A Novel Evolutionary Multitasking Feature Selection Approach for Genomic Data Classification

Yifan Yu 2200955@stu.neu.edu.cn
Dazhi Wang✉ wangdazhi1@mail.neu.edu.cn
Yanhua Chen 2000718@stu.neu.edu.cn
Hongfeng Wang hfwang@mail.neu.edu.cn
Min Huang mhuang@mail.neu.edu.cn
College of Information Science and Engineering, Northeastern University, Shenyang, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Microarray-generated genomic data has recently sparked a wave of bioinformatics and data mining research. However, such data presents significant challenges for further analysis due to its high dimensionality and small sample sizes. Feature selection is a standard approach to address this issue, as it can enhance classification performance while reducing dimensionality. This paper introduces an Improved Gray Wolf Optimization-based Evolutionary Multitasking (EMT-IGWO) feature selection approach tailored for high-dimensional classification. It adopts multi-population co-evolving searching modes that can be regarded as a typical feature selection task via a specific information-sharing mechanism. Within the proposed multitasking framework, both population diversity and global searching capabilities of EMT-IGWO are improved. Moreover, several enhancements are incorporated into the two searching modes to help stagnant individuals escape from local optima with higher probabilities. Computational results show that EMT-IGWO outperforms other compared algorithms in effectiveness and efficiency evaluated across eight public gene expression datasets.

Keywords: evolutionary multitasking, gray wolf optimization, feature selection, genomic data classification

1. Introduction

Gene expression data in bioinformatics can be obtained by microarray technology, which is used to explore the pathogenesis of certain diseases and diagnosis (Alhenawi et al., 2022).The characteristics of this type data are features (genes) with high dimensions and small sample sizes. However, some studies have indicated that only a few genes are relevant to accurately classifying classification of different classes of the problem (Golub et al., 1999). Additionally, given the limitations of microarray technology and experimental errors, there is a large amount of noise and redundant features. Therefore, selecting an optimal gene subset to achieve satisfactory classification performance is a challenging and popular topic in the domains of machine learning and data mining.

Feature selection also called attribute selection (Gandhi and Prabhune, 2017), is a pivotal data preprocessing approach. It plays an important role in analyzing the aforementioned gene expression data, which can select informative genes contributing to subsequent classification prediction (Bolón-Canedo et al., 2014). For a dataset with m features, there

are 2^m possible feature subsets to choose from, which is also categorized as an NP-Hard problem (Kılıç et al., 2021). Generally, popular feature selection methods can be classified as filter methods (Gao et al., 2016; Manikandan and Abirami, 2021; Urbanowicz et al., 2018), wrapper methods (Altarabichi et al., 2021; Li et al., 2021; Niu et al., 2018), and embedded methods (Xu and Wu, 2020; Zhang et al., 2019). Moreover, researchers have made other attempts to improve the classification performance, such as employing a meta-heuristic feature selection method to search for optimal or near-optimal feature subsets (Dokeroglu et al., 2022).

Gray Wolf Optimization (GWO), a meta-heuristic algorithm, is characterized by the division of the social hierarchy of gray wolves and the hunting mechanism (Mirjalili et al., 2014; Setiawan et al., 2021). The algorithm has fewer parameters and is more flexible, widely used in feature selection problems. However, most of the GWO-based feature selection approaches are applicable to low-dimensional classification. As the dimension increases, such algorithms are easily getting stuck in local optimum and leader wolf remaining stagnant. Considering the limitations mentioned, exploring multi-task machine learning (ML) to enhance learning is a promising direction (Wang et al., 2024). In this paper, we introduce an evolutionary multitasking (EMT) framework (Gupta et al., 2015) as an emerging paradigm to address multiple optimization tasks simultaneously.

Building on this paradigm, we propose an evolutionary multitasking feature selection approach based on an improved Gray Wolf Optimizer (GWO), termed EMT-IGWO. This method significantly enhances classification performance on high-dimensional datasets and reduces the running time. Unlike traditional EMT, which involves several problems (Gupta et al., 2015), in this study we introduce two feature selection tasks relevant to distinct subpopulations and adopt two different search modes, each performing an independent search direction. In this multi-task system, the population's diverse search modes ensure a variety of individuals. Additionally, knowledge transfer allows the tasks to guide each other's search processes effectively. Moreover, we enhance our algorithm by adjusting the convergence factor and employing a rank-based mutation approach to increase the possibility of escaping local optima.

In summary, the contributions of this paper are outlined below:

- This paper proposes an evolutionary multitasking feature selection paradigm based on gray wolf optimization for genomic data classification.
- Multi-population co-evolving searching modes, i.e., an adaptive strategy for the update process and dominance accumulation mechanism, implement knowledge transfer between the two tasks, facilitating the diversity of the population and improving the performance and robustness of the objective feature subset.
- The nonlinear multi-convergence factor and the rank-based mutation operation further enhance the distinct searching modes, which maintain the search capability as well as address the issue of being trapped in local optima.
- Computational experiments validate the effectiveness and efficiency of the EMT-IGWO on eight gene expression datasets compared with other state-of-the-art algorithms.

2. Related Work

2.1. Maximum Information Coefficient

Maximum information coefficient (MIC), an indicator for evaluating the dependence of pairs of variables (Reshef et al., 2011), is used here to capture the correlation between features and labels. The calculation of the MIC value is based on mutual information (MI). Assuming a binary input $D = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and $a - by - b$ grids, $MIC(D)$ can be obtained by Eqs. (1) and (2) (Wen et al., 2019):

$$MIC(D) = \max_{ab<B(n)} \{M(D)_{a,b}\} \tag{1}$$

$$M(D)_{a,b} = \frac{MI^*(D, a, b)}{\log(\min\{a, b\})} \tag{2}$$

where $MI^*(D, a, b)$ is the maximal available MI in $a-by-b$ grids with the limit $ab < B(n)$. Specifically, $B(n)$ is set to $n^{0.6}$ for an input with n samples. The MIC value is symmetric and normalized into a range of $[0, 1]$. Given its properties of generalization and fairness, we employ MIC to rank the features for the subsequent search.

