


ITC 2/49 Information Technology and Control Vol. 49 / No. 2 / 2020 pp. 289-301 DOI 10.5755/j01.itc.49.2.24858	Feature Selection Using Improved Forest Optimization Algorithm	
	Received 2019/12/11	Accepted after revision 2020/02/13
	 http://dx.doi.org/10.5755/j01.itc.49.2.24858	

HOW TO CITE: Xie, Q., Cheng, G., Zhang, X., Lei, P. (2020). Feature Selection Using Improved Forest Optimization Algorithm. *Information Technology and Control*, 49(1), 289-301. <https://doi.org/10.5755/j01.itc.49.2.24858>

Feature Selection Using Improved Forest Optimization Algorithm

Qi Xie, Gengguo Cheng

China School of information science and engineering, Wuhan University of Science and Technology, Wuhan, 430065, China

Xiao Zhang

Business School, University of Birmingham, Birmingham, B15 2TT, the United Kingdom

Peng Lei

Business School, University of Sydney, Sydney, 2134, Australia

Corresponding author: 85169023@qq.com

Feature selection is a very popular topic in the field of data mining and machine learning. In 2016, the feature selection using forest optimization algorithm (FSFOA) was proposed, which had a better classification performance and dimensionality reduction ability. However, there are some shortcomings in FSFOA. In this article, Feature Selection using Improved Forest Optimization Algorithm (FSIFOA) is proposed, which aims at solving the problems of FSFOA during the stages of random initialization, forming the candidate population and updating the best tree. The proposed FSIFOA is compared with some other methods including FSFOA, NSM, PSO and other algorithms. The experimental results show that FSIFOA can improve the classification accuracy of classifiers in medium and large dimension datasets. Also, the dimensionality reduction of the FSIFOA is compared with other comparable methods.

KEYWORDS: Feature selection, L1 regularization, Candidate population, Forest optimization algorithm, Updating mechanism.

1. Introduction

Feature selection is one of the popular fields in the machine learning and data mining [5]. The feature selection is an approach of selecting the most effective features from a set of features to reduce the dimension of feature space [35], [25]. Feature selection removes redundant and unrelated features during the process of data pre-processing, which can reduce the effect of dimensional disaster problem and enhance the learning performance by simplifying the task [39], [33]. In the classification, feature selection can improve the accuracy of classification, generate more efficient classifiers, and better understand the information about key features [22]. Many studies have shown that feature selection is effective [18]. Therefore, feature selection is vital in machine learning processing, which can retain useful features for following learning tasks while ignoring irrelevant and unimportant features [36].

Ghaemi et al. proposed Forest Optimization Algorithm [10] (FOA) in 2014. Ghaemi et al. proposed Feature Selection using Forest Optimization Algorithm [11] (FSFOA) in 2016. FSFOA have better performance comparing to feature selection based on hybrid genetic algorithm [14] (HGAFS), particle swarm optimization [32] (PSO), and support vector machine [21] (SVM-FuzCoc). FSFOA can improve the accuracy of feature learning, effectively remove redundant features, and also has global search capabilities. However, there are some shortcomings in FSFOA. First, the initial features of FSFOA use a random generating strategy. The random initialization strategy may fall into local optimum in the non-convex function and cannot achieve the global optimum. Second, the candidate population produced in population limiting stage will lead to the problem of category imbalance. This will affect the global seeding outcome. Third, the experiment shows that there will be trees with the same fitness but different features in the best tree update stage. FSFOA will eliminate these trees, however there are some eliminated trees with smaller dimensions or higher precision. With considering the issues above, this paper proposes a new feature selection using improved Forest Optimization Algorithm (FSIFOA). This algorithm improves the performance of FSFOA from three aspects: forest initialization, candidate population generation and the best tree updating. Fi-

nally, FSIFOA uses the same data and parameters as FSFOA to test the small, medium and large dimensional data respectively.

2. Literature Review for Feature Selection

In the field of Machine Learning and Pattern Recognition, the quality of feature selection is directly related to the capability of the classifier, therefore the method of feature selection is vital. Feature selection is divided into four parts: the search mechanism of feature subsets, the evaluation mechanism of feature subsets, the stopping criterion and the verification method [20]. The current researches focus on search mechanisms and evaluation mechanisms.

According to the different feature subset evaluation mechanisms, feature selection methods can be divided into three methods: Filter, Wrapper, and Embedding [39]. The filter method first selects the features of the data and then trains the learner [36]. The feature selection process is irrelevant to the learners. Filter feature selection methods normally apply evaluation functions to reduce the correlation among features, and to increase the correlation between categories and features [5]. The wrapper feature selection directly takes the performance of the learner as the evaluation criterion of the feature subset [39]. Researchers use different machine learning algorithms for wrapper feature selection, such as decision tree algorithm [13], genetic algorithm [3], and support vector machine [12]. The embedding feature selection integrates the feature selection process with the learner training process under the same optimization process. L1 regularization is a typical method of embedding features. Tikhonov et al. proposed a ridge regression algorithm [27], which applies L2 norm regularization to the mean squared error (MSE) loss function. Tibshirani et al. proposed the LASSO (Least Absolute Shrinkage and Selection Operator) algorithm [26], also use MSE as loss function. The difference between these two methods is LASSO uses the L1 norm regularization instead of the L2 norm regularization. The final features selected are non-ze-

ro weight obtained by the L1 regularization solution. The L1 regularization completes the learner training and the feature selection in the meantime.

