Comparative Study on Metaheuristic-Based Feature Selection for Cotton Foreign Fibers Recognition

Xuehua Zhao, Xueyan Liu, Daoliang Li, Huiling Chen, Shuangyin Liu, Xinbin Yang, Shaobin Zhan, Wenyong Zhao

Comparative Study on Metaheuristic-based Feature Selection for Cotton Foreign Fibers Recognition

Xuehua Zhao[1], Xueyan Liu[1,2], Daoliang Li[3,*], Huiling Chen[4], Shuangyin Liu[5]
Xinbin Yang[1], Shaobin Zhan[1], Wenyong Zhao[1]

[1]School of Digital Media, Shenzhen Institute of Information Technology, Shenzhen 518172, China
lcrlc@sina.com
[2]Key Laboratory of Symbolic Computation and Knowledge Engineer (Jilin University), Ministry of Education, Changchun, Jilin 130012, China
dyyzlxy@163.com
[3]College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
dliangl@cau.edu.cn
[4]College of Physics and Electronic Information, Wenzhou University, Wenzhou 325035, China
chenhuiling.jlu@gmail.com
[5]College of Information, Guangdong Ocean University, Zhanjiang Guangdong 524025, China
hdlsyxlq@126.com

Abstract. The excellent feature set or feature combination of cotton foreign fibers is great significant to improve the performance of machine-vision-based recognition system of cotton foreign fibers. To find the excellent feature sets of foreign fibers, in this paper presents three metaheuristic-based feature selection approaches for cotton foreign fibers recognition, which are particle swarm optimization, ant colony optimization and genetic algorithm, respectively. The k-nearest neighbor classifier and support vector machine classifier with k-fold cross validation are used to evaluate the quality of feature subset and identify the cotton foreign fibers. The results show that the metaheuristic-based feature selection methods can efficiently find the optimal feature sets consisting of a few features. It is highly significant to improve the performance of recognition system for cotton foreign fibers.

Keywords: Metaheuristic, Feature selection, Foreign fibers, Recognition system.

1 Introduction

The cotton foreign fibers, such as ropes, wrappers, plastic films and so on, are closely related to the quality of the final cotton textile products [1]. In the recent years, the machine-vision-based recognition systems have been widely used to assess the quality

adfa, p. 1, 2011.

of cottons [2, 3], in which classification accuracy is an key measure to validate the performance of recognition systems. To improve the classification accuracy, finding the optimal feature sets with high accuracy is an efficient way due to because it can improve the accuracy and speed of reconition systems.

Feature selection (FS) is a main approach to find the optimal feature sets by reduce the irrelevant or redundant features. Curently, FS has been used to the area of machine learning and data mining [4]. Since to find the optimum feature sets is a NP problem, the researchers begin to turn to find the near optimal feature set and have proposed many algorithms [5, 6].

Currently, metaheuristic algorithms have attracted so much attention, the representive algorithms are particle swarm optimization (PSO for short), ant colony optimization (ACO for short) and genetic algorithm (GA for short) [5, 7, 8]. For metaheuristic algorithms, they are firstly given an evaluating measure of the quality of feature sets, and iteratively improve a specific candidate set. Finally, the excellent feature sets are obtained. The metaheuristic algorithms makes few assumptions on the optimal feature sets and find the optimal feature sets in very large search spaces. This is very suitable for the FS problem.

In this paper, three metaheuristic algorithms for FS are presented to find the optimal feature combination of cotton foreign fibers, which are GA for FS (GAFS for short), ACO for FS (ACOFS for short) and PSO for FS (PSOFS for short). Two classifiers, which are the k-nearest neighbor (KNN for short) [9] and support vector machine (SVM for short) [10], are used to evaluate the quality of subsets and to identify cotton foreign fibers. The aim of our works is applying these algorithms to the data sets of cotton foreign fibers to discover the new and challenging results. The comparison analysises of these algorithms illustrate the excellent search ability of the proposed metaheuristic algorithms and the feature sets obtained by them can efficiently improve the performance of recognition systems of cotton foreign fibers.

The remainder is organized as follows. The applications context and the proposed FS methods are presented in Section 2. The results and discussion are describled in Section 3. Section 4 describles the conclusion.

