



# A novel hybrid algorithm for feature selection

Yuefeng Zheng<sup>1,2,3</sup> · Ying Li<sup>1,2</sup> · Gang Wang<sup>1,2</sup> · Yupeng Chen<sup>1,2</sup> · Qian Xu<sup>1,2</sup> · Jiahao Fan<sup>1,2</sup> · Xueting Cui<sup>1,2</sup>

Received: 28 March 2018 / Accepted: 5 May 2018

© Springer-Verlag London Ltd., part of Springer Nature 2018

## Abstract

Feature selection is an important filtering method for data analysis, pattern classification, data mining, and so on. Feature selection reduces the number of features by removing irrelevant and redundant data. In this paper, we propose a hybrid filter-wrapper feature subset selection algorithm called the maximum Spearman minimum covariance cuckoo search (MSMCCS). First, based on Spearman and covariance, a filter algorithm is proposed called maximum Spearman minimum covariance (MSMC). Second, three parameters are proposed in MSMC to adjust the weights of the correlation and redundancy, improve the relevance of feature subsets, and reduce the redundancy. Third, in the improved cuckoo search algorithm, a weighted combination strategy is used to select candidate feature subsets, a crossover mutation concept is used to adjust the candidate feature subsets, and finally, the filtered features are selected into optimal feature subsets. Therefore, the MSMCCS combines the efficiency of filters with the greater accuracy of wrappers. Experimental results on eight common data sets from the University of California at Irvine Machine Learning Repository showed that the MSMCCS algorithm had better classification accuracy than the seven wrapper methods, the one filter method, and the two hybrid methods. Furthermore, the proposed algorithm achieved preferable performance on the Wilcoxon signed-rank test and the sensitivity-specificity test.

**Keywords** Cuckoo search algorithm · Classification · Dimensionality reduction · Feature selection  
Maximum Spearman and minimum covariance

## 1 Introduction

In machine learning and data mining, many practical applications of classification include a large volume of data as well as involve a large number of features or attributes. In these data sets, there may be some redundant or irrelevant features or attributes. In order to improve the accuracy of classification, dimensionality reduction is a very effective method. Dimensionality reduction methods can be categorized into feature extraction and feature selection [1–6]. Feature extrac-

tion methods mix original features to generate a new feature subset. The new feature subset is the combination of the original ones. Some original features in new feature subset may lost [4–7]. Unlike feature extraction, feature selection finds a new set which is made up of the original features without any change. In order to improve the classification accuracy, eliminate redundant and irrelevant features, and retain the original feature information, feature selection is used to find the optimal feature subset [1–3].

Feature selection algorithms are divided into three categories: wrapper methods, filter methods, and hybrid methods.

Wrapped feature selection methods evaluate a subset of candidate features using a given learning algorithm and select the best subset of features directly from all features in the data set. Recently, metaheuristic algorithms have attracted a lot of attention due to their good performance in solving the feature selection problem [8]. Metaheuristic algorithms generally include ant lion optimization (ALO) [9], bat algorithm (BA) [10, 11], bacterial foraging optimization (BFO) [12, 13], cuckoo search (CS) [14, 15], genetic algorithm (GA) [16, 17], particle swarm optimization (PSO) [18, 19], and simulated annealing (SA) [20, 21]. Among them, the CS algorithm uses the levy

---

✉ Gang Wang  
wanggang.jlu@gmail.com

<sup>1</sup> College of Computer Science and Technology, Jilin University, Changchun 130012, People's Republic of China

<sup>2</sup> Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, People's Republic of China

<sup>3</sup> BODA College of Jilin Normal University, Siping, People's Republic of China

fight technique and has two selection operations on the nest, which can increase the effect of feature selection; therefore, this paper uses the CS algorithm as a wrapper method. However, because of the direct selection from all the features, the wrapping approach requires a high computational cost for the optimal classification accuracy.

Compared with the wrapping algorithm, the filtering algorithm has a lower time complexity. The filter method sorts the features in the data set according to certain rules and then selects the best features [3, 22, 23]. The filter algorithm is divided into univariate approach and multivariate approach according to the number of variables in the rule [24, 25]. The measurement of univariate mainly considers the relationship between features and classification labels (known as the relevance), which ignores the dependencies between features. The rules for using univariate measures are: Relief and its deformation ReliefF [26] and information gain (IG) [27]. Multivariate approach increases the measure of the relationship between features (known as the redundancy). The rules for using multivariate approach are: MDMC [28], mRMR [29] and so on. When using the multivariate method, the relevance and redundancy have the same effect in the entire filtering algorithm execution process, and there is no difference in the weights. In the filtering algorithm, after selecting fewer features, when selecting features again, the role of relevance between features and labels is greater than the role of redundancy between features; after selecting some features, when selecting features again, the role of redundancy between features is greater than the role of relevance between features and labels. In this way, the feature subset has the largest correlation and the smallest redundancy. Therefore, the weight of relevance will be higher first then lower than that of redundancy in filtering algorithm.

The wrapping method and the filtering method each have their advantages. The combination of wrapped and filtered algorithm is a hybrid algorithm. It uniquely brings together the efficiency of filters and the greater accuracy of wrappers. In the hybrid algorithm, a two-stage algorithm is known, in which the filter and wrapper models are considered as two separate steps [30, 31]. Therefore, in the wrapping algorithm, the situation where the filtered features enter the feature subset is not taken into consideration, and the solution to this situation is not considered. Another hybrid method is that the wrapper method is embedded in the filter algorithm [32] or the filter algorithm is embedded in the wrapper algorithm [33]. In this hybrid algorithm, although the boundaries between the wrapping algorithm and the filtering algorithm are unclear, the case where the filtered feature enters the feature subset is not considered.

Because the filtering method is to filter out some features according to a specific rule, some features that may play an important role in the classification result may be omitted. Therefore, in order to improve the classification accuracy and make the filtered important features enter the final feature

subset, we propose a hybrid feature selection algorithm—MSMCCS. Firstly, to select the features of high classification accuracy from data set, a new filter algorithm is proposed named as maximum Spearman minimum covariance (MSMC). Furthermore, spearman is used to measure relevance between features and labels while covariance is used to measure redundancy between features in MSMC algorithm. In addition, three parameters (*spc*, *ac*, and *dc*) are introduced in MSMC to increase the overall relevance and reduce the overall redundancy. As a screening result, MSMC calculates the score for each feature. Secondly, we have taken two strategies to pick out the missing features in CS. The score of each feature is multiplied by the weight value as grade value by which candidate feature subsets are selected by applying the first strategy. Based on the relationship between discovery probability and overall probability, the idea of crossover mutation is used to adjust candidate feature subsets by the other strategy. Therefore, some features of the candidate feature subset are removed, and some of the non-selected features enter the candidate feature subset. Finally, after two operations, we found the best classification accuracy. Hence, the experimental results confirm that MSMCCS algorithm has higher classification accuracy than the other 10 algorithms on eight data sets.

The rest of the paper is organized as follows: Section 2 introduces the basic concepts of mRMR algorithm, CS algorithm, and SVM classifier. Section 3 illustrates the two new algorithms (MSMC and MSMCCS) and their application. Section 4 describes the experimental results and analysis on feature selection. At the end, section 5 summarizes the final conclusion of this paper.

## 2 Basic theory

In this section, we represent two well-known algorithms. The first one is the mRMR filter, and the second one is a swarm intelligence-optimized algorithm based on CS strategy.

### 2.1 Maximum relevance minimum redundancy algorithm

The mRMR approach was proposed by Peng in 2005 [29]. Mutual information (MI) was used to measure the relevancy and redundancy of features. The mRMR strategy applies mutual information operations twice, in which the first MI between label and each feature is used to measure the relevancy, and the second MI between every two features is used to compute the redundancy. Next, the mRMR algorithm is illustrated as follows.

mRMR algorithm, first, calculates the mutual information value between each feature in the feature set ( $X$ ) and the label ( $C$ ); the feature corresponding to the largest mutual

information value is selected into  $S_m$ . Then, the features are selected one by one from the remaining feature sets ( $W$ ) according to the principle of maximum relevancy minimum redundancy, and the number of features in  $S_m$  is determined according to the specific problem.

$$X = \{x_1, x_2, x_3, \dots, x_n\} \quad (1)$$

where  $x_i$  is a column vector,  $i = 1, 2, \dots, n$ .  $n$  denotes the number of features in data set.

$$S_m = \{y_1, y_2, \dots, y_i\} \quad (2)$$

$y_i \in X, i = 1, 2, \dots, m, y_i$  is a column vector,  $m \leq n$ . When  $m = 0$ ,  $S_0$  is empty. When  $m = 1, S_1 = \{y_1\}$ . When  $m = 2, S_2 = \{y_1, y_2\}$ .

$$W = \{X - S_m\} = \{z_1, z_2, \dots, z_k\}, z_k \in X \quad (3)$$

$$k = 1, 2, \dots, n - m$$

$$RI = \frac{1}{m} \left[ \sum_{i=1}^m I(y_i, C) + I(z_j, C) \right] \quad (4)$$

$i = 1, 2, 3, \dots, m, j = 1, 2, \dots, n - m, y_i \in S_m, z_j \in W$ . Where  $I(y_i, C)$  and  $I(z_j, C)$  represent the value of mutual information between one feature and the label vector.  $C = \{C_1, C_2, \dots\}$ , where  $C_1, C_2, \dots$  denote labels.