Evolutionary algorithms implement random search by simulating the evolution of natural organisms. The evolutionary algorithms have high-robustness and self-adaptability. This algorithm can effectively process complex problems which traditional optimization algorithms are difficult to solve. It is also used to solve global optimal solution problems [29]. Therefore, many researchers have adopted evolutionary algorithms for feature selection [33]. Genetic algorithm (GA) is a kind of evolutionary algorithm. Yang et al. proposed using genetic algorithm [34]. Dong et al. proposed a feature algorithm combining particle information with genetic algorithm [6], which uses an improved feature granularity genetic algorithm. Dorigo proposed Ant Colony Optimization (ACO), which is a heuristic algorithm for solving combinatorial optimization problems [7, 8]. Kabir et al. proposed a hybrid ant colony optimization algorithm [15], which combines the advantages of filter feature selection and wrapper feature selection. Wan et al. proposed an improved binary-coded ant colony algorithm for feature selection [30], which uses genetic algorithms to initialize the initial information of the pheromone of the ant colony algorithm. Kenned et al. proposed particle swarm optimization [9], [16] (PSO), which uses individual information sharing to make the whole group's motion in the problem-solving space from the disorder to the orderly evolution process. Xue et al. proposed a feature selection algorithm based on particle swarm optimization in classification problems [32]. This algorithm proposes 3 new individual best and global optimal update mechanisms and 3 new initialization strategies. Zhang et al. proposed a bare-bone particle swarm optimization algorithm [36], which designed an enhanced memory strategy to update the local leader of the particle, avoiding the degradation of excellent genes in the particle. Tran et al. summarized the application of particle swarm optimization in feature selection [28]. Population extremal optimization (PEO) algorithm is also a kind of evolutionary algorithm. Zeng et al. have proposed a robust proportional-integral (PI) controller and a novel short-term traffic flow forecasting model based on PEO [19], [38]. The new proposed controller and model have good performance in practical applications.

In recent years, evolutionary algorithms have produced a new branch. Ghaemi et al. proposed a Forest Optimization Algorithm (FOA) based on the growth process of trees in the forest. The forest optimization algorithm is a bionic intelligent optimization algorithm that simulates the process of seeding in forests to search for optimal solutions to solve nonlinear continuous optimization problems [23].

In the other hand, several novel nature inspired optimization algorithms for feature selection are proposed, for example: Cuckoo search algorithm [31], grey wolf optimizer, ant lion optimization [17], crow search and cuttlefish algorithm. These new methods provide new ideas for the development of feature selection.

3. Feature Selection based on Forest Optimization Algorithm

In 2016, Ghaemi et al. proposed FSFOA [10], which uses forest optimization algorithms for feature selection. FSFOA is divided into five parts: Initialize Trees, Local Seeding, Population Limiting, Global Seeding, and Update the Best Tree.

A. Initialize Trees

Randomly generate some trees to initialize a forest, each tree consists of feature value, age and fitness value. The feature value is a randomly generated one-dimensional vector of "0" or "1". The length of the feature vector is the number of features in the data set. If the number of features is n , the feature value of the tree is a vector consisting of n "0" or "1" digits. "1" represents that the corresponding feature is selected. And "0" represents that the corresponding feature is deleted. The age of each tree is set to 0 during the initialization phase. Fitness evaluate the performance selected feature in the learning process. The fitness functions used by FSFOA are KNN [1], SVM [4], and C4.5 [24].

B. Local Seeding

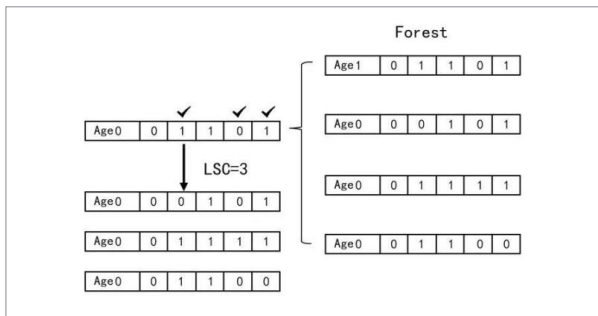
In the nature when seeding procedure of the trees begins, some seeds fall just around the parent tree and then they turn into young trees. The natural environment such as sunlight, water and soil in the new trees produced by the nearby planting is similar but slightly different as the mother tree. There will be competi-

tion between new trees and old trees, and trees suitable for the environment will survive. FSFOA simulates the phenomenon of near planting of trees in nature, calling this process Local Seeding.

In the local seeding stage, trees with an age of 0 in the forest are involved in local seeding and the remaining trees are not involved. A tree with a tree age of 0 is copied according to a parameter called Local Seeding Changes (LSC) value to generate a plurality of trees with the same feature value. Assuming an LSC value of 2, each tree in the forest with age 0 will generate two trees with the same feature values. Each newly generated tree feature value is randomly selected to be inverted, it means that: if the value is “0”, it becomes “1”, and vice versa. Finally, the Age value of all trees in the forest is increased by 1, the new tree Age value is set to 0, add the new tree to the forest afterwards, as shown in Figure 1.

Figure 1

Local seeding figure when LSC value equals to 3



Because the new tree produced by local seeding is planted nearby, the feature value of the new tree and the old tree will be similar but slightly different, they will compete with each other. The local seeding process is used for local searching.

C. Population Limiting

In nature, trees do not increase indefinitely, but in the local seeding stage, trees will continue to grow. In order to control the number of trees, FSFOA refers to the process of controlling the number of trees as the “Population Limiting”. There are two steps to limit the scale: the first one eliminates the trees with an age greater than the “life time” parameter, and put the trees into the candidate population. This way simulates the normal death of trees in nature. After the first step, if the size of the forest is larger than

the “area limit” parameter, the fitness of all the trees will be sorted in descending order, and the trees with the smallest fitness will be eliminated, and the scale of the forest will be controlled within the “area limit” value. Put the eliminated trees into the candidate population. This approach simulates the law of survival in the natural.

