A New Embedded Feature Selection Method using IBALO mixed with MRMR criteria

Yu Zhang¹, Zhuanzhe Zhao¹, Shuaishuai Zhao¹, Yongming Liu¹, Kang He²

¹School of Mechanical and Automotive Engineering, Anhui Polytechnic University, Wuhu, Anhui, 241000, China.
²School of Mechanical and Electronic Engineering, Suzhou University, Suzhou, Anhui 234000, China.
*Corresponding author’s e-mail: zhuanzhe727@ahpu.edu.cn, 523410974@qq.com

Abstract. In order to remove irrelevant data and increase classification accuracy in feature selection, this paper proposed a new Embedded feature selection method with gathering Minimal Redundancy Maximal Relevance (MRMR) criteria, Sequential Forward Selection (SFS) and Improved Binary Ant Lion Optimizer (IBALO) together. Totally, we use three different feature selection methods namely MRMR mixed with Sequential Forward Selection (MS), MS mixed with Binary Ant Lion Optimizer (MS-BALO), an improvement of MS-BALO. Experiments prove that the MS-IBALO method proposed by this paper is efficient compared to MS and MS-BALO methods.

1. Introduction

Commonly, there are three different approaches of feature selection that evaluate the quality of the selected features: Filter, Wrapper and Embedded. Filter methods first design the evaluation function to rank the features, and set the elimination threshold artificially. Wrapper methods, which mainly uses sequence search or heuristic search to find the best subset with higher classification accuracy than Filter but more time cost. The Embedded feature selection combines Filter and Wrapper to inherit the advantages of both, compared with Filter, the performance of classification is better, and the calculation speed is higher than that of Wrapper [1]. For example, The Tabu algorithm is applied for Wrapper with KNN classifier to select features [2]. Binary Ant Lion Optimizer (BALO) is mixed with the rough set and approximate entropy theory to construct feature selection method [3]. The above scholars have proposed different improvement for various feature selection methods in terms of computational speed and classification accuracy, and made vital contributions.

The aim of this paper is to make contribution to improving feature selection methods on the basis of the research of the above related scholars [4].

2. Related Theory

2.1. Ant Lion Optimizer (ALO)

The ALO algorithm is a popular swarm intelligent algorithm to be used dealing with continuous function optimization problems in recent years. This algorithm mainly imitates the hunting behavior between antlions and ants in nature. Five main steps of hunting prey such as the random walk of ants,
building traps, entrapment of ants in traps, catching preys, and rebuilding traps are implemented[5].

The algorithm is carried out under the following steps:

- Ants randomly walk around the antlions for modelling ants’ movement, Ants update their positions with random walk at every step of optimization.
- Antlions set trap bounds according to its own fitness, bounds increase adaptively with the number of iterations;
- Antlions shoot sands outwards the center of the pit once they realize that an ant is in the trap;
- If an ant is successfully preyed by antlion, it is considered that fitness value of the ant is bigger than that of the antlion, and the antlion will obtain a better position which belongs to the ant before.
- After each hunt process finished, antlions will repair the trap for the next round of hunt.

2.2. BALO

To deal with 0,1 combination optimization problem, binary variants of the ALO are proposed and used to select the optimal feature subset for classification. The BALO algorithm is proposed based on crossover, mutation operation[6]. Each ant updates position by inserting crossover operator as follows:

$$x'_i = \text{crossover}(RW_1, RW_2)$$

Where $x'_i$ represents the current position of the $i$-th individual in $t$-th iteration, and the crossover operation obtains two n-dimensional binary solutions $RW_1, RW_2$ created by ants walking around the elite and Roulette selected antlions.

$$x_d = \begin{cases} x_1^d & \text{if } r > 0.5 \\ x_2^d & \text{otherwise} \end{cases}$$

Where $x'_1$ and $x'_2$ represents the dimension $d$ of the binary solution $RW_1$ and $RW_2$, and $r$ represents a random number between 0-1. $RW$ ($RW_1$ or $RW_2$) is obtained by ants walking randomly around antlions, the calculation of each dimension requires mutation operators as equation (3):

$$x_{\text{out}}^d = \begin{cases} x_{\text{in}}^d & \text{if } r > r_0 \\ 1 - x_{\text{in}}^d & \text{otherwise} \end{cases}$$

Where $x_{\text{out}}^d$ represents the dimension $d$ of $RW$, $x_{\text{in}}^d$ represents the dimension $d$ of the antlion chosen from Roulette selected one or the elite one, $r_0$ is the mutation rate. It worth mentioning that $r_0$ is linearly decremented with iteration number ranging from 0.9 to 0 in equation (4). $r$ is a random number between 0-1 [6].

$$r_0 = 0.9 - \frac{0.9(t - 1)}{IterMax - 1}$$

2.3. IBALO

At the end of each iteration process in ALO, the individuals with poor fitness are removed from the population immediately, which could cause the decrease of population categories, as a result, the global search ability falls down. Therefore, this paper proposed an improvement method using Small-Scale Protection (SSP) mechanism. Population in original algorithm is called main pop ($mp$), population created by SSP mechanism is named sub pop ($sp$). SSP mechanism is used to choose a certain proportion of individuals from $mp$ and $sp$ at the end of each iteration by rules, $mixp$ represents the sum of $mp$ and $sp$. Based on the fitness rank of each individuals, $mixp$ can be divided into antlions ($al$) and weak individuals ($wi$), rules look the same as $mp$ and $sp$, which can be seen in equations (5),(6),(7).

$$mixp^{t+1} = mp^{t+1} + sp^{t+1} = al^{t+1} + wi^{t+1}$$
\[ mp'^{t} = mal'^{t} + mwi'^{t} \]  
\[ sp'^{t+1} = sal'^{t+1} + swi'^{t+1} \]  

(6)  
(7)  

Where the superscript indicates the iteration of algorithm process, for example, \( t \) indicates the \( t \)-th iteration. The following characters can be regarded as no superscript for convenience. \( mal \) represents antlions in \( mp \), \( mwi \) means weak individuals in \( mp \); \( sal \) represents antlions in \( sp \), \( swi \) means the weak individuals in \( sp \). Selection principles can be seen in following equations (8),(9),(10),(11).

\[ rank^{t}_c = sort(f_c(al'^{t})) \]  
\[ rank^{t}_v = sort(rmat \cdot f_v(wi'^{t})) \]  
\[ (save'^t_v, save'^t_c) = extract(rank^{t}_v, rank^{t}_c) \]  
\[ save'^t = save'^t_v + save'^t_c \]  

(8)  
(9)  
(10)  
(11)

Where \( rank_c \) and \( rank_v \) indicate the individual ranks in ascending order based on the number of features and the value of fitness function; \( f_c \) represents the statistical operator on feature number; \( f_v \) represents the statistical operator on fitness value; \( sort \) means the ascending order operator; \( rmat \) means a matrix contains random numbers between 0-1. The operator of \( extract \) can extract the high rank features. The variable of \( save \) contains \( save_c \) and \( save_v \) are stored to be ready for the next iteration of \( sp \).

3. Embedded feature selection method based on MRMR mixed with IBALO

In this paper, MRMR criteria is used as the evaluation function to rank the features[7]. SFS based on KNN classifier is used to eliminate the redundant features. Then the feature subset containing the high rank features obtained by MRMR and SFS (MS) is used as the initial starting point for IBALO. This whole method is regarded as MS-IBALO.

4. EXPERIMENTS

4.1. Data description

In order to verify the effectiveness of the proposed new algorithm, 7 different dimensional datasets from UCI are ready for experiments (see in Table 1), with dimensional bounds between 13 and 60. 75% datum for each dataset is used as training sample and the rest is regarded as test sample.

| Dataset    | Instances | Features | Classes |
|------------|-----------|----------|---------|
| Sonar      | 208       | 60       | 2       |
| Wine       | 178       | 13       | 3       |
| Lungcancer | 56        | 32       | 2       |
| Waveform21 | 5000      | 21       | 3       |
| Waveform40 | 5000      | 40       | 3       |
| Dermatology| 366       | 33       | 6       |
| WDBC       | 569       | 30       | 2       |

4.2. Experiment content and parameter settings

To highlight the advantages of MS-IBALO, MS and MS-BALO will join in the experiments with the same dataset and the same related parameter settings in order for all algorithms to perform fairly. In the ALO algorithm, population and \( IterMax \) are set to be 30 and 50, \( k \) in KNN classifier is set to be 5, each algorithm will run 20 times independently to get the best solution. The results are shown in Table 2. All experiments are implemented on a personal computer with CPU 1.8Ghz, RAM 4G and using \textit{matlab} R2016a.
4.3. Fitness function
Fitness function is proposed based on the maximum accuracy in each running. \( f_j \) represents the \( j \)-th feature, and \( tt \) means the total number of features.

\[
\text{max } \text{fitness} = \frac{1}{tt} \sum_{j=1}^{n} f_j
\]  

(12)

4.4. Results
To begin with the experiments, MS is used to rank features, we can choose Sonar as an example.

![Figure 1. Results of MS Algorithm](image)

Figure 1 poses that Accuracy is increasing generally with Feature selection number growth at first, and after Accuracy climbs to the peak of the curve (83.56%) where Feature selection number comes to 35, the curve lose its power going up. Therefore, 35 features in Sonar are used as initialization points during BALO and IBALO. The details are listed in Table 2.

| Dataset    | MS  | MS-BALO | MS-IBALO |
|------------|-----|---------|----------|
| Sonar      | 83.56 | 35 | 96.15 | 16 | 98.08 | 12 |
| Wine       | 100 | 7 | 100 | 3 | 100 | 2 |
| Lungsancer | 66.67 | 22 | 83.33 | 8 | 83.33 | 5 |
| Waveform21 | 80.73 | 21 | 79.60 | 12 | 81.15 | 15 |
| Waveform40 | 76.81 | 39 | 79.36 | 26 | 79.79 | 18 |
| Dermatology | 97.69 | 33 | 98.61 | 11 | 98.61 | 12 |
| WDBC       | 96.88 | 25 | 97.92 | 11 | 97.92 | 11 |

From Table 2, the best classification accuracy (called Best in Table 2) obtained by MS-BALO and MS-IBALO always perform better than MS except for Waveform21, which deeply confirms the effectiveness of the heuristic search; half part of the whole datasets using MS-IBALO perform the best classification accuracy among three algorithm, and the other half part perform as good as the other two, which can show MS-IBALO has certain advantages in the accuracy. With the increase of data dimension, datasets using MS-IBALO has less features than the other two.

5. Summary
This paper used MRMR to rank features based on importance, and proposed an IBALO algorithm based on SSP mechanism to remove the feature redundancy more efficient compared with MS-BALO and MS-IBALO. This paper indicates MS-IBALO can be used as a new method to deal with feature
selection problems. However, this method still has weaks in a few datasets, like getting almost same accuracy and abnormal feature removal. Therefore, we will try to study furtherly and make some improvements.

Acknowledgments
This work was supported in part by the Natural Science Foundation of Anhui Province under Grant 1808085ME127, in part by the Project of Study Abroad of Excellent Young Talents of Universities under Grant (gxgwfx2019041), in part by Suzhou College Teacher Application Ability Development Workstation (No. 2018XJYY01).

Reference
[1] Wang, S., Zhu, W. (2018) Sparse Graph Embedding Unsupervised Feature Selection. J. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48 (3): 329-341.
[2] Tahir, M.A., Bouridane, A., Kurugollu, F. (2007) Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier. J. Pattern Recognition Letters, 28 (4): 438-446.
[3] Mafarja, M. M., Mirjalili, S. (2018) Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection. J. Soft Computing.
[4] Huang, Z., Yang, C., Zhou, X., Huang, T. (2018) A hybrid feature selection method based on binary state transition algorithm and ReliefF. J. IEEE Journal of Biomedical and Health Informatics, 23(5): 1888-1898.
[5] Mirjalili, S. (2015) The Ant Lion Optimizer. J. Advances in Engineering Software, 83: 80–98.
[6] Emary, E., Zawbaa, H. M., Hassanien, A. E. (2016) Binary ant lion approaches for feature selection. J. Neurocomputing, 213: 54–65.
[7] Peng, H., Long,F., Ding, C. (2005) Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. J. IEEE Transactions on Pattern Analysis&Machine Intelligence, 27(8): 1226-1238.