2.2. Gray Wolf Optimization

2.2.1. Standard GWO

Gray Wolf Optimizer was firstly introduced by Mirjalili et al. (2014) to seek the global optimum. The hierarchical mechanism employed by GWO involves the searching and hunting of prey, with each wolf serving as a feasible solution for potentially capturing the prey. The top three optimal solutions within the population are named α, β, and δ, while the lowest-ranked candidates are referred to as ω. The ω wolves update their positions based on the positions of α, β, and δ leader wolves, which can be calculated from Eq. (3):

$$\overrightarrow{X}(t+1) = \left(\overrightarrow{X}_1 + \overrightarrow{X}_2 + \overrightarrow{X}_3\right)/3 \tag{3}$$

$$\overrightarrow{X}_1 = \overrightarrow{X}_\alpha - \overrightarrow{A} \cdot \overrightarrow{D}_\alpha, \overrightarrow{X}_2 = \overrightarrow{X}_\beta - \overrightarrow{A} \cdot \overrightarrow{D}_\beta, \overrightarrow{X}_3 = \overrightarrow{X}_\delta - \overrightarrow{A} \cdot \overrightarrow{D}_\delta \tag{4}$$

$$\overrightarrow{D}_\alpha = \left|\overrightarrow{C} \cdot \overrightarrow{X}_\alpha - \overrightarrow{X}(t)\right|, \overrightarrow{D}_\beta = \left|\overrightarrow{C} \cdot \overrightarrow{X}_\beta - \overrightarrow{X}(t)\right|, \overrightarrow{D}_\delta = \left|\overrightarrow{C} \cdot \overrightarrow{X}_\delta - \overrightarrow{X}(t)\right| \tag{5}$$

$$\overrightarrow{A} = 2a \cdot \overrightarrow{r}_1 - a \tag{6}$$

$$\overrightarrow{C} = 2\overrightarrow{r}_2 \tag{7}$$

$$a = 2 \times (1 - t/T) \tag{8}$$

where a is the convergence factor; $\overrightarrow{A}$ and $\overrightarrow{C}$ are two coefficient vectors involving random vectors between $[0, 1]$; $\overrightarrow{D}_\alpha$, $\overrightarrow{D}_\beta$ and $\overrightarrow{D}_\delta$ respectively represent the distance vectors from ω wolves to α, β, and δ leader wolves; $\overrightarrow{X}(t)$, $\overrightarrow{X}_\alpha$, $\overrightarrow{X}_\beta$, and $\overrightarrow{X}_\delta$ mean the current position vectors of ω, α, β, and δ wolves.

2.2.2. Binary GWO for feature selection

In feature selection problems, the position of each wolf consists of the value 0 or 1. Therefore, we need a transfer function to obtain X_i^d in discrete space (Pan et al., 2023).

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-10(x-0.5)}} \tag{9}$$

$$X_i^d(t) = \begin{cases} 1, \text{ if sigmoid } \left(X_i^d(t)\right) > \text{ rand} \\ 0, \text{ else} \end{cases} \tag{10}$$

Eq. (9) is the sigmoid function and $X_i^d(t)$ represents the position of wolf i in the dth dimension at iteration t.

Over the past few years, a large number of GWO-based feature selection methods have been designed for the classification problems. Transfer function is an important part of BGWO, Hu et al. (2020) tested various transfer functions and gave an updating equation for a parameter, and the experimental results validated the effectiveness of the improved BGWO algorithm. A two-phase mutation was integrated with BGWO algorithm (Abdel-Basset et al., 2020) to reduce the number of the selected features without degrading the classification performance. The experiments on 35 datasets showed its outperformance. Previous works have made significant advancements in feature selection problems. However, we have noticed that few GWO-based studies focus on high-dimensional classification tasks.

2.3. Evolutionary Multitasking

Evolutionary multitasking (EMT) addresses multiple related learning tasks simultaneously via evolutionary computation (EC) (Lin et al., 2023). In a multitasking scenario, processing one task may contribute to other search tasks because of the knowledge transferred. We can give a description of the aforementioned EMT: This paradigm learns n related tasks, denoted as $\{\mathrm{T}_i\}_{i=1}^n$, simultaneously, and performance can be enhanced by the association information between $\{\mathrm{T}_i\}_{i=1}^n$.

While EMT has been utilized in a variety of domains, applications in feature selection have still remained relatively limited. Wang et al. (2024) proposed a novel PSO-based multitask framework to achieve the information shared, which divided the initial population into two subpopulations. Extensive experiments showed the strong competitiveness of the approach compared with other algorithms. Chen et al. (2020) developed an EMT feature selection method for high-dimensional classification by ranking the importance of features and establishing two tasks according to the ranks, and two mechanisms were designed to further improve the algorithm. The computational results exhibited the effectiveness of the PSO-EMT algorithm. We note that both of the works above were performed based on the PSO algorithm. Meanwhile, in Chen et al. (2020), a knee point scheme was used to delineate whether a feature is important or not. However, the informative features may be lost via an inappropriate threshold.

3. Methodology

This section begins with an overview and fitness function of the EMT-IGWO algorithm. It then delves into the analysis of population initialization, followed by a detailed description

of pivotal procedures involved in EMT. Finally, enhancements have been made to both search modes.