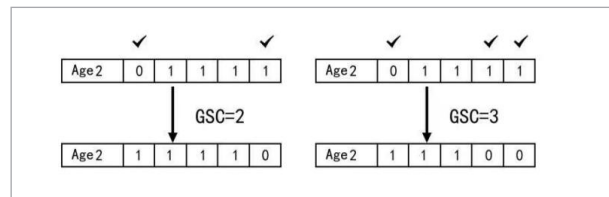
D. Global Seeding

In nature, seeds may be planted in a far distance because some external factor interfering. This phenomenon is called Global Seeding. FSFOA provides a global search method by simulating the process of global seeding, in order to avoid falling into local optimum.

FSFOA randomly selects a proportion of trees in the candidate population for global seeding based on the “transfer rate” parameter. Each tree used for global seeding is randomly selected according to the value of Global Seeding Changes (GSC), as shown in Figure 2. Global Seeding avoid the disadvantage that local seeding may fall into local optimum and provides the possibility of searching for the global optimum.

Figure 2

Global seeding figure when GSC values equals to 2 and 3



E. Update the best tree

In the stage of updating The Best Tree, select the tree with the most fitness in the forest, set the age value of these trees to 0, and put it back into the forest.

4. Feature Selection Based on Improved Forest Optimization Algorithm

A. Problem with the FSFOA

FSFOA has shortcomings in three aspects:

First: In the initialization stage, it generates the trees are completely randomized, this strategy can easily

fall into the local optimal solution in the non-convex function problem;

Second: FSFOA limit the scale of the forest through two strategies in the population limiting stage, and generates the candidate population from the eliminated trees, and uses them for global seeding proportionally. This will lead to incomplete problems of good and bad trees, which will influence the quality of global search. The first way of Population Limiting is to simulate the natural death of a high-quality tree, eliminate the tree with age to "life time" value. The tree whose age can reach the "life time" value has not been eliminated before, indicating that the fitness of such a tree will be greater than the average tree, otherwise it will be eliminated. Such a tree is a good tree. The second way is to eliminate the tree with the least fitness if the size of the forest surpasses the "Area Limit" value. This method is a process of simulating the survival of the fittest in the forest. Some trees die because the genetic or environmental problems have not grown up, indicating that the tree is a bad tree. Mix of good and bad trees to form candidate population and do global seeding, can lead to the outcome of all good trees or bad trees. We call this phenomenon the problem of category imbalance.

Third: In the best tree update stage, there will be trees with the same fitness but different features. FSFOA eliminates these trees which may have smaller dimensions or higher precision.

B. The Improvement of FSFOA

In response to the three shortcomings of FSFOA, this paper proposes three improvements:

First, the improvement of initialization strategy. Initialization strategy has two steps: Firstly, calculating Pearson correlation coefficients between all features of data and labels. The feature with positive correlation coefficient is selected as the alternative feature set. Secondly, L1 regularization feature selection method is applied to choose the feature with non-zero weight from the alternative feature set. By using Pearson correlation coefficient and L1 regularization, the feature set generated after two selections is highly correlated with the label. Compared with the random initialization strategy, the feature set selected by the new initialization strategy can converge to the extreme quickly and help to search for the optimal feature subset.

Pearson Correlation Coefficient is a linear correlation coefficient. The Pearson correlation coefficient is a statistic used to reflect the degree of linear correlation between two variables. The larger the absolute value of the correlation coefficient, the stronger the correlation. The Pearson correlation coefficient is equal to the covariance of the two vectors divided by the respective standard deviations, as shown in equation (1).

$$\rho_{xy} = \frac{Cov(X, Y)}{\sigma_x \sigma_y} = \frac{E\{(X - E(X))(Y - E(Y))\}}{\sqrt{E\{[X - E(X)]^2\}} \sqrt{E\{[Y - E(Y)]^2\}}} \quad (1)$$

In Equation (1), $Cov(X, Y)$ represents the covariance of vector X and vector Y , σ_x represents the standard deviation of vector X , σ_y represents the standard deviation of vector Y , E represents the expectation, and ρ_{xy} represents the vector X and vector Y of Pearson correlation coefficient.

The Pearson correlation coefficient ranges from -1 to 1, and a correlation coefficient of 0 indicates that the two vectors are not linearly related. If the correlation coefficient is greater than 0, the two vectors are positively correlated. If the correlation coefficient is less than 0, the two vectors are negatively correlated. The Pearson correlation coefficient feature selection method is one of filter feature selection.

$$\hat{\beta}_{lasso} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (2)$$

Tibshiran have introduced a shrinkage method to penalized least squares: L1 regularization. This method is used to shrink the coefficient toward zero. The equation of L1 regularization is showed in Equation (2). The least absolute shrinkage and selection operator method solving the following problems.

The tuning parameter is selected by cross validation, when it decreases from a large value to zero the lasso shrinkage factor increases from zero to one. The lasso can be well performed when few predictors have large coefficient while others have relative smaller coefficient. Compare to Ridge regression, the lasso introduces an absolute value for penalty form, hence it can easily reduce to zero when is small. Therefore, the model will have performed an auto-variable selection. But the Ridge regression cannot easily reach to zero. The issue with lasso is the total number of variables that the lasso variable selection procedures

is bound by the total number of samples in data set. Additionally, the lasso fails to perform grouped selection, it tends to select one variable from a group and ignore others. If there are more parameters than observations, the Lasso tend to select same number of parameters with observations.

