5]. In this method, first importance factor of each gene of experimental cancer dataset has been computed. Then initial important genes have been selected according to high importance factor of each gene. Then traditional k-means clustering algorithm has been applied on each selected gene and miss-classification errors of individual genes have been computed. Finally, the selected subset of genes has been formed by selecting most important genes with respect to less miss-classification errors. A method has proposed by Cui et al. in [6] based on the sparse maximum margin discriminant analysis (SMMDA) for dimensionality reduction of gene expression data. In this method, first the one-dimensional projection of the gene expression data has been found in the most separable direction using SMMDA. Then, the sparse representation technique has been applied to regress to the projection, and the relevance vector for the gene set has been obtained. Finally, discriminant genes have been selected according to this vector. An optimal feature selection approach for sparse linear discriminant analysis (LDA) has been proposed in [7] to be used in gene expression data. A minimization method has been used to select the important features from which the LDA will be constructed. Sun et al. [8] have proposed a feature selection algorithm for data classification that their idea is to break down a nonlinear problem into a set of locally linear ones through local learning, and then learn feature relation globally within the large margin framework. Paliwal et al. [9] have proposed an improvement of the direct linear discriminant analysis (DLDA) technique that is a well-known technique for dimensionality reduction.

In this paper, a hybrid algorithm of TLBO and mutated adaptive PSO based on fuzzy adaptation has been proposed in order to find the most influential genes in breast cancer. A multi-objective optimization problem is formed to find the optimal subset of genes relevant to breast cancer. The objectives are the highest classification performance and the lowest number of genes. The classification performance has been evaluated based on the classification *accuracy*, *sensitivity* and *specificity*. In a preprocessing stage, the genes with least useful information are discarded by using a statistical filtering method. Then the remaining genes are sent to the optimization stage. The combination of the TLBO and fuzzy adaptive PSO algorithms has been used to find the *pareto-front* solutions (all of the non-dominated gene subsets) in the breast cancer microarray dataset. It is noted that to efficiently reduce the randomness effects and to increase the robustness of the proposed algorithm, not only the folding technique has been used for selecting the test and train datasets but also the optimal selected genes have been assessed several times and the average results of the classification performance has been considered.

The rest of the paper is organized as follows. The proposed feature selection algorithm is described in Section 2. In Section 3, the results are presented. Finally, the discussion and conclusion have been provided in Sections 4 and 5, respectively.

2. Proposed optimization algorithm

In order to effectively represent the processes of the proposed optimization algorithm, each building block has been separately explained.

2.1. PSO algorithm

Particle swarm optimization algorithm is an optimization method based on population generation. It is an intelligent algorithm inspired by the social treatment of birds or fish. In the PSO algorithm, each particle moves in the search space with its corresponding velocity. Each particle pertains to its personal best solution and the global best solution among all particles which respectively denoted with ***pbest*** and ***gbest***. At the $(n+1)$-th iteration, the velocity and position of the i-th particle are updated according to the following relations [10–12].

$$v_i^{n+1} = w^n.\ \boldsymbol{v}_i^n + c_1.\ \mathrm{rand}()_1.\ (\boldsymbol{pbest}_i^n - \boldsymbol{x}_i^n) + c_2.\ \mathrm{rand}()_2.\ (\boldsymbol{gbest}^n - \boldsymbol{x}_i^n) \quad (1)$$

$$\boldsymbol{x}_i^{n+1} = \boldsymbol{x}_i^n + \boldsymbol{v}_i^{n+1} \quad (2)$$

where, c_1 and c_2 are learning factors, $\mathrm{rand}()_1$ and $\mathrm{rand}()_2$ are random values that must be in the range of $0 \le \mathrm{rand}()_{1,2} \le 1$ and w is the inertia weight.

c_1 and c_2 are positive values in the range of $0 < c_1,\ c_2 \le 2$ that is $c_1 + c_2 \le 4$ and w must be in the range of $0.4 \le w \le 1$ [13].

$\boldsymbol{v}_i^{n+1}$ and $\boldsymbol{v}_i^n$ are velocities of particle i at two successive iterations. $\boldsymbol{x}_i^n$ and $\boldsymbol{x}_i^{n+1}$are the current position and the updated position of the i-th particle (solution), respectively. Each particle can be indicated as $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, where D is the dimension of each particle. For the i-th particle, the velocity vector can be indicated as $\boldsymbol{v}_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$ [10–12]. To a large extent, the performance of the classical PSO algorithm depends on c_1, c_2 and w. In order to address this issue, in this manuscript a fuzzy adaptive method has been proposed to calculate the inertia weight of $w$and learning factors of c_1, c_2.

2.2. Mutated PSO

One of the weaknesses of the classical PSO algorithm is the possibility to trap in the local optimum solutions. In order to overcome this problem, a new mutated version of the particle swarm optimization method has been proposed in this paper. The mutation method is a strong technique to reclaim the efficiency of the PSO algorithm. In the mutated PSO algorithm and in each iteration, four particles are selected randomly from the current population. The mutated particle $(\boldsymbol{x}_{mutation,i}^n)$ is proposed as follows.

$$\boldsymbol{x}_{mutation,i}^n = \boldsymbol{gbest} + F_1(\boldsymbol{x}_{k_1}^n - \boldsymbol{x}_{k_2}^n) + F_2(\boldsymbol{x}_{k_3}^n - \boldsymbol{x}_{k_4}^n) \quad (3)$$

where ***gbest*** is the global best solution and F_1 and F_2 are the mutation parameters that must be in the range of $0.1 \le F_1, F_2 \le 0.9$ [13].

The mutation parameters are selected randomly in each iteration. For greater effectiveness, four selected particles should be as: $\boldsymbol{x}_{k_1}^n \neq \boldsymbol{x}_{k_2}^n \neq \boldsymbol{x}_{k_3}^n \neq \boldsymbol{x}_{k_4}^n \neq \boldsymbol{x}_i^n$.

Finally, in order to produce a new improved particle, the generated mutated particle $(\boldsymbol{x}_{mutation,i}^n)$ is blended with the target particle according to the following relation.

$$\boldsymbol{x}_{improved,i,p}^n = \begin{cases} \boldsymbol{x}_{mutation,i,p}^n, & \text{if } Cr > \mathrm{rand}() \\ \boldsymbol{x}_{i,p}^n & \text{otherwise} \end{cases}, \quad p = 1, 2, \ldots, D \quad (4)$$

where p is the index of the elements of each particle and Cr is the crossover constant in the range of $0 \le Cr \le 1$ [13].

It is important to note that the bounds for all elements in each new generated particle (solution) should be checked. If any element in each new generated solution exceeds its limitations, it should be replaced by its own upper or lower bound.

2.3. Fuzzy adaptive PSO

A constant or even linearly changed value of inertia weight (w) may prevent the PSO algorithm from reaching the optimum result. The influence of the previous velocity on the current velocity is specified by the inertia weight. Fuzzy tuning of the inertia weight