3.1. Overview

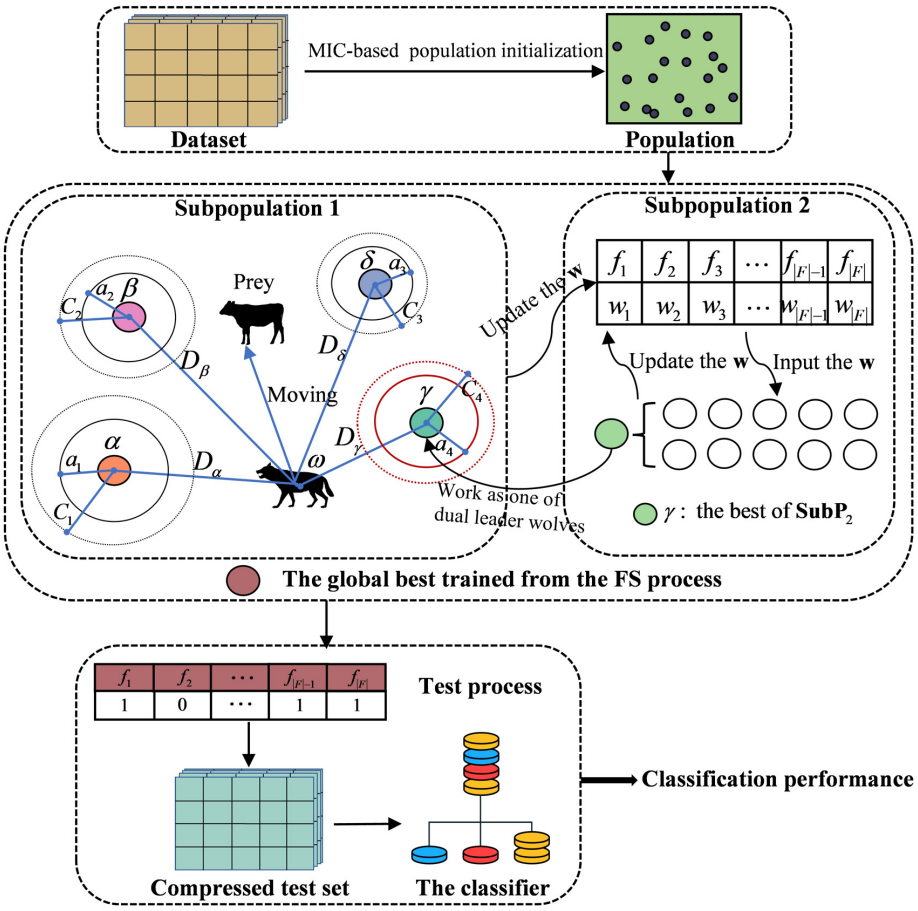

Figure 1: The architecture of the EMT-IGWO algorithm

The architecture of our proposed approach is depicted in Fig. 1. It mainly consists of two tasks: (i) Task 1 conducts one search process based on a modified GWO algorithm, which learns the knowledge from Task 2. (ii) Task 2 performs the other search via a dominance accumulation mechanism, and the mechanism utilizes the collective knowledge both from Task 1 and Task 2. Through an overall analysis, the time complexity of the evolutionary multitasking algorithm in this paper is $O(T * N * S * D)$ (T is the maximum iterations, N is the number of wolves, S is the number of samples, and D is the number of the features).

3.2. Fitness Function

A fitness function is used to evaluate the performance of the candidate feature subset. The fitness function of this paper involves two aspects: (i) classification error rate. (ii) dimension

of the optimal feature subset. Here, our target is to achieve small feature subsets without degrading classification accuracy. Therefore, the mathematical formula can be denoted by Eq. (11):

$$\text{fitness } = \theta \cdot error + (1-\theta) \cdot \frac{|S|}{|F|} \tag{11}$$

$$error = 1 - \frac{1}{c} \cdot \sum_{i=1}^{c} TPR_i \tag{12}$$

here, θ is a control parameter, a range from $[0, 1]$, which balances the error rate and the number of the selected feature subset. Since the classification performance is preferred to the feature subset size, we set $\theta = 0.99$ in this study (Pan et al., 2023); $error$ indicates the error rate of the learning algorithm; $|S|$ represents the dimension of the selected feature subset and $|F|$ is the number of features in the initial dataset. Specially, a balanced accuracy (Patterson and Zhang, 2007) is employed to handle the unbalanced data, which is given in Eq. (12). c denotes the number of classes for the classification problem, and TPR_i represents the proportion of correctly identified instances in class i.

3.3. Population Initialization

A MIC-based approach (Qu et al., 2023) is used to generate the initial population, which introduces a preference for features with higher MIC values. The detailed steps are as follows: we first calculate the MIC values between features and labels. Then the chance that one feature will be selected is obtained from Eqs. (13) and (14):

$$p(d) = \frac{MIC^d}{\sum_{d=1}^{|F|} MIC^d} \tag{13}$$

$$cp = \sum_{j=1}^{d} p(j) \tag{14}$$

where $p(d)$ is the selection probability of the dth feature. We then use a roulette wheel to select the dth feature according to the cumulative probabilities, i.e., cp. One individual can be achieved by repeating the above procedure $|F|$ times.

