Mathematical Biosciences and Engineering

MBE, 20(2): 1580–1598.
DOI: 10.3934/mbe.2023072
Received: 03 September 2022
Revised: 16 October 2022
Accepted: 20 October 2022
Published: 03 November 2022

http://www.aimspress.com/journal/MBE

Research article

An improved MOPSO approach with adaptive strategy for identifying biomarkers from gene expression dataset

Shuaiqun Wang[1,*,†], Tianshun Zhang[1,†], Wei Kong[1], Gen Wen[2] and Yaling Yu[2,3]

[1] College of Information Engineering, Shanghai Maritime University, 1550 Haigang Ave., Shanghai 201306, China
[2] Department of Orthopaedics, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China
[3] Institute of Microsurgery on Extremities, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China

* **Correspondence:** Email: wangsq@shmtu.edu.cn.

† Shuaiqun Wang and Tianshun Zhang should be regarded as joint first authors.

Abstract: Biomarkers plays an important role in the prediction and diagnosis of cancers. Therefore, it is urgent to design effective methods to extract biomarkers. The corresponding pathway information of the microarray gene expression data can be obtained from public database, which makes possible to identify biomarkers based on pathway information and has been attracted extensive attention. In the most existing methods, all the member genes in the same pathway are regarded as equally important for inferring pathway activity. However, the contribution of each gene should be different in the process of inferring pathway activity. In this research, an improved multi-objective particle swarm optimization algorithm with penalty boundary intersection decomposition mechanism (IMOPSO-PBI) has been proposed to quantify the relevance of each gene in pathway activity inference. In the proposed algorithm, two optimization objectives namely t-score and z-score respectively has been introduced. In addition, in order to solve the problem that optimal set with poor diversity in the most multi-objective optimization algorithms, an adaptive mechanism for adjusting penalty parameters based on PBI decomposition has been introduced. The performance of the proposed IMOPSO-PBI approach compared with some existing methods on six gene expression datasets has been given. To verify the effectiveness of the proposed IMOPSO-PBI algorithm, experiments were carried out on six gene datasets and the results has been compared with the existing methods. The comparative experiment results show that the proposed IMOPSO-PBI method has a higher classification accuracy and the

extracted feature genes are verified possess biological significance.

Keywords: pathway activity; multi-objective optimization; particle swarm optimization; adaptive strategy; PBI decomposition mechanism

1. Introduction

Cancer is recognized as one of the serious challenges that endanger human health due to its high incidence and mortality. Because the patient does not show obvious symptoms, cancer is not easy to be detected in the early stage, which result in a poor prognosis. Therefore, it is urgent to use scientific and effective methods to identify biomarkers of cancer for early diagnosis and treatment. With the rapid progress of sequencing technology, research in the field of cancer treatment from the gene level has become a hot topic.

Through microarray technology, genes associated with disease can be found in a short time and used as biomarkers for early diagnosis, but due to the existence of some redundant features and inherent noise in microarray data, the selection of feature genes still face great challenges [1,2]. Some researchers think that the generation and development of cancer are related to some specific genes, and have proposed several methods to find genes related to different stages of cancer development [3,4], however, due to the lack of detailed biological processes, the results obtained by gene expression-based methods cannot be proved to be completely correct [5]. Hence, pathway-based methods for feature gene extraction were proposed to obtain biomarkers from microarray data, which helps to better understand the differences between phenotypes [6,7]. Some scholars have used particle swarm optimization to infer pathway activity and achieved good results, a popular pathway activity index t-score was chosen as the objective function in [8]. [9] proposed an approach based on binary particle swarm optimization, in this approach, partial genes are chosen automatically from microarray gene data for inferring pathway activity and the remaining genes are excluded. Reference [10] developed a multi-objective particle swarm optimization technology that using protein interaction scores and two improved indicators of pathway activity as objective functions to infer pathway activity, all genes are involved in pathway activity inference with different weights in this method.

Multi-objective particle swarm optimization algorithm has become the main research direction of multi-objective optimization due to the advantages of high efficiency and speed. However, due to the lack of selection pressures, feature solutions distribute around the Pareto front, which will result in the lack of diversity of optimal solution set [11]. Some researchers use the space decomposition strategies to maintain diversity, which divide the target space into multiple regions, such as grid division method [12] and angle division method [13]. some authors put forward the idea of population decomposition, which divide the entire population into multiple subgroups, and the subgroups guide the population update in parallel according to their respective leaders [14,15]. But fixed selection pressures provided by decomposition approaches unable achieve ideal results when dealing with complex Pareto fronts problems [16]. Reference [17] developed an adaptive feature selection approach, which adjust the selection pressure of archive and particles by adjusting the parameters in the decomposition method.

Motivated by the above researches, in this article, a pathway-based multi-objective particle swarm optimization algorithm with adaptive strategy for feature gene extraction is proposed. Two weighted

pathway activity indicators are used as objective functions, and find feature solutions by optimizing them simultaneously. In addition, in order to improve the diversity of feature solutions, the adaptive penalty boundary intersection decomposition method is introduced in the process of optimization.

The remainder of this paper is organized as follows. Section 2 provides related methods contained a general description of multi-objective optimization (MOO), particle swarm optimization, and the PBI decomposition approach. The proposed adaptive feature selection method is elaborated in Section 3. In Section 4, the detailed introduction and the preprocessing of datasets are given, and the experimental results of the proposed IMOPSO-PBI approach are presented and compared with 6 existing methods. Finally, concluding remarks are given in Section 5.

2. Related methods

2.1. MOO

In a multi-objective optimization problem, multiple objective functions need to be optimized simultaneously. The optimization problem is not only used in the real-time design of production scheduling, urban transportation, network communication and other systems, but also involves intelligent planning problems such as engineering design, data mining, and capital budgeting [18]. The mathematical definition of multi-objective optimization is as Eq (1).

$$\begin{aligned} miniminze \quad & F(x) = \left(f_1(x), f_2(x), \dots, f_m(x)\right) \\ subject\ to \quad & x \in \Omega \subseteq R^n \end{aligned} \tag{1}$$

where x is a vector in an n-dimensional decision space, and $f_i(x)$ is the objective function of the multi-objective optimization problem, m is the scale of the objective function.

In recent years, evolutionary algorithms have integrated biological information into meta-heuristic algorithms. With its unique update mechanism, many breakthrough research results have been achieved in the fields of combinatorial optimization and numerical optimization [19]. Multi-objective evolutionary algorithms include: multi-objective particle swarm algorithm [20], multi-objective bee colony algorithm [21], multi-objective ant colony algorithm [22], multi-objective immune algorithm [23], multi-objective differential algorithm [24], etc.

2.2. Particle swarm optimization

Particle swarm optimization is an evolutionary computing technology whose basic idea is to find the optimal solution through the cooperation and information sharing among individuals in the group. The individuals in the population are abstracted as particles, and the particles are affected by the combined effects of themselves and the state of the population at each iteration [25]. The velocity of the ith particle at the next moment is determined by the current velocity, the personal best position ($pbest$) and the global best position ($gbest$), and the particle moves from the current position to the new position at the updated velocity. The updating process of particle velocity and position is shown in Eqs (2) and (3).

$$v_i(t+1) = \omega * v_i(t) + c_1 * r_1 * \left(pbest_i - x_i(t)\right) + c_2 * r_2 * \left(gbest_i - x_i(t)\right) \tag{2}$$

$$x_i(t+1) = x_i(t) + v_i(t+1), i = 1,2,\dots,n \tag{3}$$

where ω represent inertia weight, r_1, r_2 are the random numbers between 0 and 1, c_1, c_2 are learning factors generally set to 2.

