Applied Soft Computing 65 (2018) 58–68

Contents lists available at ScienceDirect

Applied Soft Computing

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

Multi-objective simplified swarm optimization with weighting scheme for gene selection

Chyh-Ming Lai

Institute of Resources Management and Decision Science, Management College, National Defense University, Taipei 112, Taiwan

ARTICLE INFO

Article history:
Received 24 July 2017
Received in revised form
24 November 2017
Accepted 31 December 2017
Available online 10 January 2018

Keywords:
Gene expression dataset
Gene selection
Multi-objective optimization
Aggregate filter method
Simplified swarm optimization

ABSTRACT

Gene selection can be regarded as a multi-objective problem which involves both minimizing the size of a gene subset and maximizing the prediction performance. This work proposes a hybrid filter/wrapper method for gene selection based on multi-objective optimization. In this method, an emerging aggregate filter method is adopted as a filter with which to choose the most informative genes; in addition, a multi-objective simplified swarm optimization (MOSSO) is proposed and integrated with a support vector machine as a wrapper to seek an optimal gene subset from the selected genes. Unlike most current multi-objective based methods employed to handle gene selection problems, the proposed MOSSO uses a weighting scheme to guide the search towards the interesting regions as defined by the preference, which means that not all Pareto optimal solutions are generated, but only the ones gene selection prefers. The proposed method is validated using ten gene expression datasets, and the corresponding results are compared with those obtained with existing works. Statistical analysis indicates that the proposed method is highly competitive and, can be considered a promising alternative for dealing with gene selection problems.

1. Introduction

Since the mid-1980s, microarray technology has been widely utilized for disease diagnostics [1]. It aids medical personnel and researchers to simultaneously access the expression levels of thousands of genes, and finally produce microarray data. Such data have been successfully applied to cancer classification, where the aim is to classify and predict the diagnostic category of a sample by its gene expression profile [2–4]. The challenges confronted by the development of an effective classifier are the characteristics of expression data: high dimensions, large number of irrelevant genes and small sample size which result in higher computational complexity and more prediction errors [5].

Gene selection, also called feature selection, can be considered as an efficient and effective method in enhancing the predictive performance of a model; it is a key pre-processing step in data mining. It focuses on identifying an optimal subset of genes from an expression dataset by reducing redundant, irrelevant or noisy genes [6,7]. Depending on how the relevance of each gene to the target class is evaluated, gene selection can be mainly classified into the filter, wrapper and hybrid methods [7–9].

The filter technique evaluates the relevance of a gene independently from a classifier [10–15], whereas the wrapper technique integrates a predetermined learning algorithm with a classifier to group an optimal gene subset according to the prediction accuracy [8,16–22]. Although the filter method is more efficient than the wrapper method, the classification performance of the latter is much better than that of the former [7,9]. The hybrid method has obtained promising results in a more efficient way than the wrapper method. It is a combination of the filter and wrapper techniques to take advantage of their strengths in a complementary way [23–27]. However, those methods often regard gene selection as a single-objective problem. The main drawback is the difficulty in exploring different potential trade-offs between classification accuracy and different subsets of selected genes.

Generally, gene selection is a multi-objective problem which involves both minimizing the size of a gene subset and maximizing the prediction performance. Meanwhile, optimal gene selection is a complex problem proven to be NP-hard [28]. Hence, the main focus has been on developing multi-objective methods based on evolutionary algorithms, such as particle swarm optimization (PSO) [29], genetic algorithm (GA) [30,31], simulated annealing [32] and biogeography based optimization (BBO) [26]. However, the number of genes is typically very large; most of those methods adopting the

E-mail addresses: chyh.ming.lai@gmail.com, ndmcc05001@ndu.edu.tw

https://doi.org/10.1016/j.asoc.2017.12.049
1568-4946/© 2018 Elsevier B.V. All rights reserved.

wrapper technique face the problem of intractable computational time.

In addition, those multi-objective optimization algorithms are designed to search for all Pareto optimal solutions, assuming that all non-dominated solutions are desirable [33,34]. In practice, the main purpose of gene selection is to enhance the classification performance of a classifier. Thus, gene selection may prefer to search those regions in which solutions exhibit better prediction performance instead of those with fewer genes on the Pareto front. From this point of view, those methods waste computational cost in searching undesired solutions [35,36]. Hence, more investigations are necessary into evolutionary multi-objective optimization using hybrid techniques with preference for solving gene selection problems.

The incorporation of preference in multi-objective optimization evolutionary algorithms is difficult, because the preference is vague, perceptive information and is highly dependent on the application context [37,38]. Recently, algorithms with preference-based search have been proposed [39–41]. Those algorithms introduce the preference and concentrate the search effort to solutions on the region of interest of the Pareto optimal front. As a result, the computational effort is reduced given that non-dominated solutions far from the region of interest are discarded. This work proposes a preference-based search called weight scheme (WES). It is a secondary selection criterion based on the preference and adopted to identify a non-dominated solution which has greater potential for evolving globally in the Pareto front.

In this paper, a novel hybrid method which combines an emerging aggregate filter method (AFM) [42], multi-objective simplified swarm optimization (MOSSO) and WES, is proposed to solve the gene selection problem. AFM is used to select the most informative genes, and then MOSSO with WES searches non-dominated gene subsets based on those selected genes. Finally, the performance of the proposed method is evaluated via ten benchmark datasets by a support vector machine (SVM) using leave-one-out cross validation (LOOCV).