3.4. Knowledge Transfer

In this study, we adopt two task subpopulations to search the feature space, dubbed $\mathbf{SubP}_1$ and $\mathbf{SubP}_2$. Each seeks to find an optimal feature subset independently with its own search mode. Fig. 1 depicts that the best solution in $\mathbf{SubP}_2$ serves as one of dual leader wolves γ in $\mathbf{SubP}_1$. Both updated $\mathbf{SubP}_1$ and $\mathbf{SubP}_2$ will provide the weights of features, which helps $\mathbf{SubP}_2$ to conduct its dominance accumulation. Algorithm 1 presents the pseudocode of the proposed EMT-IGWO algorithm.

Algorithm 1 EMT-IGWO for feature selection

Input: Dataset, $\boldsymbol{D}$; the number of iterations, T; population size, $2N$
Output: the optimal feature subset, *best*
$\boldsymbol{MIC} \leftarrow$ **Calculate** the MIC value for each feature on the dataset $\boldsymbol{D}$
Initialize a population with $2N$ individuals based on the MIC value of each feature
Divide the population into two subpopulations with N individuals
Set the weight vector of features to **0**
Record the global best and the best individual of $\mathbf{SubP}_2$
while $t <= T$ **do**
    Update the $\mathbf{SubP}_1$ through an adaptive strategy based on the knowledge from $\mathbf{SubP}_2$; **Upgrade** the weight vector, the leader wolves and the global best
    Update the $\mathbf{SubP}_2$ via a dominance accumulation mechanism utilizing the collective knowledge; **Upgrade** the weight vector, the best individual of $\mathbf{SubP}_2$ and the global best
    $t \leftarrow t + 1$
end while
return *best*

3.4.1. Case 1: Knowledge Transfer to $\mathbf{SubP}_1$

Standard GWO algorithm involves leader wolves of three levels. Each wolf will move a certain distance towards the three top wolves. In our study, the fourth top wolf from $\mathbf{SubP}_2$ is added to $\mathbf{SubP}_1$, which promotes the diversity of the population and avoids the stagnation of leader wolves. We propose a novel adaptive strategy for the update process to enhance the balance between the global search and local exploitation:

$$\overrightarrow{X}(t+1) = \frac{\left(\overrightarrow{X}_1 + \overrightarrow{X}_2 + \overrightarrow{X}_3 + \overrightarrow{X}_4\right)}{4} \cdot \left(1 - \frac{t}{T}\right) + \overrightarrow{X}_1 \cdot \frac{t}{T} \tag{15}$$

We then map the search space into discrete space, which is calculated as follows:

$$X_i^d(t+1) = \begin{cases} X_1^d, & \text{if } rand < (3t+T)/4T \\ X_2^d, & \text{elif } (3t+T)/4T =< rand \ \ \& \ \ rand < (2t+2T)/4T \\ X_3^d, & \text{elif } (2t+2T)/4T =< rand \ \ \& \ \ rand < (3T+t)/4T \\ X_4^d, & \text{else} \end{cases} \tag{16}$$

where X_i^d $(i = 1, 2, 3)$ has been introduced in Eqs. (4) and (10). The best solution in $\mathbf{SubP}_2$ works as one of dual leader wolves γ to obtain X_4^d. t indicates the current iteration, and T represents the maximum iteration.

3.4.2. Case 2: Knowledge Transfer to $\mathbf{SubP}_2$

Attention has paid to the influence of one feature on the entire feature subset. Therefore, a dominance accumulation mechanism is proposed in this paper, and a weight vector of features is employed to reflect the dominance of each feature. Detailed procedures are as follows: We record the error rate of each individual at successive iterations. If the error rate decreases after one evolution, the weights of features newly selected, dubbed F^+, will

improve. On the contrary, the weights of features discarded, denoted as F^-, will decrease. Simultaneously, if the error rate increases after one evolution, the weights of features newly selected will decrease. Conversely, the weights of features discarded will improve. We introduce Eqs. (17) and (18) to illustrate the steps above.

$$\mathbf{W}(F) = \begin{cases} \mathbf{W}\left(F^+\right) + ACC_i \cdot \mathbf{MIC}^+ \\ \mathbf{W}\left(F^-\right) - ACC_i \cdot \mathbf{MIC}^- \end{cases}, \text{if } error_i(t) < \text{error}_i(t-1) \tag{17}$$

$$\mathbf{W}(F) = \begin{cases} \mathbf{W}\left(F^+\right) - ACC_i \cdot \mathbf{MIC}^+ \\ \mathbf{W}\left(F^-\right) + ACC_i \cdot \mathbf{MIC}^- \end{cases}, \text{if } error_i(t) > \text{error}_i(t-1) \tag{18}$$

where $\mathbf{W}(F)$ represents the weight vector of all features; $\mathbf{W}(F^+)$ and $\mathbf{W}(F^-)$ represent that only the weights of the newly selected as well as the discarded features are updated, respectively; ACC_i indicates the classification accuracy of the current individual i; $\mathbf{MIC}^+$ and $\mathbf{MIC}^-$ are, respectively, the MIC values of the features newly selected as well as discarded. For those features unchanged, no available measure is taken. A converted function is utilized to select the features, which is given in Eq. (19):

$$\boldsymbol{\mu} = 0.5 * \left[\tanh\left(5 * \left(\frac{\mathbf{W}(F)}{\text{sum}(|\mathbf{W}(F)|)} - 0.5\right)\right) + 1\right] \tag{19}$$

where the selected probabilities $\boldsymbol{\mu}$ of features can be calculated by Eq. (19). We then compare the probability of each feature with a random value from $[0, 1]$ to achieve the subpopulation evolution in Task 2. The converted function can have small mapping values when the variable is negative, which handles the features with weights less than 0.

Task 2 perform its search via the dominance accumulation mechanism, of which the weight vector is updated based on the collective knowledge from Task1 and Task 2. The focus on exploring the dominance of each feature can locate the informative attributes in the optimal feature subset.