2.3. PBI approach

Decomposition based multi-objective evolutionary algorithm decomposed the target into a set of scalar optimization subproblems firstly, and then optimized these subproblems simultaneously. In the existing decomposition methods, weighted sum (WS) is limited to dealing with convex Pareto front problems, for non-convex Pareto frontiers, optimal solutions cannot be obtained completely. Tchebycheff approach (TCH) overcome that problem, but for a continuous multi-objective optimization problem, its aggregation function is not smooth, and the resulting Pareto front is not smooth too [26]. Compared with the above two decomposition methods, PBI approach can obtain more evenly distributed solutions [27]. The mathematical expression of PBI decomposition approach is as Eq (4).

$$\begin{aligned} &minimize\ g(x \mid \lambda, z^*) = d_1 + \theta d_2 \\ &subject\ to \quad x \in \Omega \end{aligned} \tag{4}$$

where λ is the predefined weight vector, and θ is the penalty parameter, and d_1, d_2 are shown in Eq (5).

$$d_1 = \frac{\|(F(x)^T - z^*)\lambda\|}{\| \lambda \|} \quad \text{and} \quad d_2 = \left\| F(x) - \left(z^* + d_1 \frac{\lambda}{\| \lambda \|} \right) \right\| \tag{5}$$

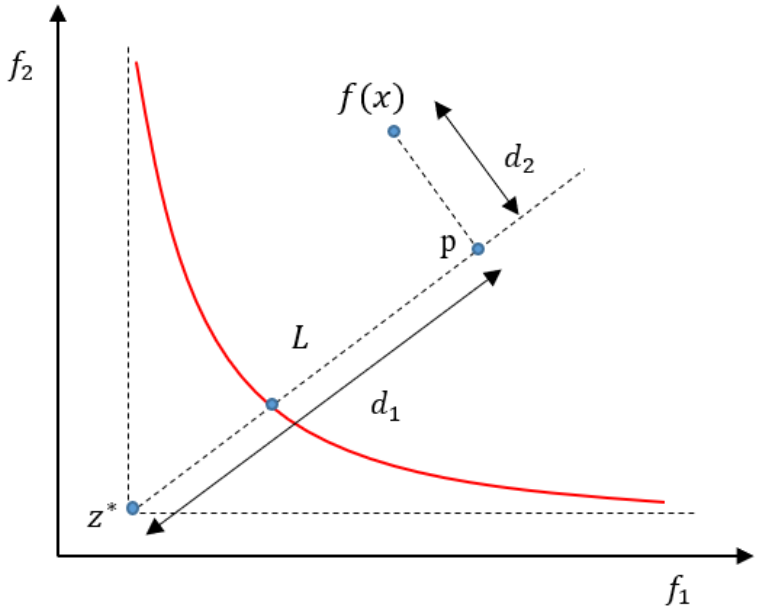

Figure 1. Distance metric of PBI method.

As shown in Figure 1, z^* is the ideal reference value, the red line is the Pareto frontier (PF), the point P is the projection of $f(x)$ on the line L, and the intersection of PF and the line L is the optimal

solution under the decomposition weight vector λ. d_1 represent the distance between the point P and the ideal point z^*, and d_2 represent the distance between $f(x)$ and the line L. It can be seen that d_1 determines the degree of convergence of the population. The smaller d_1 is, the closer its corresponding solution x is to the ideal PF, and d_2 with a penalty parameter determines the diversity of the population. Figure 2 shows the contours of g for three different penalty parameters, as can be seen from the figure, the smaller the penalty parameter θ, the larger the area of the update area, the more favorable the population is to approach the ideal PF. The larger the penalty parameter θ, the smaller the area of the update area, and the closer the updateable solution is to the weight vector, which is more conducive to the diversity of the population. That is to say, the penalty parameter θ plays an important role in the performance of penalty boundary interaction methods.

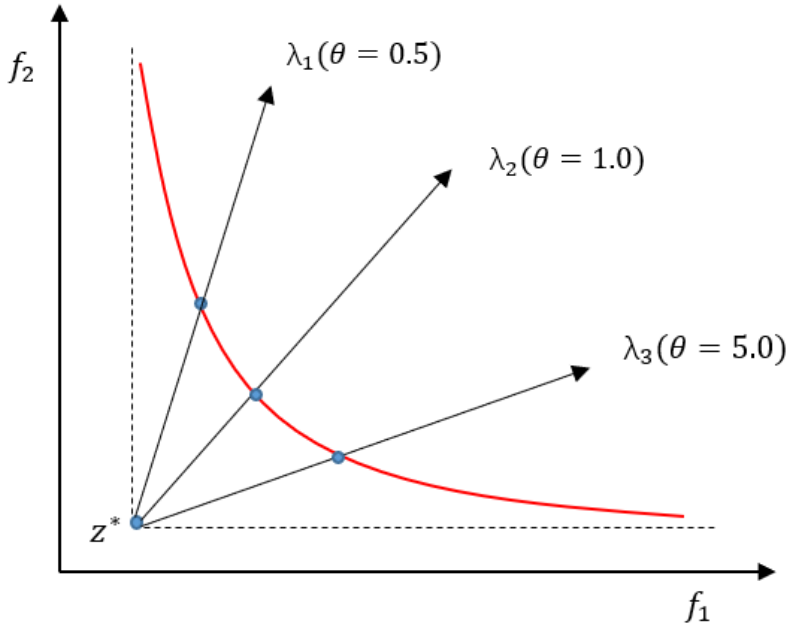

Figure 2. The contour of aggregation function under different penalty parameters.

Many scholars set the parameter to a fixed value of 5, however, it is difficult to select the solution with better convergence and diversity with a fixed penalty value. In [28], the authors introduced a subproblem-based penalty mechanism and an adaptive penalty mechanism to analyze the influence of different penalty values on the feature solution set, but this method did not consider the evolution state of each subproblem. [29] proposed a method to adjust the penalty value by the angle between the solution and the weight vector, but this increases the time complexity of the algorithm. In [17], an adaptive penalty mechanism based on PBI parameters is proposed to reduce the time complexity of calculating penalty values while maintaining diversity.

3. The proposed IMOPSO-PBI algorithm

This section mainly describes the proposed pathway-based adaptive feature gene selection algorithm. Firstly, the overall flow of the proposed algorithm is summarized and the computational complexity is described. For more comprehensibility, details about each component of IMOPSO-PBI

method are described in the following sections. The process of inferring pathway activity and particle coding is briefly introduced; two objective functions are described, and the adaptive updating mechanism of penalty value based on penalty boundary interaction is introduced.