2 Materials and methods

2.1 Application Background

The cotton foreign fibers usually fall into six groups, that are feather, hemp rope, plastic film, cloth, hair, polypropylene and, respectively. The cotton foreign fibers induce the quality of the cotton textile products [1]. The grade evaluation using machine-vision-based recognition systems is mainly an approach to solve this problem [2]. These systems in general have three key steps, which are image segmentation, feature extraction and classification of foreign fiber, respectively.

Considering real-time problem, the feature set, which includes a few number of features and has high accuracy, is important due to reduction of the detection time and improvement of classification accuracy. As a result, FS is important to the online recognition systems of cotton foreign fibers. In our work, we have applied three

metaheuristic algorithms to this area for obtaining the optimal feature sets of cotton foreign fibers.

2.2 Data Preparation

Firstly, we obtain the foreign fiber images by our test platform, and 1200 representative images including foreign fibers are selected to extract the dataset. The width of the obtained images is 4000 pixels and their height is 500 pixels. Several examples are shown in Fig. 1. These images are divided into six groups in terms of categories of foreign fibers, and every group contains 200 images.

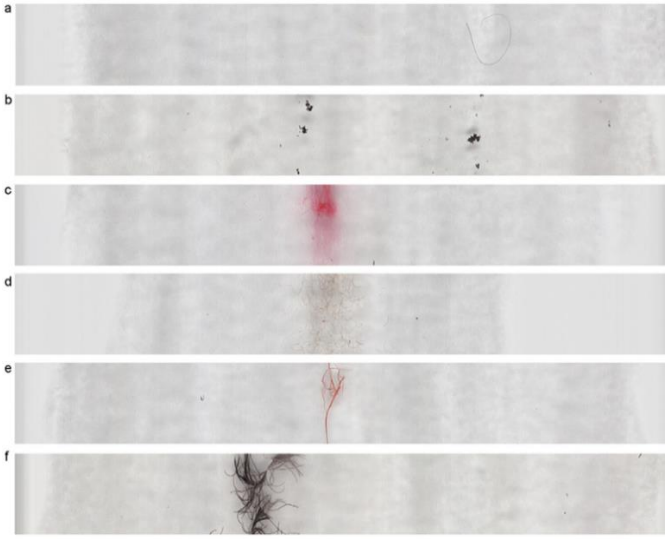


Fig. 1. Images of cotton foreign fibers: (a) hair image, (b) plastic film image, (c) cloth image, (d) hemp rope image, (e) polypropylene image, (f) feather image.

Then, we segment these images into small foreign fiber objects only including cotton foreign fiber. Finally, the 2808 objects are obtained and the number of hair, black plastic film, cloth, rope, polypropylene twines and feather objects is 204, 408, 432, 492, 528 and 744, respectively. The following step is extracting features from these objects.

The color, shape, and texture features can be extracted in cotton foreign fibers. Since accurate classification is difficult in only using one or two features [2]. Therefore we need extraction of all kinds of features including color, shape and texture features, and find the excellent feature combination by FS approaches.

In our experiment, a total of 80 features are extracted from foreign fiber objects, and the number of color, texture and shape features is 28, 42 and 10, respectively. These extracted features are used to build the 80-dimensional feature vector.

After the data is generated, normalization is implemented to reduce the impact of different dimensions.

2.3 Metaheuristic Algorithms for Feature Selection

Fitness Function. Fitness function is important for metaheuristic algorithms, it is used evaluate the quality of each subset. Considering the online classification problem, we expect that the found feature set should has small size and high accuracy. Therefore, fitness evaluation is designed to combine the accuracy of classifier with the length of feature subset. The specific fitness function in this paper is the following Eq. (1):

$$S(X) = \upsilon J(X) + \psi \mid X \mid \tag{1}$$

where X denotes the subset, $J(X)$ is the classification accuracy of subset X, $|X|$ denotes the feature number of the subset X, υ and ψ are used to adjust the relative importance of accuracy and size. In this study, two classifiers, KNN and SVM are adopted to evaluate the quality of subset.

Particle Swarm Optimization for Feature Selection. PSO belongs to population-based metaheuristics and is proposed by Kennedy and Eberhart [11]. PSO explores the search space by movements of particles with a velocity and each particle is updated based on its past best position and the current best particle. PSO can efficiently balance the exploration and exploitation and is an efficient optimization algorithm [12, 13].