Second: improvement of the candidate population. According to the two strategies of population limiting, the candidate population are divided into good trees and bad trees. And randomly select trees from good trees and bad trees are for global seeding according to the value of “transfer rate”. However, due to different population Limiting strategies, good trees and bad trees will create “category imbalances” issues, that is, the number of good trees will be much larger than the amount of bad trees, vice versa. Therefore, selecting trees with the least fitness in the forest to make up the difference between the good trees and the bad trees while the number of the good trees is greater than the number of the bad trees, vice versa.

Third: Improvements in best tree updates and selections. In FSFOA, when the maximum fitness of the new tree generated by global seeding is the same as the maximum fitness in the forest, the new tree is eliminated. In the new update method, if the maximum fitness of the new tree is equal to the maximum fitness in the forest, the new tree age is set to 0 and added to the forest. The reason is that when the new tree and the old tree have the same fitness, there is a new tree with a smaller dimension, so adding a new tree to the forest may reduce the dimension. If the fitness is the same but the dimension of the new tree is greater than or equal to the old tree. Such a new tree is also added to the forest, because such a new tree then produces a tree with better performance through seeding, thereby increasing the possibility of obtaining a global optimal solution. Finally, in the optimal tree selection phase, select the tree that the most fit. If there are multiple trees with the most fitness, select the tree with the least number of features.

5. Experiment

The experiment uses the same data set and parameters as FSFOA. The experiment obtained 11 data sets from the UCI machine learning library [2]. The

experimental program was written in python3, using the scikit-learn toolkit to write L1 embedded feature selection, Pearson correlation coefficient filtering feature selection, and other machine learning algorithms. All experiments were conducted on a MacBook Pro 3.5 GHz Intel Core i7 processor.

A. Data

The experimental data contains a total of 11 data sets, namely: “Wine”, “Ionosphere”, “Vehicle”, “Glass”, “Segmentation”, “Hepatitis”, “SRBCT”, “Heart-stat-log”, “Cleveland”, “Sonar” and “Dermatology”. FSFOA divides the experimental data set into “small dimension”, “medium dimension” and “big dimension” data sets according to the feature numbers. The corresponding number of features is: [0,19], [20,49], [50, ∞] [10]. According to the above division, the data set contains 7 small-dimensional data sets, 2 medium-dimensional data sets, and 2 large-dimensional data sets. The relevant description of the experimental data set is shown in Table 1.

Table 1

Descriptions of experimental dataset

Method name	Description	Validation method
SVM-FuzCoc	A novel SVM-based FS	70-30%
NSM	Neighbor soft margin	10-fold
FS-NEIR	Neighborhood effective information ratio based FS	10-fold
HGAFS	Hybrid genetic algorithm for FS	2-fold
PSO	Particle swarm optimization for feature selection	10-fold
SFS, SBS, SFFS	Greedy hill climbing methods	70-30%
UFSACO	Unsupervised FS algorithm based on ACO	70-30%

B. Parameters of the Experiment

We compare the results of FSIFOA, FSFOA, NSM, SVM-FuzCoc, HGAFS, FS-NEIR, UFSACO and PSO.

Table 2

Summary of methods for the comparisons

Dataset	Features	Instance	Class	Data Dimension
SRBCT	2308	63	4	Large Dimension
Sonar	60	208	2	Large Dimension
Dermatology	34	366	6	Medium Dimension
Ionosphere	34	351	2	Medium Dimension
Segmentation	19	2310	7	Small Dimension
Hepatitis	19	155	2	Small Dimension
Vehicle	18	846	4	Small Dimension
Heart-statlog	13	270	2	Small Dimension
Cleveland	13	303	5	Small Dimension
Wine	13	178	3	Small Dimension
Glass	9	214	7	Small Dimension

Table 2 shows the Summary of methods for our comparisons. The parameters of the new algorithm are consistent with the parameters of FSFOA. FSFOA has five parameters: the age limit of the tree (life time), the scale limit of the forest (area limit), the ratio of the global seeding (Transfer Rate), and the number of local seeding changes (LSC) and number of global seeding changes (GSC). The FOA algorithm indicates that the parameters “life time”, “area limit” and “transfer rate” are independent of the number of data sets [11]. FSFOA sets these three parameters to a fixed value, with “life time” of 15, “area limit” of 50, and “transfer rate” of 5% [32]. The FOA algorithm indicates that the number of parameters LSC and GSC are related to the number of features [11]. FSFOA sets the Local Seeding Changes and Global Seeding Changes values for the 11 data sets, as shown in TABLE 3.

In order to prevent over-fitting problems, experimental data is usually divided into training sets, validation sets, and test sets. Validation sets are used to reduce over-fitting problems. Considering the small amount of data in this experiment, if the validation set is increased, fewer training sets will result in under-fitting. Therefore, the experiment did not use the validation

Table 3

Parameters of experimental dataset

Dataset	Features	Local Seeding Changes	Global Seeding Changes
SRBCT	2308	460	700
Sonar	60	12	30
Dermatology	34	7	15
Ionosphere	34	7	15
Segmentation	19	4	9
Hepatitis	19	4	10
Vehicle	18	4	9
Heart-statlog	13	3	6
Cleveland	13	3	6
Wine	13	3	6
Glass	9	2	4

set, but used a 10-fold cross-validation method, a 2-fold cross-validation method, 70% for the training and 30% for the testing dataset will be implemented. 10-fold cross-validation method refers to dividing the data set into 10 folds, 9 of which are used as training set and 1 used for test, and repeat 10 times to test each fold, the final outcome will be the average of 10 results.

The performance of the evaluation algorithm uses two functions: Classification Accuracy (CA) and Dimension Reduction (DR), CA and DR as shown in Eq. (2) and (3).

$$CA = \frac{N_CC}{N_AC} \quad (3)$$

N_CC is the number of correctly classified data in the test data, and N_AC is the total number of test data.

$$DR = 1 - \frac{N_SF}{N_AF} \quad (4)$$

N_SF is the number of features selected by the algorithm, and N_AF is the total number of features.