This paper is organized as follows: multi-objective optimization is briefly descripted in Section 2. The overviews of AFM, SSO and SVM are provided in Section 3. The proposed hybrid gene selection method and its overall procedure are detailed in Section 4. Two experiments and the statistical analysis implemented for validating MOSSO are illustrated in Section 5. Finally, the conclusions are presented in Section 6.

2. Basic concepts of multi-objective optimization

Formally, an n-objective maximization problem with m inequality and p equality constraints is proposed as follows [33]:

$$\text{Maximize}\, F_i\left(\boldsymbol{x}\right), i = 1, 2, \ldots, n \tag{1}$$

$$\text{subject to}\, g_j\left(\boldsymbol{x}\right) \le 0, j = 1, 2, \ldots, m \tag{2}$$

$$h_k\left(\mathbf{x}\right) = 0, l = 1, 2, \ldots, p \tag{3}$$

where $F_i(\mathbf{x})$ is the ith objective function to be maximized, $\mathbf{x}$ is a solution vector with d decision variables. Let $\mathbf{x}$ and $\mathbf{y}$ be two feasible solutions of the above n-objective maximization problem. $\mathbf{x}$ is said to dominate $\mathbf{y}$, if the following conditions hold:

$$\forall i: F_i\left(\mathbf{x}\right) \ge F_i\left(\mathbf{y}\right) \text{ and } \exists j: F_j\left(\mathbf{x}\right) \ge F_j\left(\mathbf{y}\right) \tag{4}$$

When a feasible solution is not dominated by any other solution in the solution space, it is said to be a non-dominated solution. The set of all feasible non-dominated solutions is known as a Preto-optimal set. For a given Pareto-optimal set, the corresponding objective function values in the objective space are called the Pareto front.

3. Review of related works

3.1. Aggregate filter method (AFM)

Recently, Nguyen et al. proposed a novel aggregate filter method (AFM) which is capable of quantitatively integrating the statistical outcomes of different gene filter methods, based on analytic hierarchy process (AHP) [43]. The results show that AFM can stably yield better classification performance compared to each individual method [42,44].

Generally, filter methods rank all genes based on their own criterion, and then a subset of genes with highest ranking values are selected for supervised classification. However, the confidence in using a single criterion for identifying informative genes is not always achieved. The idea behind AFM is to achieve the synergistic effect between different filter methods through AHP [42,44]. The steps of this method are described briefly as follows:

Step 1. Choose m criteria (filter methods) with the corresponding weight vector $\boldsymbol{w} = [w_1, w_2, \ldots, w_m]^T$, calculate the score s_{ik} of each gene i for a target dataset with n genes according to each criterion k, and then produce $n \times m$ score matrix $\mathbf{S}$.

Step 2. According to Eq. (5), where $smax_k$ is the maximum score of the kth column in $\mathbf{S}$, calculate the $n \times n$ pairwise matrix $\mathbf{P}^k$ in which each element $p^k{}_{ij}$ represents the relative importance of gene i over j with respect to the kth filter method. As can be seen in the equation, $p^k{}_{ij}$ is a value limited in [1,10]; the higher it is, the more informative the ith gene is in comparison with the jth gene. If $p^k{}_{ij} = 1$, two genes are equally important.

$$p_{ij}^k = \begin{cases} c, \text{if} s_{ik} \ge s_{jk} \\ 1/c, \text{otherwise.} \end{cases} \tag{5}$$
$$c = |s_{ik} - s_{jk}| \times 9/smax_k + 1$$

Step 3. Calculate the option performance matrix $\mathbf{E} = [e_{ik}]_{n \times m}$ according to the following equation:

$$e_{ik} = \frac{1}{n}\sum\nolimits_{j=1}^{n} \frac{p_{ij}^k}{\sum_{i=1}^{n} p_{ij}^k} \tag{6}$$

Step 4. Once the option performance matrix $\mathbf{E}$ has been computed, a vector $v = [v_1, v_2, \ldots, v_n]^T$, where v_i is the comprehensive score of gene i, is obtained using the following equation:

$$v = \mathbf{E} \cdot w \tag{7}$$

3.2. Simplified swarm optimization (SSO)

Simplified swarm optimization (SSO) is a relatively new type of evolutionary computation algorithm with advantages including simplicity, efficiency and flexibility. It was originally introduced by Yeh in 2009 for overcoming the drawbacks of particle swarm optimization (PSO) in solving discrete problems [45,46], and successfully adopted in a number of applications [47–51].

Each of candidate solutions in SSO is generated randomly within the problem space and updated according to its unique update mechanism as shown in Eq. (8). Let $X_i^t = \left(x_{i1}^t, x_{i2}^t, \ldots, x_{ij}^t, \ldots x_{in}^t\right)$ be the ith solution at iteration t, x_{ij}^t be the jth variable of X_i^t, and $f\left(X_i^t\right)$ be the fitness value of X_i^t. In the update procedure of SSO, each variable x_{ij}^t is replaced successively by a value related to four different sources: the *gBest* g_j, its current *pBest* p_{ij}^{t-1}, its current value x_{ij}^{t-1}, or a random feasible value x depending on a uniform random number ρ in [0,1]. Three predetermined parameters: C_g, C_p and C_w define the probabilities of the updated variable generated from those sources. Different from PSO, the update mechanism of SSO is a simple mathematical modeling and updates each solution to