Appl Intell (2010) 33: 117–131
DOI 10.1007/s10489-008-0153-8

Sequential multi-criteria feature selection algorithm based on agent genetic algorithm

Yongming Li · Xiaoping Zeng

Published online: 15 November 2008

Abstract A multi-criteria feature selection method-sequential multi-criteria feature selection algorithm (SMCFS) has been proposed for the applications with high precision and low time cost. By combining the consistency and otherness of different evaluation criteria, the SMCFS adopts more than one evaluation criteria sequentially to improve the efficiency of feature selection. With one novel agent genetic algorithm (chain-like agent GA), the SMCFS can obtain high precision of feature selection and low time cost that is similar as filter method with single evaluation criterion. Several groups of experiments are carried out for comparison to demonstrate the performance of SMCFS. SMCFS is compared with different feature selection methods using three datasets from UCI database. The experimental results show that the SMCFS can get low time cost and high precision of feature selection, and is very suitable for this kind of applications of feature selection.

Keywords Sequential · Multi-criteria · Feature selection · Agent · Genetic algorithm

1 Introduction

In many pattern classification applications, selection of the most characterizing features (or attributes) of the observed data (such as feature selection or variable selection, among many other names) is important to maximize the classification accuracy [1–9]. This is especially important when one is required to deal with a large or even overwhelming feature set.

In all feature selection methods, evaluation criteria are crucial because it guides the search algorithm to look for the best feature subset. Different types of evaluation criteria divide feature selection methods into two categories: the filter methods and the wrapper ones (or three categories: the filter methods, the wrapper ones, and the hybrid ones) [7]. In the wrapper methods, the classification accuracy is employed to evaluate feature subsets, whereas, in the filter methods, various measurements may be used as evaluation criteria [10]. Filter methods evaluate the quality of the feature subset by using evaluation criteria [11]. Lei Wang [12] proposed one new filter method based on one new evaluation criterion, obtaining satisfying effect on some datasets. Serkan Gunal et al. [13] studied another new evaluation criterion, thereby proposing a new filter method. They are relatively computationally cheap since they do not involve the induction algorithm. However, usually they will take the risk of selecting features subsets that may not match the chosen induction algorithm. In those cases, the results are not very accurate. The wrapper methods directly use the induction algorithm to evaluate the feature subsets. Liang-Hsuan Chen et al. [14] proposed one wrapper method based on GA + SVM for diagnosis of business crisis. Zheng Jiang et al. [15] proposed the similar method. Nick J. Pizzi et al. [16] proposed another wrapper method based on a set of classifiers. From the method, we can see that if classification accuracy can be taken as evaluation criterion, this method is a multi-criteria method which is better than single criteria method. Yong Yang et al. [17] proposed similar wrapper method based on classifier ensemble. They generally outperform filter methods in terms of classification accuracy, but they are generally computationally expensive and time consuming. Thus, it is difficult for them to deal with large feature sets in practice.

Y. Li (✉) · X. Zeng
College of Communication Engineering, Chongqing University, Chongqing 400030, China
e-mail: lymcentor924924@gmail.com

Recently, some researchers studied the hybrid feature selection methods based on filter and wrapper method [10, 11, 18–20]. For example, Mohua et al. [21] proposed one improved feature selection method based on filter-wrapper method (size of feature subset and classification accuracy) and multi-objective genetic algorithm (MOGA [22]). For the feature selection of gene expression data, the comparison with some other methods shows that the method is very effective and promising. Wing W.Y. et al. [23] proposed one feature selection method with another filter-wrapper method based on generalization error; the experimental results show it is effective. However, it is only suitable for dataset with large redundancy. For dataset with features with complex interference, it removes some useful features. Huang Yuan [18] proposed another one hybrid method; he firstly did feature selection with filter method, then deal with the feature subset after filter method with wrapper method. The method can improve selection speed above classical wrapper method because the filter method can get rid of some irrelevant features before feature selection with wrapper method to decrease the time cost. However, the method still needs wrapper method and the time cost cannot be reduced a lot, especially for some applications with strict need for time cost (such as biomedical image recognition, motion target recognition, et al.). These applications have some characteristics: dimensions vary from 10 to 100, the relationships between features are very complex, and time cost required is very low. Besides, if the filter method loses some useful features, the latter wrapper method cannot get them back, the feature selection precision (i.e. classification accuracy) will be affected. Some other researchers such as Yun Li, Jinjie Huang, Cheng-Lung Huang and Blazadonakis M.E. proposed the similar hybrid feature selection method based on filter and wrapper method [19, 20, 24, 25]. The differences between their methods are different search algorithms and classifiers (induction algorithms). Yew-Soon Ong et al. [11] proposed another kind of filter-wrapper method. They adopted GA running wrapper method while one local search algorithm is embedded in the GA. The local search algorithm runs filter method. With this method, the GA has two layers iteration cycles. The outer cycle is based on wrapper method and the inner cycle is based on filter method. However, this kind of method is wrapper method using filter method, the time cost is still large.

From the discussion above, it can be seen that: (1) Wrapper method can obtain high classification accuracy, but needs high time cost too, and is not suitable for some applications with strict need for time cost. (2) Filter method can have satisfying low time cost, but any single evaluation criterion is different from classification accuracy and no single evaluation criterion can evaluate feature subset correctly. (3) Filter-wrapper method is between filter and wrapper method, but for some feature selection problems where relationship between features is complex, if the filter method loses some useful features, the latter wrapper method cannot get them back, the feature selection precision (i.e. classification accuracy) will be affected. Besides, the filter-wrapper method still uses wrapper method and still has lots of time cost. Therefore, multi-criteria feature selection method is considered which can balance time cost and precision well.