When the selected features have the maximum relevance in  $RI$  value, it is possible to have high dependency between these features. Hence, the redundancy  $Rd$  of a group of selected features is defined as

$$Rd = \frac{1}{m^2} \sum_{i=1}^m I(y_i, z_j) \quad (5)$$

$i = 1, 2, 3, \dots, m, j = 1, 2, \dots, n - m, y_i \in S_m, z_j \in W$ . Where  $I(y_i, z_j)$  is the mutual information between the  $i$ th and  $j$ th feature,  $I(y_i, z_j)$  measures the dependency of these two features.

In the process from  $S_m$  to  $S_{(m+1)}$ ,  $z_j$  is chosen to obtain the maximum value at formula (6).

$$\text{Max}(RI, Rd) = RI - Rd \quad (6)$$

This formula (6) combines two criteria, which are maximal relevancy and minimal redundancy.

In the process from  $S_m$  to  $S_{(m+1)}$ ,  $S_m$  is invariant, so  $\sum_{y_i \in S_m} I(y_i : C)$  is a constant value and can be omitted. We can get formula (7) by simplified formula (6).

$$\text{Max}(RI, Rd) = I(z_j : C) - \frac{1}{m} \sum_{y_i \in S_m, z_j \in W} I(y_i : z_j) \quad (7)$$

Our goal is to increase the prediction accuracy and reduce the number of selected feature. However, we proposed a filter method as part of CS algorithm to improve the speed and performance of the search.

## 2.2 Cuckoo search

In the famous cuckoo search (CS) algorithm, one cuckoo seeks a new host nest via Lévy flights which was proposed by French mathematician Paul Lévy, representing a model of random walk characterized by their step size which obey a power-law distribution [14, 34]. Several scholars have proven that hunters search for preys following typically the same characteristics of Lévy flights which are accepted as optimization and in many fields of science [34, 35].

CS is a metaheuristic optimization algorithm which was proposed by Xin-She Yang and Suash Deb in 2009, which delivers the solution of multimodal functions. CS is based on the following ideal states [34, 35]:

- (1) A cuckoo lays one egg each time and selects a nest randomly to brood;
- (2) The best nest with the highest quality can be passed onto the next generation;
- (3) The number of host nests is fixed, and the probability is set at  $pa \in [0, 1]$  that the egg laid by a cuckoo can be discovered by the host bird.

The following formula (8) is used to produce a new solution  $X_i^{(t+1)}$ , for a cuckoo  $i$ , by a Lévy flight:

$$X_i^{(t+1)} = X_i^{(t)} + \alpha \oplus \text{Levy}(\lambda) \quad (i = 1, 2, 3 \dots) \quad (8)$$


$$\text{Levy}(s, \lambda) \sim s^{-\lambda} \quad \lambda \in (1, 3) \quad (9)$$

where  $s$  is size of step, and  $\alpha$  ( $\alpha > 0$ ) should be associated with the scales of the problem of interest. The symbol ( $\oplus$ ) is an entry-wise multiplication.  $X_i^{(t)}$  denotes a solution at iteration  $t$  for a cuckoo  $i$ .  $X_i^{(t+1)}$  denotes a new solution at iteration  $t + 1$  for cuckoo  $i$ . The step length is the Lévy distribution by formula (9). But  $\lambda$  is not smaller than 1, and  $\lambda$  is not bigger than 3 [14]. In the iteration, CS algorithm does not consider the impact of the previous iteration on current iteration. In the third section, updated formula increased the impact of the previous iteration on current iteration.

## 3 Maximum Spearman minimum covariance cuckoo search algorithms

In this part, a hybrid algorithm named maximum Spearman minimum covariance cuckoo search (MSMCCS) was proposed. Firstly, the filter algorithm was proposed, then, the improved cuckoo search algorithm was introduced, finally, the flow chart and pseudo code of MSMCCS were shown.

**Table 1** Sort the vector X

NO	X	Change	Descending	X'
1	175	Descending Order 	216	3
2	153		181	4
3	216		175	1
4	181		167	5
5	167		153	2

### 3.1 Filter method-MSMC

#### 3.1.1 Maximum-Spearman

Spearman's rank correlation coefficient is often called non-parametric correlation coefficient, which is used to measure the connection between two vectors. Hence, Spearman's rank correlation coefficient was used to measure relevance between features and labels in maximum-spearman (MS).

The given two vectors  $X(x_1, x_2, x_3, \dots, x_n)$  and  $Y(y_1, y_2, y_3, \dots, y_n)$  are put in order of internal values from big to small, respectively, and the order of the original position were recorded after descending order. After rearrangement of vector X or Y, the order of the original position was saved in the vector  $X'(x'_1, x'_2, x'_3, \dots, x'_n)$  or  $Y'(y'_1, y'_2, y'_3, \dots, y'_n)$ , the relevancy of vectors X and Y is represented by SP. The formula (10) and formula (11) were shown as the following.

$$SP(\vec{X}, \vec{Y}) = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n} \quad (10)$$

$$d_i = x'_i - y'_i \quad (i = 1, 2, 3, \dots, n) \quad (11)$$

For example, vector  $X(175, 153, 216, 181, 167)$  and vector  $Y(182, 165, 193, 168, 184)$  were ordered by descending, respectively, then the order of the original position is saved in the vector  $X'(3, 4, 1, 5, 2)$  and  $Y'(3, 5, 1, 4, 2)$ , see Tables 1 and 2.

And obtained  $SP = 0.9$  by formulas (10) and (11), the relevancy of vectors X and Y is very high, see Table 3 and Fig. 1.

Some data appear repeatedly in the vector, and the order of these values are the same. It is necessary to modify data after rearrangement by using the average value of those data in the rearranged order during calculation. Examples are given in Table 4.


$$\max MS(l_i, C) = \text{spc}^* |SP(l_i, C)| = \text{spc}^* \left| 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n} \right| \quad (12)$$

$$\text{spc} = e^{\left(\frac{t}{T} + \frac{1}{N}\right) * \left(1 - \frac{k}{K}\right)} \quad (13)$$

In formula (12),  $l_i$  is the  $i$ th feature in data set. C is label vector in data set. The spc represents coefficient and is used to adjust the maximum correlation value. In formula (13), T is the total number of iterations, t is a current number of iterations, i represent the current feature subset number, N is the total number of feature subset, k is the amount of selected feature, and K is the number of all features.

In the above study, the  $m$  features in the descending sort of SP (formula 10) can be gained. Nevertheless, many studies confirmed that “the  $m$  best eigenvalues are not the best  $m$  eigenvalues” [29]. Some researchers have proposed reducing the redundancy of eigenvalues. So, a new method to carry out Min-Redundancy based on covariance is proposed.

**Table 2** Sort the vector Y

NO	Y	Change	Descending	Y'
1	182	Descending Order 	194	3
2	165		184	5
3	193		182	1
4	168		168	4
5	184		165	2

**Table 3** The vectors  $X'$  and  $Y'$

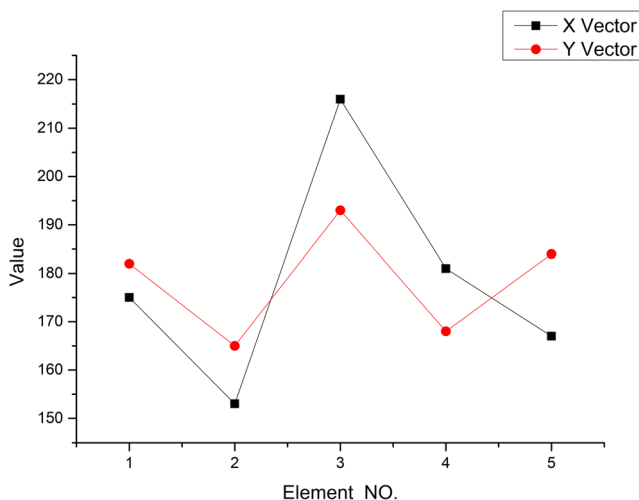
No.	$X'$	$Y'$	$d_i$	$d_i^2$
1	3	3	0	0
2	4	5	1	1
3	1	1	0	0
4	5	4	1	1
5	2	2	0	0
Sum ( $d_i^2$ ) $i = 1, 2, 3, 4, 5$				2

### 3.1.2 Minimum-covariance

The spearman was used to measure the maximum correlation, and the covariance was used to measure the minimum redundancy [36–38]. Based on the covariance calculation method of two vectors, the method of calculating the average covariance and distribution covariance are proposed. Finally, the measurement method of minimum covariance (MC) is proposed. There are two sets of  $L$  and  $S$ ,  $L = \{l_1, l_2, l_3, \dots, l_m\}$  and  $S = \{s_1, s_2, s_3, \dots, s_d\}$ . Set  $L$  represents an alternative subset of features, and there are  $m$  features in  $L$ . Set  $S$  represents a set of selected features, and  $S$  has  $n$  features. To calculate the covariance of the two vectors, the formula (14) is defined:

$$\text{cov}(x, y) = \left| \frac{\sum_{i=1}^{\text{num}} (x_i - \bar{x}) * (y_i - \bar{y})}{\text{num} - 1} \right| \quad (14)$$