3.5. Improvements for EMT-IGWO

3.5.1. Nonlinear multi-convergence factor

In traditional GWO, a fixed linear convergence factor is typically used to regulate the coefficient vector, potentially leading to premature convergence. Moreover, based on the fact that the heuristic approach for the generation of the initial population often requires combining with other strategies, especially the nonlinear convergence factor, it allows the population to maintain the global search capability. Based on Pan et al. (2023), we additionally design a nonlinear convergence factor, i.e., a_2, and the nonlinear multi-convergence factor is thus given:

$$\begin{aligned} a_1 &= 2\cos\left(\tfrac{\pi}{2} * \left(\tfrac{t}{T}\right)^2\right) \\ a_2 &= 1 + \cos\left(\pi * \tfrac{t-1}{T-1}\right) \\ a_3 &= \cos\left(\tfrac{\pi}{2} * \tfrac{t}{T}\right) \\ a_4 &= 2 - \cos\left(\tfrac{\pi}{2} * \tfrac{t}{T}\right) \end{aligned} \tag{20}$$

where t indicates the current iteration, and T represents the maximum iteration. This cosine non-linear convergence factor increases more chances for the population to jump out of the local optimum.

3.5.2. Rank-based mutation operation

Task 2 adopts a dominance accumulation mechanism to search for the optimal feature subset. More attention is paid to the features with higher weights inevitably, which may ignore the relationship between high-weight features and low-weight features. Therefore, a rank-based mutation operation is proposed to flip the features according to the following steps: (i) We divide one feature subset into two parts, the selected features and the discarded features. (ii) The selected features are sorted in ascending order by their MIC values. (iii) The discarded features are sorted in descending order by their MIC values. After that, the features in two lists get flipped by probability computed from Eq. (21):

$$\rho^d = 0.1 * e^{-d/m} + 0.01 \tag{21}$$

where ρ^d indicates the flipped possibility of the dth feature. m is the length of the selected or discarded features. From the Eq. (21), we find that the discarded features with higher MIC values have more opportunities to be flipped. On the other hand, the selected features with lower MIC values have a higher probability to get flipped. Fig. 2 depicts an example for this process.

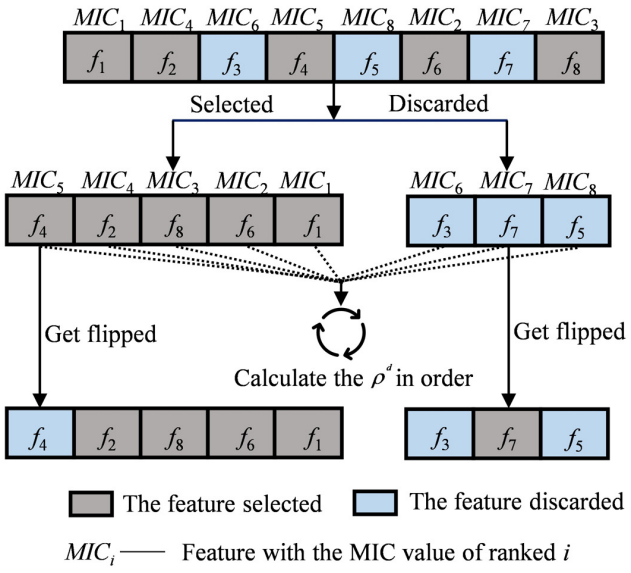

Figure 2: An example of the rank-based mutation operation

4. Experiments

4.1. Datasets

Eight gene expression datasets are employed in computational experiments, which are available on Pan et al. (2023) and Scikit Feature Selection database (Li et al., 2017). The number of features is from 3312 to 12600 among these datasets. Table 1 shows the detailed information of the involved datasets. The distribution of data is highly unbalanced. The classification on such datasets is a challenging task.

Table 1: Information of eight gene expression datasets

Datasets	Features	Instances	Classes	% Smallest class	% Largest class
lung2	3312	203	5	3	68
DLBCL	5469	77	2	25	75
TOX_171	5748	171	4	23	26
Brain Tumor 1	5920	90	5	4	67
Prostate_GE	5966	102	2	49	51
Adenoma	7457	36	2	50	50
Brain Tumor 2	10367	50	4	14	30
Lung Cancer	12600	203	5	3	68

4.2. Comparative Methods and Parameter Settings

To evaluate the classification performance of the proposed algorithm, we compare the EMT-IGWO algorithm with all available features (Full) and binary GWO (BGWO). Three state-of-the-art algorithms are also selected: the hybrid algorithm GWO and PSO (BGWOPSO) (Al-Tashi et al., 2019), the Grey-Wolf algorithm integrating a two-phase mutation (TMGWO) (Abdel-Basset et al., 2020), and the correlation-guided updating strategy and surrogate-assisted PSO (CUS-SPSO) (Chen et al., 2021). Since these methods have good theoretical analysis and show good performance in feature selection, we choose these methods as comparison methods in this study to ensure the fairness of the experiment.

This study uses the KNN as a classifier, where K is set to 5. For each dataset, 70% of the instances are employed for training sets randomly, and the remaining 30% are selected as test sets. During the training process, 5-fold cross-validation is employed. After the training, selected features will be evaluated on the test set instances to obtain the corresponding accuracy. Table 2 shows the parameter settings according to the characteristics of each method. We use uniform parameter settings for each experiment and take the average value to ensure the stability of the experimental results.

