IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 9, 2018, accepted March 18, 2018, date of publication March 27, 2018, date of current version May 16, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2818682

A Hybrid Genetic Algorithm With Wrapper-Embedded Approaches for Feature Selection

XIAO-YING LIU[1], YONG LIANG[2,3], SAI WANG[2], ZI-YI YANG[2], AND HAN-SHUO YE[2]

[1]Computer Engineering Technical College, Guangdong Polytechnic of Science and Technology, Zhuhai 519090, China
[2]Faculty of Information Technology, Macau University of Science and Technology, Taipa 999078, Macau
[3]State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Taipa 999078, Macau

Corresponding author: Yong Liang (yliang@must.edu.mo)

This work was supported in part by the Macau Science and Technology Develop Funds, Macao SAR of China, under Grant 003/2016/AFJ and in part by the China NSFC Project under Contract 61661166011.

ABSTRACT Feature selection is an important research area for big data analysis. In recent years, various feature selection approaches have been developed, which can be divided into four categories: filter, wrapper, embedded, and combined methods. In the combined category, many hybrid genetic approaches from evolutionary computations combine filter and wrapper measures of feature evaluation to implement a population-based global optimization with efficient local search. However, there are limitations to existing combined methods, such as the two-stage and inconsistent feature evaluation measures, difficulties in analyzing data with high feature interaction, and challenges in handling large-scale features and instances. Focusing on these three limitations, we proposed a hybrid genetic algorithm with wrapper−embedded feature approach for selection approach (HGAWE), which combines genetic algorithm (global search) with embedded regularization approaches (local search) together. We also proposed a novel chromosome representation (intron+exon) for global and local optimization procedures in HGAWE. Based on this ''intron+exon'' encoding, the regularization method can select the relevant features and construct the learning model simultaneously, and genetic operations aim to globally optimize the control parameters in the above non-convex regularization. We mention that any efficient regularization approach can serve as the embedded method in HGAWE, and a hybrid $L_{1/2} + L_2$ regularization approach is investigated as an example in this paper. Empirical study of the HGAWE approach on some simulation data and five gene microarray data sets indicates that it outperforms the existing combined methods in terms of feature selection and classification accuracy.

INDEX TERMS Feature selection, wrapper−embedded method, memetic framework, genetic algorithm, $L_{1/2} + L_2$ regularization.

I. INTRODUCTION

Explosive growth of data urgently requires development of new technologies and automation tools that can intelligently help us translate large amounts of data into useful information and knowledge. Indeed, it is not that all the features in the data are essential. The purpose of feature selection is, therefore, to select only a small portion of the relevant features from the original large data set so as to speed up the learning process and improve the performance of the learning model.

In recent years, various approaches have been developed for feature selection, which generally are divided into four categories: filter, wrapper, embedded and combined methods. For filter approaches, different feature selection measures have been applied to rank individual features [1], e.g., 1) information theoretic measures [2]; 2) consistency measures [3]; 3) dependency (or correlation) measures [4]; 4) distance measures [5]; 5) rough set theory [6] and 6) fuzzy set theory [7]. A major drawback of the filter methods is that they examine each feature independently, and ignore the individual performance of the feature in relation to the group, of which it is a part, despite the fact that features in a group may have a combined effect in a machine learning task.

For wrapper approaches, different machine learning algorithms have been used to evaluate the performance of selected

feature subsets, e.g., support vector machines (SVMs) [8]; K-nearest neighbors (KNN) [9]; artificial neural networks (ANNs) [10]; decision tree (DT) [11]; Naive Bayes (NB) [12]; multiple linear regression for classification [13]; extreme learning machines (ELMs) [14]; and linear discriminant analysis (LDA) [15]. Often, the results of the wrapper methods are superior to those of the filter methods, but the computational cost of the wrapper methods is high.

The third group of feature selection approaches is embedded methods, which integrate feature selection and learning procedure into a single process. Regularization methods are an important embedded technique and perform both learning model construction and automatic feature selection simultaneously. Recently, the applications of regularization approaches for feature selection have become increasingly interesting. Focusing on high dimensional feature selection problems such as gene expression microarray data, Lasso (L_1) [16], smoothly clipped absolute deviation (SCAD) [17], minimax concave penalty (MCP) [18] and $L_{1/2}$ regularization [19], [20] are popularly used regularization approaches. In gene expression studies, if genes share the same biological pathway, they are usually highly correlated and grouped [21]. Therefore, some approaches have been proposed to deal with issues of high relevance and grouping features, for example, group Lasso [22], Elastic net [23], SCAD-L_2 [21], and hybrid $L_{1/2} + L_2$ regularization (HLR) [24].

The fourth group of feature selection procedures is combined methods. Given that each feature evaluation measure has its own advantages and disadvantages, combined means that the evaluation procedure includes different types of feature selection measures such as filter and wrapper [25], [26].

Recently, evolutionary computations (EC) approaches have been widely used for feature selection because they are well known for their global optimization capabilities/potential. In the survey literature Xue *et al.* [27] mentioned that over 500 papers have been published in recent years on this topic. Based on the relevant evaluation criteria, the EC algorithms of feature selection are also divided into four categories, similar to the categorization mentioned above: 1) filter approaches: genetic algorithm (GA) [28], genetic programming (GP) [29], particle swarm optimization (PSO) [30], ant colony optimization (ACO) [31], differential evolution (DE) [32], evolutionary strategy (ES) [4]; 2) wrapper approaches: GA [33], GP [34], PSO [35], ACO [36], DE [37], ES [38], estimated distribution algorithm (EDA) [39]; 3) embedded approaches: GP [34], [40]–[42]; and 4) combined approaches: GA [8], PSO [43], ACO [44], DE [45] and memetic algorithm (MA) [46]–[51].

In the combined methods, many memetic-based feature selection approaches, which combine wrapper and filter methods, provide an opportunity for population-based optimization with local search. For example, Zhu *et al.* [46] applied GAs for wrapper feature selection and used Markov blanket approach as a local search for filter feature selection. However, such two-stage approaches have the potential limitation that filter evaluation measures may eliminate potentially useful features regardless of their performance in the wrapper approaches. In addition, the wrapper approaches usually involve a large number of assessments, and each assessment usually takes a considerable amount of time, especially when the numbers of features and instances are large. The second limitation of the existing combined feature selection methods is that they are primarily concerned with the relatively small numbers of features and instances.

Feature interaction (or grouping effect [21]) presents another difficulty in feature selection. On the one hand, a feature, which is weakly relevant to the target, could end up significantly improving the accuracy of the learning model when used together with some complementary features; on the other hand, an individually relevant feature can become redundant when used together with other features. Feature interaction occurs frequently in many areas. The third limitation of the existing combined feature selection approaches is that filter measures, which evaluate features individually, do not work well, and a subset of relevant or grouping features is required to be evaluated as a whole.

To solve these three limitations of the combined feature selection approaches, we proposed a hybrid genetic algorithm with wrapper−embedded approaches (HGAWE) to combine evolutioanry optimization (global search) and embedded regularization approaches (local search) for feature selection.

Regularization methods are an important embedded technique and perform both model learning and automatic feature selection simultaneously. Focusing on high dimensional feature selection problems, such as relevant gene selection in microarray data, many regularization approaches have been proposed in recent years, for instance, Lasso [16], SCAD [17], MCP [18] and $L_{1/2}$ [19], [20]. Since Lasso is a convex penalty function, the gradient-based coordinate descent algorithm is suitable and widely used for the global optimization of Lasso. Some efforts have also been made in response to the problem of highly correlated and grouped features, for example, Elastic net [23], SCAD-L_2 [21], and hybrid $L_{1/2} + L_2$ regularization [24]. Liu *et al.* [52] have proposed a complex harmonic regularization approach (CHR) for uncertain probabilities distribution of data. Meng *et al.* [53] have proposed a self-paced curriculum learning (SPLC) regularization approach, which significantly improves the learning efficiency when the number of instances is large. Regularization approaches are one-stage feature evaluation measures, which are suitable for complex feature selection problems with high interaction and large scales of features and instances.

However, in regularization methods, the control parameter between loss function and penalty function is very important for their performance in feature selection. The feasible value of the control parameter is generally tuned by the grid search method with k-fold cross violation approach. In recent years, many efficient regularization methods using non-convex and multimodal penalty functions have been proposed.

These regularization methods need to search across multiple parameters, which are suitable to be optimized by EC approaches, for example, GA can deal with both unimodal and multimodal search space well, and the population-based search can find the global optima of these control parameters efficiently.

Therefore, the goal of our proposed hybrid genetic algorithm with wrapper−embedded approaches (HGAWE) is to improve learning performance and accelerate the search to identify the relevant feature subsets. Particularly, the embedded method fine-tunes the population of GA solutions by selecting the signature feature, and constructs the learning model based on efficient gradient regularization approaches. The wrapper methods induce the population of GA solutions, using heuristic search strategies to globally optimize the control parameters for the non-convex regularization. Therefore, we focus on hybrid evolutionary framework, which is able to integrate feature selection and learning model construction into a single process under the global optimization of the non-convex regularization. We note that any efficient regularization approach can serve as the embedded method in HGAWE, and a hybrid $L_{1/2} + L_2$ regularization (HLR) approach is investigated as an example in this paper. Empirical study of HGAWE on some simulation data and five gene microarray data sets indicates that it outperforms existing combined feature selection methods in terms of classification accuracy and feature selection.

The rest of this paper is organized as follows. Section II describes the related works of the HGAWE method. Section III presents a hybrid genetic algorithm based on the genetic operators and the gradient regularization approach for gene selection and cancer classification. The experimental results and discussions are presented in Section IV. Finally, Section V concludes this paper.

II. RELATED WORKS

A. REGULARIZATION APPROACHES

Regularization is an important embedded feature selection approach. Suppose $\mathbf{X}$ denotes the $n \times p$ data matrix whose rows are $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $1 \le i \le n$, Y denotes the corresponding dependent variable $(y_1, y_2, \ldots, y_n)^T$.

For any control parameter λ ($\lambda > 0$), the common form of regularization is:

$$L(\lambda, \boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \{R(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta})\} \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ are the estimated coefficients, $R(\boldsymbol{\beta})$ is a loss function and $P(\boldsymbol{\beta})$ represents the regularization term. The most commonly used regularization method is the least absolute shrinkage and selection operator (Lasso, also the L_1 penalty) [16], i.e., $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|^1$. It is performing continuous shrinkage and gene selection at the same time. Some other L_1-norm type regularization methods have also been proposed. For example, the SCAD penalty [17] is symmetric, non-convex, and can produce spare solutions at the origin in the parameter space. The adaptive Lasso [54] penalizes the different coefficients with the dynamic weights in the L_1 penalty. The MCP provides the convexity of the penalized loss in sparse regions to the greatest extent, given certain thresholds for feature selection and unbiasedness. However, for large-scale feature selection problem, such as genomic data analysis, the results of the L_1 type regularization may not be sparse enough for real application. Actually, a typical gene microarray or RNA-seq data sets have many thousands of genes, and researchers often desire to select fewer but informative genes. Although the L_0 regularization, where $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|^0$, yields the sparsest solution theoretically, it has to solve an NP-hard combinatory optimization problem. In order to obtain a more concise solution and improve the predictive accuracy of the machine learning model, the researchers studied the L_p-norm ($0 < p < 1$), especially $p = \frac{1}{10}$, $\frac{1}{2}$, $\frac{2}{3}$, or $\frac{9}{10}$ [20], [55]. In the literature [19], Xu *et al.* have proposed that a $L_{1/2}$ regularization can be taken as a representative of the L_p ($0 < p < 1$) penalties, and analyzed its analytically expressive thresholding representation. Based on this thresholding representation, solving the $L_{1/2}$ regularization is much easier than solving the L_0 regularization. Moreover, the $L_{1/2}$ penalty is unbiased and has oracle properties [19], [20], [56]. These advantages make $L_{1/2}$ penalty an effective tool for high dimensional feature selection problems [57].

However, like most regularization methods, the $L_{1/2}$ penalty ignores the correlation between features, and therefore cannot analyze data with dependent structures. If there is a set of features whose correlations are relatively high, the $L_{1/2}$ method tends to select only one feature to represent the corresponding group. In order to solve the problem of highly relevant features, Zou and Hastie [23] proposed Elastic net penalty, which is a linear combination of L_1 and L_2 (the ridge technique) penalties; such a method emphasizes the grouping effect, where strongly correlated features tend to enter or leave the learning model together. Becker *et al.* [58] proposed the Elastic SCAD (or SCAD-L_2), a combination of SCAD and L_2 penalties for feature interaction. Recently, Huang *et al.* [24] proposed the hybrid $L_{1/2}+L_2$ regularization (HLR) approach to fit the logistic regression models for gene selection, where the regularization is a linear combination of the $L_{1/2}$ and L_2 penalties. For any fixed control parameter λ_1, λ_2 ($\lambda_1, \lambda_2 > 0$), the hybrid $L_{1/2} + L_2$ regularization (HLR) is defined as follows:

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \left\{ R(\boldsymbol{\beta}) + \lambda_1 |\boldsymbol{\beta}|_{1/2} + \lambda_2 |\boldsymbol{\beta}|^2 \right\} \tag{2}$$

where $|\boldsymbol{\beta}|_{1/2} = \sum_{j=1}^{p} |\beta_j|^{1/2}$, $|\boldsymbol{\beta}|^2 = \sum_{j=1}^{p} |\beta_j|^2$.

The HLR estimator $\hat{\boldsymbol{\beta}}$ is the minimizer of Eq. (3):

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \{L(\lambda, \alpha, \beta)\} \tag{3}$$

where $\lambda = \lambda_1 + \lambda_2$, and $\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2}$.

In theory, a strictly convex penalty function provides a sufficient condition for the grouping effect of features and

the L_2 penalty ensures a strict convexity [21]. Therefore, the L_2 penalty induces the grouping effect simultaneously in the HLR approach. Experimental results on artificial and real gene expression data in the literature [24] demonstrated that the HLR method is very promising.

However, many efficient regularization methods are nonconvex and need to tune across multiple penalized parameters, which are generally adjusted by the grid search method with k-fold cross violation approach. We believe that the population-based search in EC is an efficient approach to globally optimize these penalized parameters.

B. HYBRID EVOLUTIONARY APPROACHES FOR FEATURE SELECTION

In Evolutionary Computations (EC), an initial population of candidate solutions is randomly generated in the search space and iteratively updated by artificial crossover, mutation and selection operators. After several generations, the population can gradually develop high quality solutions to the optimization problems. Over the past years, local search (LS) technologies have increasingly been combined into the random search process of EC to improve the optimization efficiency [59]. These hybrid algorithms are usually called hybrid evolutionary approaches or memetic algorithms (MA). Hybrid evolutionary approaches for feature selection, which combine wrapper and filter feature evaluation measures, provide an opportunity for population-based optimization with local search. For example, Zhu *et al.* [60] proposed the filter feature ranking method in MA to balance the local and global searches for the purpose of improving the optimization quality and efficiency. Then, Zhu *et al.* [61] integrated the Markov blanket approach into MA to simultaneously identify all and part of the relevant features. Another two-stage feature selection algorithm was proposed in [62], where a Relief-F algorithm was used to rank individual features and then the top-ranked features were used as input to the memetic wrapper feature selection algorithm. Some researchers introduced heuristic mixtures that combine the filter ranking scores to guide the search processes of GA and PSO for wrapper feature selection [43], [49], [63]. Moreover, Hybrid evolutionary approaches for feature selection have already been used to solve some real application problems, such as, optimal controller design [64], motif-finding in DNA, mircoRNA and protein sequences [65], [66].

As is shown above, in most hybrid evolutionary approaches for feature selection, the EC stage is for wrapper feature selection, and the filter-based LS algorithm helps to reach a local optimal solution. However, these " wrapper+filter" two-stage hybrid evolutionary approaches do not guarantee that the selected features in the filter stage are also optimal candidates for the EC stage, since the evaluation criteria of each stage are totally different. Thus, the filter stage in hybrid evolutionary approaches may eliminate potentially useful features with no regard to their performance in the wrapper process.

III. HYBRID GENETIC ALGORITHM WITH WRAPPER–EMBEDDED APPROACHES FOR FEATURE SELECTION

Given that existing combined feature selection methods have limitations of inconsistency in feature evaluation measures, feature interactions and large scales of features and instances, in this section, we introduce a hybrid genetic algorithm with wapper−embedded approaches (HGAWE) to combine genetic operations and hybrid $L_{1/2}+L_2$ regularization (HLR) for feature selection. We propose a new chromosome representation including intron (the penalized control parameters) and exon (the coefficients of the features in the learning model) for HGAWE optimization procedure. In the first step of HGAWE, the GA population is randomly initialized with each chromosome encoded by intron and exon parts. Subsequently, the hybrid $L_{1/2}+L_2$ regularization approach (local search) is performed on the exon parts under the fixed intron parts, to reach a local optimal solution or to improve the fitness of individuals in the search population. Genetic operators such as crossovers and mutations are performed on the intron parts of the chromosomes, and the selection operator generates the next population. This process repeats itself till the stopping conditions are satisfied. Each component is explained as follows.

A. CHROMOSOME REPRESENTATION: INTRON AND EXON

In our proposed hybrid genetic algorithm with wrapper−embedded algorithm (HGAWE), a representation for the two penalized control parameters λ, α, and the coefficients $(\beta_1, \beta_2, \cdots, \beta_p)$ of the candidate feature subset can be encoded as a chromosome: $intron + exon = (\lambda, \alpha, \beta_1, \beta_2, \ldots, \beta_p)$. The length of the chromosome is denoted as $p+2$, where p is the total number of features. The chromosome is a real value string and its intron part is globally optimized by GA operators. Although the search space of the intron part is nonconvex and multimodal, GA has the global optimal ability because the dimension of the intron is quite low. On the other hand, the exon part is optimized by the regularization approach for learning model construction and feature selection synchronously. In the exon part, a nonzero value of β_i implies that the corresponding feature has been selected. In contrast, the candidate feature has been rejected if its corresponding coefficients β_i is equal to zero. The maximum allowable number of nonzero β_i in the exon of each chromosome is denoted as T. When prior knowledge about the optimal number of features is available, we may limit T to no more than the pre-defined value; otherwise T is equal to p.

B. OBJECTIVE FUNCTION

The objective function is defined by:

$$Fitness(chromosome) = Accuracy\ of\ the\ classfication\ model\ with(\lambda, \alpha, \beta_1, \beta_2, \ldots, \beta_p) \quad (4)$$

where nonzero β_i denotes the corresponding selected features subset encoded in the exon part of the chromosome. The objective function evaluates the significance of the given feature subset. In this paper, the fitness of the objective function is specified as the classification accuracy of the logistic regularization model with the chromosome $\lambda, \alpha, \beta_1, \beta_2, \ldots, \beta_p$, using the hybrid $L_{1/2} + L_2$ penalties method. Note that when two chromosomes are found to have similar fitness, i.e., the difference between their fitness is less than a small value of e ($e = 10^{-5}$ in our experiments), then the one with a smaller number of selected features is given higher chances of surviving to the next generation.

C. LS IMPROVEMENT PROCEDURE WITH HLR IN LOGISTIC REGRESSION

In this section, we consider the use of the hybrid $L_{1/2} + L_2$ penalties method with the coordinate descent algorithm as local search approach in our proposed HGAWE. In general, the coordinate descent algorithm [67] is an efficient method for solving regularization problems because its computational time increases linearly with the dimension of the feature selection problems. Therefore, HGAWE is capable of constructing the learning model and selecting the relevant features with grouping effect efficiently and synchronously.

The hybrid $L_{1/2} + L_2$ regularization (HLR) [24] in logistic model is formed as:

$$\hat{\boldsymbol{\beta}} = \arg\min \left[\hat{R}(\boldsymbol{\beta}) + \lambda \hat{P}(\boldsymbol{\beta}) \right] \tag{5}$$

where $\lambda = \lambda_1 + \lambda_2$, and $\hat{R}(\boldsymbol{\beta})$ is a loss function in logistic regression:

$$\hat{R}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} y_i \cdot X_i' \boldsymbol{\beta} + log(1 + exp(X_i' \boldsymbol{\beta})) \right\} \tag{6}$$

Here, $(y_1, y_2, \ldots, y_n)^T = Y$ denotes the decision vector of a binary value with 0 or 1 in logistic model.

$\hat{P}(\boldsymbol{\beta})$ is the HLR penalty function and it is defined as:

$$\hat{P}(\boldsymbol{\beta}) = \alpha \sum_{j=1}^{p} \sqrt{|\beta_j|} + (1 - \alpha) \sum_{j=1}^{p} |\beta_j|^2 \tag{7}$$

where $\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2}$, and $0 \leq \alpha \leq 1$.

Following Friedman *et al.* [68], Liang *et al.* [57] and Huang *et al.* [24], we use the approach of the original coordinate-wise update:

$$\beta_j \leftarrow \frac{Half(\omega_j, \lambda\alpha)}{1 + \lambda(1 - \alpha)} \tag{8}$$

where $1 \leq j \leq p$ and

$$\omega_j = \sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}) \tag{9}$$

Here, as the partial residual for fitting β_j, $\tilde{y}_i^{(j)}$ is defined as:

$$\tilde{y}_i^{(j)} = \sum_{k \neq j} x_{ik} \beta_k \tag{10}$$

Additionally, $Half(\cdot)$ is the $L_{1/2}$ thresholding operator coordinate-wise update form for the HLR approach:

$$Half(\omega_j, \lambda) = \begin{cases} \frac{2}{3}\omega_j(1 + cos(\frac{2(\pi - \varphi_\lambda(\omega_j))}{3})) \\ \qquad \text{if } |\omega_j| > \frac{\sqrt[3]{54}}{4}(\lambda)^{\frac{2}{3}} \\ 0 \qquad \text{otherwise} \end{cases} \tag{11}$$

where $\varphi_\lambda(\omega) = arccos(\frac{\lambda}{8}(\frac{|\omega|}{3})^{-\frac{3}{2}})$, $\pi = 3.14$.

Therefore, the Eq. (5) can be linearized by one-term Taylor series expansion:

$$\hat{\boldsymbol{\beta}} \approx \arg\min[\frac{1}{2n} \sum_{i=1}^{n} (Z_i - X_i' \boldsymbol{\beta})' W_i (Z_i - X_i' \boldsymbol{\beta}) + \lambda \hat{P}(\boldsymbol{\beta})] \tag{12}$$

where Z_i is the estimated response and W_i is the weight for Z_i, which can be defined as follows.

$$Z_i = X_i' \tilde{\beta} + \frac{y_i - f(X_i' \tilde{\beta})}{f(X_i' \tilde{\beta})(1 - f(X_i' \tilde{\beta}))} \tag{13}$$

$$W_i = f(X_i' \tilde{\beta})(1 - f(X_i' \tilde{\beta})) \tag{14}$$

where $f(X_i' \boldsymbol{\beta})$ is evaluated value under the current parameters:

$$f(X_i' \boldsymbol{\beta}) = \frac{exp(X_i' \boldsymbol{\beta})}{1 + exp(X_i' \boldsymbol{\beta})} \tag{15}$$

Thus, we can redefine Eq.(13) and (14) for fitting current $\tilde{\beta}$ as:

$$\tilde{Z}_i^{(j)} = \sum_{k \neq j} x_{ik} \tilde{\beta}_k \tag{16}$$

$$\omega_j = \sum_{i=1}^{n} W_i x_{ij} (Z_i - \tilde{Z}_i^{(j)}) \tag{17}$$

The procedure of the coordinate descent algorithm for the HLR penalized logistic model is described as follows.

Algorithm 1 The Coordinate Descent Algorithm for the HLR Penalized Logistic Model

1: Initialize all $\beta_j(m) \leftarrow 0 (j = 1, 2, \ldots, p)$, set $m \leftarrow 0$ and λ, α are set by GA;
2: **if** $\beta(m)$ dose not converge **then**
3: **repeat**
4: Calculate $Z(m)$ and $W(m)$ and approximate the loss function Eq. (12) based on the current $\beta(m)$;
5: **for** $j = 1$ *to* p **do**
6: Compute $\tilde{Z}_i^{(j)}(m) \leftarrow \sum_{k \neq j} x_{ik} \beta_k(m)$ and $\omega_j(m) \leftarrow \sum_{i=1}^{n} W_i(m) x_{ij} (Z_i(m) - \tilde{Z}_i^{(j)}(m))$;
7: Update $\beta_j(m) \leftarrow \frac{Half(\omega_j(m), \lambda\alpha)}{1 + \lambda(1 - \alpha)}$;
8: **end for**
9: $m \leftarrow m + 1$, $\beta(m + 1) \leftarrow \beta(m)$;
10: **until** there are no more features to be removed;
11: **end if**
12: **return** the optimal feature subset;

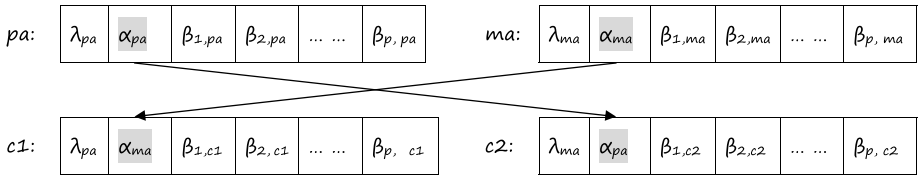


FIGURE 1. The crossover operator at α point in each chromosome.

D. GENETIC OPERATORS IN HGAWE

In the evolution process of HGAWE, standard GA operators such as fitness proportionate selection, one point crossover and uniform mutation operators can be applied. Moreover, if prior knowledge on the optimal number of features is available, the number of nonzero of β_i in each exon part of the chromosome may be constrained to a maximum of T in the evolution process.

1) CROSSOVER

We first randomly select two parents (pa, ma) from current population for later breeding. Then, the operation of crossover is used with a crossover probability p_c ($p_c = 0.85$ in our experiments) to produce offsprings that inherit characteristics from both parents. A single crossover point on the intron of both *pa* and *ma* chromosomes is generated between the penalized control parameters λ and α; then these two penalized control parameters on both sides of that point are swapped in the intron of the parent's chromosomes to create the intron part of the offsprings' chromosomes c_1 and c_2. The exon $\boldsymbol{\beta}$ of these two offsprings chromosomes are evaluated by the local optimization strategies. This procedure of crossover operation is shown in Fig 1.

2) MUTATION

The mutation operator allows diversity of populations and larger exploration of search space. During this stage, we randomly choose one of the penalized control parameters λ, α with a mutation probability p_m ($p_m = 0.1$ in our experiments) to mutate a selected chromosome. The fitness and $\boldsymbol{\beta}$ of the new chromosome generated by the mutation operation are also evaluated by the local optimization strategies.

3) SELECTION

The roulette-wheel selection [69] is used to generate the next generation from the parent and offspring populations. The selection probability $prob_c$ of the chromosome c is directly proportional to its fitness, i.e.,

$$prob_c = \frac{f(c)}{\sum f(parent) + \sum f(offspring)} \tag{18}$$

At the genetic selection stage, the candidate chromosomes with higher accuracy will be less likely to be eliminated and still have the chance to be possible.

TABLE 1. Parameters set for the EC algorithms in the seven approaches.

Parameter	Value
Population size (P)	200
Crossover probability (pc)	0.85
Mutation probability (pm)	0.1
Stopping criterion (G)	2000

IV. RESULTS AND DISCUSSION

A. ANALYSIS OF SIMULATED DATA

The goal of this section is to evaluate the performance of the HGAWE approach in the simulation study. Six approaches are compared: GA, GP, MA, Elastic net, SCAD-L_2, and the hybrid $L_{1/2} + L_2$ regularization (HLR) respectively. We simulate data from the true model

$$log(\frac{y}{1-y}) = X'\boldsymbol{\beta} + \sigma\varepsilon, \varepsilon \sim N(0, 1)$$

where $X \sim N(0, 1)$, ε is the independent random noise and σ is the control parameter for noise. Three scenarios are presented here. In every example, the dimension of features is 6000. The notation $\cdot/\cdot$ represents the number of observations in the training and test sets respectively, e.g. 100/100. Here are the details of the three scenarios.

(a) In Scenario 1, the dataset consists of 200/200 observations, we set the noise control parameter $\sigma = 0.2$ and

$$\boldsymbol{\beta} = (\underbrace{1, -1, 1, -1, \ldots, 1, -1}_{100}, \underbrace{0, \ldots, 0}_{1900},$$
$$\underbrace{2, -2, 2, -2, \ldots, 2, -2}_{100}, \underbrace{0, \ldots, 0}_{1900}, \underbrace{2, 2, \ldots, 2}_{100}, \underbrace{0, \ldots, 0}_{1900}).$$

We simulated a grouped feature situation

$$x_j = \rho \times x_1 + (1-\rho) \times x_j, j = 2, 3, \ldots, 100;$$
$$x_j = \rho \times x_{2001} + (1-\rho) \times x_j, j = 2002, 2003, \ldots, 2100;$$
$$x_j = \rho \times x_{4001} + (1-\rho) \times x_j, j = 4002, 4003, \ldots, 4100.$$

where ρ is the correlation coefficient of the grouped variables. In this example, there are three groups of correlated features. An ideal sparse regression method would select only the 300 true features and set the coefficients of the 5700 irrelevant features to zero.

(b) Scenario 2 is defined similarly to Scenario 1, except that we consider the case when there are other independent factors, which also contribute to the decision

TABLE 2. Results of the simulation.

ρ	Methods	Scenario								
		1	2	3	1	2	3	1	2	3
		Sensitivity			**Specificity**			**Accuracy**		
0.1	GA	0.902	0.584	0.527	0.991	0.956	0.862	94.32%	80.48%	77.54%
	GP	0.915	0.748	0.726	0.997	0.987	0.907	95.50%	88.47%	79.66%
	MA	0.908	0.679	0.652	0.993	0.971	0.893	94.73%	82.91%	80.65%
	Elastic net	0.910	0.726	0.724	0.994	0.975	0.904	94.53%	83.03%	79.78%
	SCAD-L_2	0.916	0.795	0.758	0.997	0.982	0.912	94.49%	82.12%	80.41%
	HLR	0.919	0.863	0.791	**0.998**	0.987	0.918	95.81%	90.15%	85.76%
	HGAWE	**0.935**	**0.906**	**0.823**	**0.998**	**0.989**	**0.926**	**97.08%**	**91.23%**	**87.81%**
0.4	GA	0.724	0.531	0.457	0.985	0.923	0.813	89.71%	76.49%	70.37%
	GP	0.798	0.712	0.674	0.992	0.957	0.866	93.64%	82.87%	77.82%
	MA	0.741	0.635	0.572	0.987	0.929	0.848	89.84%	80.04%	75.63%
	Elastic net	0.805	0.712	0.623	0.991	0.940	0.863	92.06%	82.19%	75.19%
	SCAD-L_2	0.837	0.741	0.698	0.992	0.949	0.894	92.44%	82.84%	76.51%
	HLR	0.862	0.820	0.725	0.994	0.960	0.903	93.89%	83.45%	79.22%
	HGAWE	**0.904**	**0.852**	**0.782**	**0.995**	**0.972**	**0.912**	**95.31%**	**85.79%**	**80.06%**
0.7	GA	0.563	0.467	0.417	0.961	0.891	0.775	75.08%	69.04%	62.65%
	GP	0.620	0.665	0.633	0.984	0.928	0.832	90.15%	73.94%	70.24%
	MA	0.596	0.579	0.536	0.971	0.897	0.794	89.66%	70.96%	66.30%
	Elastic net	0.675	0.637	0.561	0.977	0.905	0.816	88.17%	71.85%	65.26%
	SCAD-L_2	0.691	0.694	0.583	0.986	0.929	0.822	89.79%	74.27%	68.83%
	HLR	0.763	0.729	0.671	0.988	0.937	0.837	90.04%	77.18%	73.75%
	HGAWE	**0.820**	**0.754**	**0.724**	**0.991**	**0.943**	**0.851**	**92.34%**	**80.65%**	**76.94%**

The best performances are in bold; HLR: the hybrid $L_{1/2+2}$ regularization; HGAWE: the wrapper−embedded feature selection approach.

variable y

$$\boldsymbol{\beta} = (\underbrace{1,-1,1,-1,\ldots,1,-1}_{100},\underbrace{1.5,-2,1.7,3,-1}_{5\times20},\underbrace{0,\ldots,0}_{1800},$$
$$\underbrace{2,-2,2,-2,\ldots,2,-2}_{100},\underbrace{1.5,-2,1.7,3,-1}_{5\times20},\underbrace{0,\ldots,0}_{1800},$$
$$\underbrace{2,2,\ldots,2}_{100},\underbrace{1.5,-2,1.7,3,-1}_{5\times20},\underbrace{0,\ldots,0}_{1800}).$$

In this example, there are three groups of correlated features (similar to Scenario 1) and 300 single independent features. An ideal sparse regression method would select the 600 true features and set the coefficients of the 5400 irrelevant features to zero.

(c) In Scenario 3, the true features were added up to 1000 of the total features, $\sigma = 0.1$, and the dataset consists of 500/100 observations, and

$$\boldsymbol{\beta} = (\underbrace{1,-1,1,-1,\ldots,1,-1}_{100},\underbrace{1.5,-2,1.7,3,-1}_{5\times20},\underbrace{0,\ldots,0}_{1800},$$
$$\underbrace{2,-2,2,-2,\ldots,2,-2}_{100},\underbrace{1.5,-2,1.7,3,-1}_{5\times20},\underbrace{0,\ldots,0}_{1800},$$
$$\underbrace{2,2,\ldots,2}_{100},\underbrace{1.5,-2,1.7,3,-1}_{5\times20},\underbrace{1,1,\ldots,1}_{400},\underbrace{0,\ldots,0}_{1400}).$$

$$x_j = \rho \times x_1 + (1-\rho) \times x_j, \quad j = 2, 3, \ldots, 100;$$
$$x_j = \rho \times x_{2001} + (1-\rho) \times x_j, \quad j = 2002, 2003, \ldots, 2100;$$

TABLE 3. The detailed information of five real gene expression datasets used in the experiments.

Dataset	No. samples	No. genes	Classes
AML	6283	116	High risk / Low risk
DLBCL	7399	240	High risk / Low risk
Lymphoma	7129	77	DLBCL / FL
Prostate	12600	102	Normal / Tumor
Lung cancer	22401	164	Normal / Tumor

$$x_j = \rho \times x_{4001} + (1-\rho) \times x_j, \quad j = 4002, 4003, \ldots, 4100;$$
$$x_j = 0.1 \times x_{4201} + 0.9 \times x_j, \quad j = 4202, 4203, \ldots, 4600.$$

In this example, there are three groups of correlated features (similar to Scenario 1), 400 correlated features(the corrected parameter is 0.1) and 300 independent features. An ideal sparse regression method would select only the 1000 true features and set the coefficients of the 5000 irrelevant features to zero.

In our experiment, we set the correlation coefficient ρ of features to 0.1, 0.4, 0.7 respectively. The learning model in GA, MA, Elastic net, SCAD-L_2, HLR and HGAWE is the logistic classification approach. In GP, the multitree classifier is used. For each iteration of GA and MA, the number of selected features based on the filter of information gain is set to 2000. The configuration parameters used by EC algorithms in these seven approaches are listed in Table 1.

TABLE 4. Results of empirical datasets.

	Methods	Training accuracy	Test accuracy	No. selected genes
AML	GA	95.93%	91.87%	32
	GP	97.02%	92.82%	21
	MA	96.35%	91.13%	25
	Elastic net	96.67%	92.04%	28
	SCAD-L_2	96.62%	92.94%	23
	HLR	97.46%	93.78%	22
	HGAWE	**97.84%**	**94.32%**	**19**
DLBCL	GA	91.97%	88.58%	24
	GP	95.34%	91.22%	14
	MA	93.40%	90.34%	18
	Elastic net	94.62%	92.54%	21
	SCAD-L_2	95.39%	92.03%	16
	HLR	97.21%	93.15%	17
	HGAWE	**97.28%**	**93.73%**	**13**
Lymphoma	GA	95.43%	91.56%	63
	GP	96.14%	93.37%	38
	MA	96.08%	92.64%	54
	Elastic net	95.93%	92.17%	41
	SCAD-L_2	96.42%	91.65%	28
	HLR	98.18%	93.26%	29
	HGAWE	**98.51%**	**94.03%**	**27**
Prostate	GA	95.79%	90.34%	42
	GP	97.26%	93.81%	27
	MA	95.07%	90.83%	34
	Elastic net	96.52%	92.51%	31
	SCAD-L_2	95.82%	92.89%	26
	HLR	97.15%	92.63%	23
	HGAWE	**98.32%**	**94.17%**	**22**
Lung cancer	GA	97.14%	90.73%	51
	GP	98.25%	92.11%	45
	MA	97.42%	91.59%	49
	Elastic net	96.94%	90.85%	34
	SCAD-L_2	97.63%	92.27%	36
	HLR	98.16%	92.48%	34
	HGAWE	**98.83%**	**93.61%**	**33**

The best performances are in bold; HLR: the hybrid $L_{1/2+2}$ regularization; HGAWE: the wrapper−embedded feature selection approach.

In the regularization algorithms of these seven approaches, the control parameters of Elastic net, SCAD-L_2, and HLR approaches are tuned by the 10-fold cross-validation (CV) approach in the training set. Note that, the Elastic net and HLR methods are tuned by the 10-CV approach on the two-dimensional parameter surfaces. The SCAD-L_2 is tuned by the 10-CV approach on the three-dimensional parameter surfaces. Then, different classifiers are built by these seven feature selection approaches. Finally, the obtained classifiers are applied to the test set for classification and prediction. We repeat the simulations 100 times for each method and compute the mean classification accuracy on the test sets. To evaluate the quality of the selected features for these approaches, the sensitivity and specificity of the feature selection performance [70] are defined as follows:

$$\begin{aligned}
TruePositive(TP) &:= \left|\beta. * \hat{\beta}\right|_0, \\
TrueNegative(TN) &:= \left|\bar{\beta}. * \bar{\hat{\beta}}\right|_0, \\
FalsePositive(FP) &:= \left|\bar{\beta}. * \hat{\beta}\right|_0, \\
FalseNegative(FN) &:= \left|\beta. * \bar{\hat{\beta}}\right|_0, \\
Sensitivity &:= \frac{TP}{TP+FN}, \quad Specificity := \frac{TN}{TN+FP}.
\end{aligned}$$

TABLE 5. The 10 top genes in the AML dataset.