$\text{Cov}(x, y)$  is a covariance value of vector  $x$  and  $y$ ; every vector has  $\text{num}$  observations;  $x_i$  is the  $i$ th value in vector  $x$ ;  $y_i$  is the  $j$ th value in vector  $y$ .  $\bar{x}$  and  $\bar{y}$  are respectively the mean value of vector  $x$  and  $y$ .  $\text{Cov}(l_j, s_i)$  is covariance value of the  $j$ th feature in the alternative feature subset and the  $i$ th feature in the selected feature subset. Where  $l_j \in L$ ,  $s_i \in S$ ,  $j = 1, 2, 3, \dots, n$ ,  $i = 1, 2, 3, \dots, m$ . According to the calculation method of covariance value between two features, the average covariance calculation method between one feature in the alternative feature



**Fig. 1** The trend chart of vectors  $X$  and  $Y$

**Table 4** The rank of the same numerical value in the vector

No.	Variate $x_i$	Position of descending sort	Rank $x'_i$
1	60.58	5	5
2	57.08	1	1
3	57.08	2	(2 + 3)/2 = 2.5
4	51.58	3	(2 + 3)/2 = 2.5
5	80.25	4	4

The value of Max Spearman (MS) corresponding feature in data set is found according to formula (12)

set ( $L$ ) and the selected feature set ( $S$ ) is presented. The formula (15) as following:

$$\text{averagecov}(l_j, S) = \frac{\sum_{i=1}^d \text{cov}(l_j, s_i)}{d} \quad (15)$$

$l_j$  is the  $j$ th feature in the alternative feature subset,  $S$  is the selected feature subset,  $s_i$  is the  $i$ th feature in  $S$ , which has  $d$  features.  $\text{Cov}(l_j, s_i)$  is calculated from formula (14).

When a feature is selected from the alternative feature set ( $L$ ), assume that there are three features in the selected feature set ( $S$ ), which are recorded as  $s_1, s_2$ , and  $s_3$ . There are some features in  $L$  set, two of which are recorded as  $l_1$  and  $l_2$ . Average covariance values between all features in  $S$  and  $l_1$  as well as  $l_2$  are illustrated in the following table:

From the Table 5, it is possible to analyze this in view of average covariance,  $\text{averagecov}(l_1, S) = 0.18$ ,  $\text{averagecov}(l_2, S) = 0.2$ ,  $0.18 < 0.2$ . According to the thought of minimum covariance, we select feature  $l_1$ . The covariance between  $s_1$  and  $s_2$  is very low; covariance between  $l_1$  and  $s_3$  is high. The total covariance will be increased if we select  $l_1$ . The mean of covariance value in feature  $l_2$  and set  $S$  is higher than that of  $l_1$  and set  $S$ , but covariance value in feature  $l_2$  and set  $S$  is relatively even, which means that covariance between  $l_2$  and  $s_1, s_2$ , and  $s_3$  is not too high. So, we select feature  $l_2$ . In this paper, we propose a new concept of distributive covariance, see formula (16).

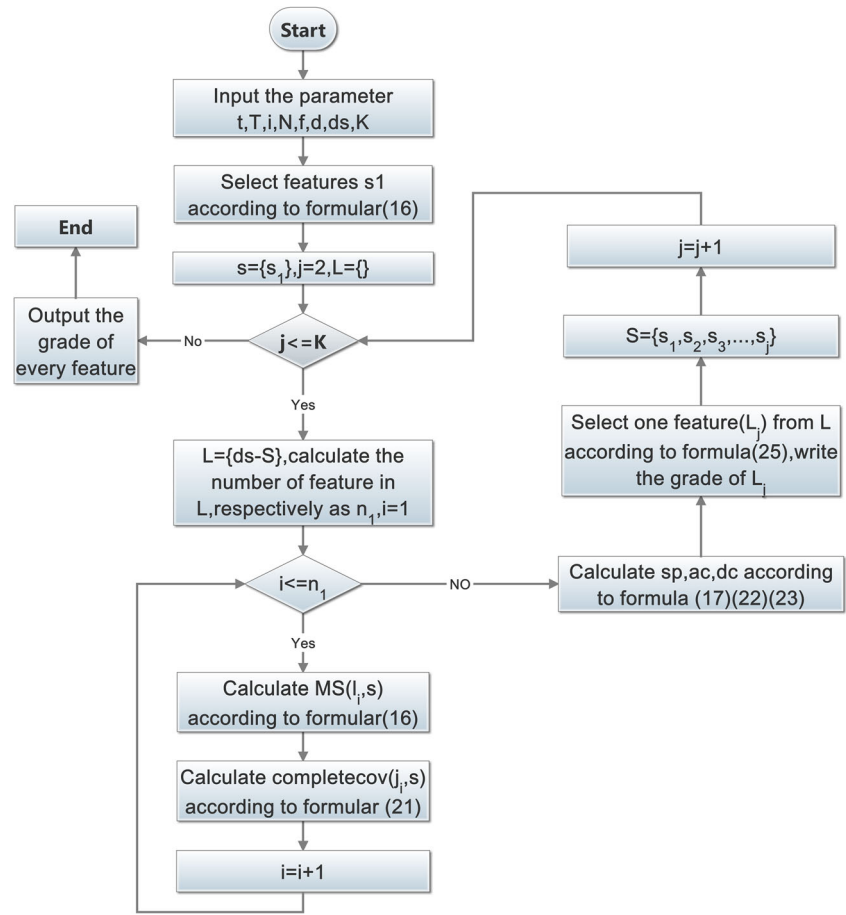
$$\begin{aligned} \text{distributecov}(l_j, S) \\ = \frac{\sum_{i=1}^d \left( \left| \text{cov}(l_j, s_i) - \text{averagecov}(l_j, S) \right| \right)}{d} \end{aligned} \quad (16)$$

**Table 5** the average covariance values of  $L_1$  and  $L_2$

	$l_1$	$l_2$
$s_1$	$\text{Cov}(l_1, s_1) = 0$	$\text{Cov}(l_2, s_1) = 0.1$
$s_2$	$\text{Cov}(l_1, s_2) = 0$	$\text{Cov}(l_2, s_2) = 0.2$
$s_3$	$\text{Cov}(l_1, s_3) = 0.54$	$\text{Cov}(l_2, s_3) = 0.3$
Averagecov( $l_j, S$ )	Averagecov( $l_1, S$ ) = 0.18	Averagecov( $l_2, S$ ) = 0.2



Fig. 2 Flowchart of MSMC



In the hypothesis, S set has features but is empty at beginning. Therefore, in the choice of features, we must not only consider  $\text{averagecov}(l_j, s)$  and  $\text{distributeconv}(l_j, s)$  but also consider their weights, see formula (17).

$$\text{completecov}(l_j, S) = \text{ac} * \text{averagecov}(l_j, S) + \text{dc} * \text{distributeconv}(l_j, S) \quad j = 1, 2, 3, \dots, m \quad (17)$$

$$\text{ac} = \exp\left(\frac{K-k+1}{K}\right) * \exp\left(\cos\left(\left(\frac{1}{T} + \frac{1}{N}\right) * \pi\right)\right) \quad (18)$$

$$\text{dc} = \tan\left(\sin\left(\frac{k * \pi}{2 * K}\right)\right) * \exp\left(\frac{\text{lenbest} * \text{lenF} + \frac{1}{N}}{K}\right) * \frac{k}{K} \quad (19)$$

In formulas (18) and (19), the six variables ( $T, t, i, N, k, K$ ) are the same as the variables in (13).  $\text{lenbest}$  is the length of feature subset which is selected and is the global optimal feature subset;  $\text{lenF}$  is the number of feature in data set.

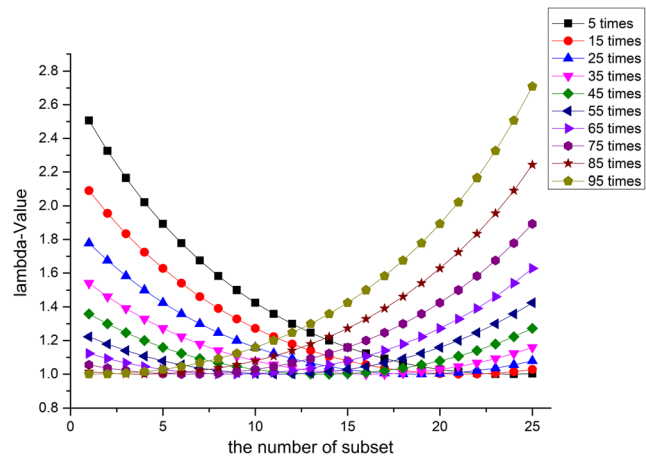
From the point of redundancy, the feature of minimum covariance value is selected. In the L set, the feature of the minimum value of  $\text{completecov}(l_j, S)$  is selected into S.

$$\text{MC}(l_j, S) = \min(\text{completecov}(l_j, S)) \quad (20)$$

S set is a set of selected features.  $l_j \in L, j = 1, 2, 3, \dots, m$ .