Son Doan et al. proposed one multi-criteria feature selection algorithm for text classification [26]. The algorithm makes use of multiple evaluation criteria and gets improved performance. However, firstly the algorithm does not consider the correlation between features, practically, the correlation is very important, so the search algorithm needs to be improved; secondly, the threshold value is very hard to be made certain in practice; thirdly, time cost is still high and is k times than that of feature selection algorithm with one evaluation criterion (k means the number of evaluation criteria introduced) supposing filter method with each different evaluation criterion needs same time cost; fourthly, the correct processing on all the feature subsets with different evaluation criteria is hard to decide. Here, for Son Doan et al., the processing is based on this formula: $S \leftarrow S_1 \cup S_2 \cup \cdots \cup S_k$. But this processing is possible to involve some weak features (unimportant features). Hanchuan Peng et al. [27] proposed the two criteria feature selection method; the two criteria are relevancy and redundancy respectively. They incorporate the two criteria into one formula and use one search algorithm to look for optimal feature subset based on this formula. After that, one wrapper method is adopted to find compact feature subset. In our opinion, this method is still a filter-wrapper method. Christm Emmanouilidis et al. [28] applied two "evaluation criteria" into GA for feature selection. The two criteria are to minimize the number of features in the subset and to maximize classification accuracy. Since the classification accuracy is used to evaluate the quality of feature subset, we think this method as wrapper method or filter-wrapper method. As we know, the method will lead to lots of time cost. Benjamin Auffarth et al. [29] propose similar method as Christm Emmanouilidis, the difference is that the different classifiers are used.

Since the multi-criteria feature selection methods belong to filter method, they have the advantage of feature selection speed (i.e. low time cost required). With multi-criteria, different evaluation criteria can evaluate the feature subset in different angles, thereby improving the feature selection precision. Therefore, this kind of feature selection method is preferable. Hence, one new feature selection method is proposed here, it is sequential multi-criteria feature selection method. The major procedure is that several evaluation criteria are used to guide feature selection. The first evaluation criterion is used to guide feature selection for a period. Afterwards, the second evaluation criterion is used to guide feature selection for a period. The similar process continues until all the evaluation criteria are used. As we know, with this method, when one criterion is used to guide feature selection, it is based on the feature subsets guided by its previous criteria, the quality of the feature subsets is quite better than that of feature subsets generated at random, naturally the time cost will be reduced. So it is faster than parallel multi-criteria feature selection method. This method has the following advantages:

1. With more evaluation criteria used, the population (i.e. one group of feature subsets) becomes better. The bad and weak features are removed gradually. This process is quite different from the filter-wrapper method. With the filter-wrapper method, when filter method is over, the wrongly lost features can not be taken back. But with this method, with genetic processing (selection, crossover, mutation), the wrongly lost features can be possibly taken back.
2. This method does not apply wrapper method; therefore, the time cost needed is quite low.
3. This method is faster than parallel multi-criteria feature selection method.

Therefore, for the feature selection problems with strict need for low time cost and high precision, this method is attractive.

The paper is organized as follows: Sect. 2 describes the analysis and realization of the sequential multi-criteria feature selection method. In Sect. 3, by several datasets from UCI database, we evaluate the efficiency of this method for feature selection through comparing it with filter methods with single evaluation criterion, wrapper method, filter-wrapper method and other multi-criteria method. Finally, some conclusions are offered in Sect. 4 and future work is involved in the same section.

2 Analysis and realization of algorithm

2.1 Sequential multi-criteria idea

2.1.1 Consistency and otherness of evaluation criteria

The evaluation criteria within filter method have very strong consistency with classification accuracy (i.e. the evaluation criterion within wrapper method) and have consistent direction to guide feature selection. It is the major reason why these evaluation criteria can be used for feature selection. Here, we list and analyze three evaluation criteria to show their consistency with classification accuracy (the analysis is based on classification of two classes, for the classification with multi-class, the principle is similar).

(1) evaluation criterion 1 and its fitness function

The corresponding fitness function of evaluation criterion 1 is: $\mathit{fitness}_1 = \sum_{i=1}^{N}(S_b/S_w)_i - corr2$. Here, $S_b = (m_1 - m_2)^2$, $S_w = (\sigma_{M\_1})^2 + (\sigma_{M\_2})^2$, N is number of features, m_1 is mean of specimens within class 1 under some feature, m_2 is mean of specimens within class 2 under same feature, $corr2$ is correlation between features. σ_{M_1} is variation of specimens within class 1 under some feature, σ_{M_2} is variation of specimens within class 2 under same feature.

From the fitness function above, the S_b means the variance between class 1 and class 2 (the variance is called between-class variance), the variance is 2-Euclidean distance. If it is larger, then the classification of two classes is possible to be easier, and the classification accuracy is possible to be better. S_w means the variance between specimens under same feature and is called within-class variance. Since the specimens belong to same class, naturally less S_w is preferable. The two parameters (S_b and S_w) are satisfied better, the corresponding feature subset is better. Better feature subset will be possible to lead to higher classification accuracy. $corr2$ is correlation between features, it means the interference between features. As we know, less interference between features is preferable. Ideal case is the features are orthogonal each other (i.e. the features do not interfere each other). Besides, less $corr2$ is possible to lead to fewer features, thereby reducing the dimensional complexity of classifier. High classification accuracy and fewer features mean good feature subset.

(2) evaluation criterion 2 and its fitness function

The corresponding fitness function of evaluation criterion 2 is: $\mathit{fitness}_2 = \sum_{i=1}^{N}(\frac{S'_b}{S'_w})_i - corr2$, here, $S'_b = |m_1 - m_2|$, $S'_w = ((\sigma_{M\_1})^2 + (\sigma_{M\_2})^2)^{1/2}$.

From the formula above, it can be seen that the S_b means the variance between class 1 and class 2 (the variance is called between-class variance), the variance is 1-Euclidean distance. If it is larger, then the classification of two classes is possible to be easier, and the classification accuracy is possible to be better. Besides, S_w means the variance between specimens under same feature and is called within-class variance. Apparently, according to this formula, the feature subset with high fitness value will be possible to lead to high classification accuracy and less complex classifier.

(3) evaluation criterion 3 and its fitness function

The corresponding fitness function of evaluation criterion 2 is: $\mathit{fitness}_3 = \sum_{i=1}^{N}(\mathrm{tr}[S_b])_i - corr2$, here, $\mathrm{tr}[S_b] = \sum_{i=1}^{c} n_i \|m_i - m\|^2$, N is number of features, m_i is mean of specimens within ith class under some feature, m is mean of all specimens under same feature, n_i is number of specimens within ith class.