3.1. Framework of the proposed IMOPSO-PBI algorithm

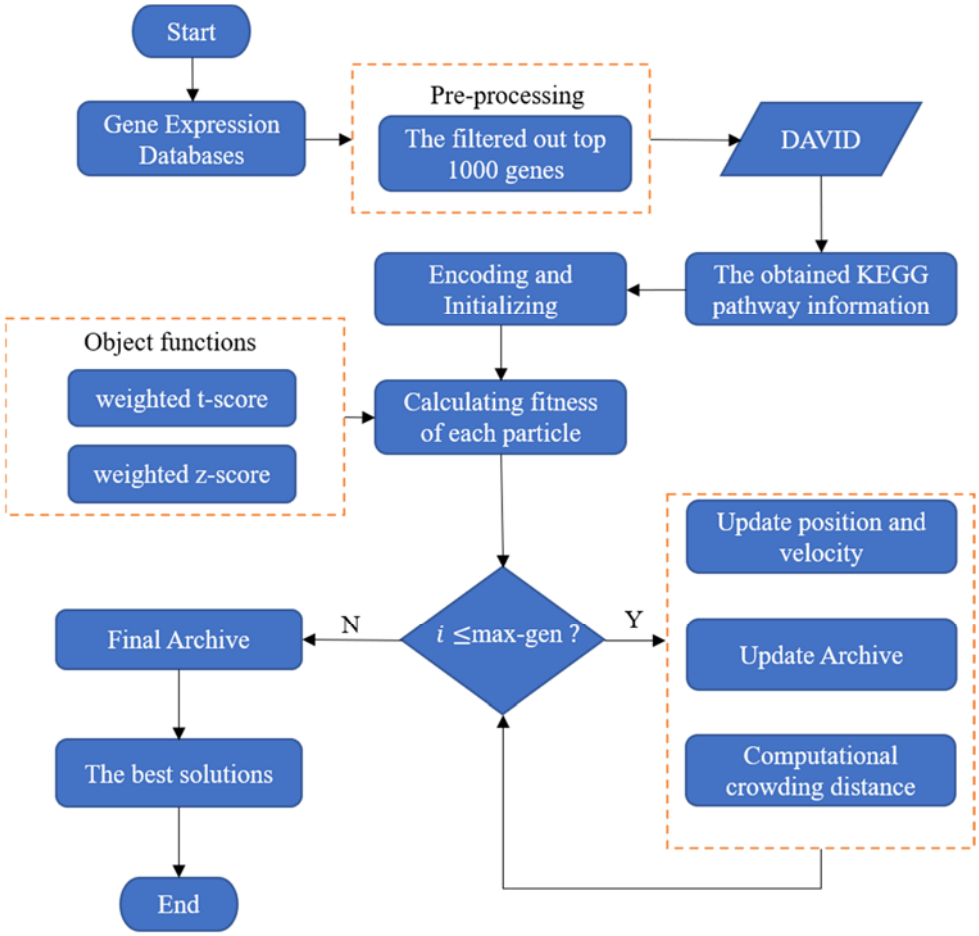

Figure 3. Flowchart of our proposed IMOPSO-PBI method.

The flowchart of our proposed IMOPSO-PBI method is shown in Figure 3, and the whole process of IMOPSO-PBI method is shown in Table 1. Firstly, the original datasets are preprocessed, the top 1000 feature genes are selected and the corresponding pathway has been obtained through the David database. Then, the pathway is encoded according to the encoding strategy introduced in Section 3.2, and an initial population containing S particles has been obtained. Next, the population is initialized, the position of each particle in the population is randomly initialized between 0 and 1, the target value of each particle is calculated through the objective functions f_1 and f_2, and the speed and position of the particle are determined by Eqs (1) and (2) to update. The leader archive is updated through Algorithm 1, and the penalty value is also updated during the archive update process, and the crowding distance between particles is calculated. Finally, the solution in the lead archive is the desired result.

Table 1. Framework of the proposed IMOPSO-PBI method.

Algorithm 1 Framework of the proposed algorithm

Input: Pre-processed Gene Expression Profile(G), Swarm size(S),
maximal generation number(*max-gen*);
Output: Final archive ← A set of Non-dominated Solutions
$\{p_1, p_2, \ldots p_N\}$ ← Obtain KEGG pathways from DAVID
for k = 1 to S **do**
GENERATE (P_k) ← Initialize the particles
f_1, f_2 ← Calculating fitness of each particle
while i <= *max-gen* **do**
Update position and velocity of particles by Eqs (1) and (2);
Update Archive;
Computational crowding distance;
end while
Obtain solutions from the final archive;
end for

3.2. Inferring pathway activity and encoding strategy

At first, the top 1000 genes taken from the preprocessed data (described in Section 4.1) are put into the DAVID Bioinformatics Resources [30], then we can obtain the corresponding KEGG pathway. Because each pathway contains several genes, so the pathway activity can be calculated by Eq (6).

$$\alpha(p_i) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(e_{ij} * \omega_i)}{\sqrt{\sum_{i=1}^{M}\omega_i}} \tag{6}$$

Here, M is the number of genes in the pathway, and N is the number of samples of each gene.

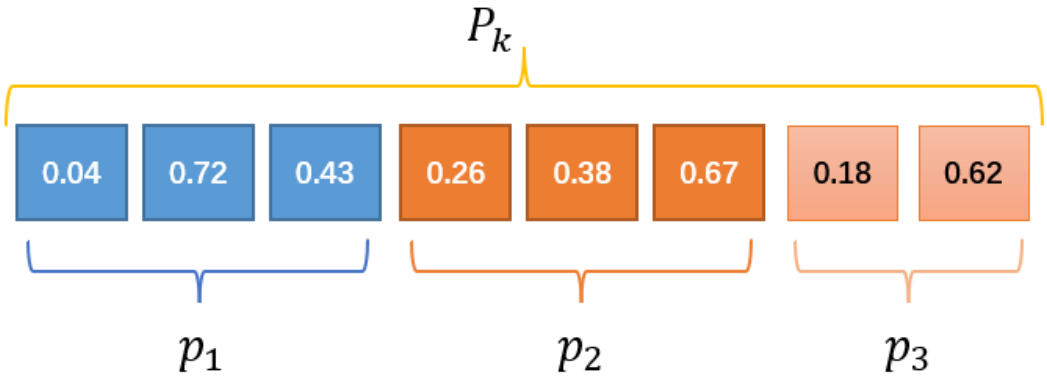

Figure 4. Particle encoding technique.

After calculating the pathway activity, we encode particles based on the pathway. As shown in Figure 4, particle P_k contains three pathways, $p_1 = \{g_1^1、 g_2^1、 g_3^1\}$, $p_2 = \{g_1^2、 g_2^2、 g_3^2\}$, $p_3 = \{g_1^3、 g_2^3\}$, g_i^j represents the ith gene on the jth pathway. The numerical value on each cell represents the degree of relevance of that gene in inferring pathway activity.