### 3.1.3 Maximum Spearman and minimum covariance

In the above introduction of maximum Spearman and minimum covariance, in order to calculate the score of each feature


 Fig. 3 The changes of  $\lambda$

in dataset, maximum value and minimum value need to be arranged into a formula. There are two methods of taking the maximum value, one of which is  $MS(l_j, C)$  minus the  $MC(l_j, S)$ , and the other is  $MS(l_j, C)$  divided by the  $MC(l_j, S)$ . To reduce the complexity of computation, we use the first one.

$$feagrade(s_i) = \begin{cases} \max(spc * |sp(l_j, C)|) & i = 1 \\ \max(spc * |SP(l_j, C)| - ac * averagecov(l_j, S) - dc * distributecov(l_j, S)) & i > 1 \end{cases} \quad (21)$$

$s_i \in S, l_j \in L, i = 1, 2, 3 \dots n$ ,  $L$  is the alternative feature subset,  $C$  is column labels and  $S$  is the selected feature subset. The three parameters  $spc$ ,  $ac$ , and  $dc$  are coefficients for relevance and redundancy.  $Feagrade(s_i)$  expressions are the score of  $i$ th feature in dataset.

In the MSMC algorithm, there are three parameters  $spc$ ,  $ac$ , and  $dc$ , which are used to adjust the  $MS$  value, the  $averagecov$  value and the  $distributecov$  value. When the features in  $S$  are relatively small, the correlation is measured. When the numbers of feature are larger, the redundancy is measured. Therefore, the value of the coefficients  $spc$  and  $ac$  will gradually decrease to 1 as the number of selected features increases, and the value of the coefficient  $dc$  will gradually increase from 0. The trend of three coefficients can provide guarantee for the maximum correlation and minimum redundancy of feature subset.

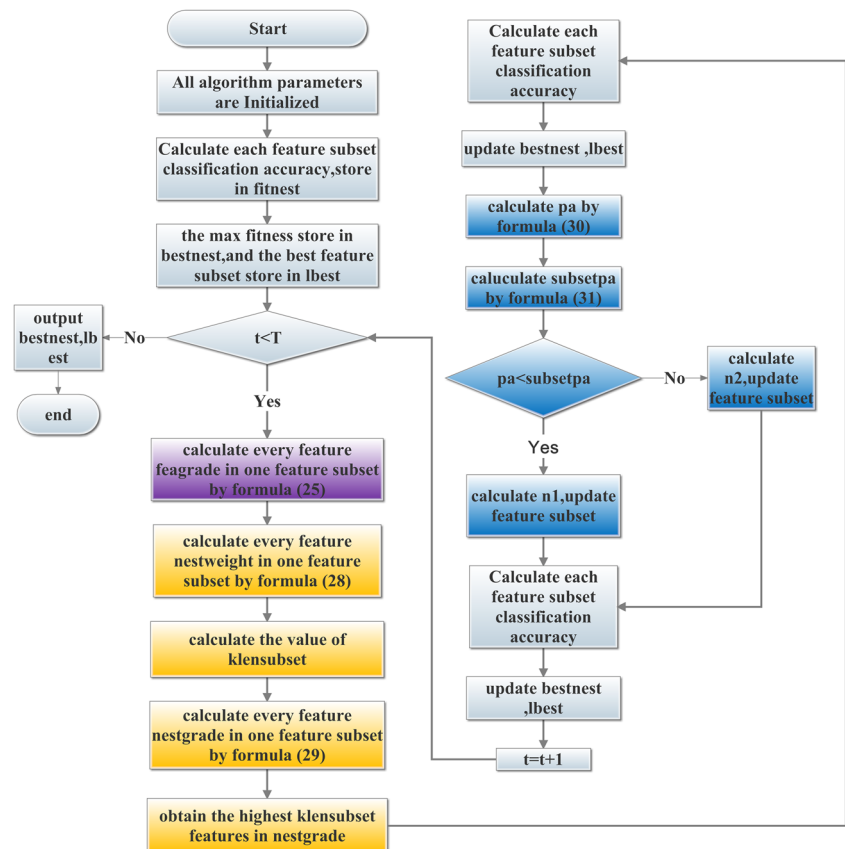
When the feature subset is empty,  $averagecov$  and  $distributecov$  cannot be calculated. Formula (21) calculates the feature score. When the feature subset is not empty, according to formula (21), the scores of other features were calculated one by one.

In Fig. 2, some parameters for MSMC algorithm are provided, including the number of iterations and feature subset information. In this algorithm, there are two loops. The score between one feature in the alternative feature set ( $L$ ) and the selected feature set ( $S$ ) is calculated in inner loop. The feature is selected one by one according to the score in outer loop. The score of each feature is computed in MSMC algorithm.

### 3.2 Improved cuckoo search algorithms

The MSMCCS algorithm combines the MSMC algorithm with the ICS algorithm. The cuckoo search (CS) algorithm is improved through three stages. This section describes the improvement and their implication.

Fig. 4 MSMCCS flow chart



**Table 6** Description of data sets used in the experiments

No.	Data sets	No. of classes	No. of instances	No. of features	Abbreviation
1	German	2	1000	24	Ger
2	Ionosphere	2	351	34	Ion
3	Segment	7	2310	19	Seg
4	Sonar	2	208	60	Son
5	Vehicle Silhouettes	4	846	18	Veh
6	Vowel	11	990	13	Vow
7	Wine	3	178	13	Win
8	Zoo	7	101	17	Zoo

### 3.2.1 Enhance the speed of convergence

In CS,  $\lambda$  is fixed [14, 15]. In order to increase the optimization space, parameter  $\lambda$  must be updated according to formula (22).

$$\lambda = \frac{3 - \sin\left(\frac{\pi^* t}{2^* T} + \frac{\pi^* i}{2^* n}\right)}{1 + \sin\left(\frac{\pi^* i}{2^* n} + \frac{\pi^* t}{2^* T}\right)} \quad (22)$$

where  $i$  is the number of current feature subset,  $n$  is the number of all feature subset,  $t$  is times of current iteration, and  $T$  is the total times of iteration. According to formula (9) ( $1 < \lambda \leq 3$ ), when the value of  $t$  is very small, the changing trend of  $\lambda$  decreases gradually. With the increasing of iteration, the change of  $\lambda$  is augmented gradually. At the beginning of parameter  $\lambda$ 's change, feature subset is close to the local optimum and local optimum is found. Then, feature subset continuously finds the global optimal which is better than local optimal. The variety of parameter  $\lambda$  helps to obtain a global optimal solution rather than a local optimum one. Figure 3 shows the changes of  $\lambda$ .

CS was proposed to solve optimization problem based on biological behavior and physical systems in nature. Generally, CS can find the global optimum with long period of time and slowness in convergence [34, 35]. So, in order to enhance the speed of convergence, the updated formula for every feature subset has been modified, increasing the impact of last iteration's

optimal value over current feature subset. The updated formula is as follows:

$$s = s + \lambda^* \text{step}(\text{bestnest-s}) + (\text{Ibest-s}) \quad (23)$$

where  $s$  is one feature subset,  $\text{step}$  is calculated by formula (9),  $\text{bestnest}$  is the globally best feature subset at present, and  $\text{Ibest}$  is the best feature subset that is generated in last iteration;  $\lambda$  is calculated by formula (22).

The calculation methods of parameter  $\lambda$  and location update formula are modified to provide guarantees for obtaining the optimal feature subset.

### 3.2.2 Selected candidate feature subset

After the above modifications, the value as weight for each feature is obtained. But some weights are negative and not normalize. In order to normalize the weight value, formula (24) is introduced.

$$\text{nestweight}(j) = \frac{1}{1 + \exp^{s(j)}} \quad (j = 1, 2, \dots, n) \quad (24)$$

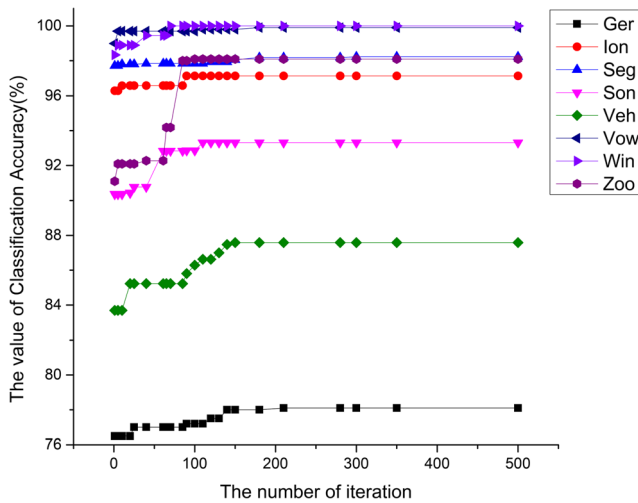
where  $s(j)$  represents the original weight value of the  $j$ th feature in data set,  $\text{nestweight}(j)$  represents the  $j$ th feature weight value after the standard. There are  $n$  features in data set.  $\text{nestweight}(j) \in (0, 1)$ ,  $j = 1, 2, \dots, n$ .