Table 2: Parameter settings

Method	Population size	MaxIter	Parameter values
BGWO	30	100	$a = 2 - 2 * (iter/MaxIter)$, $A = [0, 2]$
BGWOPSO	30	100	$c_1 = c_2 = c_3 = 0.5$, $w = 0.5 + rand/2$, $A = [0, 2]$
TMGWO	30	100	$a = 2 - 2 * (iter/MaxIter)$, $A = [0, 2]$, $M_p = 0.5$
CUS-SPSO	30	100	$c_1 = c_2 = 1.5$, $w = 0.9 - (iter/MaxIter)/2$, $n_c = 2$, $A = 0.15$, $B = 0.05$
EMT-IGWO	$40/Num\_Task$	100	$a_1 = 2\cos\left(\frac{\pi}{2} * \left(\frac{t}{T}\right)^2\right)$, $a_2 = 1 + \cos\left(\pi * \frac{t-1}{T-1}\right)$, $a_3 = \cos\left(\frac{\pi}{2} * \frac{t}{T}\right)$, $a_4 = 2 - \cos\left(\frac{\pi}{2} * \frac{t}{T}\right)$, $Num\_Task = 2$

4.3. Computational Results and Discussions

We run the computational experiments on Python 3.8 with an Intel Xeon Platinum 8474C vCPU and 80GB of memory. The following metrics are used for the evaluation of the performance: 1) the size of the feature subset obtained (Size); 2) the average training time

(Time); 3) the best classification accuracy (Best); 4) the average classification accuracy (AVG) on ten independent runs and corresponding standard deviation (Std).

Table 3: Test results on datasets with high dimensions

Dataset	Method	Size	Time (s)	Best	AVG±Std
lung2	Full	3312.00	—	98.36	92.95±3.20
	BGWO	1636.00	58.25	98.36	95.08±3.28
	BGWOPSO	159.33	56.71	95.08	91.26±3.83
	TMGWO	147.70	3167.06	96.72	91.64±3.91
	CUS-SPSO	1853.30	86.59	96.72	94.59±2.08
	EMT-IGWO	369.50	56.50	**100.00**	**96.23±2.32**
DLBCL	Full	5469.00	—	**95.83**	87.50±4.93
	BGWO	2740.00	36.15	**95.83**	90.97±4.87
	BGWOPSO	358.83	43.37	91.67	84.03±6.13
	TMGWO	130.63	4454.04	91.67	80.73±7.02
	CUS-SPSO	2809.90	68.70	**95.83**	86.67±6.67
	EMT-IGWO	544.40	73.21	**95.83**	**92.92±3.25**
TOX_171	Full	5748.00	—	78.85	68.08±8.08
	BGWO	3360.33	75.55	78.85	69.23±8.60
	BGWOPSO	306.17	65.08	76.92	68.59±5.25
	TMGWO	943.13	6337.86	75.00	66.35±4.82
	CUS-SPSO	3750.60	128.93	80.77	72.69±7.13
	EMT-IGWO	1077.30	88.16	**86.54**	**73.85±6.50**
Brain Tumor1	Full	5920.00	—	82.14	71.43±5.98
	BGWO	2928.50	39.48	82.14	73.21±7.41
	BGWOPSO	57.33	42.06	85.71	73.81±9.76
	TMGWO	151.50	4968.04	**96.43**	79.91±9.53
	CUS-SPSO	2643.20	77.62	85.71	76.07±6.97
	EMT-IGWO	536.10	97.44	89.29	**83.21±6.20**
Prostate_GE	Full	5966.00	—	90.32	79.03±8.06
	BGWO	2969.50	53.91	**100.00**	79.57±11.48
	BGWOPSO	39.00	54.05	90.32	82.80±6.34
	TMGWO	461.25	5473.78	90.32	81.85±7.30
	CUS-SPSO	3089.60	91.53	90.32	81.29±6.58
	EMT-IGWO	750.00	95.20	96.77	**86.77±5.09**
Adenoma	Full	7457.00	—	**100.00**	87.27±10.91
	BGWO	3644.83	36.41	**100.00**	93.94±4.69
	BGWOPSO	2.83	40.45	**100.00**	83.33±13.38
	TMGWO	7.25	5664.25	**100.00**	92.05±5.83
	CUS-SPSO	3069.80	71.85	**100.00**	88.18±5.82
	EMT-IGWO	340.40	73.77	**100.00**	**98.18±3.63**
Brain Tumor2	Full	10367.00	—	80.00	61.33±12.58
	BGWO	5165.67	41.69	80.00	66.67±10.33
	BGWOPSO	111.17	48.36	80.00	68.89±10.89
	TMGWO	400.88	8858.13	73.33	68.33±5.91
	CUS-SPSO	5620.10	107.91	80.00	61.33±10.67
	EMT-IGWO	1127.30	134.14	**86.67**	**74.67±13.27**
Lung Cancer	Full	12600.00	—	93.44	88.69±2.37
	BGWO	6251.83	117.77	93.44	90.16±2.93
	BGWOPSO	567.67	60.76	91.80	86.89±4.64
	TMGWO	237.50	9900.34	93.44	87.70±5.26
	CUS-SPSO	6449.80	233.81	93.44	88.20±3.93
	EMT-IGWO	1484.30	177.07	**98.36**	**91.31±3.74**

4.3.1. Results of Classification performance on Training and Test Datasets

The convergence curves during the training process are depicted in Fig. 3. It can be observed that EMT-IGWO method exhibits a significant advantage over the compared algorithms on multiple training datasets. In addition, we can also find that EMT-IGWO still has excellent search capacity in the late iterative stage. On the other hand, the performance of the classification accuracy on the test datasets is an important measure of the robustness of one approach. Table 3 shows the classification accuracy on 8 examined datasets. Remarkably, EMT-IGWO achieves the highest average and best classification accuracy on almost all test datasets. Specifically, the average classification accuracy of EMT-IGWO on the eight test datasets is 87.14%, a significant improvement of 6.01% compared to the second-ranked CUS-SPSO algorithm. Furthermore, the red lines are distributed mostly outside the entire spider web in Fig. 4, which intuitively illustrates that our proposed approach performs better on test datasets compared with other algorithms. It is noted that EMT-IGWO exhibits strong performance on test datasets such as lung2, Adenoma, and DLBCL, and that the outer perimeter of the spider web is well-represented by the red line.