3.3. Objective functions

In this paper, weighted t-score and weighted z-score are selected as the objective function of the multi-objective optimization algorithm. In general, the t-score is used to measure the ability to differentiate the cumulative expression of the constituent genes of a given pathway [31]. The mathematical expression of the t-score is as follows.

$$t(p_i) = \frac{\mu(x) - \mu(y)}{\sqrt{\frac{\sigma(x)}{s_1} + \frac{\sigma(y)}{s_2}}} \tag{7}$$

where p_i indicates the pathway activity level of a given pathway, $\mu(x)$ and $\mu(y)$ represent the mean of pathway activity for classes C_1 and C_2 respectively, $\sigma(x)$ and $\sigma(y)$ represent the standard deviation of pathway activity for classes C_1 and C_2 respectively. s_1 and s_2 represent the number of samples in the two classes. In this paper, the concept of weighted t-score is introduced, which assumes that all genes in the pathway participate in the inference of pathway activity with a certain weight. The formula of weighted t-score is defined as Eq (8).

$$t_\omega(p_i) = \frac{\mu_\omega(x) - \mu_\omega(y)}{\sqrt{\frac{\sigma_\omega(x)}{s_1} + \frac{\sigma_\omega(y)}{s_2}}} \tag{8}$$

where $\mu_\omega(x) = \frac{\sum_{i=1}^{s_1}(x_i^k * \omega_i)}{\sum_{i=1}^{s_1}\omega_i}$, $\mu_\omega(y) = \frac{\sum_{i=1}^{s_2}(y_i^k * \omega_i)}{\sum_{i=1}^{s_2}\omega_i}$ represent the weighted mean of pathway activity for classes C_1 and C_2 respectively, and the weighted standard deviation of pathway activity for classes C_1 and C_2 are described as

$$\sigma_\omega(x) = \sqrt{\frac{s_1}{s_1 - 1} \frac{\sum_{i=1}^{s_1}(x_i^k - \mu x_k)^2 * \omega_i}{\sum_{i=1}^{s_1}\omega_i}},\ \sigma_\omega(x) = \sqrt{\frac{s_2}{s_2 - 1} \frac{\sum_{i=1}^{s_2}(x_i^k - \mu x_k)^2 * \omega_i}{\sum_{i=1}^{s_2}\omega_i}}$$

Suppose a particle consists of n pathways, then the weighted t-score of the particle should be expressed as Eq (9).

$$t_w(P_k) = \frac{\sum_{i=1}^{\mathrm{n}} t_\omega(p_i)}{\mathrm{n}} \tag{9}$$

The higher the weighted t-score is, the greater the differentiation ability is. Therefore, the weighted t-score of the particle should be maximized. The proposed algorithm is designed to be a minimization problem, so the first objective function can be expressed as Eq (10).

$$f_1 = \frac{1}{t_w(P_k)} \tag{10}$$

Weighted z-score is chosen as another objective function in this paper, the z-score is a measure of the distance between a data point and its overall mean, a positive z-score means that the data point is greater than the weighted mean, and a negative z-score means it is less than the weighted mean. The weighted z-score of a particle is described as Eq (11).

$$Z(P_k) = \frac{\alpha_w(P_k) - \mu_w(P_k)}{\sigma_w(P_k)} \tag{11}$$

in the formula, $\alpha_w(P_k)$, $\mu_w(P_k)$, $\sigma_w(P_k)$ represent the pathway activity, weighted mean and weighted standard deviation of particle P_k respectively, here $\alpha_w(P_k) = \frac{\sum_{i=1}^{n} \alpha(p_i)}{n}$. The smaller the absolute value of the z-score, the closer it is to the overall mean, so the second objective function can be expressed as Eq (12).

$$f_2 = |Z(P_k)| \tag{12}$$

3.4. The mechanism for archive updating

Since the capacity of the archive is limited, it is necessary to update the archive in real time during the feature selection process. This paper updates the archive by adaptively changing the penalty value of the PBI parameter. According to Section 2.3, we can know that when the value of d_2 corresponding to the current weight vector is small, increasing the penalty value helps to obtain a solution closer to the direction of the current weight vector. Therefore, this section proposes a method that change the penalty value of d_2 adaptively. When the d_2 value of the solution obtained under a specific weight vector in the archive increases, the penalty value decreases exponentially. Its specific mathematical expression is as Eq (13).

$$\theta_i(t) = 1 + (genx * e^{-Nd_i}) \tag{13}$$

where, d_i is the value of d_2 of the corresponding weight vector at the ith iteration, the outer part of the formula is set to 1 to maintain the original balance, the coefficient N of d_i is set to 30. Moreover, in order to avoid too small d value to make the convergence worse, an adaptive penalty value that varies with the number of iterations is introduced, the $genx$ value is added to Eq (13) to ensure that the penalty value varies within the normal range, $genx$ is expressed as Eq (14).

$$\text{gen x} = \theta_{\min} + (\theta_{\max} - \theta_{\min}) * \frac{gen}{Maxgen} \tag{14}$$

To satisfy the condition that the initial penalty value should be greater than 1, the value of $\theta_{\min}$ should be set to 1, and for the value of $\theta_{\max}$, 100 is a good choice.

The process of updating archive and penalty values is shown in Table 2. Firstly, the optimal solution is selected from the former leader archive under the current weight vector. Then, remove the optimal solution from the former leader archive and add it to the updated archive. At last, calculate the d_2 value corresponding to the current weight vector and update the penalty value.

Table 2. The pseudocode of archive updating.

Algorithm 1 Archive Updating

Input: leader archive (*LA*), size of leader archive (*LAS*)
Output: updated leader archive (LA')
penalty values for each weight vector (θ)
for i =1 to *LAS* **do**
$p \leftarrow$ Select the best particle from *LA* under θ_i by Eq (4);
$LA \leftarrow LA \setminus p$;
$LA' \leftarrow LA' \cup p$;
$d_i \leftarrow$ Calculate d_2 value of ith weight vector by Eq (5);
$\theta_i \leftarrow$ Update θ_i by Eq (13);
end for

3.5. Computational crowding distance

The set of non-dominated solutions distributed on the Pareto front is called the pareto optimal set, the optimal solutions can be stored in the archive. Here, the process of updating the archive and calculating the crowding distance is described. Assume that the maximum number of non-dominated solutions that the archive can accommodate at each iteration is N, suppose that by calculating the fitness in each iteration, M non-dominated solutions are obtained, if $M < N$, keep the optimal solution in the archive, else if $M > N$, calculate the crowding distance for all solutions and add solutions with higher crowding distances to the updated archive.

4. Experiments results

4.1. Datasets and preprocessing

Six real gene expression datasets were used in this paper, which are available from the public website (www.biolab.si/supp/bi-cancer/projections/info/).

Prostate: Gene expression measurements for samples of prostate tumors and adjacent prostate tissue not containing tumor were used to build this classification model. The number of genes in this dataset are 12,533 and the number of samples are 102, including 50 normal tissues and 52 prostate tumor samples.

DLBCL: Diffuse large B-cell lymphomas (DLBCL) and follicular lymphomas (FL) are two B-cell lineage malignancies that have very different clinical presentations, natural histories and response to therapy. The number of genes in this dataset are 7070 and the number of samples are 77, including Diffuse large B-cell lymphoma 58 examples and Follicular lymphoma 19 examples.

GSE412(Child-ALL): The childhood ALL data set (GSE412) includes gene expression information on 110 childhood acute lymphoblastic leukemia samples. The number of genes in this dataset are 8280 and the number of samples are 110, including Diffuse large B-cell lymphoma 50 examples and Follicular lymphoma 60 examples.