**Table 7** Average classification accuracy via eight algorithms

No.	DS	ALO	GA	BBA	SA	PSO	CS	BFO	mRMR	mGA	mPSO	MSMCCS
1	Ger	75.76	74.94	74.98	74.72	72.40	76.70	76.76	74.59	75.96	77.10	77.88
2	Ion	92.67	95.41	95.48	95.85	86.35	87.87	96.03	92.81	95.96	93.92	97.14
3	Seg	97.26	96.62	96.21	97.81	97.85	93.29	94.55	96.84	97.80	97.06	98.29
4	Son	71.98	71.11	69.00	81.60	76.68	70.10	78.23	87.07	90.39	90.90	93.30
5	Veh	83.69	84.64	82.98	83.80	80.60	76.24	75.77	84.38	84.16	72.70	87.52
6	Vow	74.04	71.47	70.10	78.46	71.92	58.99	78.69	82.51	98.04	95.89	99.80
7	Win	98.37	98.36	98.04	97.06	97.37	97.81	99.44	94.26	98.94	98.33	100.00
8	Zoo	94.04	94.13	94.30	96.39	89.30	93.37	93.70	95.25	96.74	97.09	98.02

DS data set, mGA mRMR + GA, mPSO mRMR + PSO





**Fig. 5** The classification accuracy via 500 iterations on eight data sets

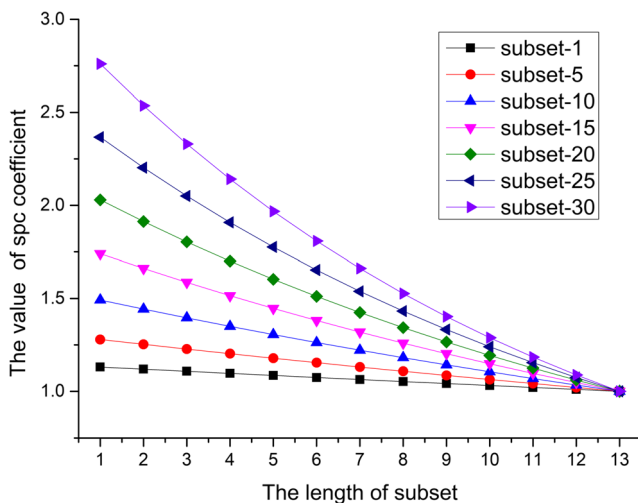
In order to take advantage of the wrapper method and the filter method, the two algorithms are combined. The weight obtained by the wrapper algorithm is multiplied by the score obtained by the filter method, and then, a grade value is obtained for each feature in the data set, see (25). The candidate feature subsets are composed of features corresponding to the first klensubset grade values.

$$\text{nestgrade}(j) = \text{nestweight}(j) * \text{msmcgrade}(j) \quad (25)$$

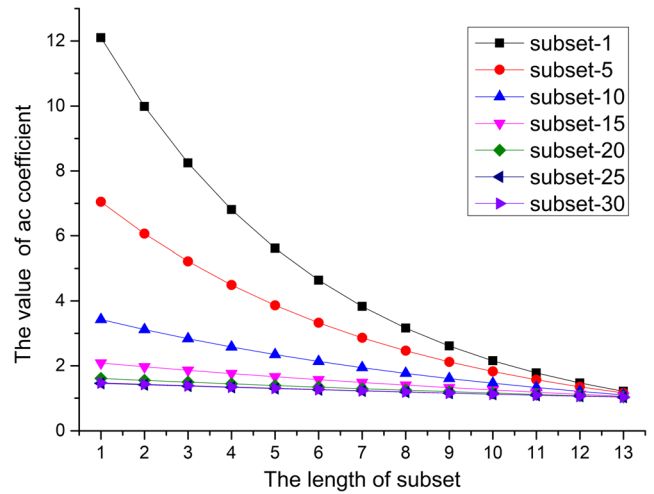
where  $j$  is  $j$ th feature in data set,  $j = 1, 2, \dots, n$ .

### 3.2.3 Adjusted candidate feature subset

After the combination of CS and MSMC, some features are selected as candidate feature subset, and some features are eliminated in data set. In order to improve the accuracy of classification, the features in the candidate feature subset are adjusted by the relationship between the probability of discovery and the probability of feature subset.



**Fig. 6** The value of spc when feature subsets is 1, 5, 10, 15, 20, 25, and 30 on Win data set

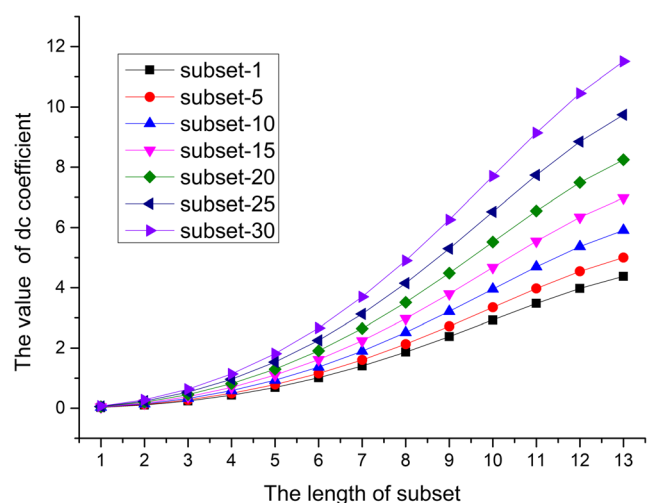


**Fig. 7** The value of ac when feature subsets is 1, 5, 10, 15, 20, 25, and 30 on Win data set

In standard CS,  $pa$  signifies probability of being discovered, which is a fixed value [14, 35]. With the increasing of iteration, probability of being discovered is increased in reality. So,  $pa$  cannot be a fixed value. The value of  $pa$  is modified and increased with the number of iterations by formula (26).

$$pa = \frac{e\left(\frac{8^*t}{T}\right) - 1}{e\left(\frac{4^*t}{T}\right) + 1} \quad (26)$$

In standard CS, one feature is discovered when the probability of the feature is bigger than  $pa$  (discovery probability). Then the value of feature is modified [14, 35]. In this paper, the probability value of any feature cannot represent the probability value of whole feature subset. The average probability value of all features in feature subset is seen as the probability value of the whole feature subset. The average probability value of all



**Fig. 8** The value of dc when feature subsets is 1, 5, 10, 15, 20, 25, and 30 on Win data set

features is written to subsetpa which is calculated by formula (27). According to the idea of cross-mutation, the candidate features subset produced in 3.2.2 have been taken variation operation. If  $subset \geq pa$ ,  $n1$  is calculated by formula (28) ( $1 \leq n1 \leq lensubset$ ). Then, change the  $n1$  feature values to zero where the value is one and the feature probability is maximal. Otherwise,  $n2$  is calculated by formula (29) ( $1 \leq n2 \leq d - lensubset$ ). Then, change the  $n2$  feature values to one where the value is zero and the feature probability is minimal.

$$subsetpa = \frac{\sum_{i=1}^n featurepa(i)}{n} \quad (27)$$

$$n1 = \left\lceil \frac{pa * klensubset}{subsetpa} \right\rceil \quad (28)$$

$$n2 = \left\lceil \frac{subsetpa * (n - klensubset)}{pa} \right\rceil \quad (29)$$

In formula (27), where  $featurepa(i)$  is probability of the  $i$ th feature in data set,  $featurepa(i)$  is produced by random,  $n$  is the number of feature in data set. In formula (28) and (29),  $pa$  is the discovery probability,  $klensubset$  is the number of selected feature,  $subsetpa$  is the probability of whole feature subset and it is calculated by formula (27).

### 3.3 MSMCCS algorithm

In Fig. 4, the part with gray background is the original CS, and the parts in other colors represent the improved CS. The pink rectangle is charted as the result of MSMC filtering algorithm. Four red rectangles are the formation process of the candidate feature subsets. Five blue boxes represent the process of adjusting the candidate feature subsets. After initialization, the optimal classification accuracy is calculated, and then, the loop is started. The loop consists of two parts which are the formed candidate feature subsets and the adjusted candidate feature subsets. In the first part, each feature is scored by MSMC algorithm (Fig. 2), then candidate feature sets are selected and their classification accuracy is calculated, leading to the update of optimal classification accuracy. In the second part, the probability of discovery is calculated, and features of candidate feature subsets are adjusted. Then, the classification accuracy of candidate feature sets is calculated, leading to the update of optimal classification accuracy. After exiting the loop, the optimal classification accuracy is output.