From the formula above, the third evaluation criterion maximizes between-classes variance and minimizes within-classes variance. Apparently, according to this formula, the

Table 1 The principle of consistency between filter method and wrapper method

	Feature subsets (individuals)																Fitness value based on evaluation criterion 1
The individuals in *i*th generation with filter method	0	1	0	0	0	0	1	1	1	0	1	1	1	1	1	1	17.3667
	0	1	0	0	0	0	1	1	1	1	1	1	1	0	1	1	15.0876
	0	0	0	0	1	0	1	1	1	0	1	1	1	0	1	1	15.2921
	0	0	0	0	1	0	1	1	1	1	1	0	1	0	1	1	15.0242
	0	0	0	0	1	1	1	1	1	1	1	1	1	0	1	1	15.2692
	0	0	0	0	1	1	1	1	1	0	1	1	1	0	1	1	15.1325
	1	0	0	0	1	1	1	1	0	0	1	1	1	1	1	1	12.7416
	1	0	0	0	1	1	1	0	1	0	1	0	1	0	1	0	12.9325
	1	0	0	0	1	0	1	0	1	1	1	0	1	1	1	1	16.1011
	0	0	0	0	1	0	1	1	1	1	1	0	1	1	1	1	17.4400
	0	0	0	0	1	0	1	1	1	0	1	0	1	0	1	1	14.8875
	0	0	0	0	1	0	1	1	1	0	1	0	0	0	1	1	15.1138
	0	1	0	0	1	0	1	1	1	0	1	1	0	1	0	0	13.6880
	0	0	0	0	1	1	1	1	1	0	1	0	1	0	1	1	14.7278
	0	1	0	0	1	0	1	1	1	0	1	1	1	1	1	1	17.0234
	0	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	16.0764
	0	1	0	0	0	0	1	1	1	0	1	1	1	1	1	1	17.3667
	0	0	0	0	1	0	1	1	1	0	1	1	1	0	1	1	15.2921
	0	1	0	1	0	0	1	1	1	0	1	0	1	0	1	1	13.6251
	0	1	0	1	0	0	1	1	1	0	1	0	1	1	1	1	16.0409
Final individual with wrapper method	0	0	0	1	0	0	1	1	1	1	0	0	0	0	1	1	

feature subset with high fitness value will be possible to lead to high classification accuracy and less complex classifier.

Through the analysis above, the evaluation criteria within filter method have very strong consistency with classification accuracy and have consistent direction to guide feature selection. The characteristic is called principle of consistency. From the Table 1, the principle of consistency can be seen more clearly.

From Table 1, we can see that with filter method, the individuals with high fitness value have lots of same genes with the optimal individual with wrapper method. It means that the evaluation criteria with filter method have very strong consistency with classification accuracy with wrapper method.

However, different evaluation criteria are designed based on different understanding of people on pattern classification problems. From the fitness functions of different evaluation criterion, they are different and evaluate feature subset from different angle. For example, distance evaluation criteria are based on distance idea. People design this kind of evaluation criteria based on the thought that some feature with high between-class variance and long within-class variance is a good feature. Information evaluation criteria are another kind of evaluation criteria. People design this kind of evaluation criteria based on the thought that some feature with high relationship with class is a good feature. With different evaluation criteria, corresponding fitness functions are different, they can be called $f_i(x)$, i means ith evaluation criterion. The feature subsets through these evaluation criteria guide classification and obtain corresponding classification accuracy, therefore there is mapping relationship between fitness function and classification accuracy based on same specimens. The mapping relationship can be called $h(f_i(x))$, where the mapping operator $h \cdot f_i()$ can be called $q_i()$. Suppose feature subset is independent variable, the corresponding true classification accuracy is attributive variable y, the mapping relationship is $g()$, then $y = g(x)$. Apparently, $g(x)$ is different from any of $q_1(x), q_2(x), \ldots, q_n(x)$. It is why any evaluation criterion with filter method is different from classification accuracy with wrapper method. It is also why the feature subset from filter method is worse than that from wrapper method usually. Therefore, if the ith evaluation criterion is used to do feature selection, even the global optimal feature subset based on the ith evaluation criterion (corresponding fitness function is $f_i(x)$) is obtained, it is just near global optimal solution based on $g(x)$, the corresponding classification accuracy is not satisfying naturally. Seen from Fig. 1, $q_1(x)$

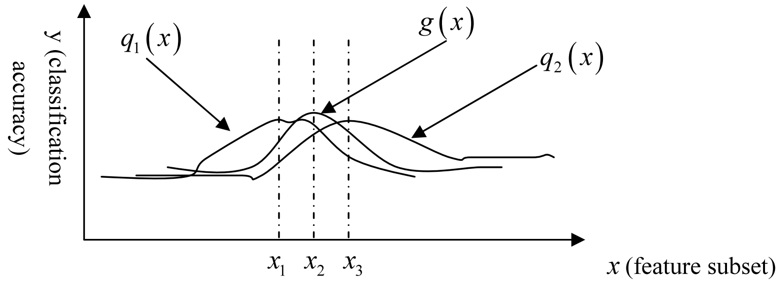

Fig. 1 The principle of otherness of evaluation criteria

and $q_2(x)$ mean the relationship between two feature subsets from two evaluation criteria and classification accuracy, $g(x)$ means the relationship between the feature subset and the true classification accuracy. x_1, x_2, x_3 mean the corresponding optimal feature subset. Seen from the Fig. 1, the $q_1(x)$ and $q_2(x)$ have otherness with $g(x)$. Compared with the $g(x_2)$, the $q_1(x_1)$ and $q_2(x_3)$ is lower. The characteristic is called principle of otherness.

2.1.2 *Sequential multi-criteria idea*

As the discussion above, multi-criteria feature selection method should be a good choice. However, as the discussion in Sect. 1, the parallel multi-criteria feature selection method has several drawbacks. In contrast, the serial multi-criteria feature selection method can overcome these drawbacks. The major idea of the serial multi-criteria feature selection is that it uses more than one evaluation criterion serially (one after another), when one criterion is used to guide feature selection, it is based on the feature subsets guided by its previous criteria, the quality of the feature subsets is quite better than that of feature subsets generated at random, naturally the time cost will be reduced. The time cost is lower than that in parallel multi-criteria feature selection method. Besides, it is not necessary to consider the correct processing on all the feature subsets with different evaluation criteria, because the processing is serial.