The value range of CA and DR is [0,1]. CA is the accuracy of the classification algorithm. The larger the CA value is, the better the classification performance. DR is the feature dimension selection ability of the feature selection algorithm. The larger the DR value, the smaller the algorithm dimension.

The fitness function uses KNN, C4.5, and SVM, and the parameters are shown in TABLE 4.

Table 4

Fitness functions and parameters

Fitness functions	Parameters
KNN	K=1, k=3, k=5
C4.5	J48
SVM	The core function is rbf

C. Data Analysis

The experimental results of FSIFOA and other algorithms in different data sets, different fitness functions and different verification methods have been shown in the following 10 tables. In the data sets "Sonar", "Dermatology", "Tonosphere", "Segmentation"

and "Vehicle", the test accuracy and dimensional reduction ability of the FSIFOA algorithm are improved under the same test conditions comparing to FSFOA. In the "SRBCT" dataset, the FSIFOA algorithm and FSFOA have the same test accuracy, and the FSIFOA algorithm has better performance in reducing dimensions of the features. In the "Heart-statlog" and "Wine" datasets, the FSIFOA algorithm has improved test accuracy and the dimension reduction capability under some conditions. In the "Cleveland" and "Glass" data sets, the test accuracy is improved, and the dimensional reduction ability is weakened under limited conditions.

In Table 5 to Table 14, FSIFOA is compared to other algorithms such as SFS, SBS, SFFS, NSM, SVM-FuzCoc, HGAFS, FS-NEIR, UFSACO and PSO. The results show that FSIFOA performs well in dimension

Table 5

Comparison between algorithms of FSIFOA and other algorithms on SRBCT

Algorithm	CA (%)	DR (%)	Classifier	Validation Method
FSFOA	94.73%	49.06%	1-NN	70%-30%
FSIFOA	94.73%	61.48%	1-NN	70%-30%
SVM-FuzCoc	89.47%	56.46%	1-NN	70%-30%

Table 6

Comparison between algorithms of FSIFOA and other algorithms on Sonar

Algorithm	CA (%)	DR (%)	Classifier	Validation Method
FSFOA	74.60%	56.67%	1-NN	70%-30%
FSIFOA	76.19%	76.67%	1-NN	70%-30%
SVM-FuzCoc	73.17%	68.33%	1-NN	70%-30%
FSFOA	82.69%	52.45%	J48	10-fold
FSIFOA	85.18%	65%	J48	10-fold
FS-NEIR	75.97%	91.66%	J48	10-fold
FSFOA	71.43%	60%	5-NN	70%-30%
FSIFOA	74.60%	68.33%	5-NN	70%-30%
PSO	72.22%	90%	5-NN	70%-30%
FSFOA	72.11%	63.33%	svm	2-fold
FSIFOA	75.94%	78.33%	svm	2-fold
HGAFS	73.65%	65%	svm	2-fold

Table 7

Comparison between algorithms of FSIFOA and other algorithms on Vehicle

Algorithm	CA (%)	DR (%)	Classifier	Validation Method
FSFOA	73.04%	31.57%	J48	10-fold
FSIFOA	77.74%	44.44%	J48	10-fold
FS-NEIR	70.98%	50%	J48	10-fold
FSFOA	73.98%	50%	5-NN	70%-30%
FSIFOA	76.77%	55.56%	5-NN	70%-30%
PSO	85.30%	69.40%	5-NN	70%-30%
FSFOA	62.41%	47.22%	svm	2-fold
FSIFOA	69.03%	66.67%	svm	2-fold
HGAFS	76.36%	38.89%	svm	2-fold

Table 8

Comparison between algorithms of FSIFOA and other algorithms on Ionosphere

Algorithm	CA (%)	DR (%)	Classifier	Validation Method
FSFOA	93.16%	68.57%	J48	10-fold
FSIFOA	96.62%	61.76%	J48	10-fold
FS-NEIR	92.59%	82.35%	J48	10-fold
FSFOA	92.30%	61.76%	3-NN	10-fold
FSIFOA	93.83%	76.47%	3-NN	10-fold
NSM	92.00%	88.23%	3-NN	10-fold
FSFOA	89.43%	54.28%	5-NN	10-fold
FSIFOA	93.23%	79.41%	5-NN	10-fold
PSO	87.27%	90.41%	5-NN	10-fold
FSFOA	94.28%	57.14%	svm	2-fold
FSIFOA	95.16%	58.82%	svm	2-fold
FSFOA	89.52%	54.28%	1-NN	70%-30%
FSIFOA	95.16%	61.76%	1-NN	70%-30%
SVM-FuzCoc	89.46%	88.23%	1-NN	70%-30%
FSFOA	95.12%	47.05%	J48	70%-30%
FSIFOA	99.06%	67.65%	J48	70%-30%
UFSACO	88.61%	11.17%	J48	70%-30%

Table 9

Comparison between algorithms of FSIFOA and other algorithms on Segmentation

Algorithm	CA (%)	DR (%)	Classifier	Validation Method
FSFOA	96.20%	30%	3-NN	10-fold
FSIFOA	96.88%	52.63%	3-NN	10-fold
NSM	95%	63.15%	3-NN	10-fold