Algorithm 1 MSMCCS pseudo-code

```

1: Input: Data set, number of iterations T, number of cuckoo n,
2: Output: bestnest, lbnest
3: Initialize every nests(i)  $i=1,2,...,n$ 
4: Calculate every fitness(i) using fitness function  $i=1,2,...,n$ 
5:  $fitness(k) = \text{maximum} \{fitness(n)\}$ , the corresponding nest is k,  $bestnest = fitness(k)$ ,  $lbnest = nest(k)$ ,  $1 \leq k \leq n$ 
6: While  $t < T$  do
7:   Calculate  $nestweight(i)$  of every cuckoo nest with formula (28)  $i=1,2,3,...,n$ 
8:   Calculate  $feagrade(i)$  of every cuckoo nest with formula (25)  $i=1,2,3,...,n$ 
9:   Calculate  $nestgrade(i)$  of every cuckoo nest with formula (29)  $i=1,2,3,...,n$ 
10:  Selected features from  $nestgrade(i)$  to constitute candidate feature subset  $i=1,2,3,...,n$ 
11:  Calculate each feature subset classification accuracy
12:  Update  $bestnest, lbnest$ 
13:  Calculate  $pa$  by formula (30)
14:  Calculate  $subsetpa$  by formula (31)
15:  If  $pa < subsetpa$  then
16:    calculate  $n1$ , update feature subset
17:  Else
18:    calculate  $n2$ , update feature subset
19:  Endif
20:  Calculate each feature subset classification accuracy
21:  Update  $bestnest, lbnest$ 
22: End While
23: Output  $bestnest, lbnest$ 

```

In this part, first of all, the new filter method named MSMC is proposed based on Spearman and covariance. Three parameters  $spc$ ,  $ac$ , and  $dc$  are introduced in MSMC to increase the

overall relevance and reduce the overall redundancy. Then, in CS algorithm, the effect of convergence speed is raised via modifying the calculation method of parameter  $\lambda$  and location

updating formula. Then, through combination of MSMC and ICS algorithm, the candidate feature subsets are found, and the optimal classification accuracy is calculated. Finally, the features in the candidate feature subset are adjusted by the relationship between the probability of discovery and the probability of feature subset. As a result, MSMCCS algorithm is proposed to find the best classification accuracy.

## 4 Experiment results and analysis

### 4.1 Data sets introduction

In order to achieve the superiority of the presented MSMCCS algorithm, a series of tests are conducted by eight data sets which are selected from the UCI machine learning databases [39]. The excellence of the proposed algorithm is evaluated based on the following data sets: German, Ionosphere, Segment, Sonar, Vehicle Silhouettes, Vowel, Wine, and Zoo. Table 6 describes the detailed information of these datasets in which three data sets are two-classification and one data set is three-classification. Vehicle Silhouettes is four-classification. Segment and Zoo are seven-classification. Vowel is eleven-classification. The number of instances in data sets is between 101 and 2310. The number of features in data sets is between 13 and 60.

### 4.2 Algorithm parameter setting

In the following experiment, wrapper-type feature selection algorithm, filter-type feature selection algorithm, and hybrid-type feature selection algorithm were compared with proposed MSMCCS algorithm. The feature selection algorithms in wrapper depend on classifier and the parameter setting. In MSMCCS algorithm, iteration is 100 times, the number of subset is 30, every data set was tested five times, the classification accuracy is the average of five times. Wrapper-type

feature selection algorithms and filter-type feature selection algorithms are studied by some researchers. However, it is every difficult to find other algorithms that is the exactly same setting as MSMCCS algorithm. In order to make fair comparisons, we reproduce the three types of algorithms (wrapper, filter, hybrid). In these algorithms, the times of iteration, the number of individual, the calculative methods of average classification accuracy are the same as the MSMCCS.

In wrapper method, the initial parameters and actual methods of PSO, GA, BBA, SA, and CS come from references [11, 21, 32, 35, 40]. ALO algorithm is altered in binary structure so that it can solve feature selection problem. For ALO algorithm, the range of position for ant and ant lion is zero or one. When the number produced by random is bigger than 0.5 or as the same as 0.5, the corresponding position value is 1 which means the corresponding feature is selected. Otherwise, the corresponding position value is 0 which means the corresponding feature is not selected.

The detailed parameter values of each wrapper-type algorithm are introduced as follows. For PSO, the number of particles is 30, maximum value of weight ( $w_{\max}$ ) is 09, minimum value of weight ( $w_{\min}$ ) is 0.4, coefficient ( $c1$  and  $c2$ ) is 2, and the maximum number of iterations ( $i$ ) is 100. For GA, the number of chromosomes ( $N$ ) is 30, crossover probability ( $P_c$ ) is 0.7, mutation probability ( $P_m$ ) is 0.02, and maximum number of iterations ( $i$ ) is 100. For SA, initial temperature ( $T_0$ ) is 0.8, stop temperature ( $T_f$ ) is  $0.8^{30}$ , cooling factor ( $f$ ) is 0.8, and maximum number of iterations is 3000; it means that the number of particles and the number of iteration are multiplied. For BBA, number of bats ( $N$ ) is 30, loudness ( $L$ ) is 1.5, pulse rate is 0.5, maximum value of frequency ( $Q_{\max}$ ) is 1, minimum value of frequency ( $Q_{\min}$ ) is 0, maximum number of iterations ( $i$ ) is 100. For ALO, the number of ants ( $N$ ) is 30, minimum of all variables ( $lb$ ) is 0, maximum of all variables ( $ub$ ) is 1, and maximum number of iterations ( $i$ ) is 100. For CS, the number

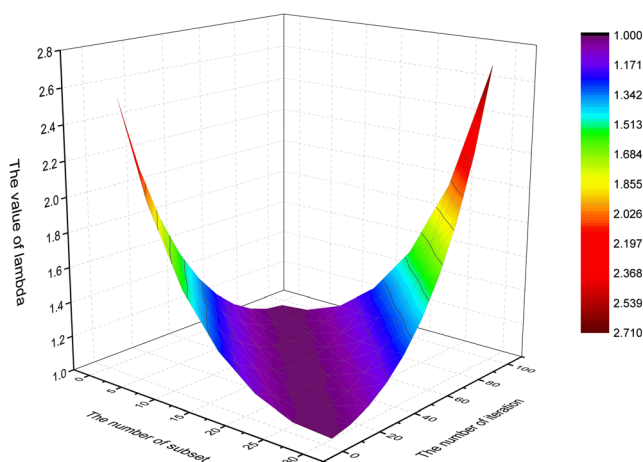


Fig. 9 The trend chart of lambda on Win data set

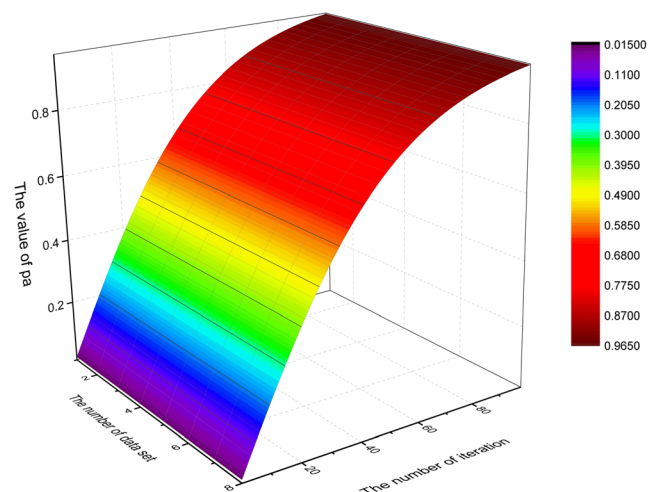


Fig. 10 the trend chart of discovery probability on every data set

**Table 8** Sensitivity via eleven algorithms

No.	DS	ALO	GA	BBA	SA	PSO	CS	BFO	mRMR	<i>mGA</i>	<i>mPSO</i>	MSMCCS
1	Ger	0.89	0.90	0.91	0.91	0.90	0.90	0.90	0.90	0.91	0.90	0.91
2	Ion	0.95	0.98	0.98	0.98	0.98	0.98	0.97	0.98	0.98	0.97	0.99
3	Son	0.75	0.85	0.80	0.81	0.76	0.66	0.75	0.88	0.86	0.82	0.90

DS data set, *mGA* mRMR + GA, *mPSO* mRMR + PSO

of nests ( $N$ ) is 30, discovery rate of alien eggs/solution ( $pa$ ) is 0.25, levy exponent and coefficient ( $\beta$ ) is 1.5, step size ( $\alpha$ ) is 0.01, and maximum number of iterations ( $i$ ) is 100.

In filter method, the maximum relevance minimum redundancy (mRMR) algorithm is chosen. The mid-method was selected in mRMR algorithm to order the features. Because the data set selected is not microarray, and in order to find the best classification accuracy, the numbers of selected feature ( $K$ ) are variant.  $K = 1, 2, 3, \dots, N$ ,  $N$  is the number of features in data set. The corresponding feature subset whose classification accuracy is the highest as the candidate feature subset. The best classification accuracy in candidate feature subset is searched according to backward search strategy [29]. After the abovementioned operation, the optimal classification accuracy for every data set is gained.

In hybrid algorithm, mRMR algorithm is used as filter method; PSO algorithm and GA algorithm are used as wrapper method. At the same time, the relevant parameter setting conditions do not change.