GSE2535 (Chronic Myeloid leukemia Treatment): Imatinib induces complete cytogenetic response in the majority of patients with chronic myeloid leukemia (CML) in chronic phase. The number of

genes in this dataset are 12,625 and the number of samples are 28, including non-responder to imatinib treatment 12 examples and responder to imatinib treatment 16 examples.

GSE2443 (Prostate Cancer Treatment): After an initial response to androgen ablation, most prostate tumors recur, ultimately progressing to highly aggressive androgen-independent cancer. The number of genes in this dataset are 12,627 and the number of samples are 20, including androgen dependent tumor 10 examples and androgen - independent tumor 10 examples.

GSE116959 (Lung adenocarcinoma): Based on GPL17077 platform（Agilent-039494 Sure Print G3 Human GE v28x60K Microarray 039381）, the number of samples are 68, including normal Lung tissue 11 examples and LUAD 57 examples.

The above original datasets can be obtained as matrix format, whose columns are genes and rows are samples. firstly, the signal-to-noise ratio for each gene(column) needs to be calculated by Eq (15).

$$|SNR| = \left| \frac{mean\ (\ class\ 1) - \ mean\ (\ class\ 2)}{S.D.\ (\ class\ 1) + \ S.D.\ (\ class\ 2)} \right| \tag{15}$$

Compute the mean and standard deviation (S.D.) of classes 1 and 2 and put them into signal-to-noise equation, and then according to the calculated signal-to-noise ratio, the genes (column) are sorted in descending order. A high SNR indicates that these genes are distributed over a wide range of values, some genes with low signal-to-noise ratio may be considered unimportant for class labels, so we take the top 1000 genes and used them for further experiments. In order to eliminate the adverse effects caused by singular sample data, normalize the filtered data using min-max normalization method.

4.2. Related parameter settings

In this section, the proposed algorithm is compared with some existing algorithms, to ensure the objectivity and fairness of the comparison results, the parameters of this experiment are consistent with the parameter settings in the existing algorithm. The specific experimental parameters are listed in Table 3.

Table 3. Parameters used in the algorithm.

Parameters	Values
Number of Iterations	100
Swarm size	25
Initialization of degree of correlation of each gene	Rand (0,1)
Weighting factor (c_1)	2
Weighting factor (c_2)	2
$r_1,\ r_2$	Rand (0,1)

4.3. Performance metrics

Performance metrics play an important role in the process of verifying and comparing algorithm performance. Six classic performance metrics are chosen in this paper, they are sensitivity, specificity, F-score, accuracy, G-mean and AUC, and the expressed as follows respectively.

$$\text{sensitivity} = \frac{tp}{tp + fn} \tag{16}$$

$$\text{specificity} = \frac{tn}{tn + fp} \tag{17}$$

$$\text{F-score} = \frac{2tp}{2tp + fn + fp} \tag{18}$$

$$\text{accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \tag{19}$$

$$Gmean = \sqrt{recall * specificity} = \sqrt{sensitivity * specificity} \tag{20}$$

tp, tn, fp, fn are represented true positive, true negative, false positive and false negative, respectively. In order to avoid the contingency of experimental results and improve the reliability, 10-fold cross-validation method was used in this experiment. Each dataset is divided into ten parts, of which nine parts are used for training and the remaining one is used for testing, and the results are demonstrated in Tables 4–9. In addition, the Pareto fronts of the proposed algorithm and existing algorithms on the experimental datasets are shown in Figure 5(a)–(f), respectively.

4.4. Comparative analysis

In order to verify the performance of the proposed algorithm, some existing algorithms are selected for comparison with it. For example, a method for identifying genes using protein interaction scores [10], the sequential forward search method [5], T-test [32], feature selection algorithm based on feature clearness [33], a correlation-based feature selection method [34], and a multi-objective particle swarm algorithm [3]. Next, we will analyze the comparison results.

Table 4. Performance comparison of different methods for Prostate dataset.

	Sensitivity	Specificity	F-score	Accuracy	G-mean	AUC
Proposed	**0.96275**	**0.9758**	**0.96239**	**0.96472**	**0.9693**	**0.9728**
MOPSO-PPI	0.95384	0.9617	0.95195	0.95738	0.9578	0.9526
MOPSO	0.93459	0.912	0.9265	0.9235	0.9232	0.9385
SFS	0.89998	0.864	0.88697	0.88234	0.8818	0.9169
T-set	0.9269	0.816	0.88132	0.87256	0.8697	0.9154
CFS	0.9131	0.9201	0.9211	0.9112	0.9166	0.9215
CBFS	0.8558	0.93	0.8971	0.8971	0.8921	0.9138

Table 5. Performance comparison of different methods for DLBCL dataset.

	Sensitivity	Specificity	F-score	Accuracy	G-mean	AUC
Proposed	**0.95376**	**0.98572**	**0.92592**	**0.97582**	**0.9696**	**0.9794**
MOPSO-PPI	0.94512	0.97369	0.90916	0.96465	0.9593	0.9713
MOPSO	0.92222	0.9379	0.87635	0.9332	0.93	0.9655
SFS	0.74445	0.9655	0.79647	0.9131	0.8478	0.8966
T-set	0.83335	0.9172	0.8035	0.89738	0.8743	0.9540
CFS	0.5556	0.9355	0.6667	0.8684	0.7209	0.9308
CBFS	0.1944	0.9555	0.2966	0.7829	0.431	0.9354

Table 6. Performance comparison of different methods for GSE412 (Child-All) dataset.

	Sensitivity	Specificity	F-score	Accuracy	G-mean	AUC
Proposed	**0.9236**	**0.98674**	**0.89742**	**0.90691**	**0.9546**	**0.9462**
MOPSO-PPI	0.91456	0.98680	0.85621	0.88423	0.951	0.9258
MOPSO	0.716	0.89667	0.77904	0.81453	0.8013	0.9040
SFS	0.68	0.9067	0.75478	0.80363	0.7852	0.9027
T-set	0.672	0.82668	0.71157	0.7563	0.7453	0.8840
CFS	0.6400	0.9133	0.7442	0.7990	0.7645	0.8827
CBFS	0.7100	0.6359	0.7427	0.7773	0.6719	0.8994

Table 7. Performance comparison of different methods for GSE2535 (Chronic Myeloid Leukemia Treatment) dataset.

	Sensitivity	Specificity	F-score	Accuracy	G-mean	AUC
Proposed	**1**	**0.9135**	**0.9216**	**0.89537**	**0.9558**	**0.9382**
MOPSO-PPI	**1**	0.89163	0.90762	0.87532	0.9443	0.9025
MOPSO	**1**	0.44447	0.8585	0.80357	0.6667	0.83334
SFS	0.84375	0.62499	0.7897	0.74998	0.7261	0.7083
T-set	0.71875	0.625	0.69905	0.6786	0.6702	0.8333
CFS	0.5900	0.8771	0.6967	0.7143	0.7194	0.8358
CBFS	0.6250	0.7343	0.7143	0.7143	0.6774	0.7708

Table 8. Performance comparison of different methods for GSE2443 (Prostate Cancer Treatment) dataset.