Table 10

Comparison between algorithms of FSIFOA and other algorithms on Dermatology

Algorithm	CA (%)	DR (%)	Classifier	Validation Method
FSFOA	96.99%	21.42%	J48	10-fold
FSIFOA	97.81%	73.53%	J48	10-fold
FS-NEIR	68.53%	22.22%	J48	10-fold
FSFOA	97.27%	45.71%	1-NN	70%-30%
FSIFOA	99.07%	58.82%	1-NN	70%-30%
SBS	91.78%	58.23%	1-NN	70%-30%
SFFS	93.70%	62.35%	1-NN	70%-30%
FSFOA	90.09%	44.11%	J48	70%-30%
FSIFOA	98.15%	70.59%	J48	70%-30%
UFSACO	95.28%	26.47%	J48	70%-30%

Table 11

Comparison between algorithms of FSIFOA and other algorithms on Heart-statlog

Algorithm	CA (%)	DR (%)	Classifier	Validation Method
FSFOA	85.15%	48.07%	J48	10-fold
FSIFOA	84.07%	61.54%	J48	10-fold
FS-NEIR	75.97%	91.66%	J48	10-fold
FSFOA	85.18%	35.71%	3-NN	10-fold
FSIFOA	83.33%	53.85%	3-NN	10-fold
NSM	84%	69.23%	3-NN	10-fold
FSFOA	84.07%	50%	svm	2-fold
FSIFOA	84.81%	76.92%	svm	2-fold
HGAFS	82.59%	76.92%	svm	2-fold

Table 12

Comparison between algorithms of FSIFOA and other algorithms on Cleveland

Algorithm	CA (%)	DR (%)	Classifier	Validation Method
FSFOA	55.55%	71.42%	1-NN	70%-30%
FSIFOA	62.22%	61.54%	1-NN	70%-30%
SVM-FuzCoc	61.01%	46.10%	1-NN	70%-30%

Table 13

Comparison between algorithms of FSIFOA and other algorithms on Wine

Algorithm	CA (%)	DR (%)	Classifier	Validation Method
FSFOA	96.06%	21.42%	J48	10-fold
FSIFOA	97.25%	53.85%	J48	10-fold
FS-NEIR	95.04%	61.53%	J48	10-fold
FSFOA	98.87%	42.58%	3-NN	10-fold
FSIFOA	95.61%	61.54%	3-NN	10-fold
NSM	98%	53.84%	3-NN	10-fold
FSFOA	98.07%	50%	1-NN	70%-30%
FSIFOA	95.61%	61.54%	1-NN	70%-30%
SVM-FuzCoc	97.12%	53.84%	1-NN	70%-30%
SFS	97.69%	35.38%	1-NN	70%-30%
SBS	94.77%	46.15%	1-NN	70%-30%
SFFS	96.56%	36.92%	1-NN	70%-30%
FSFOA	96%	57.14%	J48	70%-30%
FSIFOA	96.73%	61.54%	J48	70%-30%
UFSACO	95.08%	61.53%	J48	70%-30%
FSFOA	99.20%	30.76%	5-NN	70%-30%
FSIFOA	95.70%	38.46%	5-NN	70%-30%

Table 14

Comparison between algorithms of FSIFOA and other algorithms on Glass

Algorithm	CA (%)	DR (%)	Classifier	Validation Method
FSFOA	75.70%	50%	J48	10-fold
FSIFOA	78.13%	33.33%	J48	10-fold
FS-NEIR	93.95%	70.58%	J48	10-fold
FSFOA	71.88%	40%	1-NN	70%-30%
FSIFOA	75.38%	55.56%	1-NN	70%-30%
SFFS	71.77%	37.77%	1-NN	70%-30%
FSFOA	68.22%	60%	svm	2-fold
FSIFOA	68.69%	33.33%	svm	2-fold
HGAFS	65.51%	44.44%	svm	2-fold

reduction and classification accuracy. For example: in Sonar data set, the classification accuracy of FSIFOA algorithm is higher than that of SVM-FuzCoc algorithm and FS-NEIR algorithm Method, PSO algorithm and HGAFS algorithm. The dimension reduction value of FSIFOA algorithm is higher than that of SVM-FuzCoc algorithm and HGAFS algorithm.

In the Ionosphere data set, the classification accuracy of FSIFOA algorithm is bigger than that of FS-NEIR algorithm, NSM algorithm, PSO algorithm, SVM-FuzCoc algorithm and UFSACO algorithm. The dimension reduction value of FSIFOA algorithm is higher than that of UFSACO algorithm.

In the Dermatology data set, the classification accuracy of FSIFOA algorithm is higher than that of FS-NEIR algorithm, SBS algorithm, SFFS algorithm and UFSACO algorithm. The dimension reduction value of FSIFOA algorithm is higher than that of FS-NEIR algorithm, SBS algorithm and UFSACO algorithm.

In Wine data set, the classification accuracy of FSIFOA is higher than FS-NEIR algorithm, SBS algorithm and UFSACO algorithm. The dimension reduction of FSIFOA algorithm is higher than NSM algorithm, SVM-FuzCoc algorithm, SFS algorithm, SBS algorithm, SFFS algorithm and UFSACO algorithm, and only lower than FS-NEIR algorithm.

In conclusion, the FSIFOA performs better than all the other algorithms in predicting accuracy in the SRBCT, Sonar, Ionosphere, Segmentation, Dermatology, and Cleveland data sets.

But in the other data sets, such as Vehicle, Heart-statlog, Wine, and Glass, the FSIFOA algorithm showed better predictive accuracy than only some other algorithms.

D. Analysis of Experiment Results

First, in the "SRBCT" dataset, the FSIFOA algorithm and FSFOA have the same test accuracy because the number of features in the "SRBCT" data set is 2308 but the data is only 63, and the number of features is much larger than the number of data. Such a data set is easy to fall into overfitting issue. The "SRBCT" dataset uses 70% for the training and 30% for the testing dataset. A total of 19 test data, 94.73% of the test accuracy, 18 of the 19 data are classified correctly, and only one classification error. The test accuracy of 94.73% reached the limit of the classification accuracy rate. If the accuracy rate is increased, the classification accuracy rate will reach 100%.

Second, in the “Heart-statlog” and “Wine” data sets, the classification accuracy rate has decreased in some situation. In the “Cleveland” and “Glass” data sets, the dimensional reduction ability has declined in some cases. The main reason is that the four datasets (“Heart-statlog”, “Wine”, “Cleveland”, and “Glass”) have the smallest feature dimensions compared to other datasets, with 13 features and 9 features, respectively. It indicates FSIFOA has limited classification performance and dimensional reduction ability on data sets with too small dimensions.

Third, in the data sets “Sonar”, “Dermatology”, “Ionosphere”, “Segmentation”, and “Vehicle”, the FSIFOA algorithm has improved classification performance and dimensional reduction ability. The FSIFOA algorithm has good performance in the data set of medium and large dimensions.

Fourth, the disadvantage of this algorithm is the random initialization of the algorithm initialization process. Now the algorithm complexity is increased by using Pearson correlation coefficient and L2 regularization. The algorithm divides the candidate forest into high quality trees and low quality trees, which further increases the cost of the algorithm. Furthermore, although the FSIFOA performs better than FS-

FOA in large and middle dimension datasets, it does not perform better than FSFOA in low dimension dataset. The dimensionality reduction capability of FSIFOA algorithm is also not significantly better than other algorithms.

6. Conclusion

The FSIFOA algorithm proposes three improvements: firstly, the Pearson correlation coefficient and L1 regularization are used instead of the random initialization problem in the initialization stage. Secondly, the good trees and the bad trees are separated, and the quantity gap between them will be filled to solve the problem of category imbalance. Thirdly, in the update stage, trees with the same precision but different dimensions are added to the forest. The data sets with small, medium and large dimensions are tested through experiments.

The new algorithm is compared with FSFOA, NSM, PSO and other algorithms. The results show that the FSIFOA is better than FSFOA in medium and large dimension datasets. The FSIFOA also has advantages over other algorithms in 6 datasets out of 10 datasets.

References

- Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, 1992, 46(3), 175-185. <https://doi.org/10.1080/00031305.1992.10475879>
- Blake, C., Keogh, E., Merz, C. J. UCI Repository of Machine Learning Databases. University of California, Irvine, 1995. Available: [http://www.ics.uci.edu/\\$mlearn/MLRepository.html](http://www.ics.uci.edu/$mlearn/MLRepository.html)
- Chiang, L. H., Pell, R. J. Genetic Algorithms Combined with Discriminant Analysis for Key Variable Identification. *Journal of Process Control*, 2004, 14(2), 143-155. [https://doi.org/10.1016/S0959-1524\(03\)00029-5](https://doi.org/10.1016/S0959-1524(03)00029-5)
- Cortes, C., Vapnik, V. Support-Vector Networks. *Machine Learning*, 1995, 20(3), 273-297. <https://doi.org/10.1023/A:1022627411411>
- Diao, R., Chao, F., Peng, T., Snooke, N., Shen, Q. Feature Selection Inspired Classifier Ensemble Reduction. *IEEE Transactions on Cybernetics*, 2017, 44(8), 1259-1268. <https://doi.org/10.1109/TCYB.2013.2281820>
- Dong, H., Li, T., Ding, R., Sun, J. A Novel Hybrid Genetic Algorithm with Granular Information for Feature Selection and Optimization. *Applied Soft Computing*, 2018, 65, 33-46. <https://doi.org/10.1016/j.asoc.2017.12.048>
- Dorigo, M., Birattari, M. *Ant Colony Optimization*. Springer, New York, 2010.
- Dorigo, M., Stützle, T. *Ant Colony Optimization: Overview and Recent Advances*. *Handbook of Metaheuristics*, Springer, 2019, 311-351. https://doi.org/10.1007/978-3-319-91086-4_10
- Eberhart, R., Kennedy, J. *Particle Swarm Optimization*. *Proceedings of ICNN'95 - International Conference on Neural Networks*, Perth, WA, Australia, December 1, 1995. <http://dx.doi.org/10.1109/ICNN.1995.488968>.
- Ghaemi, M., Feizi-Derakhshi, M. R. Feature Selection using Forest Optimization Algorithm. *Pattern Recognition*, 2016, 60, 121-129. <https://doi.org/10.1016/j.patcog.2016.05.012>
- Ghaemi, M., Feizi-Derakhshi, M. R. Forest Optimization Algorithm. *Expert Systems with Applications*, 2014, 41(15), 6676-6687. <https://doi.org/10.1016/j.eswa.2014.05.009>

12. Guyon, I., Weston, J., Barnhill, S., Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 2002, 46(1-3), 389-422. <https://doi.org/10.1023/A:1012487302797>
13. Hsu, W. H. Genetic Wrappers for Feature Selection in Decision Tree Induction and Variable Ordering in Bayesian Network Structure Learning. *Information Sciences*, 2004, 163(1-3), 103-122. <https://doi.org/10.1016/j.ins.2003.03.019>
14. Huang, J., Cai, Y., Xu, X. A Hybrid Genetic Algorithm for Feature Selection Wrapper Based on Mutual Information. *Pattern Recognition Letters*, 2007, 28(13), 1825-1844. <https://doi.org/10.1016/j.patrec.2007.05.011>
15. Kabir, M. M., Shahjahan, M., Murase, K. A New Hybrid Ant Colony Optimization Algorithm for Feature Selection. *Expert Systems with Applications*, 2012, 39(3), 3747-3763. <https://doi.org/10.1016/j.eswa.2011.09.073>
16. Kennedy, J. Particle Swarm Optimization. *Encyclopedia Machine Learning*, 2010, 760-766. <http://dx.doi.org/10.1109/ICNN.1995.488968>
17. Ksiazek K, Polap D, Wozniak M, et al. Radiation Heat Transfer Optimization by the Use of Modified Ant Lion Optimizer. 2017 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2017. <https://doi.org/10.1109/SSCI.2017.8280853>
18. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. L., Liu, H. Feature Selection: A Data Perspective. *ACM Computing Surveys*, 2016, 50(6), 94. <https://doi.org/10.1145/3136625>
19. Lu, K., Zhou, W., Zeng, G., Zheng, Y. Constrained Population Extremal Optimization-Based Robust Load Frequency Control of Multi-Area Interconnected Power System. *International Journal of Electrical Power & Energy Systems*, 2019, 105, 249-271. <https://doi.org/10.1016/j.ijepes.2018.08.043>
20. Molina, L. C., Belanche, L., Nebot, À. Feature Selection Algorithms: A Survey and Experimental Evaluation. *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002)*, 9-12 December 2002, Maebashi City, Japan. IEEE.
21. Moustakidis, S. P., Theocharis, J. B. Svm-fuzcoc: A Novel Svm-Based Feature Selection Method using a Fuzzy Complementary Criterion. *Pattern Recognition*, 2010, 43(11), 3712-3729. <https://doi.org/10.1016/j.patcog.2010.05.007>
22. Nag, K., Pa, N. R. A Multiobjective Genetic Programming-Based Ensemble for Simultaneous Feature Selection and Classification. *IEEE Transactions on Cybernetics*, 2015, 46(2):499-510. <https://doi.org/10.1109/TCYB.2015.2404806>
23. Nie, D. G. Improvement of Forest Optimization Algorithm and Discrete Study. (M.S. thesis). Lanzhou University, Lanzhou, China, 2016.
24. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, 2014.
25. Tang, B., Kay, S., He, H. B. Toward Optimal Feature Selection in Naive Bayes for Text Categorization. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 28(9). <https://doi.org/10.1109/TKDE.2016.2563436>
26. Tibshirani, R. Regression Shrinkage and Selection Via The Lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, 1996, 58(1), 267-288. <https://doi.org/10.1111/j.1467-9868.2011.00771.x>
27. Tikhonov, A. N., Arsenin, V. I. Solutions of Ill-posed Problems. John Wiley & Sons, New York, Toronto, London, Sydney, 1977.
28. Tran, B., Xue, B., Zhang, M. Overview of Particle Swarm Optimisation for Feature Selection in Classification. *Applied Soft Computing*, 2014, 18, 261-276. <https://doi.org/10.1016/j.asoc.2013.09.018>
29. Trivedi, A., Srinivasan, D., Sanyal, K., Ghosh, A. A Survey of Multiobjective Evolutionary Algorithms based on Decomposition. *IEEE Transactions on Evolutionary Computation*, 2017, 21(3), 440-462. <http://dx.doi.org/10.1109/TEVC.2016.2608507>
30. Wan, Y., Wang, M., Ye, Z., Lai, X. A Feature Selection Method based on Modified Binary Coded Ant Colony Optimization Algorithm. *Applied Soft Computing*, 2016, 49, 248-258. <https://doi.org/10.1016/j.asoc.2016.08.011>
31. Woźniak, M., Połap, D., Napoli, C., Tramontana, E. Graphic Object Feature Extraction System Based on Cuckoo Search Algorithm. *Expert Systems with Applications*, 2016, 66, 20-31. <https://doi.org/10.1016/j.eswa.2016.08.068>
32. Xue, B., Zhang, M., Browne, W. N. Particle Swarm Optimisation for Feature Selection in Classification: Novel Initialisation and Updating Mechanisms. *Applied Soft Computing*, 2014, 18, 261-276. https://doi.org/10.1007/978-3-319-13563-2_51
33. Xue, B., Zhang, M., Browne, W. N., Yao, X. A Survey on Evolutionary Computation Approaches to Feature Selection. *IEEE Transactions on Evolutionary Computation*, 2015, 20(4), 606-626. <https://doi.org/10.1109/TEVC.2015.2504420>

34. Yang, J., Honavar, V. Feature Subset Selection using a Genetic Algorithm. *Feature Extraction, Construction and Selection*, Springer, New York, 1998, 117-136. https://doi.org/10.1007/978-1-4615-5725-8_8
35. Zhang, F., Chan, P. P. K., Biggio, B., Yeung, D. S., Roli, F. Adversarial Feature Selection against Evasion Attacks. *IEEE Transactions on Cybernetics*, 2014, 6(3), 766-777. <https://doi.org/10.1109/TCYB.2015.2415032>
36. Zhang, X., Mei, C. L., Chen, D. G., Li, J. Feature Selection in Mixed Data: A Method using a Novel Fuzzy Rough Set-Based Information Entropy. *Pattern Recognition*, 2016, 56(1), 1-15. <https://doi.org/10.1016/j.patcog.2016.02.013>
37. Zhang, Y., Gong, D., Hu, Y., Zhang, W. Feature Selection Algorithm based on Bare Bones Particle Swarm Optimization. *Neurocomputing*, 2015, 148, 150-157. <https://doi.org/10.1016/j.neucom.2012.09.049>
38. Zhao, F., Zeng, G., Lu, K. EnLSTM-WPEO: Short-Term Traffic Flow Prediction by Ensemble LSTM, NNCT Weight Integration and Population Extremal Optimization. *IEEE Transactions on Vehicular Technology*, 2019, 1-1. <https://doi.org/10.1109/TVT.2019.2952605>
39. Zhou, Z. H. *Machine Learning*. Tsinghua University Press, Beijing, China, 2016.