In this research, the fitness function has been substituted by SVM classifier in three-type algorithms (wrapper, filter, hybrid). The feature subset with the highest classification accuracy based on SVM classifier is the solution to the problem. In other words, the best feature subset is found. The radial base function (RBF) is used as the kernel function of the SVM model. Penalty parameter  $C$  and RBF parameter  $\gamma$  are selected by the grid search method.

In testing the classification accuracy of the eight data sets which are mentioned in Table 3, tenfold cross validation technique is accepted. This technique is made up of 10 cycles. In every cycle, the data set is split tenfold. Furthermore, ninefolds is used for training, and the last onefold is used for testing in the tenfolds. There is one classification accuracy in each fold. At last, the average classification accuracy of 10 cycles is the result of fitness function.

The average classification accuracy of the 11 algorithms on each data set is shown in Table 7. Sensitivity and specificity are statistical measures of the performance of a binary classification test [41]. Sensitivity measures the proportion of positives that are correctly identified. Specificity measures the proportion of negatives that are correctly identified.

### 4.3 Results and discussion

From Table 7, it shows that MSMCCS achieves the highest average classification accuracy in eight data sets. The classification accuracy of MSMCCS algorithm reaches 100% in Win data set. MSMCCS algorithm is at least bigger 1% than other algorithms in Son and Veh data set.

In MSMCCS algorithm, first of all, the score of every feature in data set is calculated by maximum Spearman and minimum covariance algorithm. After combining the score with the weight of CS, the feature in data set is ordered by descending. Finally, some features that rank in front will be selected as candidate feature subsets. After this operation, the purpose of the elimination redundancy features is achieved. The features in candidate feature subset are adjusted according probabilistic relationship, so the higher classification accuracy of feature subset is found.

From Fig. 5, it shows that horizontal axis represents iterations from 0 to 500, and vertical axis is classification accuracy for each data set. The gap of classification accuracy is very small when iteration is 100 and 500. The gap is 0 in Ion, Se, Son, Vow, Win, and Zoo data set, and the gap is 0.3 in German data set, while the gap is 0.4 or so in Veh data set. It tells us that this algorithm is convergent before 100 iterations. This algorithm combines wrapper-type with filter-type feature selection. In filter-type feature selection, value of each feature represents their significance in feature subsets by maximum Spearman and minimum covariance. In wrapper-type feature selection, the features without redundancy are positioned in front in

**Table 9** Specificity via eleven algorithms

No.	DS	ALO	GA	BBA	SA	PSO	CS	BFO	mRMR	<i>mGA</i>	<i>mPSO</i>	MSMCCS
1	Ger	0.47	0.38	0.44	0.40	0.31	0.40	0.46	0.45	0.43	0.47	0.47
2	Ion	0.88	0.90	0.91	0.90	0.64	0.70	0.79	0.90	0.91	0.91	0.92
3	Son	0.72	0.86	0.88	0.93	0.76	0.70	0.77	0.89	0.90	0.85	0.94

DS data set, *mGA* mRMR + GA, *mPSO* mRMR + PSO

**Table 10** The comparison based on Wilcoxon signed-rank test on Veh data set

A1	C*	C*	C*	C*	C*	C*	C*	C*	C*	C*
A2	ALO	GA	BBA	SA	PSO	CS	BFO	mRMR	mGA	mPSO
Z	-2.807	-2.090	-2.670	-2.807	-2.805	-2.803	-2.803	-2.397	-2.666	-2.803
P	0.005	0.037	0.005	0.008	0.005	0.005	0.005	0.017	0.008	0.005

A1: Algorithm 1, A2: Algorithm 2, C\*: MSMCCS, *mGA* mRMR + GA, *mPSO* mRMR + PSO

descending order of weight and grade so as to realize minimum and maximum of classification accuracy in 100 iterations.

Figures 6, 7, and 8 show the change of three parameters *spc*, *ac*, and *dc* with increase in the number of feature subset. We choose 7 feature subsets (1, 5, 10, 15, 20, 25, and 30) from 30 feature subsets based on Win data set when the number of iteration is 10. From the above figures, the initial value of *spc* is between 1 and 3; the last value of *spc* is 1. The initial value of *ac* declines from 9 to 1.3; the last value of *ac* is 1. The initial value of *dc* is around 0.5; the last value of *dc* is raised from 4.3 to 11.5.

*Spc* is the weight of maximum Spearman. In the selection process, Spearman represents relevance between label and feature. It is an important role when the number of features is less in feature subset. The effect of Spearman does not change much in the increase of the feature number. *Ac* is the weight to measure average covariance between features. When the selected feature subset (*S*) is little, the redundancy between features mainly relies on the average covariance, so the initial value of *ac* is bigger. *Dc* is the weight to measure distribution covariance between features. When the selected feature subset (*S*) is empty, distribution covariance is not used for feature selection. With the increase of selected features, the average covariance and distribution covariance both measure redundancy between features. In the process, the role of average covariance is gradually reduced. Meanwhile, the role of distribution covariance is increased gradually. Therefore, in the selection process of feature subset, *ac* is the descent curve and *dc* is the rising curve. With the increase in the number of feature subsets, the average covariance and distribution covariance value have accumulated more and more. The distribution covariance is more effective than the average covariance in discriminating redundancy between features. Therefore, the initial value of *ac* is decreased gradually, and the final value of *dc* is increased gradually.

Figure 9 shows that X axis represents the number of feature subsets in the wrapped method, from 1 to 30; Y axis represents the number of iterations, from 1 to 100; Z axis indicates the change of lambda, from 1.0 to 3.0. At the start of iteration, the small-numbered feature subset has relatively high  $\lambda$  value which is 2.5. With the increasing number of feature subsets,  $\lambda$  gradually decreased to 1. When the number of iterations is about 5, with the increasing number of feature subsets,  $\lambda$  decreased firstly and then increased, from 1.2 down to 1 and then from 1 up to about 1.4. When the number of iterations is

bigger than 90, with the increasing number of feature subsets,  $\lambda$  increases from 1 to about 2.70 and close to the maximum value of 3.0. On the whole,  $\lambda$  changes from big to small, and then becomes bigger. Starting with larger space for optimal classification accuracy, MSMCCS search gradually and more carefully. In order to prevent falling into the local optimal value, MSMCCS search with bigger step sizes. The change of  $\lambda$  is good for searching the optimal classification accuracy, and will not fall into local optimum.

Figure 10 shows that with the increase of the number of iterations, the discovery probability becomes 1 from 0 little by little. Although the discovery probability is changed, the trends are the same for each data set. As can be seen from the formula (30), the change of discovery probability and the number of iteration are related, and the change has nothing to do with data set. Therefore, the trend of the change for discovery probability in figure is reasonable.

Sensitivity and specificity are statistical measures, which are suitable for binary classification test. However, there are five data sets (Seg, Veh, Vow, Win, and Zoo) are not binary classification data sets; therefore, the sensitivity and specificity of each algorithm are tested only on three UCI binary data sets (Ger, Ion, and Son). From Table 8, it shows that MSMCCS algorithm gains the

**Table 11** Descriptive statistics of eleven algorithms based on Veh data set

Algorithm	Descriptive statistics				
	N	Mean	Std. deviation	Minimum	Maximum
ALO	10	83.69	3.06	76.57	88.10
GA	10	84.64	3.90	78.57	91.67
BBA	10	82.98	3.25	75.38	87.06
SA	10	83.80	3.61	76.47	89.41
PSO	10	80.60	3.73	75.00	81.18
CS	10	76.24	3.49	71.43	83.33
BFO	10	75.77	4.43	71.43	94.12
mRMR	10	84.38	4.24	77.38	90.59
mGA	10	84.16	3.54	79.76	89.41
mPSO	10	72.70	6.17	63.10	82.14
MSMCCS	10	86.41	2.95	80.00	89.41

*mGA* mRMR + GA, *mPSO* mRMR + PSO



**Table 12** The average classification accuracy via three classifiers

No.	Data set	MSMCCS SVM	MSMCCS KNN	MSMCCS NB	MSMCCS RF	MSMCCS Adaboost
1	Ger	77.88	75.36	74.78	74.36	75.16
2	Ion	96.87	95.82	94.54	95.33	94.53
3	Seg	98.29	95.7	96.73	96.86	82.99
4	Son	93.10	90.55	91.60	87.75	87.53
5	Veh	86.41	85.27	84.30	80.71	77.06
6	Vow	99.80	97.52	97.71	94.85	91.12
7	Win	100.00	98.79	98.37	97.98	96.02
8	Zoo	98.02	95.39	95.78	94.76	87.06

highest sensitivity in Ion and Son data sets. Moreover, MSMCCS is most sensitive to BBA, SA, and mRMR + GA on Ger data sets. Table 9 shows that MSMCCS reaches the highest specificity in two data sets (Ion and Son). Moreover, the specificity of MSMCCS, mRMR + PSO, and ALO are highest on the Ger data set. From the results, it can be seen that the improved  $\lambda$  and  $pa$  obviously enhance the classification accuracy. Therefore, the improved lambda and pa make a big contribution to sensitivity and specificity.