On the other hand, as we know, different evaluation criteria have different feature selection capability according to same or different feature selection problems. Therefore, if the evaluation criteria are used in random order without considering their different capability, the optimal feature selection performance can not be assured for different datasets.

Now, the three evaluation criteria (described in Sect. 2.1.1) are used for feature selection according to some datasets. If the different capabilities of the three criteria are not considered, these criteria are used simply one by one without considering their feature selection capability. For example, firstly, evaluation criterion 1 is used; secondly evaluation criterion 2 is used; finally evaluation criterion 3 is used. The corresponding dataset is about wave (the dataset will be described in Sect. 3). The feature selection performance of this kind of random multi-criteria feature selection method is shown in Table 2.

Table 2 Performance of random multi-criteria feature selection method

	Random multi-criteria (f1-f2-f3)	Evaluation criterion f1	Evaluation criterion f2	Evaluation criterion f3
Average generation	58.0	36.5	39.4	42.9
Average time (s)	21.8321	11.8204	18.5467	25.0562
Classification accuracy	0.8314	0.8291	0.8303	0.8360

In Table 2, f1, f2 and f3 mean the fitness functions based on evaluation criterion 1, 2 and 3 respectively. Table 2 shows that according to the dataset, this random multi-criteria feature selection method has similar time cost with filter method with single evaluation criterion. The random multi-criteria feature selection method has higher classification accuracy than filter method with evaluation criterion 1 and 2. However, the random multi-criteria feature selection method has lower classification accuracy than filter method with evaluation criterion 3. Theoretically, the multi-criteria method uses the three criteria one by one, the first used criterion is evaluation criterion 1. So this first used criterion is very important since it is the beginning point of feature selection (i.e. searching). Based on this analysis, we rearrange the order of the use of these three criteria based on their different capability. According to their feature selection precision on this dataset, the evaluation criterion 3 is best, the evaluation criterion 1 is worst, therefore, the preferable order is f3-f2-f1 (i.e. firstly the criterion 3 is used; finally the criterion 1 is used). The processing is called sequential multi-criteria feature selection method, please see Table 3. Accordingly, the classification accuracy of the feature subset from this method is improved (Sect. 3 will give detailed description and experimental results).

2.2 Study of search algorithm

In order to distinguish the different method (single evaluation criterion, multi-criteria, wrapper) better, a search algorithm with high precision is necessary. Here, chain-like

Table 3 The idea of sequential multi-criteria feature selection method

Step 1: adopt evaluation criterion 1 to guide feature selection, obtaining the feature selection capability of this evaluation criterion, here classification accuracy stands for the feature selection capability
Step 2: go back to Step 2, obtaining the feature selection capability of all the evaluation criteria
Step 3: rearrange them according to their evaluation criteria, obtaining the new order $Seq(i)$
Step 4: do feature selection using the first evaluation criterion $Seq(1)$, if the loose stop condition is met, then go to Step 5
Step 5: repeat Step 4 until last evaluation criterion $Seq(n)$, if the strict stop condition is met, then output final feature subset

*Here, loose stop condition means not too strict stop condition, its purpose is to avoid losing some important features, so the loose stop condition is enough

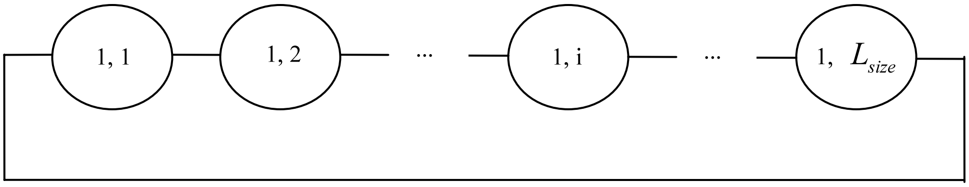

Fig. 2 Chain-like agent structure

agent genetic algorithm (CAGA) is adopted, the agent structure can be seen in Fig. 2; the reason is that it can obtain good solution in high dimensional and multi-peak search space with high precision [30].

2.2.1 *Structure of population*

In the chain-like agent genetic algorithm, all the agents live in a chain-like environment, L which is called an agent chain. The size of L is $1 \times L_{\text{size}}$, where L_{size} is an integer, 1 means one dimensional agent structure. Each agent is fixed on a chain-point and it can only interact with its neighbors.

Definition 1 Assuming that the agent that is located at $(1, i)$ is represented as $L_{1,i}, i = 1, 2, \ldots, L_{\text{size}}$, the neighbors of $L_{1,i}$, $Neibors_{1,i}$ are defined as follows:

$$Neibors_{1,i} = \{L_{1,i_1}, L_{1,i_2}\} \tag{1}$$

where

$$i_1 = \begin{cases} i-1 & i \neq 1 \\ L_{\text{size}} & i = 1 \end{cases}, \qquad i_2 = \begin{cases} i+1 & i \neq L_{\text{size}} \\ 1 & i = L_{\text{size}} \end{cases}.$$

The agent chain can be described as the one in Fig. 1. Each circle represents an agent, the data in a circle represents its position in the chain, and the agent can interact with the left neighboring one and the right neighboring one.

In traditional *GA*s, those individual that will generate children are usually selected from all individuals according to their fitness value. But in nature, a global selection does not exist, and the real natural selection only occurs in a local environment, and each individual can only interact with the neighboring ones. That is, the natural evolution is like a kind of local phenomenon. The information can be shared globally only after a process of diffusion. For description, the search algorithm is called as CAGA because of its chain-like agent structure.

2.2.2 *Neighborhood competition selection operator*

The neighborhood competition selection operator is described as follows: suppose the order of competition selection is from left to right, the current agent is $L^t_{1,i}$, the neighbors are $Nbs_{1,i}$, $Nbs_{1,i} = \{L^t_{1,i1}\ L^t_{1,i2}\}$, $i = 1, 2, \ldots, popsize$. Updating of $L^t_{1,i}$ is as the following formula:

$$\left\{\begin{array}{l} L^t_{1,i} = L^t_{1,i} \\ L^t_{1,i} = L^t_{1,i} \circ L^t_{1,i1} \\ L^t_{1,i} = L^t_{1,i} \circ L^t_{1,i2} \end{array}\right. \quad \left.\begin{array}{c} fitness(L^t_{1,i}) > fitness(\max(L_{1,i1}, L_{1,i2})) \\ \max(L_{1,i1}, L_{1,i2}) = L_{1,i1}\,\&\,fitness(L_{1,i1}) > fitness(L^t_{1,i}) \\ \max(L_{1,i1}, L_{1,i2}) = L_{1,i2}\,\&\,fitness(L_{1,i2}) > fitness(L^t_{1,i}) \end{array}\right\}. \tag{2}$$

In the formula (2), $\circ$ means competition selection between agent $L^t_{1,i}$ and $L^t_{1,i1}$, the two agents consist of lots of genes:

$$\begin{array}{l} L^t_{1,i} = \left(c^t_{i,1} \quad c^t_{i,2} \quad \cdots \quad c^t_{i,j} \quad \cdots \quad c^t_{i,\text{length}}\right), \\ L^t_{1,i1} = \left(c^t_{i1,1} \quad c^t_{i1,2} \quad \cdots \quad c^t_{i1,j} \quad \cdots \quad c^t_{i1,\text{length}}\right), \end{array} \tag{3}$$

$c^t_{i,j}$ means jth gene of $L^t_{1,i}$, $c^t_{i1,j}$ means jth gene of $L^t_{1,i1}$, length means number of genes of single agent. The competition selection between agent $L^t_{1,i}$ and $L^t_{1,i1}$ can be called $L^t_{1,i} \circ L^t_{1,i1}$, the processing is as follows:

$$\begin{cases} c^t_{i,j} = c^t_{i,j} & c^t_{i,j} = c^t_{i1,j} \\ c^t_{i,j} = \mathrm{U}(0,1) & c^t_{i,j} \neq c^t_{i1,j} \end{cases}. \tag{4}$$

U(0, 1) means random number generator and is within the domain [0, 1].

The procedures are as follows:

Step 1: define one register *temp* with space of (1 × 2), $temp \leftarrow (L^t_{1,i}, \max(L_{1,i1}, L_{1,i2}))$;
Step 2: update $L^t_{1,i}$ according to formula (2);
Step 3: judge if $i = popsize$, if true, go to crossover processing, or not, $i \leftarrow i + 1$, turn to Step 1.

Dynamic competition strategy During competition process, the $Max_{1,i} = \max(L_{1,i_1}, L_{1,i_2})$. The competition process is done in ascending order, after the competition of the 1st agent, the 1st agent is updated. Assuming the ith agents before competition and after competition are $L^{pre}_{1,i}$ and $L^{post}_{1,i}$ respectively, so $Max_{1,i}$ is determined by

$$Max_{1,i} = \begin{cases} \max(L^{pre}_{1,L_{\text{size}}}, L^{pre}_{1,i+1}) & i = 1, \\ \max(L^{post}_{1,L_{\text{size}}-1}, L^{post}_{1,1}) & i = L_{\text{size}}, \\ \max(L^{post}_{1,i-1}, L^{pre}_{1,i+1}) & \text{else}. \end{cases} \quad (5)$$

2.2.3 *Neighborhood adaptive crossover operator*

In the crossover process, the crossover probability $p_{c,i}$ is calculated adaptively. The corresponding formula is as follows:

$$p_{c,i} = \begin{cases} (\frac{f_{\max} - f'_i}{f_{\max} - f_{ave}})^{\frac{1}{GH(i,i')}} & f' \geq f_{ave}, \\ 1 & f' < f_{ave}. \end{cases} \quad (6)$$

Here, $p_{c,i}$ means the probability of crossover about the crossover operation between the $L_{1,i}$ and $Max_{1,i}$, $GH(i, i')$ means the distance between the $L_{1,i}$ and $Max_{1,i}$, f' means the maximum value of both the individuals, $f_{\max}$ means the maximum value of all the individuals in the current population, f_{ave} means the average fitness value of all the individuals. The crossover procedure is as follows:
if U(0, 1) $< p_{c,i}$
do single point crossover processing between $L_{1,i}$ and $Max_{1,i}$
else keep $L_{1,i}$ no change.

2.2.4 *Adaptive mutation operator*

In the crossover process, the mutation probability p_m is calculated adaptively based on the length of chromosome. The p_m is determined by: $p_m = \frac{1}{n}$, where n means number of genes, namely the length of chromosome.

The crossover procedure is as follows:
if U(0, 1) $< p_m$
do single point mutation processing between $L_{1,i}$ (namely, some gene changes its value from 1 to 0 or vice versa randomly)
else keep $L_{1,i}$ no change.

2.2.5 *Stop criterion (stop condition)*

f_{ave} can reflect the evolution of the current population. f_{best} stands for the best average fitness value since beginning. k_{stop} means a counter, it counts the number that f_{best} has no change. If $k_{stop} > k$, the search stops.

2.2.6 *Elitism strategy*

Agents have knowledge which is related with the problem that they are designed to solve. With elitism strategy, the agent can inherit the good solution from the former generation. This method can make the best solution within ith generation better than or equals to the best solution in the former $(i - 1)$ generations.

2.2.7 *Realization of search algorithm (CAGA)*

The search Algorithm—dynamic chain-like agent genetic algorithm (CAGA) is described as the following Table 4.

2.3 Realization of SMCFS algorithm

Combining the idea of sequential multi-criteria idea and search algorithm (CAGA), the sequential multi-criteria feature selection algorithm (SMCFS) is proposed in the Table 5.

From the procedure of SMCFS, more than one evaluation criteria are adopted in sequential order, and the sequential order is made certain by feature selection capability of each evaluation criterion. Therefore, this kind of multi-criteria method is better than random multi-criteria feature selection method. Steps 9–11 make best use of the consistency of multi-criteria, former criterion provides better group of feature subsets (population) for latter criterion, thereby reducing time cost needed by latter criterion. Besides, the Steps 9–11 make best use of the otherness of multi-criteria, the latter criterion complements the former criterion, in favor of jumping out of local optima that former criterion easily falls into.

2.4 Computational complexity

Compared with filter method with single evaluation criterion, the SMCFS just use multiple criteria one after another one. As we know, with this method, when one criterion is used to guide feature selection, it is based on the feature subsets guided by its previous criterion, the quality of the feature subsets is quite better than that of random feature subsets, naturally the time cost will be reduced. So it is faster than parallel multi-criteria feature selection method. Suppose the different evaluation criterion has similar time cost and the computational complexity of filter method with single evaluation criteria is $O(pg)$, the computational complexity of the parallel multi-criteria feature

Table 4 Procedures of CAGA

Step 1: initialize L^0, update pop_{best}^{0}, and $t \leftarrow 0$

Step 2: do dynamic neighboring competitive selection processing and update L^t, obtaining $L^{t+1/3}$

Step 3: for each agent in $L^{t+1/3}$, do crossover processing on it, obtaining $L^{t+2/3}$

Step 4: for each agent in $L^{t+2/3}$, do mutation processing on it, obtaining L_{end}^{t}

Step 5: find ind_{best}^{ct} in L_{end}^{t}, and compare ind_{best}^{ct} and ind_{best}^{t-1}, if $Eng(ind_{best}^{ct}) > Eng(ind_{best}^{t-1})$, then $ind_{best}^{t} \leftarrow ind_{best}^{ct}$, else, $ind_{best}^{t} \leftarrow ind_{best}^{t-1}$, $L^{t+1} \leftarrow L_{end}^{t}$

Step 6: if stop criterion is satisfied, then output ind_{best}^{t} and stop, else $t \leftarrow t+1$, go to Step 2

* L^t represents the agent chain in the tth generation, and $L^{t+1/3}$ and $L^{t+2/3}$ are the mid-chains between L^t and L^{t+1}, L_{end}^{t} is the agent chain after mutation processing in the tth generation. ind_{best}^{t} is the best agent among $\{L^0, L^1, \ldots, L^t\}$, and ind_{best}^{ct} is the best agent in L^t. p_c and p_m are the probabilities to perform the neighboring crossover processing and the mutation processing

Table 5 Procedures of this algorithm (SMCFS)

Step 1: initialize population, construct the agent structure; %each agent stands for one feature subset, 1 means not chosen, 0 means chosen

Step 2: design the corresponding fitness function according to evaluation criterion 1, calculate the fitness value of all the agents

Step 3: do selection processing on all the agents with neighboring competition selection operator

Step 4: do crossover processing on all the agents with neighboring adaptive crossover operator

Step 5: do mutation processing on all the agents with adaptive mutation operator

Step 6: judge if the stop condition is met; if true, output the optimal feature subset, and send it to classifier, obtaining corresponding classification accuracy (i.e. feature selection capability); if not, turn to Step 2

Step 7: repeat Steps 1–6, obtaining the feature selection capability of all the evaluation criteria

Step 8: according to feature selection capability of evaluation criteria, obtain new order of usage of evaluation criteria $Seq(i)$

Step 9: set up cycle of $i = 1 : n$ (n is number of evaluation criteria) to do feature selection

Step 10: judge if the current used evaluation criterion is last one (i.e. $Seq(n)$); if true, judge if the strict stop condition is met; if true, turn to Step 11; if false, turn to Steps 2–5, then turn to Step 10. If the current used evaluation criterion is not last one, judge if the loose stop condition is met, if true, turn to Step 11, if false, turn to Steps 2–5, then turn to Step 10

Step 11: judge if the current used evaluation criterion is last one, if true, output optimal feature subset; if false, use next evaluation criterion (i.e. $Seq(i+1)$), then turn to Steps 2–5, then turn to Step 10

selection method is $O(npg)$, and the computational complexity of SMCFS is $O(kpg)$, where p is size of population, g is number of search generations and n is number of evaluation criteria. The k is one coefficient and meets the formula: $1 < k < n$. Through lots of experiments, usually the domain of k is $[1, 2]$.

3 Experiments and analysis of results

In order to verify the performance of SMCFS proposed in this paper, the authors realized the algorithm with MATLAB and organized five groups of experiments. In the first group of experiments, several different GAs are used to do feature selection based on evaluation criterion 1 (described in Sect. 2.1.1) according to several datasets. The purpose of the experiments is to show the search algorithm used in SMCFS has good searching capability. In the second group of experiments, SMCFS is compared with filter method with single evaluation criterion. The purpose of the experiments is to show SMCFS can have better feature selection precision than filter method with single criterion. In the third group of experiments, SMCFS is compared with wrapper method with several different classifiers. The purpose of the experiments is to show SMCFS can have close feature selection precision with wrapper method and have lower time cost than wrapper method. In the fourth group of experiments, SMCFS is compared with filter-wrapper feature selection method. The purpose of the experiments is to show the satisfying feature selection capability of SMCFS. In the fifth group of experiments, SMCFS is compared with other multi-criteria feature selection methods. The purpose of the experiments is to show the advantage of SMCFS over them.

The datasets are selected from popular international UCI database. The dataset 1 is letter- recognition dataset; the number of features is 16. The class a and b are used, there are 789 specimens belonging class a, 766 specimens belonging class b. The dataset 2 is waveform dataset; the number of features is 40. The class wave 0 and wave 1 are used, there are 2000 specimens belonging to class wave 0 and

Table 6 Some information about datasets

Datasets	Number of features	Number of specimens	Number of classes	Population size	Evaluation of feature subset	Stop criterion
Letter	16	1555	2	30	5-fold CV	$k > 10$ or 1000
Wave	40	2000	2	30	5-fold CV	$k > 10$ or 1000
Sonar	60	208	2	30	5-fold CV	$k > 10$ or 1000

*Here the stop criterion is $k > 10$ or maximum number of iteration is more than 1000

wave 1 respectively. The dataset 3 is Sonar dataset. There are 208 specimens belonging to two classes; the number of features is 60. The corresponding experimental condition is: CPU: 3 GHz, memory: 504 MHz; for comparison, the size of the populations of the four algorithms is 30; for SGAE, initial $p_c = 0.65$, initial $p_m = 0.05$; for SFGA, initial $p_m = 0.05$, the p_c is adaptive; for AGA and CAGA, the p_c and p_m are adaptive. The stop criterion is $k > 10$. For clarity, Table 6 shows some information about the experiments. For different groups of experiments, some information changes and is not involved in this table. The fitness value of a chromosome or selected feature subset is evaluated using some kind of classifier and 5 fold cross validation (5-fold CV).

3.1 Feature selection experiments by search algorithm

We know that the feature selection means the searching of the optimal features combination through optimization method. Therefore, good search algorithm is essential for feature selection method. Here, four genetic algorithms including AGA [31], MAGA [32], SFGA [33] and SGAE [34] are adopted to be compared with CAGA within SMFCS. The reasons for choosing these genetic algorithms are: firstly, SGAE is a traditional genetic algorithm with elitism strategy, and it has been used in various areas and performs well, so it is suitable to be compared with other improved genetic algorithms. Secondly, AGA is a representational adaptive genetic algorithm and can keep the diversity of population effectively, the comparison with it can shows CAGA has more powerful searching capability, can keep the diversity of population more effectively and avoid premature convergence to get near-global optima. Thirdly, FSGA is another improved genetic algorithm with adaptive crossover operator. It can adaptively adjust its probability of crossover to keep the diversity of population, and performs well for over ten benchmark functions. Fourthly, MAGA is an improved genetic algorithm with lattice-like agent population structure proposed recently. In [32], the MAGA is described to perform better than some well-known algorithms such as OGA/Q, AEA, FEP, BGA, so the comparison with it can show CAGA better performance over those genetic algorithms indirectly. This group of experiments is conducted to show the satisfying search capability of CAGA for feature selection.

The evaluation criterion 1 (described in Sect. 2.1.1) is used here, and its corresponding fitness function is made up of two parts here, it is: *fitness* = discriminability-correlation. The fitness function below is adopted for feature selection here: Fitness function (evaluation criterion): $fitness = \sum_{i=1}^{N}(\frac{S_b}{S_w})_i - corr2$, where, N is the number of the features; S_b means between-classes variance, $S_b = (m_1 - m_2)^2$; m_1 means the first class specimens under some feature; m_2 means the second class specimens under the same feature; $S_w = (\sigma_{class1})^2 + (\sigma_{class2})^2$; $corr2$ is correlation between features selected, and is called as within-classes variance. Here, the $corr2$ is to calculate the correlation of the features matrices of the two classes. $corr2$ is initialized as 0. The first step is to calculate the correlation of the first feature vector and the second feature vector within the first class p_corr1, then calculate the correlation of the first feature vector and the second feature vector within the second class p_corr2, after that, the correlation of the first feature and the second feature p_corr can be obtained with the formula: p_corr = (p_corr1 + p_corr2)/2. The p_corr is added to $corr2$. With the same processing, the correlation of the second feature and the third feature can be gotten and be added to the $corr2$. The processing lasted until the correlation of the $(N-1)$th feature and the Nth feature is obtained and is added to the $corr2$. At this time, the $corr2$ is obtained. For the evaluation 2 and 3 (described in Sect. 2.1.1), the processing is similar.

Figure 3 shows that the searching capability of the five search algorithms according to the dataset 1. For showing the searching capability of them, the initial population is same. In the figure, sga means SGAE, zsypc means AGA, haimin means SFGA, chain means CAGA, mage means MAGA. Abscissa means number of generations, Vertical axis means fitness value. From the figure, it can be seen that CAGA is the fastest one to get the near global optima, and the search result through CAGA is most precise. The corresponding number of generations needed is the smallest.

Based on the three datasets, the five GAs are used for feature selection respectively for 10 times respectively. Here, the used network is BP network.

The definition of the classification rate is described in Definition 2.

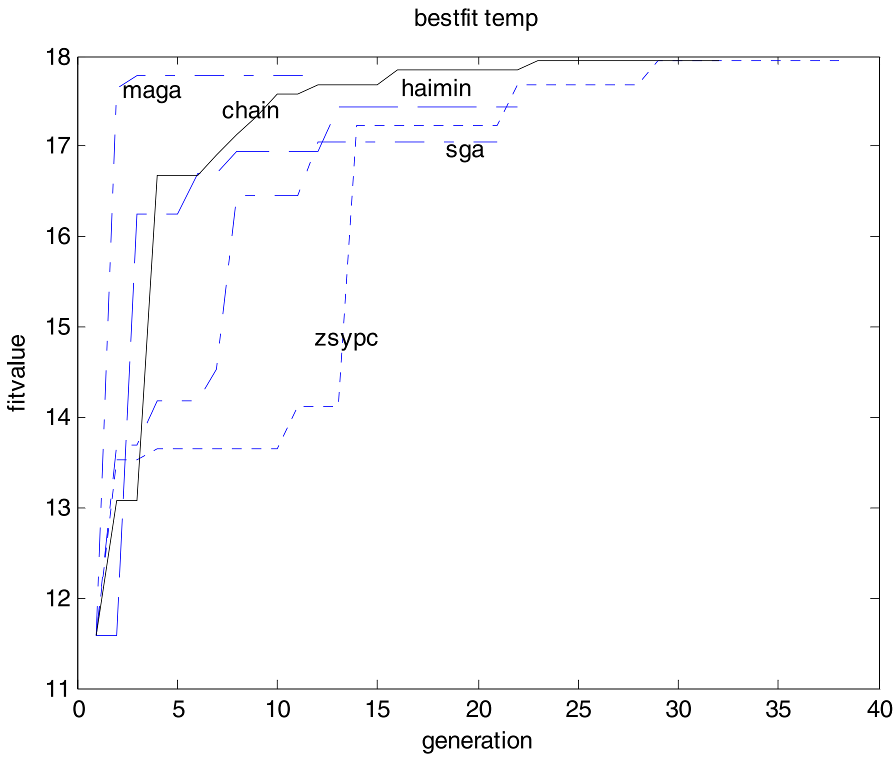

Fig. 3 The comparison of the searching capability for dataset 1

Definition 2 [1 0] and [0 1] is represented as the output object of the two classes. $N_i(1)$ stands for the ith specimens belonging to the first class, $N_i(2)$ stands for the ith specimens belonging to the second class. Suppose the output of the ith specimens is P_i, P_i is row vector with 1×2, then the output is as follows:

$$N_i(1) = \begin{cases} 1 & (P_i(1,1) \geq 0.8) \text{ and } (P_i(1,2) \leq 0.2), \\ 0 & \text{else}, \end{cases}$$

$$N_i(2) = \begin{cases} 1 & (P_i(1,1) < 0.2) \text{ and } (P_i(1,2) > 0.8), \\ 0 & \text{else}, \end{cases}$$

where, 1 means correction, 0 means error, then the classification rate of the specimens of the two classes is defined respectively as follows:

$$CA = \frac{\sum_{i=1}^{n\_class} \sum_{j=1}^{n\_test(i)} N_j(i)}{\sum_{i=1}^{n\_class} n\_test(i)}$$

where, CA means classification accuracy; i means sequence number of classes; j means sequence number of specimen with some class; $n\_test(i)$ means the number of tested specimens belonging to ith class; n_class means number of classes, here, $n\_class = 2$; $N_j(i)$ means the jth specimen with the ith class.

Table 7 lists experimental results of two of the 10 times experiments based on the former two datasets.

From Table 7, for letter dataset, the number of features from SGAE, AGA and CAGA is similar, but the number of features from CAGA is not big and more stable. Besides, the running time of CAGA is longer than SGAE, AGA and SFGA, it is because the CAGA adopts neighboring genetic operators and increases extra computational cost, but for dynamic feature selection problem, the additional time cost is acceptable. Compared with MAGA, the time cost of CAGA is lower than that of MAGA slightly because CAGA adopts more efficient agent structure and genetic operators than MAGA. The fitness value of feature subset from CAGA is the highest and most stable (always same during the two times experiments), it is because CAGA always is able to find the global optimal or near global optimal feature subset. The best advantage of CAGA is the classification accuracy; the classification accuracy of the feature subset from CAGA is better than that from other four GAs. As we know, the classification accuracy is a very essential parameter for feature selection. For wave dataset, the advantage of CAGA over other four GAs is more apparent.

3.2 Comparison of SMCFS and filter methods with single evaluation criterion

In order to compare the feature selection capability of SMCFS and filter methods with single evaluation criterion, the three evaluation criteria are considered in this group of ex-

Table 7 Comparison of feature selection capability of five GAs

DS	ET	CP	SGAE	AGA	SFGA	MAGA	CAGA
Letter	1	NF	10	9	10	11	10
		RT	13.7970	41.031	29.9220	28.109	26.2650
		BF	17.6629	17.6897	17.9449	17.7852	17.9449
		CA	0.9275	0.95	0.95	0.935	0.98
	2	NF	9	10	8	10	10
		RT	7.6570	50.078	21.2970	27.641	21.5150
		BF	17.8082	17.9449	16.7284	17.6629	17.9449
		CA	0.9375	0.94	0.9225	0.94	0.94
Wave	1	NF	18	21	18	15	14
		RT	18.4060	18.531	32.0630	55.314	52.371
		BF	−0.1269	−0.5034	0.4300	1.5769	1.5467
		CA	0.8025	0.7425	0.8025	0.77	0.8275
	2	NF	22	21	18	14	14
		RT	20.1400	19.859	30.5460	57.492	51.782
		BF	−0.1569	0.0130	0.2469	1.5443	1.6144
		CA	0.7825	0.8025	0.7950	0.79	0.8625

*DS means dataset, ET means experimental times, CP means compared parameters, NF means number of selected features, RT means running time, BF means the best fitness value, and CA means classification accuracy of selected feature subset

Table 8 Comparison of SMCFS with filter methods with single evaluation criterion (two criteria)

DS	CP	SMCFS (f3-f1)	Evaluation criterion 1 f1	Evaluation criterion 3 f3
Letter	ANF	9.5	10	9.5
	ART	5.9341	5.0033	6.9373
	ACA	0.9649	0.9384	0.9587
Wave	ANF	13	14.5	13.3
	ART	20.1118	11.8204	25.0562
	ACA	0.8998	0.8291	0.8360
Sonar	ANF	24	25.7	25.5
	ART	27.7675	32.5687	33.4545
	ACA	0.9654	0.9123	0.9320

*ANF means average number of features; ART means average running time for feature selection; ACA means average classification accuracy

Table 9 Comparison of SMCFS with filter methods with single evaluation criterion (three criteria)

DS	CP	SMCFS (f3-f2-f1)	Evaluation criterion 1 f1	Evaluation criterion 2 f2	Evaluation criterion 3 f3
Letter	ANF	9.5	10	9.4	9.5
	ART	5.9341	5.0033	4.9969	6.9373
	ACA	0.9799	0.9484	0.9495	0.9587
Wave	ANF	13	14.5	14.4	13.3
	ART	20.1118	11.8204	18.5467	25.0562
	ACA	0.9098	0.8291	0.8303	0.8360
Sonar	ANF	24	25.7	24.6	25.5
	ART	27.7675	32.5687	35.6743	33.4545
	ACA	0.9741	0.9123	0.9210	0.9320

periments. Here, four methods are compared. They are SMCFS, filter feature selection method with evaluation criterion 1 (f1), filter feature selection method with evaluation criterion 1 (f2), and filter feature selection method with evaluation criterion 1 (f3). In terms of multi-criteria, there are two cases: one case is two criteria (f1 and f3) are used to construct multi-criteria; another case is three criteria (f1, f2 and f3) are used to construct multi-criteria. For each method and each case, the same experiments are done for ten times, the final results are statistical average results and are involved in the Tables 8 and 9.

In the case of two criteria, the sequential multi-criteria method (SMCFS-(f3-f1)) is obtained. Since the criterion 3 is better than criterion 1 for the three datasets, the former is used to be first evaluation criterion, then criterion 1. Seen from the Table 8, the SMCFS has smaller features than f1 and f3 averagely. Besides, the SMCFS has similar time cost as the f1 and f3. Most importantly, the SMCFS has highest classification accuracy which is essential for feature selection.

In the case of three criteria, the sequential multi-criteria method (SMCFS-(f3-f2-f1)) is obtained. