*animals*

*Article*

# A Modified Memetic Algorithm with an Application to Gene Selection in a Sheep Body Weight Study

**Maoxuan Miao [1], Jinran Wu [2], Fengjing Cai [1,*] and You-Gan Wang [2]**

[1] College of Mathematics and Physics, Wenzhou University, Wenzhou 325035, China; 194611073127@stu.wzu.edu.cn
[2] School of Mathematical Sciences, Queensland University of Technology, Brisbane 4001, Australia; j73.wu@qut.edu.au (J.W.); you-gan.wang@qut.edu.au (Y.-G.W.)
* Correspondence: caifj7704@wzu.edu.cn

**Simple Summary:** Due to lacking exploitation capability, traditional genetic algorithm cannot accurately identify the minimal best gene subset. Thus, the improved splicing method is introduced into a genetic algorithm to enhance exploitation capability for achieving balance between exploitation and exploration of GA. It can effectively identify true gene subsets with high probability. Furthermore, a dataset of the body weight of Hu sheep has been used to show that the proposed method can obtain a better minimal subset of genes with a few iterations, compared with all considered algorithms including genetic algorithm and adaptive best-subset selection algorithm.

**Abstract:** Selecting the minimal best subset out of a huge number of factors for influencing the response is a fundamental and very challenging NP-hard problem because the presence of many redundant genes results in over-fitting easily while missing an important gene can more detrimental impact on predictions, and computation is prohibitive for exhaust search. We propose a modified memetic algorithm (MA) based on an improved splicing method to overcome the problems in the traditional genetic algorithm exploitation capability and dimension reduction in the predictor variables. The new algorithm accelerates the search in identifying the minimal best subset of genes by incorporating it into the new local search operator and hence improving the splicing method. The improvement is also due to another two novel aspects: (a) updating subsets of genes iteratively until the no more reduction in the loss function by splicing and increasing the probability of selecting the true subsets of genes; and (b) introducing *add* and *del* operators based on backward sacrifice into the splicing method to limit the size of gene subsets. Additionally, according to the experimental results, our proposed optimizer can obtain a better minimal subset of genes with a few iterations, compared with all considered algorithms. Moreover, the mutation operator is replaced by it to enhance exploitation capability and initial individuals are improved by it to enhance efficiency of search. A dataset of the body weight of Hu sheep was used to evaluate the superiority of the modified MA against the genetic algorithm. According to our experimental results, our proposed optimizer can obtain a better minimal subset of genes with a few iterations, compared with all considered algorithms including the most advanced adaptive best-subset selection algorithm.

**Keywords:** gene selection; sheep weight; memetic algorithm; modifications; local search operator

**Citation:** Miao, M.; Wu, J.; Cai, F.; Wang, Y.-G. A Modified Memetic Algorithm with an Application to Gene Selection in a Sheep Body Weight Study. *Animals* **2022**, *12*, 201. https://doi.org/10.3390/ani12020201

Academic Editor: Martino Cassandro

Received: 26 November 2021
Accepted: 14 January 2022
Published: 15 January 2022

## 1. Introduction

In data mining, feature selection is a fundamental strategy to handle "the curse of dimensionality" [1]. With an effective feature selection procedure, the redundant and irrelevant features are eliminated to improve the performance of the learning process [2,3]. Further, the feature selection approaches can identify small subsets of biologically important genes, which are the most relevant to the target trait, such as genetic diseases and angiotensin-converting enzyme 2 [4,5]. Thus, in this paper, considering a new application to gene selection in a sheep body weight study, we propose a modified memetic algorithm to effectively select the most important genes from more than 52,000 genes.

The filter algorithm and hybrid feature selection algorithm are the main feature selection methods. The filter algorithm is based on data characteristics, such as distance [6], correlation [7], and statistical distribution [8], to select subsets of genes [9]. Although gene selection using filter algorithms is fast and simple, the top k genes contain some redundant and irrelevant genes for not considering correlation between genes and unreliable feature evaluation principle. The hybrid feature selection method is usually utilized to select a few important genes out of a huge number of genes [10]. In the hybrid algorithm, the filter algorithm is firstly utilized to eliminate many genes, then the wrapper algorithm is used to further compact the selected subset of genes [11]. It is worthwhile to note that the filter algorithm may eliminate many useful genes in initial step, and the wrapper method can learn subsets of gene interact with learning algorithm [12]. One of the most typical methods is the hybrid dragonfly black hole algorithm for gene selection for the RNA-seq COVID-19 data, and the authors achieved a good performance for the investigated data [4]. However, the algorithm usually encounters nest effect and produces a sub-optimal subset. Different from two above-mentioned types of feature-selection methods, heuristic methods can effectively overcome nest effect [13–15]. Genetic algorithm (GA) is one of heuristic methods that can widely be applied to gene selection [16], which effectively searches the entire gene subset space by combining exploration with exploitation [17]. Exploration can provide a promising subset of genes of the entire gene subset space, while exploitation can guarantee that the promising subset of genes move toward to the best subset of genes. However, GA lacks exploitation capability [18]; this means that GA cannot obtain the core subset of genes. Therefore, local search is incorporated into GA to enhance exploitation capability, and Memetic Algorithm (MA) is proposed [19].

MA is an improvement of GA, which can undergo self-improvement [20]. Currently, various local search operators have been incorporated into GA to improve exploitation capability. For example, the filter algorithm is embedded into GA for fast identification of the important subset of features [21]. The great deluge algorithm is combined with GA to improve fine-tuning capability of GA [22]. Lamarckian learning is incorporated into GA to utilize the most appropriate local search method among local search methods [23]. Moreover, a hybrid method based on dynamic GA and random forest was developed by Pashaei and Pashaei [5] to distinguish a small meaningful set of genes for cancer classification. They have been proven to outperform other state-of-the art feature selection methods, such as filter and heuristic search algorithms, but they cannot maximize exploitation capability as the local search operator is utilized to obtain the best subset of genes as possible. Splicing method [24] can recover the true subset of genes with high probability but cannot limit the size of gene subsets. Therefore, in our present work, a new local search operator together with an improved splicing method, is embedded into GA, which is known as GA based on an improved splicing method, for significant single-nucleotide polymorphisms (SNPs) identification. There exist some advantages in the proposed method: (1) Compared with filter algorithm and hybrid algorithm, it can effectively explore the entire space of gene subsets and find minimal best gene subsets; (2) Compared with GA, it can enhance exploitation capability of GA for achieving balance between exploitation and exploration of GA in term of improved splicing method; and (3) Compared with traditional MA, it can recover the true subset of genes with high probability and limit the size of gene subsets.

The three main contributions in the paper can be given as follows: (1) A new local search operator along with an improved splicing method, is proposed in the paper. The *add* and *del* operators are embedded into the splicing method. It can recover the true subset of genes with high probability, and limit the sizes of subsets of genes; (2) A modified memetic algorithm, GA based on an improved splicing method, is proposed in this paper. It can accelerate search to identify the minimal best subset of genes. The improved splicing method is utilized to improve starting points to enhance efficiency of search. The improved splicing method replaces the mutation operator to enhance exploitation capability for achieving balance between exploitation and exploration of GA; and (3) In projects where identification of SNPs for body weights is required, the GA based on an improved splicing method can find the minimal best subset of genes compared with other heuristic methods, including GA [25], $\beta$-hill climbing [26], salp swarm algorithm [27], artificial bee colony algorithm [28], sine cosine optimization algorithm [29], and binary particle swarm optimization [30], and adaptive best-subset selection from Zhu et al. [24].

## 2. Methods

In this section, a modified memetic algorithm is proposed by combining an improved splicing method with genetic algorithm. In other words, the improved splicing method is incorporated into GA to accelerate the search for identifying the minimal best subset of genes. The main advantages of the proposed method are two-fold: (1) it can provide the promising subset of genes of the whole gene subset space based on selection and crossover operators; and (2) compared with traditional GA, it has strong exploitation capability and recovers the minimal best subset of genes with high probability based on an improved splicing method.

### *2.1. The Genetic Algorithm*

Genetic algorithm is a heuristic algorithm, which is mainly based on selection and crossover operators [31]. Selection operator directs GA to find the most promising subset of genes of the entire gene subset space. The crossover operator has an exploration capability, which can direct GA to escape from sub-optimal locations. The crossover operator, individual representation, initialization, and selection operator are discussed below.

**a. Individual Representation:**

For feature selection, each individual is represented as a subset of genes. Each individual is encoded by a binary vector, i.e., $a_i(t) = [f_1, f_2, \cdots, f_q]$, $f_k \in \{0, 1\}$, where $f_k = 0$ denoted as the $k$-th gene is not selected while $f_k = 1$ denoted as the $k$-th gene is selected.

**b. Initialization:**

$N$ individuals are randomly generated, which consist of a population $P(0)$, i.e., $P(0) = [a_1(0), \cdots, a_i(0), \cdots, a_N(0)]$. The procedure of randomly generating an individual $a_i(t)$ is shown in Algorithm 1. An initial probability $P_{\text{initial}} = s/q$ is defined as a gene $f_i = 1$, where $s$ is the expected number of a subset of genes. A gene is randomly assigned as 0 or 1 by the following way: $f_i = 1$ if $U(0, 1) < P_{\text{initial}}$; else, $f_i = 0$. The entire procedure is repeated until $q$ genes are all assigned.

**Algorithm 1** Initialization of individual

```
Input: individual $a_i(t) = []$, probability $P_{\text{initial}}$
for $i = 1, 2, \cdots, q$ do
    if $U(0, 1) < P_{\text{initial}}$ then
        $a_i(t) = [1, a_i(t)]$
    else
        $a_i(t) = [0, a_i(t)]$
    end if
end for
Output: individual $a_i(t)$
```

**c. Selection Operator:**

Generally, the higher the quality of gene subsets, the more likely they are within the most promising region of entire search space. Therefore, the survival probability of high quality gene subsets should be set higher. Based on this principle [32], a proportional roulette wheel selection is used in the paper. Since error indicators are considered as fitness function in the paper, $F_i$ is lower and the individual $a_i(t)$ is more likely to survive. The survival probability is formulated as,

$$P_i = \frac{\frac{1}{F_i}}{\sum_{i=1}^{N} \frac{1}{F_i}}, \tag{1}$$

where $F_i$ is the fitness value of the $i$-th individual $a_i(t)$. Each subset of genes is allocated with a survival probability. Then, roulette wheel sampling-based survival probability is utilized to select individuals $a_i(t)$.

**d. Crossover Operator:**

The crossover operator has two distinct characteristics [33]: (1) Genes common to parents are preserved in the offspring. It means that some important genes common to parents improved by the improved splicing method can still be retained to the next generation; (2) Produces offspring that are contained in a region of the search space spanned by parents. It means that the produced offspring remain in the promising area when the parents are located in the promising region of the entire search space. The uniform crossover operator is a general form of single-point or two-point crossover, which is used in the paper. It can be flexible to adjust disruptive effect in term of crossover disruption probability $P_0$, which is defined as using the bit value of the first parents [34].

**e. The Improved Splicing Method:**

Splicing method [24], also called the polynomial method, is where subsets of genes can be updated iteratively until the loss function cannot be improved by splicing. It can guarantee separation of unimportant genes from a subset of genes. However, the size of subsets of genes gradually increases over generations. It leads to the minimal best gene subset, which, potentially, cannot be found in higher generations. Therefore, we propose an improved splicing method in this paper. The advantages have three points, as follows: (1) The *add* and *del* operators are introduced into the splicing method to limit the maximal size of subsets of genes $S_{max}$. The del operator is used to delete some insignificant genes in active sets to achieve the desirable maximal size of subsets of genes, while the add operator is used to add deleted genes for *del* operator into inactive sets; (2) It can recover the true subset of genes with high probability; (3) It has strong exploitation capability. Since the mutation operator has less exploitation capability and may introduce some unimportant or irrelevant genes into the subset of genes, mutation operator is replaced by the improved splicing method. The entire process of the improved splicing method mainly is consisted of five parts, i.e., individual segmentation, evaluation, *add* and *del* operators, swap, and merge. We illustrate the details of the improved splicing method with an example, as shown in Figure 1, and the algorithm is shown in Algorithm 2.

**Step 1 Individual Segmentation:**

The individual, i.e., $[f_1, f_2, \cdots, f_q]$, is divided into active set $\mathcal{A} = \{f_i | f_i = 1, i = 1, \cdots, q\}$ and inactive set $\mathcal{I} = \{f_i | f_i = 0, i = 1, \cdots, q\}$. For example, the individual is represented as $[1, 0, 1, 0, 1, 1, 0]$, shown in Figure 1a. It is divided into active set $\mathcal{A} = \{f_1, f_3, f_5, f_6\}$ and inactive set $\mathcal{I} = \{f_2, f_4, f_7\}$, as shown in Figure 1b.

**Step 2 Judge:**

If the size of active set $\mathcal{A}$ is greater than the maximal size of the subset of genes $S_{max}$, go into **Step 3**; otherwise, go into **Step 5**.

**Step 3 Evaluation:**

The backward sacrifice is used to evaluate the score of each gene in active set $\mathcal{A}$. The score of each gene is evaluated by

$$\xi_j = \frac{X_j^T X_j}{2n} (\beta_j)^2, \tag{2}$$

where $n$ is the number of samples and $\beta_j$ is coefficient of the $j$-th gene. In Figure 1c, score of the gene $f_3$ is the highest and score of the gene $f_6$ is the lowest.

**Step 4** ***add*** **and** ***del*** **Operators:**

Some genes with the lowest score in active set $\mathcal{A}$ are deleted, and then added into inactive set $\mathcal{I}$. The *del* operator is used to delete $|\mathcal{A}| - S_{max}$ genes with the lowest score in active set $\mathcal{A}$. Deleted $|\mathcal{A}| - S_{max}$ genes are added into inactive set $\mathcal{I}$ for *add* operator. In Figure 1d, $S_{max}$ is set as 3. Since the score of the gene $f_6$ is the lowest, the gene $f_6$ in active set $\mathcal{A}$ is deleted, and then added into inactive set $\mathcal{I}$.

**Step 5 Evaluation:**

The score of each gene in active set $\mathcal{A}$ is calculated by backward sacrifice and the score of each gene in inactive set $\mathcal{I}$ is calculated by forward sacrifice. Backward sacrifice is formulated as Equation (2), while forward sacrifice is formulated as,

$$\zeta_j = \frac{X_j^T X_j}{2n}\left(\frac{d_j}{X_j^T X_j / n}\right)^2, \tag{3}$$

where $n$ is the number of samples and $d_j = X_j^T(Y - X_{\mathcal{A}}\beta_{\mathcal{A}})/n$. They are both filter methods based on change in linear loss function. The larger the change in loss function, the more significant the gene is.

**Step 6 Swap:**

$k$ genes of the lowest scores in active set $\mathcal{A}$ are consisted of set $\mathcal{A}_k$; $k$ genes of the highest scores in inactive set $\mathcal{I}$ are consisted of set $\mathcal{I}_k$. Then, sets $\mathcal{A}_k$ and $\mathcal{I}_k$ have swapped each other. Parameter $k$ is less than or equal $min(|\mathcal{A}|, |\mathcal{I}|)$. In order to find the best minimal subset of genes, we should search the optimal parameter $k$ from range $\{1, 2, \cdots, min(|\mathcal{A}|, |\mathcal{I}|)\}$ by using grid search. For example, in Figure 1, the parameter $k$ is 1. The score of the gene $f_5$ in active sets is the lowest and the score of the gene $f_2$ in inactive sets is the highest in terms of **Step 5**. Therefore, the sets $\mathcal{A}_k = \{f_5\}$ and $\mathcal{I}_k = \{f_2\}$ are obtained. Then, the sets $\mathcal{A}_k$ and $\mathcal{I}_k$ have swapped with each other. Finally, the new active set $\{f_1, f_2, f_3\}$ and inactive set $\{f_4, f_5, f_6, f_7\}$ are obtained, as shown in Figure 1e.

**Step 7 Update:**

The active set $\mathcal{A}$ is updated by repeating **Steps 5–6** until the loss function $L = \frac{1}{2n}||Y - X_{\mathcal{A}}\beta_{\mathcal{A}}||_2^2$ cannot be improved. Then, go to **Step 8**.

**Step 8 Merge:**

The updated active set $\mathcal{A}$ and inactive set $\mathcal{I}$ are merged to form a new individual, as shown in Figure 1f.

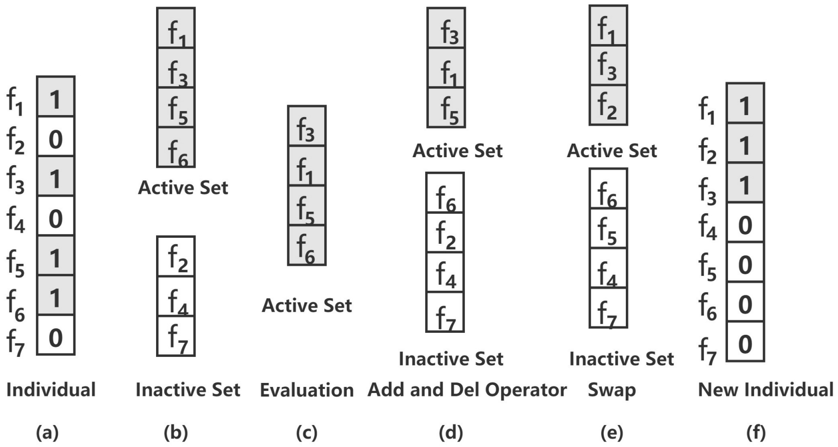

**Figure 1.** Example of the improved splicing method (**a**) Individual. (**b**) Individual Segmentation. The individuals are divided into active sets and inactive sets. (**c**) Evaluation. The score of each gene in an active set is evaluated in terms of backward sacrifice. (**d**) *add* and *del* Operators. Some genes with the lowest scores in active sets are deleted, then added into inactive sets. (**e**) Swap. A gene with the lowest score ($f_5$) in an active set and a gene with the highest score ($f_2$) in an inactive set swap each other. (**f**) Merge. The active set and inactive set are merged into a new individual.

**Algorithm 2** Improved splicing method

```
Input: An individual $A$, i.e., $A = [f_1, f_2, \cdots, f_p]$, $f_j \in \{0, 1\}$, Sample $X \in n \times p$, $Y \in n \times 1$, threshold $\tau$; Maximal size of the subset of genes $S_{max}$;
The individual $A$ is divided into active set $\mathcal{A} = \{f_i | f_i = 1, i = 1, \cdots, p\}$ and inactive set $\mathcal{I} = \{f_i | f_i = 0, i = 1, \cdots, p\}$;
if $|\mathcal{A}| > S_{max}$ then
    Calculate the score of each gene in active set $\mathcal{A}$ in terms of backward sacrifice.
    Delete $|\mathcal{A}| - S_{max}$ genes of the lowest score in active set $\mathcal{A}$ to obtain a new active set $\mathcal{A}$; Then deleted $|\mathcal{A}| - S_{max}$ genes are added into inactive set $\mathcal{I}$ to obtain a new inactive set $\mathcal{I}$.
    $\beta_{\mathcal{A}} = (X^T_{\mathcal{A}} X_{\mathcal{A}})^{-1} X^T_{\mathcal{A}} Y, d_{\mathcal{I}} = X^T_{\mathcal{I}}(Y - X_{\mathcal{A}}\beta_{\mathcal{A}})/n$, where $X_{\mathcal{A}} \in n \times |\mathcal{A}|$, $X_{\mathcal{I}} \in n \times |\mathcal{I}|$;
else
    $\beta_{\mathcal{A}} = (X^T_{\mathcal{A}} X_{\mathcal{A}})^{-1} X^T_{\mathcal{A}} Y, d_{\mathcal{I}} = X^T_{\mathcal{I}}(Y - X_{\mathcal{A}}\beta_{\mathcal{A}})/n$, where $X_{\mathcal{A}} \in n \times |\mathcal{A}|$, $X_{\mathcal{I}} \in n \times |\mathcal{I}|$;
end if
repeat
    Calculate Loss function $L = L_0 = \frac{1}{2n}||Y - X_{\mathcal{A}}\beta_{\mathcal{A}}||^2_2$;
    Calculate the score $\xi_j$ of each gene in active set $\mathcal{A}$ in terms of backward sacrifice.
    Calculate the score of each gene $\zeta_j$ in inactive set $\mathcal{I}$ in terms of forward sacrifice.
    for $k = 1, 2, \cdots, min(|\mathcal{A}|, |\mathcal{I}|)$ do
        $\mathcal{A}_k = \{j \in \mathcal{A} : \sum_{i \in \mathcal{A}} I(\xi_j \geq \xi_i) \leq k\}, \mathcal{I}_k = \{j \in \mathcal{I} : \sum_{i \in \mathcal{I}} I(\zeta_j \leq \zeta_i) \leq k\}$;
        $\tilde{\mathcal{A}} = (\mathcal{A} \backslash \mathcal{A}_k) \bigcup \mathcal{I}_k$, $\tilde{\mathcal{I}} = (\mathcal{I} \backslash \mathcal{I}_k) \bigcup \mathcal{A}_k$, $\tilde{d}_{\tilde{\mathcal{I}}} = X^T_{\tilde{\mathcal{I}}}(Y - X_{\tilde{\mathcal{A}}}\beta_{\tilde{\mathcal{A}}})$, $\tilde{\beta}_{\tilde{\mathcal{A}}} = (X^T_{\tilde{\mathcal{A}}} X_{\tilde{\mathcal{A}}})^{-1} X^T_{\tilde{\mathcal{A}}} Y$;
        Calculate Loss function $L_n = \frac{1}{2n}||Y - X_{\tilde{\mathcal{A}}}\tilde{\beta}_{\tilde{\mathcal{A}}}||^2_2$;
        if $L_n < L$ then
            $\beta_{\mathcal{A}} = \tilde{\beta}_{\tilde{\mathcal{A}}}$, $d_{\mathcal{I}} = \tilde{d}_{\tilde{\mathcal{I}}}$, $\mathcal{A} = \tilde{\mathcal{A}}$, $\mathcal{I} = \tilde{\mathcal{I}}$, $L = L_n$;
        end if
    end for
until $L_0 - L < \tau$
Merge active set $\mathcal{A}$ with inactive set $\mathcal{I}$ to generate a new individual $A$.
Output: A new individual $A$.
```

### *2.2. The Proposed Method*

A modified memetic algorithm, genetic algorithm based on an improved splicing method, is a modified version of GA. To accelerate the search to identify the minimal best subset of genes, a new local search operator, improved splicing method, is utilized to improve starting points and mutation operator is replaced with the improved splicing method to enhance exploitation capability for achieving balance between exploitation and exploration of GA. In addition, the elitist operator is introduced into GA to prevent loss of the important subset of genes. The proposed algorithm is consisted of the following steps and the flow chart is shown in Figure 2.

**Step 1 Initialization:**

Set $t = 1$ and Maximal number of generation $T$; $N$ subsets of genes $a_i(t), i = 1, \cdots, N$, is randomly generated, consisted of a population, i.e., $P(t) = [a_1(t), \cdots, a_N(t)]$. Each subset of genes is randomly generated in terms of Algorithm 1;

**Step 2 Improved Splicing Method:**

The improved splicing method shown in Algorithm 2 is used to improve each subset of genes $a_i(t)$ in $P(t), i = 1, 2, \cdots, N$;

**Step 3 Evaluation:**

Calculate fitness value $F_i$ of each subset of genes $a_i(t), i = 1, 2, \cdots, N$, in initial population $P(t)$;

**Step 4 New Population:**

Set empty list of new population $NP$, i.e., $NP = []$;

**Step 5 Elitist Operator:**

The best subset of genes $B(t)$ in the current generation is directly added into new the population $NP$, i.e., $NP = [B(t), NP]$;

**Step 6 Selection Operator:**

Proportional roulette wheel selection is utilized to select parents $a_i(t)$ and $a_j(t)$ in the current population $P(t)$;

**Step 7 Crossover Operator:**

Parents $a_i(t)$ and $a_j(t)$ are recombined to produce offspring $o_i(t)$ and $o_j(t)$ by a uniform crossover;

**Step 8 Improved Splicing Method:**

The improved splicing method, shown in Algorithm 1, is used to improve offspring $o_i(t)$ and $o_j(t)$. Then, improved offspring $o_i(t)$ and $o_j(t)$ are added into the new population $NP$, i.e., $NP = [o_i(t), o_j(t), NP]$;

**Step 9 Repeat:**

Repeat **Steps 6–8** until producing $(N + 1)$ individuals;

**Step 10 Evaluation:**

$t = t + 1$; $P(t) = NP$ in population of the next generation; calculate fitness value $F_i$ of each subset of genes $a_i(t), i = 1, 2, \cdots, N$, in the new population $P(t)$;

**Step 11 Stopping Criterion:**

Repeat **Steps 4–10** until maximum number of generations $T$ is satisfied or no improvement over 5 generations continuously. Then, go to **Step 12**;

**Step 12 Output:**

Output the best subset of genes.

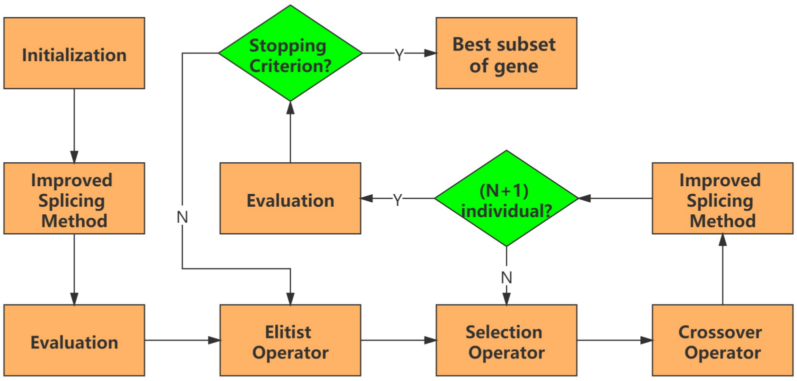

**Figure 2.** The flow chart of the proposed optimizer.

## 3. Materials

To check the performance of the proposed optimizer, seven well-known feature selection methods are compared with it, including adaptive best-subset selection (ABSS) [24], genetic algorithm (GA) [25], binary particle swarm optimization (BPSO) [30], binary salp swarm algorithm (BSWA) [27], sine cosine optimization algorithm (SCA) [29], artificial bee colony algorithm (ABC) [28], and $\beta$-hill climbing [26].

### *3.1. The Investigated Data*

In meat production, the body weight (BW) of sheep is a key economic trail [35]. As pointed out by Cao et al. [36], BWs measured at birth and other life stages are major indicators for productivity, health, and preventive management. In genetics, some SNPs are associated with BW, and the identification of these SNPs can improve the efficiency of sheep breeding programs. However, selecting the significant SNPs among sheep genes is a NP-hard problem [37].

Here, to investigate the effectiveness of the proposed method, we conduct experiments on the dataset of three measures of body weights of 240 Hu sheep, including birth weight, six-month weight, and weaning weight. The dataset is available in the GEO accession number GSE152717 [36]. We exclude some genes containing missing values. Then, a brief dataset characteristic for three types of body weight of Hu sheep is shown in Table 1.

**Table 1.** The characteristics of a dataset for three measures of BWs from Hu sheep.

| Type of BW | Number of Genes | Number of Instances |
|---|---|---|
| Birth weight | 54,183 | 240 |
| Six-month weight | 54,183 | 240 |
| Weaning weight | 54,183 | 240 |

### *3.2. The Fitness Function Setting*

The mean squared error (MSE) is used as fitness function in the paper, and support vector regression (SVR) [38] is used as a regression model to predict body weights at three different times in this paper.

### *3.3. The Hyper-Parameter Setting*

For the body weight measured on each occasion, the hyper-parameters of the proposed method are set as follows: Number of generations $T = 30$; Number of individuals $N = 20$; Threshold $\tau = 0.01|\mathcal{A}|\log(p)\log(\log n)/n$ [24], where $|\mathcal{A}|$ is the size of the active set, $n$ is the number of training sets and $p$ is the number of features; Parameters of SVR, including penalty parameter $C$, $\epsilon$ in $\epsilon$-insensitive loss function and $\sigma$ in Gaussian kernel function are shown in Table 2; The expected size of the subset of genes $s = 2000$; For maximal size of the subset of genes $S_{max}$ and crossover disruption probability $P_0$, the grid search is utilize to search the optimal combination $S_{max}$ and $P_0$. Range of $P_0$ is set as $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and range of $S_{max}$ is set as $\{10, 20, 30, 40, 50\}$. The result is shown in Figure 3. In the dataset of the birth weight, when $S_{max} = 30, P_0 = 0.5$, MSE is smaller; In the dataset of the six-month weight, when $S_{max} = 40, P_0 = 0.5$, MSE is smaller; In the dataset of the weaning weight, when $S_{max} = 50, P_0 = 0.2$, MSE is smaller. The other parameter settings of the proposed method for the body weights on each occasion are reported in Table 2. Additionally, the parameter settings for the considered benchmark methods are recorded in Appendix A.

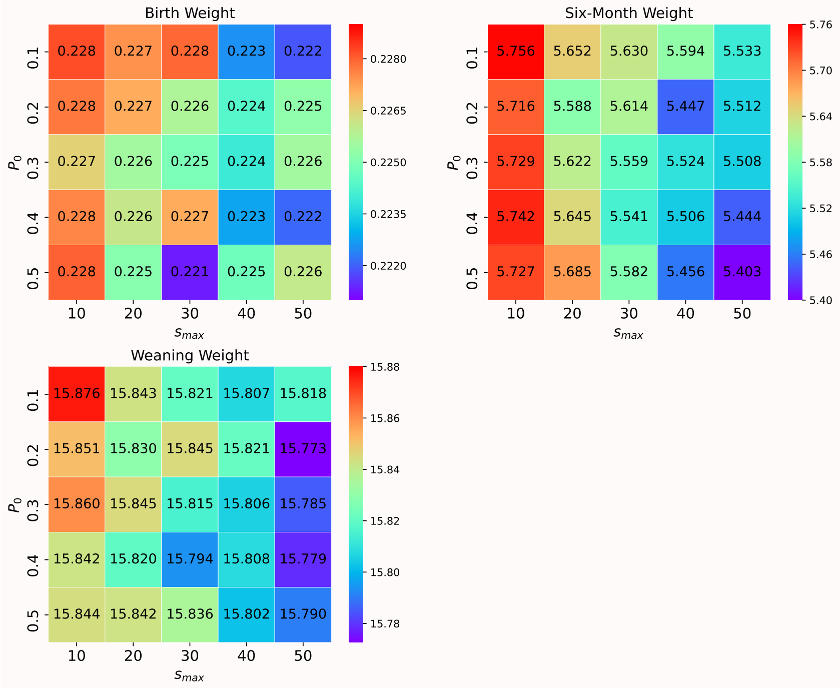

**Figure 3.** Mean squared error with different combinations $s_{max}$ and $P_0$ for analyzing the body weights on three occasions.

**Table 2.** The parameters values used in the proposed method for analyzing the body weights.

| Type of BW | $S_{max}$ | $P_0$ | C | $\epsilon$ | $\sigma$ | T | N | s |
|---|---|---|---|---|---|---|---|---|
| Birth Weight | 30 | 0.5 | 0.1 | 0.01 | $10^{-5}$ | 30 | 20 | 2000 |
| Six-Month Weight | 50 | 0.5 | 1 | 0.01 | 0.01 | 30 | 20 | 2000 |
| Weaning Weight | 50 | 0.2 | 0.1 | 0.01 | $10^{-5}$ | 30 | 20 | 2000 |

## 4. Results and Discussion

To test the performance of feature selection methods, for predicting the body weights on each occasion, the instances are divided into training set (170 samples) and test set (70 samples). The training set is used to obtain the relevant subset of genes while the test set is used to evaluate MSE of SVR [39]. The initialization of individuals uses Algorithm 1, the fitness function uses MSE, and the regression model uses SVR in the other heuristic methods for fair comparison. The experiment results are shown in Table 3. Some main points can be obtained from our experiment as follows.

**Table 3.** Performance of feature selection methods for predicting the body weights on three occasions.

| | Birth Weight | | Six-Month Weight | | Weaning Weight | |
|---|---|---|---|---|---|---|
| **Method** | **MSE** | **NumF** | **MSE** | **NumF** | **MSE** | **NumF** |
| SVR | 0.2393 | All | 5.6800 | All | 15.9099 | All |
| SWA | 0.2392 | 2053 | 5.6424 | 1947 | 15.8819 | 2018 |
| ABC | 0.2390 | 1890 | 5.6264 | 2025 | 15.8762 | 1947 |
| SCA | 0.2391 | 1954 | 5.6433 | 2022 | 15.8815 | 1964 |
| GA | 0.2393 | 1995 | 5.6470 | 1905 | 15.8831 | 1952 |
| BPSO | 0.2392 | 1923 | 5.6464 | 2080 | 15.8835 | 1922 |
| $\beta$-hill climbing | 0.2389 | 1928 | 5.6385 | 1995 | 15.8799 | 1984 |
| ABSS | 0.2292 | 9 | 5.8182 | 9 | 15.8993 | 9 |
| **Proposed method** | **0.2213** | **28** | **5.4026** | **50** | **15.7727** | **48** |

Note: NumF-number of feature.

### *4.1. SVR vs. Other Heuristic Methods*

In Table 3, the SVR models with heuristic methods, including $\beta$-Hill Climbing, SWA, ABC, SCA, BPSO, can significantly reduce the candidatures of gene to nearly 2000 genes. Furthermore, according to the error indicator (MSE), the performance of these SVR models is very similar. For example, all indicators for the birth weight are 0.239, thus we can confirm that more than 52,000 genes are irrelevant to the birth weight.

### *4.2. Proposed Method vs. ABSS*

In Table 3, the performance of SVR combined with the proposed method is more outstanding compared with SVR combined with ABSS for the body weight of the Hu sheep on each occasion. For the six-month weight, the number of selected genes is only a few by ABSS, and the performance of SVR significantly worsens. This means that ABSS may be limited by correlation feature method in the initial step of ABSS. It cannot provide the promising subset of genes of the entire gene subset space while the proposed method can provide the promising subset of genes of the entire gene subset space by selection and crossover operators.

### *4.3. Proposed Method vs. GA*

In Figures 4 and 5, mean and min fitness value of each generation in the proposed method are less than in GA over 5 generations or more for the dataset of body weight at each of the three time points. Both the mean and min fitness value of each generation gradually decreases (in the proposed method) and increases (in GA) over generations for the body weight at 6 and 12 months, and at weaning, respectively. It shows that the mutation operator lacks exploitation capability while the improved splicing method has strong exploitation capability. In addition, in Figure 5, a minimal min fitness value is found in a few generations. It means that the initial start points, and offspring improved by the improved splicing method, can enhance the efficiency of search. Furthermore, in Table 3, the quality of the selected subset of genes by the proposed method not only is better than by GA, but also the size of the selected gene subset by the proposed method is less than by GA. It shows that the improved splicing method is embedded into GA to find the minimal best size of subset of genes.

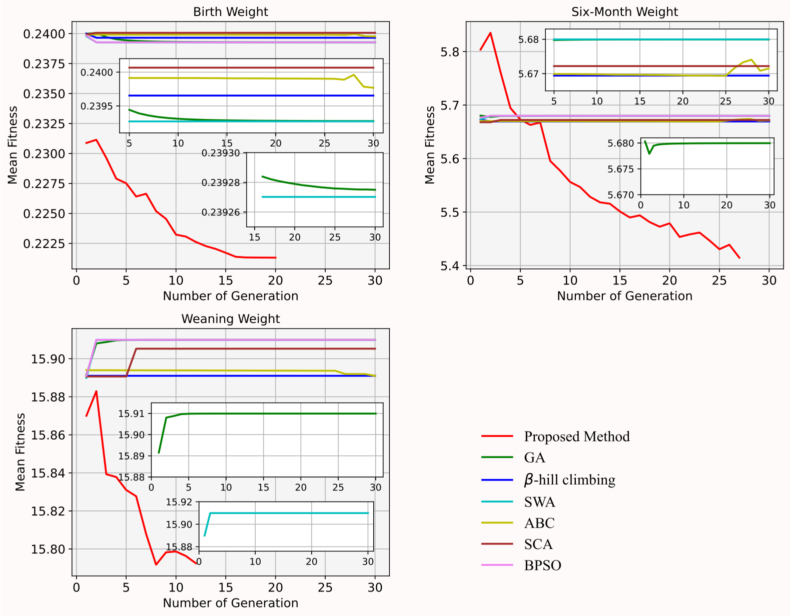

**Figure 4.** Mean fitness value of GA and the proposed method for the dataset of the body weights on three occasions.

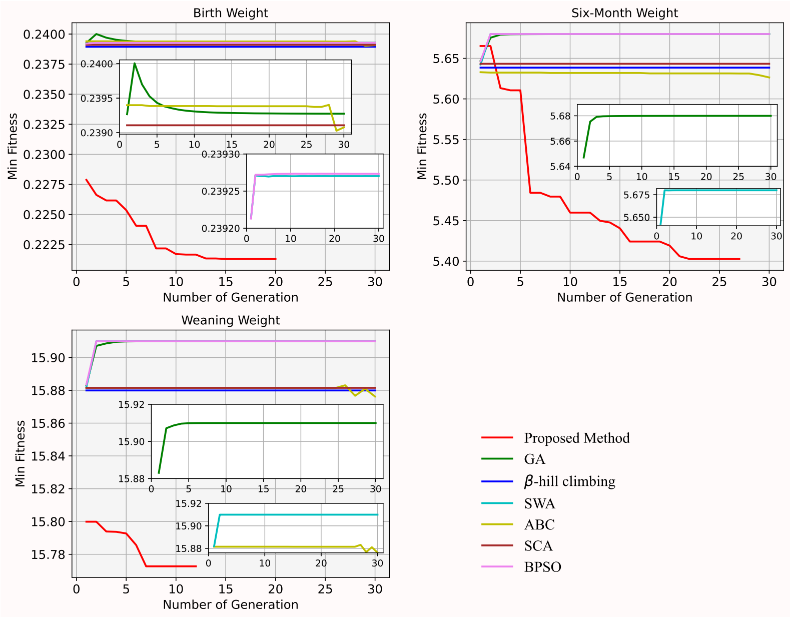

**Figure 5.** Min fitness value of GA and the proposed method for the body weights on each of the three occasions.

### 4.4. Proposed Method vs. Other Heuristic Algorithms

In Figure 4, mean fitness almost remain unchanged over 5 generations in other state-of-arts optimization algorithms while mean fitness almost continuously decreases over 5 generations in the proposed method. It shows that the improved splicing method can prevent premature convergence to identify best gene subsets while the other optimization algorithms easily encounter sub-optimal gene subsets. Furthermore, in Table 3, the performance of SVR combined with the proposed method outperforms SVR combined with other heuristic methods, including $\beta$-hill climbing, SWA, ABC, SCA, and BPSO. The size of the selected gene subset by the proposed method is smaller than by heuristic algorithm, including $\beta$-hill climbing, SWA, ABC, SCA, and BPSO. Here, we can conclude that the proposed method is a successful heuristic algorithm to find the minimal best subset of genes for gene selection problems.

### *4.5. Selected Genes*

The experimental results exhibit that the proposed method can yield minimal best gene subsets compared with other feature selections. Thus, Table 4 shows the best subset of genes obtained by using the proposed method for the body weight on the three occasions. Figure 6 displays a heatmap created for the identified best subset of genes for the body weight at the three occasions. The heatmap describe degree of similar and dissimilar among selected genes and sheep.

**Table 4.** Selected genes by using the proposed method for the body weight on the three occasions.

| Type of BW | Selected Genes |
|---|---|
| Birth Weight | OAR1_103547224.1,OAR1_156571804.1,OAR1_174344110.1,OAR1_200004020.1,OAR1_246870800.1,OAR1_254887296.1,OAR11_46573882.1,OAR13_25820059.1,OAR18_41747912.1,OAR19_10882333.1,OAR25_30336172.1,OAR5_63048172.1,OAR6_31217798.1,OAR6_73014482.1,OAR9_1294177.1,OARX_10743378.1,OARX_26036045.1,OARX_8338946.1,s02618.1,s12587.1,s15089.1,s16141.1,s17567.1,s44639.1,s50831.1,s64861.1,s70488.1,s75788.1 |
| Six-Month Weight | OAR1_103051402.1,OAR1_138627292.1,OAR1_214050298.1,OAR1_225069797.1,OAR1_252270534.1,OAR1_57838218.1,OAR1_72149006.1,OAR11_43264793_X.1,OAR13_22347021.1,OAR13_9894722.1,OAR16_60244426.1,OAR17_77852604.1,OAR19_33417302.1,OAR2_149404956.1,OAR2_16857269.1,OAR2_208372724.1,OAR20_12878785.1,OAR21_26401940.1,OAR23_40774711.1,OAR23_60556779.1,OAR25_1902197.1,OAR25_41478486.1,OAR26_24866987.1,OAR26_42053565.1,OAR26_48607412.1,OAR3_113183544.1,OAR3_119620209.1,OAR3_27184388.1,OAR3_88091256.1,OAR5_89062571.1,OAR7_73093180.1,OAR7_97719696.1,OAR8_73226726.1,OARUn.2169_4737.1,s01688.1,s01826.1,s06354.1,s07270.1,s13965.1,s14962.1,s17349.1,s35998.1,s36469.1,s52321.1,s56088.1,s64103.1,s70105.1,s71447.1,s72138.1,s72816.1 |
| Weaning Weight | OAR1_174220716.1,OAR1_193099978.1,OAR1_208929906.1,OAR1_285395930.1,OAR1_286637130.1,OAR1_89153973_X.1,OAR11_8041122.1,OAR13_58349162.1,OAR13_63513054.1,OAR16_67669492.1,OAR17_12809597.1,OAR17_28751326.1,OAR17_32705384.1,OAR17_37807906.1,OAR18_55245057.1,OAR19_41399545.1,OAR2_127646604_X.1,OAR2_130068033.1,OAR2_161930539.1,OAR2_75375830_X.1,OAR20_5632451.1,OAR22_28398167.1,OAR25_25782002.1,OAR26_10130406.1,OAR3_158876220.1,OAR3_182544365.1,OAR3_195768950.1,OAR3_235746854.1,OAR3_88835595_X.1,OAR4_125875837.1,OAR4_93485963.1,OAR4_97984717.1,OAR5_42553589.1,OAR5_93445511_X.1,OAR6_121519387.1,OAR7_78566470.1,OAR8_36682621.1,OAR9_31965185.1,s07941.1,s26017.1,s44731.1,s48924.1,s50855.1,s56042.1,s56962.1,s59822.1,s65507.1,s65803.1 |

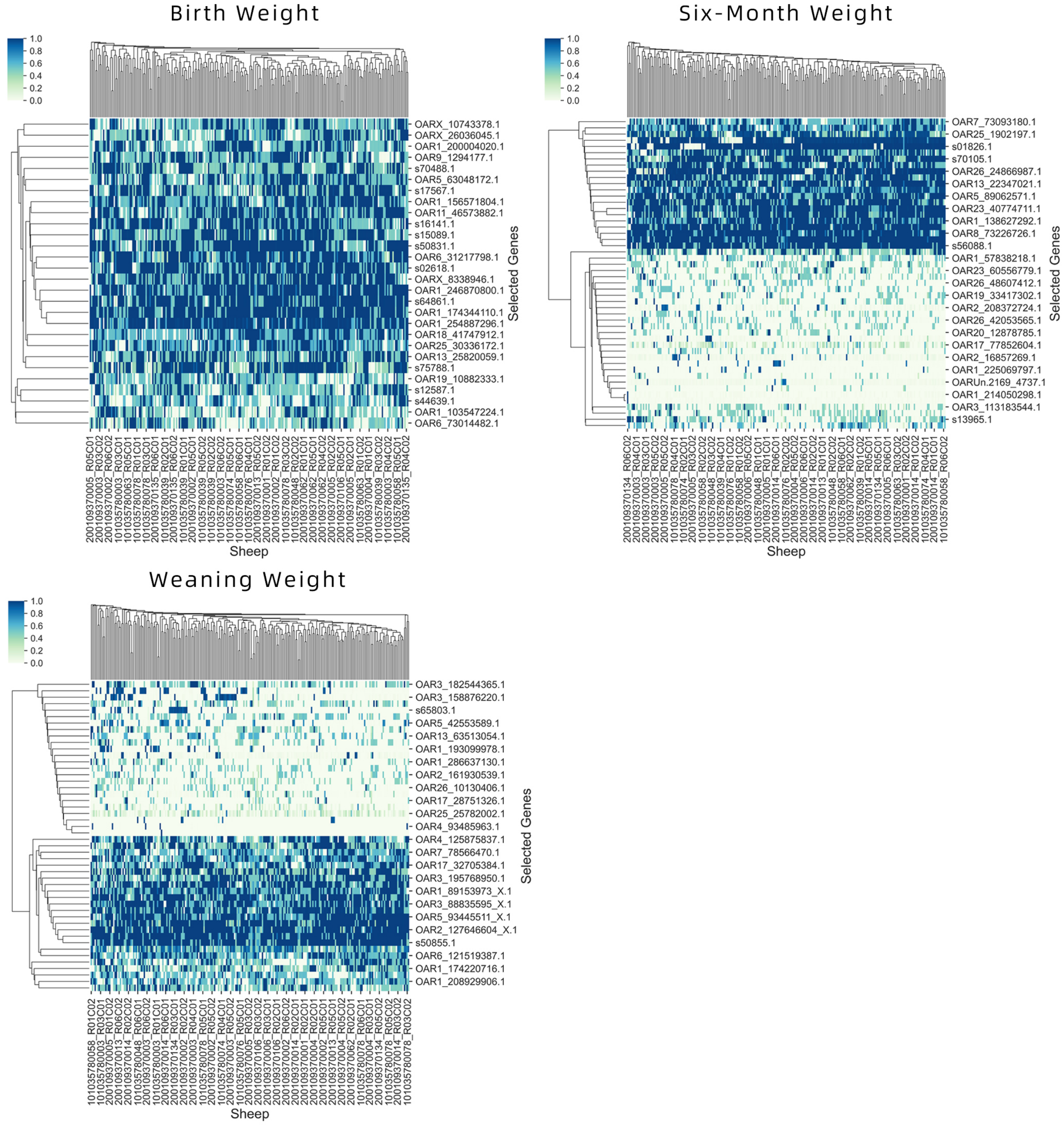

**Figure 6.** The heatmap of the actual expression profiles for the best subset of genes obtained from the proposed method.

### 4.6. Statistical Analysis

The non-parametric Friedman test is used to show whether there exists any statistically significant difference among 8 feature selection methods. The average rank of each algorithm is shown in Table 5. As shown in Table 5, the proposed method has placed in rank one. The Iman and Davenport statistic $F_F$ [40] is calculated as 5.579. The result is much larger than critical values ($F_{0.1}(7, 14) = 2.19$). This means that the null hypothesis is rejected, i.e., there are significant differences among the eight feature selection methods. Then, the Nemenyi test in post hoc Holm test [40] further is employed to show significant difference between the proposed method and other feature selection methods. As shown in Table 6, there are significant difference between the proposed method and GA and the proposed method and BPSO.

**Table 5.** Average rankings of MSE among 8 algorithms on three datasets using Friedman test.

| | Proposed | SWA | ABC | SCA | GA | BPSO | $\beta$-Hill Climbing | ABSS |
|---|---|---|---|---|---|---|---|---|
| Rank | 1 | 5.17 | 2.67 | 4.67 | 7 | 6.5 | 3 | 6 |

**Table 6.** Post hoc Holm test (0.1).

| Comparison | $p$-Values | Result |
|---|---|---|
| Proposed vs. SWA | 0.428 | $H_0$ is not rejected |
| Proposed vs. ABC | 0.900 | $H_0$ is not rejected |
| Proposed vs. SCA | 0.583 | $H_0$ is not rejected |
| Proposed vs. GA | 0.055 | $H_0$ is rejected |
| Proposed vs. BPSO | 0.108 | $H_0$ is rejected |
| Proposed vs. $\beta$-Hill Climbing | 0.900 | $H_0$ is not rejected |
| Proposed vs. ABSS | 0.195 | $H_0$ is not rejected |

## 5. Conclusions

In this paper, a modified memetic algorithm, a genetic algorithm based on an improved splicing method, has been proposed for gene selection problems. Different from traditional genetic algorithm, the optimizer can accelerate search to identify the minimal best subset of genes. It can absorb characteristics of crossover and selection operators to provide the promising subset of genes of the entire gene subset space. Furthermore, the improved splicing method can reduce the size of the promising subset of gene and recover the true subset of genes with a high probability. The initial points are improved by the improved splicing method to enhance efficiency of search, and the mutation operator is replaced by the improved splicing method to enhance exploitation capability for achieving balance between exploration and exploitation of GA. Therefore, the proposed optimizer can effectively achieve the best minimal subset of genes out of thousands of genes. Moreover, by using the body weights on each of the three occasions, we have demonstrated that our modified memetic algorithm can find the best minimal subset of genes compared with all considered algorithms, including ABSS. In addition, the proposed optimizer can be generalized to other high dimensional optimization problems [41,42].

**Author Contributions:** M.M.: software, visualization, formal analysis, writing—original draft.; J.W.: visualization, formal analysis, writing—review and editing, data curation; F.C.; investigation, project administration; Y.-G.W.: supervision, project administration, investigation, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the Australian Research Council Discovery Project (DP160104292), and the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) (Grant No: CE140100049), and Zhejiang Provincial Natural Science Foundation of China (Grant No: LY19A010014). A demo of the proposed method in this paper can be obtained by sending a request to the author (you-gan.wang@qut.edu.au).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data used to support the findings of this study are available from Wang, Y.-G. upon reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A

The parameter settings of SVR for analyzing the body weights are set as follows: In Birth weight, $C = 0.1$, $\epsilon = 0.01$, $\sigma = 10^{-5}$; In six-month weight, $C = 1$, $\epsilon = 0.01$, $\sigma = 0.01$; In weaning weight, $C = 0.1$, $\epsilon = 0.01$, $\sigma = 10^{-5}$; crossover disruption probability $P_0$

settings of GA for the body weights are set as follows: $P_0 = 0.5$ in birth weight; $P_0 = 0.5$ in six-month weight; $P_0 = 0.2$ in weaning weight. The parameters settings of other feature selection methods for the body weights are reported in Table A1.

**Table A1.** The parameters settings of other feature selection methods in analyzing the body weights.

| Method | Parameter |
|---|---|
| GA [25] | $T = 30, N = 20, s = 2000, p_m = 0.047$ |
| BPSO [30] | $T = 30, N = 20, s = 2000, v_{max} = 6,$<br>$w = 0.7298, c_1 = 1.49618, c_2 = 1.49618$ |
| SCA [29] | $T = 30, N = 20, s = 2000, \alpha = 1$ |
| ABC [28] | $T = 30, N = 20, s = 2000,$ Trials limit $= 50,$<br>Flips limit $= 0.1$ |
| SWA [27] | $T = 30, N = 20, s = 2000$ |
| $\beta$-hill climbing [26] | $T = 30, N = 20, s = 2000, \mathcal{N} = 0.9, \beta = 0.5$ |
| ABSS [24] | $\tau = 0.01\lvert\mathcal{A}\rvert \log(p) \log(\log n)/n,$<br>$s_{max} = [\frac{n}{\log p \log(\log n)}], k_{max} = \lvert\mathcal{A}\rvert$ |

## References

1. Mafarja, M.; Mirjalili, S. Whale optimization approaches for wrapper feature selection. *Appl. Soft Comput.* **2018**, *62*, 441–453.
2. Liu, H.; Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012; Volume 454.
3. Li, B.; Zhang, N.; Wang, Y.G.; George, A.W.; Reverter, A.; Li, Y. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet.* **2018**, *9*, 237.
4. Pashaei, E.; Pashaei, E. Gene selection using hybrid dragonfly black hole algorithm: A case study on RNA-seq COVID-19 data. *Anal. Biochem.* **2021**, *627*, 114242.
5. Pashaei, E.; Pashaei, E. Gene selection using intelligent dynamic genetic algorithm and random forest. In Proceedings of the 2019 11th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey, 28–30 November 2019; pp. 470–474.
6. Palma-Mendoza, R.J.; Rodriguez, D.; De-Marcos, L. Distributed reliefF-based feature selection in spark. *Knowl. Inf. Syst.* **2018**, *57*, 1–20.
7. Su, Q.; Wang, Y.; Jiang, X.; Chen, F.; Lu, W.C. A cancer gene selection algorithm based on the KS test and CFS. *BioMed Res. Int.* **2017**, *2017*, 1645619.
8. Ullah, I.; Paul, S.; Hong, Z.; Wang, Y.G. Significance tests for analyzing gene expression data with small sample sizes. *Bioinformatics* **2019**, *35*, 3996–4003.
9. Nakariyakul, S. A hybrid gene selection algorithm based on interaction information for microarray-based cancer classification. *PLoS ONE* **2019**, *14*, e0212333.
10. Alzubi, R.; Ramzan, N.; Alzoubi, H.; Amira, A. A hybrid feature selection method for complex diseases SNPs. *IEEE Access* **2017**, *6*, 1292–1301.
11. Pirgazi, J.; Alimoradi, M.; Abharian, T.E.; Olyaee, M.H. An efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets. *Sci. Rep.* **2019**, *9*, 18580.
12. Mustaqeem, A.; Anwar, S.M.; Majid, M.; Khan, A.R. Wrapper method for feature selection to classify cardiac arrhythmia. In Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea, 11–15 July 2017; pp. 3656–3659.
13. Lv, J.; Peng, Q.; Chen, X.; Sun, Z. A multi-objective heuristic algorithm for gene expression microarray data classification. *Expert Syst. Appl.* **2016**, *59*, 13–19.
14. Shukla, A.K.; Tripathi, D.; Reddy, B.R.; Chandramohan, D. A study on metaheuristics approaches for gene selection in microarray data: algorithms, applications and open challenges. *Evol. Intell.* **2020**, *13*, 309–329.
15. Wu, J.; Wang, Y.G.; Burrage, K.; Tian, Y.C.; Lawson, B.; Ding, Z. An improved firefly algorithm for global continuous optimization problems. *Expert Syst. Appl.* **2020**, *149*, 113340.
16. Motieghader, H.; Najafi, A.; Sadeghi, B.; Masoudi-Nejad, A. A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata. *Inform. Med. Unlocked* **2017**, *9*, 246–254.
17. Črepinšek, M.; Liu, S.H.; Mernik, M. Exploration and exploitation in evolutionary algorithms: A survey. *ACM Comput. Surv. (CSUR)* **2013**, *45*, 1–33.
18. Oh, I.S.; Lee, J.S.; Moon, B.R. Hybrid genetic algorithms for feature selection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 1424–1437.
19. Rezapoor Mirsaleh, M.; Meybodi, M.R. Balancing exploration and exploitation in memetic algorithms: A learning automata approach. *Comput. Intell.* **2018**, *34*, 282–309.

20. Ghosh, M.; Begum, S.; Sarkar, R.; Chakraborty, D.; Maulik, U. Recursive memetic algorithm for gene selection in microarray data. *Expert Syst. Appl.* **2019**, *116*, 172–185.
21. Zhu, Z.; Ong, Y.S.; Dash, M. Wrapper–filter feature selection algorithm using a memetic framework. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2007**, *37*, 70–76.
22. Guha, R.; Ghosh, M.; Kapri, S.; Shaw, S.; Mutsuddi, S.; Bhateja, V.; Sarkar, R. Deluge based genetic algorithm for feature selection. *Evol. Intell.* **2021**, *14*, 357–367.
23. Ghosh, M.; Malakar, S.; Bhowmik, S.; Sarkar, R.; Nasipuri, M. Feature selection for handwritten word recognition using memetic algorithm. In *Advances in Intelligent Computing*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 103–124.
24. Zhu, J.; Wen, C.; Zhu, J.; Zhang, H.; Wang, X. A polynomial algorithm for best-subset selection problem. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 33117–33123.
25. Desale, K.S.; Ade, R. Genetic algorithm based feature selection approach for effective intrusion detection system. In Proceedings of the 2015 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 8–10 January 2015; pp. 1–6.
26. Al-Betar, M.A.; Hammouri, A.I.; Awadallah, M.A.; Abu Doush, I. Binary $\beta$-hill climbing optimizer with S-shape transfer function for feature selection. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *12*, 7637–7665.
27. Rizk-Allah, R.M.; Hassanien, A.E.; Elhoseny, M.; Gunasekaran, M. A new binary salp swarm algorithm: development and application for optimization tasks. *Neural Comput. Appl.* **2019**, *31*, 1641–1663.
28. Santana, C.J., Jr.; Macedo, M.; Siqueira, H.; Gokhale, A.; Bastos-Filho, C.J. A novel binary artificial bee colony algorithm. *Future Gener. Comput. Syst.* **2019**, *98*, 180–196.
29. Hafez, A.I.; Zawbaa, H.M.; Emary, E.; Hassanien, A.E. Sine cosine optimization algorithm for feature selection. In Proceedings of the 2016 International Symposium on Innovations in Intelligent Systems and Applications (INISTA), Sinaia, Romania, 2–5 August 2016; pp. 1–5.
30. Xue, B.; Zhang, M.; Browne, W.N. New fitness functions in binary particle swarm optimisation for feature selection. In Proceedings of the 2012 IEEE Congress on Evolutionary Computation, Brisbane, QLD, Australia, 10–15 June 2012; pp. 1–8.
31. Mirjalili, S. Genetic algorithm. In *Evolutionary Algorithms and Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 43–55.
32. Razali, N.M.; Geraghty, J. Genetic algorithm performance with different selection strategies in solving TSP. In *Proceedings of the World Congress on Engineering*; International Association of Engineers Hong Kong: Hong Kong, Chian, 2011; Volume 2, pp. 1–6.
33. Merz, P.; Freisleben, B. Fitness landscapes and memetic algorithm design. *New Ideas Optim.* **1999**, *1999*, 245–260.
34. De Jong, K.A.; Spears, W.M. A formal analysis of the role of multi-point crossover in genetic algorithms. *Ann. Math. Artif. Intell.* **1992**, *5*, 1–26.
35. Gebreselassie, G.; Berihulay, H.; Jiang, L.; Ma, Y. Review on genomic regions and candidate genes associated with economically important production and reproduction traits in sheep (Ovies aries). *Animals* **2020**, *10*, 33.
36. Cao, Y.; Song, X.; Shan, H.; Jiang, J.; Xiong, P.; Wu, J.; Shi, F.; Jiang, Y. Genome-wide association study of body weights in Hu sheep and population verification of related single-nucleotide polymorphisms. *Front. Genet.* **2020**, *11*, 588.
37. Abdulqader, D.M.; Abdulazeez, A.M.; Zeebaree, D.Q. Machine learning supervised algorithms of gene selection: A review. *Mach. Learn.* **2020**, *62*, 233–244.
38. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222.
39. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–45.
40. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
41. Ding, Z.; Tian, Y.C.; Tang, M.; Li, Y.; Wang, Y.G.; Zhou, C. Profile-guided three-phase virtual resource management for energy efficiency of data centers. *IEEE Trans. Ind. Electron.* **2019**, *67*, 2460–2468.
42. Wu, J.; Ding, Z. Improved grey model by dragonfly algorithm for chinese tourism demand forecasting. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 199–209.