	Sensitivity	Specificity	F-score	Accuracy	G-mean	AUC
Proposed	**1**	**0.986**	**0.9903**	**0.9896**	**0.993**	**0.9463**
MOPSO-PPI	**1**	0.98	0.9867	0.9857	0.9899	0.9137
MOPSO	**1**	0.96	0.981818	0.98	0.9798	0.80
SFS	0.84	0.92	0.8723	0.88	0.8791	0.7891
T-set	0.92	0.88	0.9094	0.9	0.8998	0.7677
CFS	**1**	0.8010	0.9091	0.9021	0.895	0.7715
CBFS	0.80	**1**	0.8889	0.9	0.8944	0.860

Table 9. Performance comparison of different methods for GSE116959 dataset.

	Sensitivity	Specificity	F-score	Accuracy	G-mean	AUC
Proposed	**0.9681**	**0.9846**	**0.9862**	**0.9822**	**0.9763**	**0.9431**
MOPSO-PPI	0.9426	0.9638	0.9752	0.9683	0.9531	0.9347
MOPSO	0.9238	0.9582	0.9613	0.9584	0.9408	0.9218
SFS	0.8947	0.9326	0.9427	0.9482	0.9135	0.9152
T-set	0.9021	0.9288	0.9296	0.9353	0.9154	0.9238
CFS	0.8572	0.8914	0.9192	0.9136	0.8741	0.9175
CBFS	0.8324	0.8761	0.8783	0.893	0.8540	0.8926

Table 4 shows the performance comparison of different methods for Prostate dataset, the sensitivity, specificity, F-score, accuracy, G-mean and AUC of the proposed IMOPSO-PBI algorithm are 0.96275, 0.9758, 0.96239, 0.96472, 0.9693, 0.9728 respectively, it is obvious that the performance of the proposed IMOPSO-PBI algorithm is better than that of the comparison algorithm. And as Figure 5(a) shows, when f_2 is the same, the non-dominant solution obtained by the proposed IMOPSO-PBI algorithm has a smaller f_1. Table 5 shows the performance comparison of different methods for DLBCL dataset, the sensitivity, specificity, F-score, accuracy, Gmean and AUC of the proposed IMOPSO-PBI algorithm are 0.95376, 0.98572, 0.92592, 0.97582 ,0.9696, 0.9794 respectively, better than that of the comparison algorithm. And as Figure 5(b) shows, no matter for f_1 or f_2, the proposed IMOPSO-PBI algorithm has better performance. Table 6 shows the performance comparison of different methods for GSE412 (Child-All) dataset, the sensitivity, specificity, F-score, accuracy, Gmean and AUC of the proposed algorithm are 0.9236, 0.98674, 0.89742, 0.90691, 0.9546, 0.9462 respectively, the specificity of proposed IMOPSO-PBI method is slightly lower than the MOPSO-PPI algorithm, but the difference can be ignored. And as Figure 5(c) shows, the proposed IMOPSO-PBI method can get a better Pareto front.

For GSE2535 (Chronic Myeloid Leukemia Treatment) dataset, the performance comparison results are shown in Table 7, the sensitivity of the proposed IMOPSO-PBI method, MOPSO-PPI and the MOPSO algorithm are the same value as 1, the proposed IMOPSO-PBI algorithm has better performance on other metrics, and it can be seen from Figure 5(d) that most of the solutions on the Pareto front of the proposed IMOPSO-PBI method outperform other algorithms. For GSE2443 (Prostate Cancer Treatment) dataset, Table 8 describes the outcome of the proposed IMOPSO-PBI method and other techniques, the specificity of CBFS higher than the proposed IMOPSO-PBI method, apart from this, the proposed IMOPSO-PBI method has a better performance on other metrics, the proposed IMOPSO-PBI algorithm has a better Pareto front, which can be seen from the Figure 5(e). For GSE116959 (Lung adenocarcinoma) dataset, although the Pareto front of MOPSO-PPI locally is better than the proposed IMOPSO-PBI as shown in Figure 5(f), our proposed algorithm is better on the whole. From Table 9, we can know that the proposed IMOPSO-PBI algorithm has better performance on the metrics. In general, the proposed IMOPSO-PBI algorithm has better performance than the comparison algorithms on all experimental data, which proves that the IMOPSO-PBI algorithm in this paper is useful and efficient.

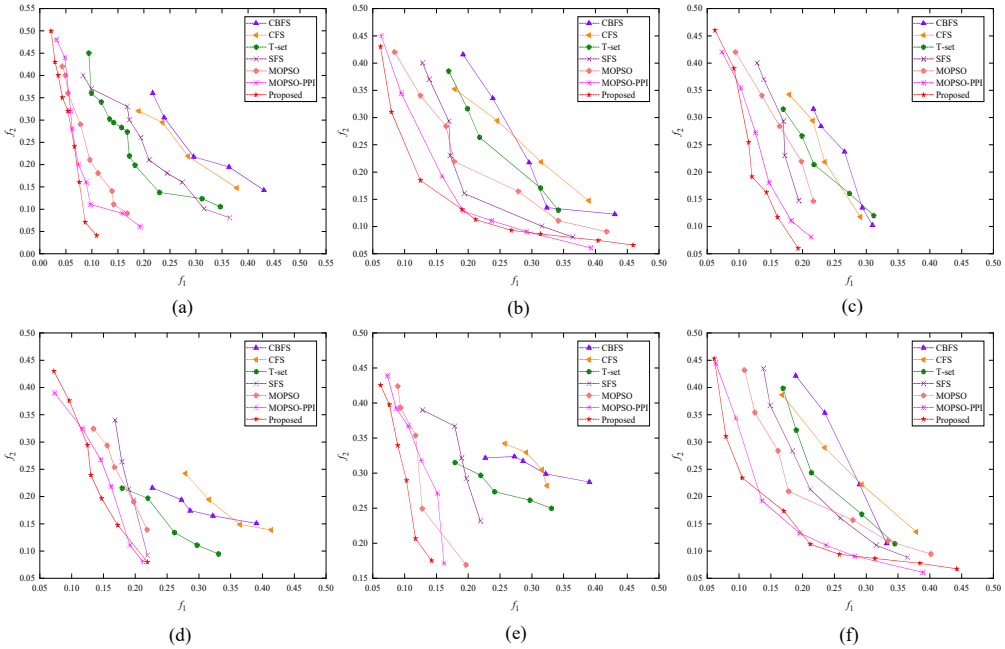

Figure 5. The Pareto fronts of contrast algorithm on the experiment datasets.

4.5. Biological relevance

Table 10. Disease associated with the resultant pathway markers.