The Wilcoxon signed-rank test is proposed by Frank Wilcoxon as a non-parametric statistical hypothesis test [42]. This strategy is applied to contrasting two related samples. We can decide whether the corresponding data distributions are identical based on this test. In this paper, the Wilcoxon signed-rank test is executed by SPSS software. The data information in Tables 10 and 11 are the result of applying SPSS software. In Table 10, ten pairs of Wilcoxon signed-rank tests are made on Veh data set. With the significant level 0.05, it can be seen from Table 10 that the performance of MSMCCS is better than other ten algorithms. In other words, the classification accuracy of each fold is raised.

The descriptive statistics of eleven algorithms are described in Table 11.  $N$  represents the processing times. Mean measures the central tendency of the data set. The standard deviation is a measure that is used to quantify the amount of variation or dispersion of a set of data values. From Table 11, it can be seen that the mean value of MSMCCS is higher than the other ten algorithms. The results mean the central tendency the MSMCCS is the best. They represent that the MSMCCS algorithm is optimized by introducing the selected probability of each feature in improved classification accuracy rate. Thus, the test results indicate the proposed strategy very effective.

Eight public data sets of UCI machine learning repository are tested by SVM, K-Nearest Neighbor (KNN), Naive Bayesian (NB), Random Forest (RF), and Adaboost classifiers, respectively [43]. From Table 12, we can see that MSMCCS based on SVM classifier reaches the highest average classification accuracy in all data sets. Furthermore, it is

obvious that the performance of MSMCCS applied with SVM classifier is better than using other classifiers. As a result, the SVM is the most suitable classifier for MSMCCS.

## 5 Conclusion and future work

In order to solve the stability problem, this paper proposes a hybrid feature selection algorithm-MSMCCS, which is to embed the filter method into the wrapper method. The process of the algorithm is to select some features from all the features as candidate feature subsets according to the maximum Spearman and minimum covariance strategy, and then select the optimal feature subset according to the probability relationship. From the experimental process, we can see that the proposed algorithm have fast convergence. Finally, experimental results show that the classification accuracy of the proposed method is significantly better than the other ten algorithms. Moreover, the sensitivity, specificity, and Wilcoxon signed-rank test were used to assess the statistical significance of the differences between the proposed method and the other methods. The performance of MSMCCS applied with SVM classifier is better than using other four classifiers.

Microarray data is a significant representation. In the future work, we will use the microarray data as a research object. Through the process of selecting features and adjusting feature subsets, we will find optimal feature subsets with higher classification accuracy and shorter length from the data set.

**Acknowledgments** This research is supported by the National Natural Science Foundation of China (NSFC) under grant no. 61602206.

## References

1. Armanfard N, Reilly JP, Komeili M (2016) Local feature selection for data classification. *IEEE Trans Pattern Anal Mach Intell* 38: 1217–1227
2. Zeng H, Cheung YM (2011) Feature selection and kernel learning for local learning-based clustering. *IEEE Trans Pattern Anal Mach Intell* 33:1532–1547



3. Wang D, Nie F, Huang H (2015) Feature selection via global redundancy minimization. *IEEE Trans Knowl Data Eng* 27:2743–2755
4. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans Pattern Anal Mach Intell* 19:711–720
5. Lu H, Plataniotis KN, Venetsanopoulos AN (2008) MPCA: multilinear principal component analysis of tensor objects. *IEEE Trans Neural Netw* 19:18–39
6. He X, Yan S, Hu Y, Niyogi P, Zhang HJ (2005) Face recognition using laplacianfaces. *IEEE Trans Pattern Anal Mach Intell* 27:328–340
7. Belkin M, Niyogi P (2003) Laplacian Eigenmaps for dimensionality reduction and data representation. *Neural Comput* 15:1373–1396
8. Miguel GT, Ruben A, Concha B, Pedro L (2013) Comparison of metaheuristic strategies for peakbin selection in proteomic mass spectrometry data. *Inf Sci* 222:229–246
9. Mirjalili S (2015) The ant lion optimizer. *Adv Eng Softw* 83:80–98
10. Yang XS, He X (2013) Bat algorithm: literature review and applications. *Int J Bio-Inspir Com* 5:141–149
11. Rodrigues D, Pereira LAM, Nakamura RYM, Costa KAP, Yang XS, Souza AN, Papa JP (2014) A wrapper approach for feature selection based on Bat Algorithm and Optimum-Path Forest. *Expert Syst Appl* 41:2250–2258
12. Passino KM (2002) Biomimicry of bacterial foraging for distributed optimization and control. *IEEE Control Syst* 22:52–67
13. Chen YP, Li Y, Wang G, Zheng YF, Xu Q, Fan JH, Cui XT (2017) A novel bacterial foraging optimization algorithm for feature selection [J]. *Expert Syst Appl* 83(C):1–17
14. Yang XS, Deb S (2009) Cuckoo search via Lévy flights. *World Congress on Nature & Biologically Inspired Computing*, 210–214
15. Mohapatra P, Chakravarty S, Dash PK (2015) An improved cuckoo search based extreme learning machine for medical data classification. *Swarm Evol Compu* 24:25–49
16. Tsai CF, Eberle W, Chu CY (2013) Genetic algorithms in feature and instance selection. *Knowl-Based Syst* 39:240–247
17. Wang Z, Shao YH, Wu TR (2013) A GA-based model selection for smooth twin parametric-margin support vector machine. *Pattern Recogn* 46:2267–2277
18. Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: *Proceedings of the conference on neural networks*, IEEE Perth, Australia, 1942–1948
19. Vieira SM, Mendonça LF, Farinha GJ, Sousa JMC (2013) Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. *Appl Soft Comput* 13:3494–3504
20. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680
21. Lin SW, Lee ZJ, Chen SC, Tseng TY (2008) Parameter determination of support vector machine and feature selection using simulated annealing approach. *Appl Soft Comput* 8:1505–1512
22. Sebban M, Nock R (2002) A hybrid filter/wrapper approach of feature selection using information theory. *Pattern Recogn* 35: 835–846
23. Freeman C, Dana, Basir O (2015) An evaluation of classifier-specific filter measure performance for feature selection. *Pattern Recogn* 48:1812–1826
24. Sardana M, Agrawal RK, Kaur B (2015) An incremental feature selection approach based on scatter matrices for classification of cancer microarray data. *Int J Comput Math* 92(2):277–295
25. Mohamed NS, Zainudin S, Othman ZA (2017) Metaheuristic approach for an enhanced mRMR filter method for classification using drug response microarray data. *Expert Syst Appl* 90:224–231
26. Yang P, Ho JW, Yang YH, Zhou BB (2011) Gene-gene interaction filtering with ensemble of filters. *Bmc Bioinf* 12:2901–2917
27. Dai J, Xu Q (2013) Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Appl Soft Comput* 13(1):211–221
28. Chembumroong S, Shuang C, Yu H (2015) Maximum relevancy maximum complementary feature selection for multi-sensor activity recognition [J]. *Expert Syst Appl* 42(1):573–583
29. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 27:1226–1238
30. Akadi AE, Amine A, Ouadighi AE, Aboutajdine D (2011) A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. *Knowl Inf Syst* 26:487–500
31. Alshamlan H, Badr G, Alohal Y (2015) mRMR-abc: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. *Biomed Res Int* 2015(4):1–15
32. Unler A, Murat A, Chinnam RB (2011) Mr(2)PSO: a maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. *Inf Syst* 181:4625–4641
33. Moradi P, Gholampour M (2016) A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy [J]. *Appl Soft Comput* 43:117–130
34. Yang XS, Deb S (2014) Cuckoo search: recent advances and applications. *Neural Comput Applic* 24(1):169–174
35. Ouassab A, Ahiod B, Yang X-S (2014) Discrete cuckoo search algorithm for the travelling salesman problem. *Neural Comput & Applic* 24(7–8):1659–1669
36. Turhal ÜÇ, Duysak A (2015) Cross grouping strategy based 2DPCA method for face recognition. *Appl Soft Comput* 29:270–279
37. Katrutsa AM, Strijov VV (2015) Stress test procedure for feature selection algorithms. *Chemom Intell Lab Syst* 142:172–183
38. Berrendero JR, Cuevas A, Torrecilla JL (2014) Variable selection in functional data classification: a maxima-hunting proposal. *Stat Sin* 619–638. <https://doi.org/10.5705/ss.202014.0014>
39. Li SY, Li TR, Liu D (2013) Incremental updating approximations in dominance-based rough sets approach under the variation of the attribute set. *Knowl Based Syst* 40:17–26
40. Huang CL, Wang CJ (2006) A GA-based feature selection and parameters optimization for support vector machines. *Expert Syst Appl* 31:231–240
41. Kane MD, Jatke TA, Stumpf CR, Lu J, Thomas JD, Madore SJ (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* 28:4552–4557
42. Conover WJ (1973) On methods of handling ties in the Wilcoxon signed-rank test. *J Am Stat Assoc* 68:985–988
43. Soria D, Garibaldi JM, Ambrogio F, Biganzoli EM, Ellis IO (2011) A ‘non-parametric’ version of the naive Bayes classifier. *Knowl Based Syst* 24:775–784