Rank	GA	GP	MA	Elastic net	SCAD-L_2	HLR	HGAWE
1	VEGF	SNRPN	MEIS1	GSTM1	DNMT1	MLH1	FLT3
2	PRDX2	FHIT	ALOX12	SFRP2	CDH13	CCDC69	CDKN2B
3	TP73	PNLIP	CDKN2A	GRAF	SFRP5	JUNB	INK4B
4	SFRP5	INK4B	CDHI3	GSTM1	ABCA8	GLIPR1	GSTM1
5	FHIT	A4GALT	SNRPN	PTPN6	RARA	PTPN6	SFRP1
6	GLIPR1	SFRP1	GSTM1	FHIT	GSTM1	RUNX3	NPM1
7	GRK5	PRDX2	GRAF	JUNB	WNT5A	GSTM1	SFRP2
8	TNXB	SLN	SFRP5	SFRP5	PNLIP	MEIS1	SFRP5
9	GRAF	PTPN6	PDLIM2	ALOX12	DAPK1	ALOX12	CEBPA
10	PTRF	DAPK1	PTRF	CDKN2A	HOXA9	SFRP5	GLIPR1

where the $.*$ is the element-wise product, and $|\cdot|_0$ calculates the numberof non-zero elements in a vector, $\bar{\beta}$ and $\bar{\hat{\beta}}$ are the logical ''not'' operators on the true coefficients vector β and the simulated $\hat{\beta}$.

Table 2 shows the feature selection and classification performances of different methods in the different parameter settings with Scenarios 1-3. We found that with the decrease of the correlation coefficient ρ, the models' performances can be better. In Table 2, the HGAWE approach always selects the most correct relevant features in different data environment with Scenarios 1-3. The highest sensitivities and specificities of feature selection obtained by HGAWE means that HGAWE selects most relevant features and deletes most irrelevant features respectively. Thus, the classification accuracy obtained by the HGAWE approach also outperforms other EC and regularization methods.

B. ANALYSIS OF REAL DATA

In this section, we use five publicly available gene expression microarray datasets: AML, DLBCL, Prostate, Lymphoma and Lung cancer, to further evaluate the effectiveness of our proposed HGAWE method. The AML dataset, first mentioned by Bullinger *et al.* [71], has 116 patients, which contain 6,283 genes. The DLBCL contains about 240 samples' information, which was first published in [72] by Rosenwald. Each sample includes the expression data of 8,810 genes. The Prostate dataset was originally proposed by Singh *et al.* [73]; it contains the expression profiles of 12,600 genes for 50 normal tissues and 52 prostate tumour tissues. The Lymphoma dataset [74] contains 77 microarray gene expression profiles of the 2 most prevalent adult lymphoid malignancies: 58 samples of diffuse large B-cell lymphomas and 19 follicular lymphomas (FL). The original data contains 7,129 gene expression values. The Lung cancer dataset [75] contains 164 samples with 87 lung adenocarcinomas and 77 adjacent normal tissues with 22401 microarray gene expression profiles. The Lung cancer dataset can be downloaded at www.ncbi.nlm.nih.gov/geowith through access number (GSE40419). A brief introduction of these datasets is summarized in Table 3.

In order to accurately assess the performance of the seven different feature selection approaches, the real datasets are randomly divided into two pieces: two thirds of the samples are put in the training set used for the model estimation, and the remaining one third of data are used to test the estimation performance. For regularization approaches, the penalized parameters are tuned by the 10-fold cross validation. For each real dataset, the procedures using different methods are repeated over 100 times respectively.

Table 4 describes the averaged training accuracies (10-CV) and test accuracies obtained by different feature selection approaches regularization models in the five datasets. It is obvious that the performance of the HGAWE approach is better than the other six approaches. The relevant gene selection performances of different approaches in the five real datasets are also shown in Table 4. The number of genes selected by our proposed HGAWE model is the smallest compared to the other six feature selection approaches. In regularization approaches with grouping effect, such as Elastic net, SCAD-L_2, and HLR, the performance of HLR is better than that of Elastic net and SCAD-L_2 in gene selection. On the contrary, in EC approaches, such as GA, GP and MA, the performance of GP is better than that of GA and MA in gene selection and classification. Comparing the performances of the seven feature selection algorithms, Table 4 proves that our proposed HGAWE approach has better performances in both gene selection and predictive classification.

C. DISCUSSION

For biological analysis of the results, 10 top-ranked selected genes obtained by the different methods in the AML dataset are shown in Table 5. Compared with the other feature selection methods, the HGAWE approach selects some unique genes, such as SFRP1 and SFRP2, which are members of the Sfrp family, a kind of signal transduction proteins. The Sfrp family proteins play a key role in transmitting the TGF-beta signals from the cell-surface receptor to cell nucleus, mutation or deletion of AML disease, which has been proved to lead to pancreatic cancer [73]. We think the Sfrp family may be strongly associated with AML diseases. In the other genes selected by the HGAWE approach, the gene

FLT3 can stimulate the motility of AML diseases. The expression of FLT3 has been found to be up regulated in some different kinds of AML diseases [71]. The protein encoded by the gene NPM1 is said to be very similar to the tumor suppressor of drosophila, which is a highly relevant gene to AML diseases [76]. Moreover, some relevant genes selected by other regularization models using Elastic net, SCAD-L_2, and HLR approaches are also found by the HGAWE, for example, SFRP5 and GSTM1. They are significantly associated to AML diseases, which has been discussed in [77].

We also obtain similar experimental results from the analysis of the other four real gene expression datasets. The biological analysis shows that the HGAWE approach not only can find the relevant genes that are selected by other feature selection methods, but also can find some unique genes, which are not selected by other models but are significantly associated to diseases. Hence, the HGAWE approach may identify the relevant genes accurately and efficiently.

V. CONCLUSION

In this paper, we developed a hybrid genetic algorithm with wrapper−embedded (HGAWE) to combine genetic operations and hybrid $L_{1/2+2}$ regularization approaches for feature selection in learning model construction, cancer classification and gene selection. Genetic operators such as crossover, mutation and selection for global optimization and an efficient regularization method for local search are designed to complete this HGAWE approach. The experiment results show that the HGAWE approach outperforms some existing feature selection regularization estimation approaches. It can effectively select the relevant features of bio-mark genes, predict the patients' class, and construct the learning model accurately in high dimensional biological datasets. The HGAWE approach is proved a more practical tool for feature selection and learning prediction.