Disease	Gene Symbol (#PMID)
Prostatic neoplasms Prostate carcinoma	GSTP1 (25), KLK3 (82), IGF1 (20), GSTP1 (25), ERG (54), TGFBR2 (5) BCL2 (120), KLK3 (1857), IGF1R (42), BCL2 (120), ESR2 (66), MSMB (68), HOXB13 (54), PCAT1 (43), FHIT (7), MKI67 (6)
Malignant neoplasm of prostate	ERG (441), GSTP1 (134), IGF1 (111), HNF1B (22), PIK3CA (127), EPHB2 (39), TP53 (310), CHEK2 (27)
Anemia	HAMP (147), GATA1 (12), TNF (29), CSF3 (11), IFNA2 (7)
Carcinoma	ESR1 (70), PTGS2 (57), ABCB1 (29), HIF1A (17), CDK4 (15), BRCA1 (81), PTEN (60)
Lymphoma	BCL2 (262), PTEN (13), CDK4 (9), TP53 (156), KRAS (5), ATM (22) RHOA (4), EPHX1 (3)
Leukemia	KMT2A (385), NRAS (19), JAK2 (38), STAT3 (26), CSF3R (20), GSTM1 (13), BRAF (12), ATM (17)
Acute lymphocytic leukemia	KMT2A (202), IKZF1 (81), ABL1 (264), PBX1 (67), FLT3 (48), PETN (7), LMO2 (5)
Lung adenocarcinoma	ITGB4 (182), CDC20 (115), MMP9 (96), CALM1 (73), RRAS (62), ID1 (56)

In order to verify the biological relevance of the selected pathway marker genes, the top 60% pathway markers are chosen and searched in the disease-gene association database (http://www.disgenet.org/). From this database we can obtain the number of PubMed citations of the disease-gene association, and a part is shown in Table 10. The first column of the Table 10 is the disease name and the second column is the corresponding gene symbol, in addition, the numbers in parentheses indicate the number of PubMed citations. From the results, we can know that the marker genes are related to particular diseases, that is to say, the selected marker genes are biologically related.

To explore the potential role of marker genes on overall survival rate, the Kaplan-Meier (KM) survival curve was presented based on the partial marker genes in this study. In Figure 6(a)–(f), the horizontal axis represents the patient's survival time and the vertical axis represents the patient's survival rate. The red and blue lines represent the high and low expression groups, respectively. Although there was no significant difference in survival rate between the two groups during the initial period, with the increase of time, the survival rate of the high expression group decreased rapidly. The low expression group had higher survival rate and longer survival time, which demonstrates that the cancer is related to some specific genes. Through identifying the biomarkers of cancers can help prediction and diagnosis cancers, which can result in a good prognosis.

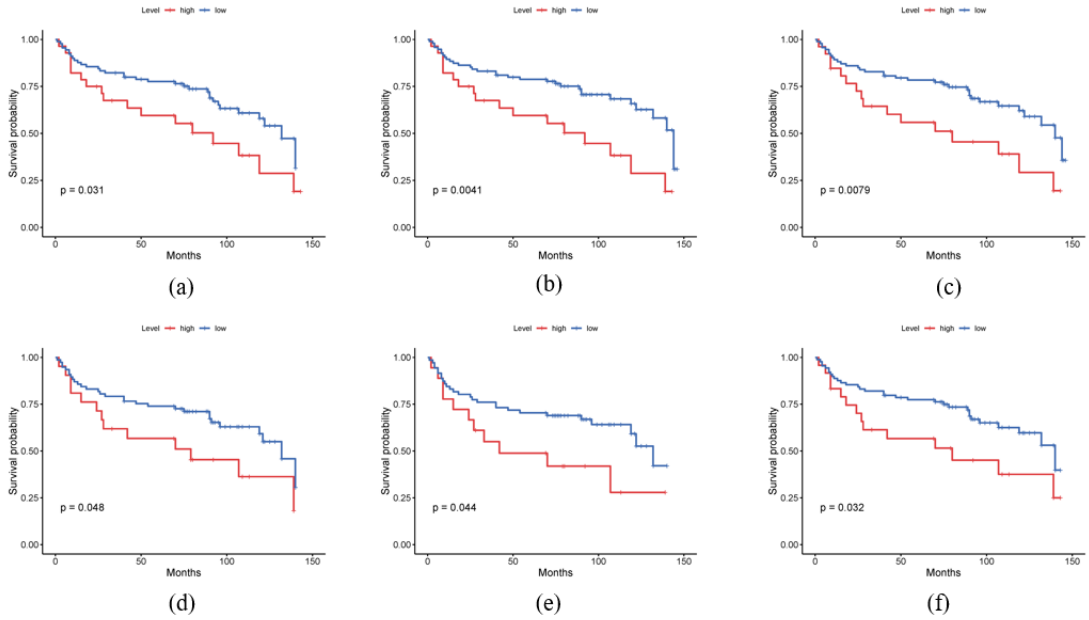

Figure 6. Survival curves of cancer patients in different datasets.

5. Conclusions

Biomarkers play an important role in the diagnosis and treatment of diseases, in this paper, a decomposition-based multi-objective optimization algorithm is proposed to screen pathway marker genes. Consider in the process of inferring pathway activity, the contribution of each gene should be different. In this research, the relevance of each gene in pathway activity inference is quantified. Moreover, in order to avoid the excessive concentration of feature solutions in most multi-objective optimization algorithms, an adaptive mechanism based on PBI decomposition is introduced. The

proposed IMOPSO-PBI method is applied to six real datasets and compared with some existing algorithms, the results show that the method achieves better performance. In addition, the biological relevance of the screened marker genes was proved by biological analysis. In the future, we can optimize on this framework, such as choosing different objective functions and performance metrics, and we can also use this method on other datasets, such as HNSCC, KIRC, besides, some public available single-cell RNA-seq along with the bulk RNA-seq are also good choice.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61803257) and Natural Science Foundation of Shanghai (No. 18ZR1417200).

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. M. Mandal, A. Mukhopadhyay, A graph-theoretic approach for identifying non-redundant and relevant gene markers from microarray data using multiobjective binary PSO, *PLoS One*, **9** (2014), 13. https://doi.org/ 10.1371/journal.pone.0090949
2. S. Bandyopadhyay, S. Mallik, A. Mukhopadhyay, A survey and comparative study of statistical tests for identifying differential expression from microarray data, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **11** (2014), 95–115. https://doi.org/ 10.1109/TCBB.2013.147
3. A. Mukhopadhyay, M. Mandal, Identifying non-redundant gene markers from microarray Data: A multiobjective variable length PSO-based approach, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **11** (2014), 1170–1183. https://doi.org/10.1109/TCBB.2014.2323065
4. Y. Saeys, I. Inza, P. Larraaga, A review of feature selection techniques in bioinformatics, *Bioinformatics*, **23** (2007), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344
5. S. Bouatmane, M. A. Roula, A. Bouridane, S. Al-Maadeed, Round-Robin sequential forward selection algorithm for prostate cancer classification and diagnosis using multispectral imagery, *Mach. Vision Appl*., **22** (2011), 865–878. https://doi.org/10.1007/s00138-010-0292-x
6. S. Ma, M. R. Kosorok, Identification of differential gene pathways with principal componentanalysis, *Bioinformatics*, **25** (2009), 882–889. https://doi.org/10.1093/bioinformatics/btp085
7. J. J. Su, B. J. Yoon, E. R. Dougherty, Accurate and reliable cancer classification based on probabilistic inference of pathway Activity, *PLoS One*, **4** (2009), 10. https://doi.org/10.1371/journal.pone.0008161
8. N. M. Borisov, N. V. Terekhanova, A. M. Aliper, L. S. Venkova, P. Y. Smirnov, S. Roumiantsev, et al., Signaling pathways activation profiles make better markers of cancer than expression of individual genes, *Oncotarget*, **5** (2014), 10198–10205. https://doi.org/10.18632/oncotarget.2548