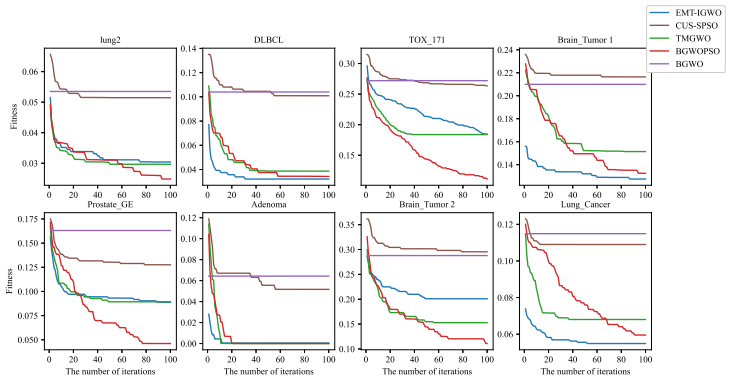

Figure 3: Convergence curves of different methods during the training process

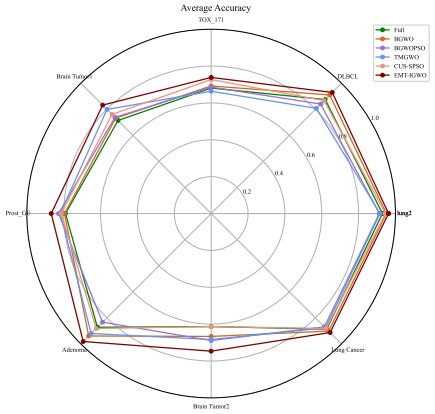

Figure 4: Classification accuracies of different methods on test datasets

4.3.2. Results of the number of selected features

Table 3 shows the average feature subset size of all the methods. BGWOPSO ranks first in terms of the dimension reduction. However, numerous informative features are lost, which leads to its poor performance. In contrast, the average number of features selected by EMT-IGWO is 778.66, second only to BGWOPSO and TMGWO. However, EMT-IGWO achieves an average subset size of only approximately 11% of all available features on all the examined datasets, which demonstrates the satisfactory dimension reduction capability of EMT-IGWO. In general, EMT-IGWO has better tradeoff of the classification performance and the dimension reduction capability compared to other approaches.

4.3.3. Results of running time

The running time of EMT-IGWO and the compared approaches are indicated in the fourth column of Table 3. The training time of EMT-IGWO ranks third among all the five methods on all the test datasets. Compared to the BGWOPSO, the cumulative training time of EMT-IGWO is approximately 5 minutes longer than that of BGWOPSO. The main reason is that further enhancement strategies are incorporated into EMT-IGWO for better classification performance, the running time of which is acceptable. Obviously, the running time of EMT-IGWO is less than 100s on over half of the examined datasets, which has exhibited the efficiency of EMT-IGWO.

4.4. Statistical Significance Test

To further analyze the performance difference among all the methods, we use the Wilcoxon signed-rank test to conduct the statistical significance test. Table 4 reveals the results of the pairwise comparison. Here, the symbols '+', '≈', and '−' illustrate that there are +, ≈ , and − examples of EMT-IGWO that are superior to, similar to, and inferior to the compared algorithms, respectively. From Table 4, the p-value obtained by the average classification accuracy is all less than 0.05 under 95% confidence. Therefore, we can conclude that our proposed EMT-IGWO can provide significant results compared to other algorithms.

Table 4: Wilcoxon signed-rank test analysis (AVG)

Comparison	+	≈	−	p-value
EMT-IGWO vs. BGWO	8	0	0	0.0078
EMT-IGWO vs. BGWOPSO	8	0	0	0.0078
EMT-IGWO vs. TMGWO	8	0	0	0.0078
EMT-IGWO vs. CUS-SPSO	8	0	0	0.0078

5. Conclusion

This paper introduces EMT-IGWO, a novel feature selection method for classifying large-scale genomic data. EMT-IGWO employs a GWO-based evolutionary multitasking paradigm,

offering several notable advantages. Each subpopulation operates in an independent searching pattern, enabling diverse searching directions for each task. The transferred knowledge between tasks increases population diversity and enhances searching capabilities. Additionally, improvements in the search patterns prevent individuals from becoming trapped in local optima. The experimental results showed that EMT-IGWO can effectively improve classification accuracy while consuming acceptable runtime compared to state-of-the-art methods. However, the dimension reduction capability of EMT-IGWO cannot provide a significant advantage. In our future work, we aim to further improve the dimensionality reduction capability of the proposed algorithm and extend our multi-task approach for addressing multiple related feature selection tasks that share common features simultaneously.

References

Mohamed Abdel-Basset, Doaa El-Shahat, Ibrahim El-Henawy, Victor Hugo C De Albuquerque, and Seyedali Mirjalili. A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection. *Expert Systems with Applications*, 139:112824, 2020.

Qasem Al-Tashi, Said Jadid Abdul Kadir, Helmi Md Rais, Seyedali Mirjalili, and Hitham Alhussian. Binary optimization using hybrid grey wolf optimization for feature selection. *Ieee Access*, 7:39496–39508, 2019.