REFERENCES

[1] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, nos. 1–4, pp. 131–156, 1997.
[2] H. Xia, J. Zhuang, and D. Yu, "Multi-objective unsupervised feature selection algorithm utilizing redundancy measure and negative epsilon-dominance for fault diagnosis," *Neurocomputing*, vol. 146, pp. 113–124, Dec. 2014.
[3] N. Spolaôr, A. C. Lorena, and H. D. Lee, "Multi-objective genetic algorithm evaluation in feature selection," in *Proc. Int. Conf. Evol. Multi-Criterion Optim.*, 2011, pp. 462–476.
[4] C.-M. Wang and Y.-F. Huang, "Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5900–5908, 2009.
[5] M. Mandal and A. Mukhopadhyay, "A graph-theoretic approach for identifying non-redundant and relevant gene markers from microarray data using multiobjective binary PSO," *PLoS ONE*, vol. 9, no. 3, p. e90949, 2014.
[6] M. Banerjee, S. Mitra, and H. Banka, "Evolutionary rough feature selection in gene expression data," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 4, pp. 622–632, Jul. 2007.
[7] B. Chakraborty, "Genetic algorithm with fuzzy fitness function for feature selection," in *Proc. IEEE Int. Symp. Ind. Electron.*, vol. 1. Feb. 2002, pp. 315–319.
[8] A. M. P. Canuto and D. S. C. Nascimento, "A genetic-based approach to features selection for ensembles using a hybrid and adaptive fitness function," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2012, pp. 1–8.
[9] L. D. Vignolo, D. H. Milone, and J. Scharcanski, "Feature selection for face recognition based on multi-objective evolutionary wrappers," *Expert Syst. Appl.*, vol. 40, no. 13, pp. 5077–5084, 2013.
[10] S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 2052–2064, 2014.
[11] S. F. da Silva, M. X. Ribeiro, J. D. E. S. B. Neto, C. Traina-Jr., and A. J. M. Traina, "Improving the ranking quality of medical image retrieval using a genetic feature selection method," *Decision Support Syst.*, vol. 51, no. 4, pp. 810–820, 2011.
[12] P. Sousa, P. Cortez, R. Vaz, M. Rocha, and M. Rio, "Email spam detection: A symbiotic feature selection approach fostered by evolutionary computation," *Int. J. Inf. Technol. Decision Making*, vol. 12, no. 4, pp. 863–884, 2013.
[13] R. Leardi, R. Boggia, and M. Terrile, "Genetic algorithms as a strategy for feature selection," *J. Chemometrics*, vol. 6, no. 5, pp. 267–281, 1992.
[14] D. Chyzhyk, A. Savio, and M. Graña, "Evolutionary ELM wrapper feature selection for Alzheimer's disease CAD on anatomical brain MRI," *Neurocomputing*, vol. 128, pp. 73–80, Mar. 2013.
[15] T.-C. Chen, Y.-C. Hsieh, P.-S. You, and Y.-C. Lee, "Feature selection and classification by using grid computing based evolutionary approach for the microarray data," in *Proc. 3rd IEEE Int. Conf. Comput. Sci. Inf. Technol. (ICCSIT)*, vol. 9. Jul. 2010, pp. 85–89.
[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B (Statist. Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.
[17] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.
[18] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, 2010.
[19] Z. Xu, H. Zhang, Y. Wang, X. Chang, and Y. Liang, "$L_{1/2}$ regularization," *Sci. China Inf. Sci.*, vol. 53, no. 6, pp. 1159–1169, Jun. 2010.
[20] Z. Xu, X. Chang, F. Xu, and H. Zhang, "$L_{1/2}$ regularization: A thresholding representation theory and a fast solver," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1013–1027, Jul. 2012.
[21] L. Zeng and J. Xie, "Group variable selection via SCAD-L_2," *Statistics*, vol. 48, no. 1, pp. 49–66, 2014.
[22] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc., B (Statist. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.
[23] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., B (Statist. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.
[24] H.-H. Huang, X.-Y. Liu, and Y. Liang, "Feature selection and cancer classification via sparse logistic regression with the hybrid $L_{1/2+2}$ regularization," *PLoS ONE*, vol. 11, no. 5, p. e0149675, 2016.
[25] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.
[26] B. De La Iglesia, "Evolutionary computation for feature selection in classification problems," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 3, no. 6, pp. 381–407, 2013.
[27] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016.
[28] P. L. Lanzi, "Fast feature selection with genetic algorithms: A filter approach," in *Proc. IEEE Int. Conf. Evol. Comput.*, Apr. 1997, pp. 537–540.
[29] K. Neshatian and M. Zhang, "Improving relevance measures using genetic programming," in *Proc. Eur. Conf. Genet. Program.*, 2012, pp. 97–108.
[30] L. Cervante, B. Xue, M. Zhang, and L. Shang, "Binary particle swarm optimisation for feature selection: A filter based approach," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2012, pp. 1–8.
[31] R. Jensen and Q. Shen, "Finding rough set reducts with ant colony optimization," in *Proc. UK Workshop Comput. Intell.*, 2003, vol. 1. no. 2, pp. 15–22.
[32] X. Liu, C. Yu, and Z. Cai, "Differential evolution based band selection in hyperspectral data classification," in *Proc. Int. Symp. Intell. Comput. Appl.*, 2010, pp. 86–94.
[33] U. Kamath, K. De Jong, and A. Shehu, "Effective automated feature construction and selection for classification of biological sequences," *PLoS ONE*, vol. 9, no. 7, p. e99982, 2014.

[34] D. P. Muni, N. R. Pal, and J. Das, "Genetic programming for simultaneous feature selection and classifier design," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 1, pp. 106–117, Feb. 2006.

[35] A. Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems," *Eur. J. Oper. Res.*, vol. 206, no. 3, pp. 528–539, 2010.

[36] Z. Yan and C. Yuan, "Ant colony optimization for feature selection in face recognition," in *Biometric Authentication*. Berlin, Germany: Springer, 2004, pp. 221–226.

[37] L. Wang, H. Ni, R. Yang, V. Pappu, M. B. Fenn, and P. M. Pardalos, "Feature selection based on meta-heuristics for biomedicine," *Optim. Methods Softw.*, vol. 29, no. 4, pp. 703–719, 2014.

[38] I. Vatolkin, W. Theimer, and G. Rudolph, "Design and comparison of different evolution strategies for feature selection and consolidation in music classification," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, May 2009, pp. 174–181.

[39] K. Shelke, S. Jayaraman, S. Ghosh, and J. Valadi, "Hybrid feature selection and peptide binding affinity prediction using an EDA based algorithm," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2013, pp. 2384–2389.

[40] A. Purohit, N. S. Chaudhari, and A. Tiwari, "Construction of classifier with feature selection based on genetic programming," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2010, pp. 1–5.

[41] Q. Chen, B. Xue, L. Shang, and M. Zhang, "Improving generalisation of genetic programming for symbolic regression with structural risk minimisation," in *Proc. Genet. Evol. Comput. Conf.*, 2016, pp. 709–716.

[42] Q. Chen, M. Zhang, and B. Xue, "Feature selection to improve generalization of genetic programming for high-dimensional symbolic regression," *IEEE Trans. Evol. Comput.*, vol. 21, no. 5, pp. 792–806, Oct. 2017.

[43] H. B. Nguyen, B. Xue, I. Liu, and M. Zhang, "Filter based backward elimination in wrapper based PSO for feature selection in classification," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2014, pp. 3111–3118.

[44] M. M. Kabir, M. Shahjahan, and K. Murase, "A new hybrid ant colony optimization algorithm for feature selection," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3747–3763, 2012.

[45] R. N. Khushaba, A. Al-Ani, A. AlSukker, and A. Al-Jumaily, "A combined ant colony and differential evolution feature selection algorithm," in *Proc. Int. Conf. Ant Colony Optim. Swarm Intell.*, 2008, pp. 1–12.

[46] Z. Zhu, Y.-S. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognit.*, vol. 40, no. 11, pp. 3236–3248, Nov. 2007.

[47] Z. Zhu, S. Jia, and Z. Ji, "Towards a memetic feature selection paradigm," *IEEE Comput. Intell. Mag.*, vol. 5, no. 2, pp. 41–53, May 2010.

[48] Z. Zhu and Y.-S. Ong, "Memetic algorithms for feature selection on microarray data," in *Advances in Neural Networks—ISNN*. Berlin, Germany: Springer, 2007, pp. 1327–1335.

[49] J. Huang, Y. Cai, and X. Xu, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," *Pattern Recognit. Lett.*, vol. 28, no. 13, pp. 1825–1844, 2007.

[50] M. A. Esseghir, G. Goncalves, and Y. Slimani, "Memetic feature selection: Benchmarking hybridization schemata," in *Proc. Int. Conf. Hybrid Artif. Intell. Syst.*, 2010, pp. 351–358.

[51] Y. Wen and H. Xu, "A cooperative coevolution-based pittsburgh learning classifier system embedded with memetic feature selection," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2011, pp. 2415–2422.

[52] X.-Y. Liu, S. Wang, and Y. Liang, "Novel regularization method for biomarker selection and cancer classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, Jan. 2018.

[53] D. Meng, Q. Zhao, and L. Jiang. (2015). "What objective does self-paced learning indeed optimize?" [Online]. Available: https://arxiv.org/abs/1511.06049

[54] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1418–1429, Dec. 2006.

[55] Q. Lyu, Z. Lin, Y. She, and C. Zhang, "A comparison of typical ℓ_p minimization algorithms," *Neurocomputing*, vol. 119, pp. 413–424, Nov. 2013.

[56] J. Zeng, S. Lin, Y. Wang, and Z. Xu, "$L_{1/2}$ regularization: Convergence of iterative half thresholding algorithm," *IEEE Trans. Signal Process.*, vol. 62, no. 9, pp. 2317–2329, May 2014.

[57] Y. Liang *et al.*, "Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification," *BMC Bioinf.*, vol. 14, no. 1, p. 198, 2013.

[58] N. Becker, G. Toedt, P. Lichter, and A. Benner, "Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data," *BMC Bioinf.*, vol. 12, no. 1, p. 138, 2011.

[59] D. Sudholt, "The impact of parametrization in memetic evolutionary algorithms," *Theor. Comput. Sci.*, vol. 410, no. 26, pp. 2511–2528, 2009.

[60] Z. Zhu, Y.-S. Ong, and M. Dash, "Wrapper-filter feature selection algorithm using a memetic framework," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 1, pp. 70–76, Feb. 2007.

[61] Z. Zhu, Y. S. Ong, and J. M. Zurada, "Identification of full and partial class relevant genes," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 7, no. 2, pp. 263–277, Apr. 2010.

[62] C.-S. Yang, L.-Y. Chuang, Y.-J. Chen, and C.-H. Yang, "Feature selection using memetic algorithms," in *Proc. 3rd Int. Conf. Converg. Hybrid Inf. Technol. (ICCIT)*, vol. 1. Nov. 2008, pp. 416–423.

[63] S. S. Kannan and N. Ramaraj, "A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm," *Knowl.-Based Syst.*, vol. 23, no. 6, pp. 580–585, 2010.

[64] A. Arab and A. Alfi, "An adaptive gradient descent-based local search in memetic algorithm applied to optimal controller design," *Inf. Sci.*, vol. 299, pp. 117–142, Apr. 2015.

[65] A. Zibakhsh and M. S. Abadeh, "Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function," *Eng. Appl. Artif. Intell.*, vol. 26, no. 4, pp. 1274–1281, 2013.

[66] C. Bi, "Memetic algorithms for de novo motif-finding in biomedical sequences," *Artif. Intell. Med.*, vol. 56, no. 1, pp. 1–17, 2012.

[67] R. Mazumder, J. H. Friedman, and T. Hastie, "SparseNet: Coordinate descent with nonconvex penalties," *J. Amer. Statist. Assoc.*, vol. 106, no. 495, pp. 1125–1138, 2011.

[68] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statist. Softw.*, vol. 33, no. 1, p. 1, 2010.

[69] A. Lipowski and D. Lipowska, "Roulette-wheel selection via stochastic acceptance," *Phys. A, Statist. Mech. Appl.*, vol. 391, no. 6, pp. 2193–2196, 2012.

[70] W. Zhang, Y.-W. Wan, G. I. Allen, K. Pang, M. L. Anderson, and Z. Liu, "Molecular pathway identification using biological network-regularized logistic models," *BMC Genomics*, vol. 14, no. 8, p. S7, 2013.

[71] L. Bullinger *et al.*, "Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia," *New England J. Med.*, vol. 350, no. 16, pp. 1605–1616, 2004.

[72] A. Rosenwald *et al.*, "The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma," *New England J. Med.*, vol. 346, no. 25, pp. 1937–1947, 2002.

[73] D. Singh *et al.*, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[74] M. A. Shipp *et al.*, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Med.*, vol. 8, no. 1, pp. 68–74, 2002.

[75] J.-S. Seo *et al.*, "The transcriptional landscape and mutational profile of lung adenocarcinoma," *Genome Res.*, vol. 22, no. 11, pp. 2109–2119, 2012.

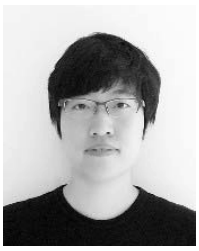
[76] B. A. Boone *et al.*, "Loss of SMAD4 staining in pre-operative cell blocks is associated with distant metastases following pancreaticoduodenectomy with venous resection for pancreatic cancer," *J. Surg. Oncol.*, vol. 110, no. 2, pp. 171–175, 2014.

[77] H. Vuong, F. Cheng, C.-C. Lin, and Z. Zhao, "Functional consequences of somatic mutations in cancer using protein pocket-based prioritization approach," *Genome Med.*, vol. 6, no. 10, p. 81, 2014.

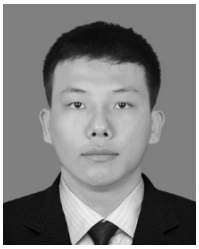
XIAO-YING LIU received the B.Sc. and M.Sc. degrees in automation control from the Taiyuan University of Technology, Taiyuan, China, and the Ph.D. degree in computer technology and application from the Macau University of Science and Technology, Macau. She is currently an Associate Professor with the Guangdong Polytechnic of Science and Technology. She has authored or co-authored over 10 technical papers and international patents, and published two computer programming books. Her current research interests include computer science and artificial intelligence.

YONG LIANG received the B.Sc. and M.Sc. degrees in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 1996 and 1999, respectively, and the Ph.D. degree in computer science from the Chinese University of Hong Kong, Hong Kong, in 2003. From 2004 to 2006, he was a Post-Doctoral Researcher with the Department of Computer Science and Engineering, Chinese University of Hong Kong, and later joined Shantou University, China. He joined the Faculty of Information Technology, Macau University of Science and Technology, in 2007, where he is currently a Professor. He has authored and co-authored over 50 papers. His research interests are in computational intelligent, including machine learning, big data analysis, evolutionary computation, and bioinformatics.

SAI WANG was born in Taiyuan, China, in 1988. He received the B.Sc. degree in computer science and technology and the M.Sc. degree in systems engineering from Shanxi University, Taiyuan, in 2011 and 2014, respectively. He is currently pursuing the Ph.D. degree of computer technology and application with the Macau University of Science and Technology. His current interests include bioinformatics and machine learning.



ZI-YI YANG received the B.Sc. degree from the Tongji Zhejiang College, China, in 2013, and the M.Sc. degree from the Macau University of Science and Technology, Macau, in 2015, where she is currently pursuing the Ph.D. degree. From 2015 to 2017, she was a Bioinformatics Analysis Engineer with the Beijing Genomics Institute, China. Her current research interests include machine learning, feature selection, and bioinformatics.



HAN-SHUO YE was born in Guangzhou, China, in 1993. He received the B.Sc. degree in electronic information technology from the Macau University of Science and Technology, Macau, in 2016, where he is currently pursuing the master's degree. His main research interests are in the areas of data research feature selection for genetic test.