9. M. Mandal, J. Mondal, A. Mukhopadhyay, A PSO-based approach for pathway marker identification from gene expression data, *IEEE Trans. Nanobiosci.*, **14** (2015), 591–597. https://doi.org/10.1109/TNB.2015.2425471
10. P. Dutta, S. Saha, S. Naskar, A multi-objective based PSO approach for inferring pathway activity utilizing protein interactions, *Multimed. Tools Appl.*, **80** (2021), 30283–30303. https://doi.org/10.1007/s11042-020-09269-8
11. A. Trivedi, D. Srinivasan, K. Sanyal, A. Ghosh, A survey of multiobjective evolutionary algorithms based on decomposition, *IEEE Trans. Evol. Comput.*, **21** (2017), 440–462. https://doi.org/10.1109/tevc.2016.2608507
12. S. X. Yang, M. Q. Li, X. H. Liu, J. H. Zheng, A grid-based evolutionary algorithm for many-objective optimization, *IEEE Trans. Evol. Comput.*, **17** (2013), 721–736. https://doi.org/10.1109/tevc.2012.2227145
13. C. H. Liang, C. Y. Chung, K. P. Wong, X. Z. Duan, Parallel optimal reactive power flow based on cooperative co-evolutionary differential evolution and power system decomposition, *IEEE Trans. Power Syst.*, **22** (2007), 249–257. https://doi.org/10.1109/tpwrs.2006.887889
14. Z. H. Zhan, J. J. Li, J. N. Cao, J. Zhang, H. S. H. Chung, Y. H. Shi, Multiple populations for multiple objectives: A coevolutionary technique for solving multiobjective optimization problems, *IEEE Trans. Cybern.*, **43** (2013), 445–463. https://doi.org/10.1109/tsmcb.2012.2209115
15. Y. C. Yang, T. X. Zhang, W. Yi, L. J. Kong, X. L. Li, B. Wang, et al., Deployment of multistatic radar system using multi-objective particle swarm optimisation, *IET Radar Sonar Navig.*, **12** (2018), 485–493. https://doi.org/10.1049/iet-rsn.2017.0351
16. J. F. Qiao, H. B. Zhou, C. L. Yang, S. X. Yang, A decomposition-based multiobjective evolutionary algorithm with angle-based adaptive penalty, *Appl. Soft. Comput.*, **74** (2019), 190–205. https://doi.org/10.1016/j.asoc.2018.10.028
17. Y. Xue, B. Xue, M. J. Zhang, Self-adaptive particle swarm optimization for large-scale feature selection in classification, *ACM Trans. Knowl. Discov. Data*, **13** (2019), 27. https://doi.org/10.1145/3340848
18. X. M. He, S. H. Dong, N. Zhao, Research on rush order insertion rescheduling problem under hybrid flow shop based on NSGA-III, *Int. J. Prod. Res.*, **58** (2020), 1161–1177. https://doi.org/10.1080/00207543.2019.1613581
19. Z. Zheng, J. Y. Long, X. Q. Gao, Production scheduling problems of steelmaking-continuous casting process in dynamic production environment, *J. Iron Steel Res. Int.*, **24** (2017), 586–594. https://doi.org/10.1016/s1006-706x(17)30089-4
20. C. Dai, Y. P. Wang, M. Ye, A new multi-objective particle swarm optimization algorithm based on decomposition, *Inf. Sci.*, **325** (2015), 541–557. https://doi.org/10.1016/j.ins.2015.07.018
21. R. Akbari, R. Hedayatzadeh, K. Ziarati, B. Hassanizadeh, A multi-objective artificial bee colony algorithm, *Swarm Evol. Comput.*, **2** (2012), 39–52. https://doi.org/10.1016/j.swevo.2011.08.001
22. N. S. Sani, M. Manthouri, F. Farivar, A multi-objective ant colony optimization algorithm for community detection in complex networks, *J. Ambient Intell. Humaniz. Comput.*, **11** (2020), 5–21. https://doi.org/10.1007/s12652-018-1159-7
23. J. Xu, Z. L. Nie, J. J. Zhu, Characterization and selection of probability statistical parameters in random slope pwm based on uniform distribution, *IEEE Trans. Power Electron.*, **36** (2021), 1184–1192. https://doi.org/10.1109/tpel.2020.3004725

24. Q. Z. Lin, Y. P. Ma, J. Y. Chen, Q. L. Zhu, C. A. C. Coello, K. C. Wong, et al., An adaptive immune-inspired multi-objective algorithm with multiple differential evolution strategies, *Inf. Sci.*, **430** (2018), 46–64. https://doi.org/10.1016/j.ins.2017.11.030
25. A. R. Jordehi, Enhanced leader PSO (ELPSO): A new PSO variant for solving global optimisation problems, *Appl. Soft. Comput*., **26** (2015), 401–417. https://doi.org/10.1016/j.asoc.2014.10.026
26. S. Y. Jiang, S. X. Yang, Y. Wang, X. B. Liu, Scalarizing functions in decomposition-based multiobjective evolutionary algorithms, *IEEE Trans. Evol. Comput*., **22** (2018), 296–313. https://doi.org/10.1109/tevc.2017.2707980
27. Q. F. Zhang, H. Li, MOEA/D: A multiobjective evolutionary algorithm based on decomposition, *IEEE Trans. Evol. Comput.*, **11** (2007), 712–731. https://doi.org/10.1109/tevc.2007.892759
28. S. X. Yang, S. Y. Jiang, Y. Jiang, Improving the multiobjective evolutionary algorithm based on decomposition with new penalty schemes, *Soft Comput.*, **21** (2017), 4677–4691. https://doi.org/10.1007/s00500-016-2076-3
29. Y. R. Zhou, Y. Xiang, Z. F. Chen, J. He, J. H. Wang, A scalar projection and angle-based evolutionary algorithm for many-objective optimization problems, *IEEE Trans. Cybern.*, **49** (2019), 2073–2084. https://doi.org/10.1109/tcyb.2018.2819360
30. D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat. Protoc.*, **4** (2009), 44–57. https://doi.org/10.1038/nprot.2008.211
31. K. Wang, M. Y. Li, M. Bucan, Pathway-based approaches for analysis of genomewide association studies, *Am. J. Hum. Genet.*, **81** (2007), 1278–1283. https://doi.org/10.1086/522374
32. P. Baldi, A. D. Long, A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes, *Bioinformatics*, **17** (2001), 509–519. https://doi.org/10.1093/bioinformatics/17.6.509
33. M. Seo, S. Oh, CBFS: High performance feature selection algorithm based on feature clearness, *PLoS One*, **7** (2012), 10. https://doi.org/10.1371/journal.pone.0040419
34. J. Cai, J. W. Luo, S. L. Wang, S. Yang, Feature selection in machine learning: A new perspective, *Neurocomputing*, **300** (2018), 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

AIMS Press