Esra'a Alhenawi, Rizik Al-Sayyed, Amjad Hudaib, and Seyedali Mirjalili. Feature selection methods on gene expression microarray data for cancer classification: A systematic review. *Computers in Biology and Medicine*, 140:105051, 2022.

Mohammed Ghaith Altarabichi, Sławomir Nowaczyk, Sepideh Pashami, and Peyman Sheikholharam Mashhadi. Surrogate-assisted genetic algorithm for wrapper feature selection. In *2021 IEEE congress on evolutionary computation (CEC)*, pages 776–785. IEEE, 2021.

Verónica Bolón-Canedo, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, José Manuel Benítez, and Francisco Herrera. A review of microarray datasets and applied feature selection methods. *Information sciences*, 282:111–135, 2014.

Ke Chen, Bing Xue, Mengjie Zhang, and Fengyu Zhou. An evolutionary multitasking-based feature selection method for high-dimensional classification. *IEEE Transactions on Cybernetics*, 52(7):7172–7186, 2020.

Ke Chen, Bing Xue, Mengjie Zhang, and Fengyu Zhou. Correlation-guided updating strategy for feature selection in classification with surrogate-assisted particle swarm optimization. *IEEE Transactions on Evolutionary Computation*, 26(5):1015–1029, 2021.

Tansel Dokeroglu, Ayça Deniz, and Hakan Ezgi Kiziloz. A comprehensive survey on recent metaheuristics for feature selection. *Neurocomputing*, 494:269–296, 2022.

Swati S Gandhi and SS Prabhune. Overview of feature subset selection algorithm for high dimensional data. In *2017 International Conference on Inventive Systems and Control (ICISC)*, pages 1–6. IEEE, 2017.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Variational information maximization for feature selection. *Advances in neural information processing systems*, 29, 2016.

Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

Abhishek Gupta, Yew-Soon Ong, and Liang Feng. Multifactorial evolution: Toward evolutionary multitasking. *IEEE Transactions on Evolutionary Computation*, 20(3):343–357, 2015.

Pei Hu, Jeng-Shyang Pan, and Shu-Chuan Chu. Improved binary grey wolf optimizer and its application for feature selection. *Knowledge-Based Systems*, 195:105746, 2020.

Fatih Kılıç, Yasin Kaya, and Serdar Yildirim. A novel multi population based particle swarm optimization for feature selection. *Knowledge-Based Systems*, 219:106894, 2021.

An-Da Li, Bing Xue, and Mengjie Zhang. A forward search inspired particle swarm optimization algorithm for feature selection in classification. In *2021 IEEE Congress on Evolutionary Computation (CEC)*, pages 786–793. IEEE, 2021.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

Jiabin Lin, Qi Chen, Bing Xue, and Mengjie Zhang. Evolutionary multitasking for multi-objective feature selection in classification. *IEEE Transactions on Evolutionary Computation*, 2023.

G Manikandan and S Abirami. An efficient feature selection framework based on information theory for high dimensional data. *Applied Soft Computing*, 111:107729, 2021.

Seyedali Mirjalili, Seyed Mohammad Mirjalili, and Andrew Lewis. Grey wolf optimizer. *Advances in engineering software*, 69:46–61, 2014.

Ben Niu, Xuesen Yang, Hong Wang, Kaishan Huang, and Sung-Shun Weng. Feature subset selection using a self-adaptive strategy based differential evolution method. In *Advances in Swarm Intelligence: 9th International Conference, ICSI 2018, Shanghai, China, June 17-22, 2018, Proceedings, Part I 9*, pages 223–232. Springer, 2018.

Hongyu Pan, Shanxiong Chen, and Hailing Xiong. A high-dimensional feature selection method based on modified gray wolf optimization. *Applied Soft Computing*, 135:110031, 2023.

Grant Patterson and Mengjie Zhang. Fitness functions in genetic programming for classification with unbalanced data. In *AI 2007: Advances in Artificial Intelligence: 20th Australian Joint Conference on Artificial Intelligence, Gold Coast, Australia, December 2-6, 2007. Proceedings 20*, pages 769–775. Springer, 2007.

Litao Qu, Weibin He, Jianfei Li, Hua Zhang, Cheng Yang, and Bo Xie. Explicit and size-adaptive pso-based feature selection for classification. *Swarm and Evolutionary Computation*, 77:101249, 2023.

David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

Qisthina Syifa Setiawan, Zuherman Rustam, and Jacub Pandelaki. Comparison of naive bayes and support vector machine with grey wolf optimization feature selection for cervical cancer data classification. In *2021 international conference on decision aid sciences and application (DASA)*, pages 451–455. IEEE, 2021.

Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189–203, 2018.

Xubin Wang, Haojiong Shangguan, Fengyi Huang, Shangrui Wu, and Weijia Jia. Mel: Efficient multi-task evolutionary learning for high-dimensional feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Tao Wen, Deyi Dong, Qianyu Chen, Lei Chen, and Clive Roberts. Maximal information coefficient-based two-stage feature selection method for railway condition monitoring. *IEEE Transactions on Intelligent Transportation Systems*, 20(7):2681–2690, 2019.

Xueyuan Xu and Xia Wu. Feature selection under orthogonal regression with redundancy minimizing. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3457–3461. IEEE, 2020.

Huaqing Zhang, Jian Wang, Zhanquan Sun, Jacek M Zurada, and Nikhil R Pal. Feature selection for neural networks using group lasso regularization. *IEEE Transactions on Knowledge and Data Engineering*, 32(4):659–673, 2019.