Supposing the particle i is denoted as $\vec{X}_i = (x_{i,1}, x_{i,2}, \mathrm{L}\ , x_{i,d})$, which has the velocity $\vec{V}_i = (v_{i,1}, v_{i,2}, \mathrm{L}\ , v_{i,d})$. $\vec{P}_i = (p_{i,1}, p_{i,2}, \mathrm{L}\ , p_{i,d})$ denotes the past best position of the particle i. $\vec{P}_g = (p_{g,1}, p_{g,2}, \mathrm{L}\ , p_{g,d})$ denotes the best particle.

Each bit of the particle only lies in one of two states, i.e. zero or one, which will be changed according to probabilities. To change the velocity from continuous space to probability space, the following sigmoid function is using in algorithm:

$$sig(v_{i,j}) = \frac{1}{1 + \exp(-v_{i,j})}, \quad j = 1, 2, \mathrm{L}\ , d \tag{2}$$

The velocity is recalculated in terms of Eq. (3):

$$v_{i,j}^{t+1} = w \times v_{i,j}^{t} + c_1 \times r_1\ (p_{i,j}^{t} - x_{i,j}^{t}) + c_2 \times r_2\ (p_{g,j}^{t} - x_{i,j}^{t}) \tag{3}$$

where w denotes inertia weight and is updated at the iteration t according to Eq. (4):

$$w_t = w_{\min} + (w_{\max} - w_{\min}) \frac{(t_{\max} - t)}{t_{\max}} \tag{4}$$

where w_{max} denotes respectively the maximal value of the inertia weight, w_{min} denotes the minimum of the inertia weight. The t_{max} is the maximal times of iterations. The parameters c_1 and c_2 denote the acceleration coefficients. The parameters r_1 and r_2 are

random numbers varying from 0 to 1. $x_{i,j}$, $p_{i,j}$ and $p_{g,j}$ belong to zero or one. v_{max} is the maximum velocity.

The new particle position is updated by Eq. (5):

$$x_{i,j}^{t+1} = \begin{cases} 1, & if\ rnd < sig(v_{i,j}) \\ 0, & if\ rnd \geq sig(v_{i,j}) \end{cases}, j = 1,2,\mathrm{L}\ ,d \tag{5}$$

where and *rnd* denotes the random number in [0, 1] from uniform distribution.

Ant Colony Optimization for Feature Selection. ACO is proposed based on the idea of ants finding food by the shortest path between food source and nest [14]. In ACO, the addressed problem is modelled as a graph, in which the ants search a minimum cost path. The good paths mean the emergent result of the global cooperation among ants. In each iteration finding the optimal solutions, many ants construct their solutions by heuristic information and trail pheromone [15].

In the ACOFS, every candidate solution is mapped into an ant represented by a binary vector where the bit one or zero respectively means that the corresponding feature is selected or not.

Heuristic information: Heuristic information generally represents the attractiveness of each feature. If heuristic information is not used, the algorithm would be greedy, and the better solution is not found [5]. To evaluate the heuristic information [16], the information gain is calculated in this study.

Feature selection: In each iteration, the ant k determines whether the feature i is selected or not according to transition probability p_i. The transition probability p_i is given as follows:

$$p_i^k(t) = \begin{cases} \dfrac{[\tau_i(t)]^{\alpha}[\eta_i(t)]^{\beta}}{\sum_{u=J^k}[\tau_u]^{\alpha}[\eta_u]^{\beta}} & \text{if } i \in J^k \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where J^k represents the feasible feature set, η_i is the heuristic desirability of the feature i, τ_i is the pheromone value of the feature i. α and β is used to adjust the relative importance of the heuristic information and pheromone.

Pheromone update: After all ants have constructed their feature sets, the algorithm trigger the pheromone evaporation, and according to Eq. (7) each ant k deposits a quantity of pheromone $\Delta\tau_i^k(t)$, which is calculated according the following equation:

$$\Delta\tau_i^k(t) = \begin{cases} \gamma\left(s^k(t)\right) & \text{if } i \in s^k(t) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $s^k(t)$ denotes the found feature set by ant k at iteration t. $\gamma(x)$ denotes fitness function.

The pheromone can be updated according to the following rule:

$$\tau_i(t+1) = (1-\rho)\tau_i(t) + \sum_{k=1}^{m} \Delta\tau_i^k(t) + \Delta\tau_i^g(t) \tag{8}$$

where $\rho \in (0,1)$ denote the pheromone decay coefficient which can avoid stagnation. m denotes the number of ants and g is the best ant. For all ants, the pheromone is updated according to Eq. (8)

Genetic Optimization for Feature Selection. GA is proposed by Holland, which is a metaheuristic based on the idea of genetic and natural selection. GA has been used to solve the FS tasks [7]. In evolution, each species has to change its chromosome combination to adapt to the complicated and changing environment for surviving in the world. In GA, a chromosome denotes a potential solution of problem, which is evaluated by fitness function. In GA, a new generation with better survival abilities is generated by crossover and mutation. GA usually includes coding, selection, crossover, mutation.

Encoding: In the GAFS, a chromosome represents a candidate feature set which consist of genes, and a gene is a feature which are encoded in the form of binary strings. If the gene is coded into '1', the corresponding feature is selected, otherwise, the gene is coded into '0'. The bit i in the chromosome denotes the feature i. Each chromosome is initialized randomly.

Selection: The selection is to select the chromosomes into the new generation among the current population. A certain number of chromosomes are probabilistically selected into the next generation, where the probability of selecting chromosomes i is Eq. (9)

$$\Pr(i) = \frac{S(i)}{\sum_{j=1}^{m} S(j)} \tag{9}$$

where $S(i)$ denotes the fitness value of the chromosome i, m is the number of the chromosomes.

Crossover: The crossover operator is used to exchange genes between two chromosomes. The representative crossover operators are multi-point crossover, double-point crossover and single-point crossover [7]. In this study, double-point crossover is performed. After some chromosomes have been selected into the next generation's population, additional chromosomes are generated by using a crossover operation. Crossover takes two parent individuals, which are chosen from the current generation by using the probability function given by Eq. (9), and creates two offspring by recombining portions of both parents.

Mutation: mutation operator is used to determine the variety of the chromosomes, which let local variations into the chromosomes and keeps the diversity of the popula-

tion, This can help to find the good solution in search space. Here, the number of chromosomes depends on the mutation rate. Then a gene of the mutating chromosome is selected at random and its value is changed from '1' to '0' or '0' to '1', respectively.

3 Results and Discussion

In our experiments, the configuration of computer is as follows: CPU 2.66GHz, main memory 4.0GB and Windows 7 system. All the algorithms are coded and run in the Matlab development environment. Two different classifiers, KNN and SVM, are taken for evaluating the quality of solution. To efficiently evaluate the methods, the 10-fold cross validation is used in our experiments.

The parameters of PSOFS, ACOFS and GAFS is set according to Table 1. These parameters are selected in terms of experiences.

Table 1. Parameter settings for three FS algorithms

Parameters	PSOFS	ACOFS	GAFS
υ	0.9	0.9	0.9
ψ	0.1	0.1	0.1
m	50	50	50
$Iter_{\max}$	100	50	100
c_1	2		
c_2	1.5		
$w_{\max}$	0.95		
$w_{\min}$	0.4		
$v_{\max}$	3		
α		1	
β		1.5	
ρ		0.3	
p_r			0.9
p_m			0.05

Table 2. Comparison of performance of the five algorithms

Methods	Classifiers	Average accuracy (%)	Average number of feature
ACOFS	KNN	**92.7**	**20**
PSOFS		92.3	26
GAFS		91.8	27
ACOFS	SVM	91.5	**25**
PSOFS		**91.9**	31
GAFS		91.4	35

Table 2 shows the results of performance comparisons of three algorithms. As we can see, for KNN classifier, ACOFS has the best result among these methods, the selected subset has the smallest size and highest classification accuracy. For SVM,

PSOFS can obtain the subset with highest accuracy, but the subset obtained by ACOFS includes the least features.

Fig. 2 intuitively shows the accuracy of the different subsets obtained by three algorithms and the original set with KNN and SVM, respectively. As shown in Fig. 2, for KNN and SVM, all the optimal subsets obtained by three metaheuristic-based algorithms achieve much higher accuracy than the original set without feature selection.

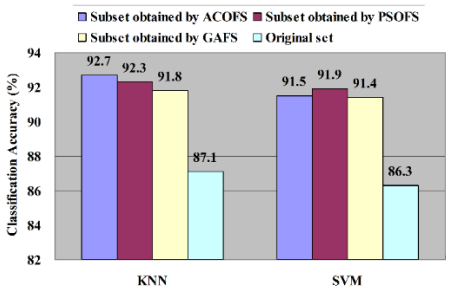


Fig. 2. Results of classification accuracy with KNN and SVM

Table 3. Classification results of the optimal set obtained by ACOFS

Classes	Results						Sample number	Accuracy (%)
	(1)	(2)	(3)	(4)	(5)	(6)		
(1) plastic film	396	0	0	0	4	8	408	97.06
(2) cloth	0	408	0	0	0	24	432	**94.44**
(3) hemp rope	0	8	416	0	28	40	492	84..55
(4) hair	0	0	0	196	0	8	204	**96.08**
(5) Polypropylene	0	0	4	20	504	0	528	95.45
(6) feather	8	48	16	0	12	660	744	88.71
Average								**92.72**

Table 4. Classification results of the optimal set obtained by PSOFS

Classes	Results						Sample number	Accuracy (%)
	(1)	(2)	(3)	(4)	(5)	(6)		
(1) plastic film	396	0	0	0	0	12	408	97.06
(2) cloth	0	408	0	0	0	24	432	**94.44**
(3) hemp rope	0	16	412	0	28	36	492	83.74
(4) hair	0	0	0	192	4	8	204	**94.12**
(5)Polypropylene	0	0	16	8	504	0	528	95.45
(6) feather	12	44	0	24	0	664	744	89.25
Average								**92.34**

Table 5. Classification results of the optimal set obtained by GAFS

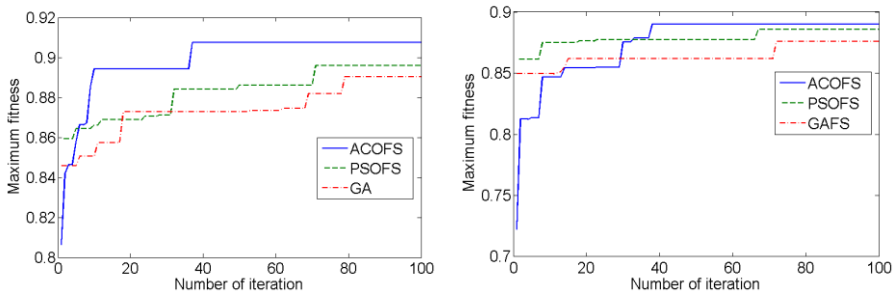
Classes	Results	Sample	Accuracy

	(1)	(2)	(3)	(4)	(5)	(6)	number	(%)
(1) plastic film	396	0	0	0	0	12	408	97.06
(2) cloth	0	408	0	0	0	24	432	**94.44**
(3) hemp rope	0	12	408	0	24	48	492	82.93
(4) hair	0	0	0	192	12	0	204	**94.12**
(5) Polypropylene	0	0	24	12	492	0	528	93.18
(6) feather	12	48	0	12	12	660	744	88.71
Average								**91.81**

Table 6. Classification results of the original set

Classes	Results						Sample number	Accuracy (%)
	(1)	(2)	(3)	(4)	(5)	(6)		
(1) plastic film	384	0	0	0	0	24	408	94.12
(2) cloth	0	372	20	0	0	40	432	**86.11**
(3) hemp rope	0	12	408	24	12	36	492	82.93
(4) hair	0	0	32	160	0	12	204	**78.43**
(5) Polypropylene	0	0	28	0	496	4	528	93.94
(6) feather	24	60	0	12	0	648	744	87.10
Average								**87.10**

Tables 3-6 show the detailed classification results of the subset obtained by three meta-heuristic algorithms and the original set using KNN classifier. As shown in Tables 3-6, the classification accuracy of the cloth and hair is efficiently improved by using the subset selected by three meta-heuristic algorithms, at least increased by 8% and 15%, respectively.



(a) KNN (b) SVM

Fig. 3. Fitness of three methods with KNN and SVM

The curves of fitness of the three metaheuristic-based algorithms with KNN, SVM in a certain run are shown in Fig. 3, respectively. The curves shown in Fig. 3 are representative according to our preliminary experiments. As we can see in Fig. 3, for KNN and SVM, ACOFS is more efficient and faster than PSOFS and GAFS. On average, PSOFS and GACOFS need about 70 iterations to find the optimum subset, while ACOFS only needs less than 40 iterations.

4 Conclusions

A key issue in machine-vision-based recognition system of cotton foreign fibers is to find the optimal feature set. In this study, FS based on metaheuristic optimization, namely ACOFS, PSOFS and GAFS, have been proposed to address this FS problem. Two different classifiers, KNN and SVM are taken for evaluating the quality of solution and classification. The experimental results show the presented methods have the excellent ability of finding a reduced set of features with high accuracy in the dataset of cotton foreign fiber. The selected feature set is great significant for machine-vision-based recognition system for cotton foreign fibers. In our future work, we will focus on improving the performance of classifiers used in the recognition rate of recognition systems of cotton foreign fibers.

Acknowledgments: This study is funded by the National Natural Science Foundation of China (61402195, 61471133 and 61571444), Guangdong Natural Science Foundation (2016A030310072), Guangdong Science and Technology Plan Project (2015A070709015 and 2015A020209171), the Science and Technology Plan Project of Wenzhou, China (G20140048), Shenzhen strategic emerging industry development funds (JCYJ20140418100633634.

References

1. Yang, W., Li, D., Zhu, L., Kang, Y., Li, F.: A New Approach for Image Processing in Foreign Fiber Detection. Comput. Electron. Agr. 68(1), 68-77 (2009)
2. Li, D., Yang, W., Wang, S.: Classification of Foreign Fibers in Cotton Lint Using Machine Vision and Multi-Class Support Vector Machine. Comput. Electron. Agr. 74(2), 274-279(2010)
3. Yang, W., Lu, S., Wang, S., Li, D.: Fast Recognition of Foreign Fibers in Cotton Lint Using Machine Vision. Math. Comput. Model. 54(3), 877-882 (2011)
4. Lin, J. Y., Ke, H. R., Chien, B. C., Yang, W. P.: Classifier Design with Feature Selection and Feature Extraction Using Layered Genetic Programming. Expert. Syst. Appl. 34(2), 1384-1393(2008)
5. Bolón-Canedo, V., Sánchez-Maro?o, N., Alonso-Betanzos, A.: A Review Of Feature Selection Methods on Synthetic Data. Knowl. Inf. Syst. 34(3), 483-519 (2013)
6. Sun, Z., Bebis, G., Miller, R.: Object Detection Using Feature Subset Selection. Pattern Recogn. 37(11), 2165-2176 (2004)
7. Pedergnana, M., Marpu, P. R., Dalla Mura, M., Benediktsson, J. A., Bruzzone, L.: A Novel Technique for Optimal Feature Selection in Attribute Profiles Based on Genetic Algorithms. IEEE T. Geosci. Remot. 51(6), 3514-3528 (2013)
8. Chen, H. L., Yang, B., Wang, G., Liu, J., Xu, X., Wang, S. J., Liu, D. Y.: A Novel Bankruptcy Prediction Model based on an Adaptive Fuzzy K-Nearest Neighbor Method. Knowl-Based Syst. 24(8), 1348-1359 (2011)
9. Cover, T., Hart, P.: Nearest Neighbor Pattern Classification. IEEE T. Inform. Theory, 13(1), 21-27 (1967)
10. Xuegong, Z.: Introduction to Statistical Learning Theory and Support Vector Machines. Acta Automatica Sinica, 26(1), 32-42 (2000)
11. Eberhart, R. C., Kennedy, J.: A New Optimizer Using Particle Swarm Theory. In Proceedings of the sixth international symposium on micro machine and human science, Vol. 1, pp. 39-43 (1995).

12. Chen, H. L., Yang, B., Wang, G., Wang, S. J., Liu, J., Liu, D. Y.: Support Vector Machine Based Diagnostic System for Breast Cancer Using Swarm Intelligence. J Med. Syst. 36(4), 2505-2519 (2012).
13. Kennedy, J., Eberhart, R. C.: A Discrete Binary Version of the Particle Swarm Algorithm. In IEEE International Conference on Systems, Man, and Cybernetics 1997, Vol. 5, pp. 4104-4108 (1997)
14. Dorigo, M., Maniezzo, V., Colorni, A.: Ant System: Optimization by a Colony of Cooperating Agents. IEEE T Syst. Man Cy. B, 26(1), 29-41(1996)
15. Zhao, X., Li, D., Yang, B., Ma, C., Zhu, Y., Chen, H.: Feature Selection Based on Improved Ant Colony Optimization for Online Detection of Foreign Fiber in Cotton. Appl. Soft Comput. 24, 585-596.24 (2014)
16. Forsati, R., Moayedikia, A., Jensen, R., Shamsfard, M., Meybodi, M. R.: Enriched Ant Colony Optimization and its Application in Feature Selection. Neurocomputing, 142, 354-371 (2014)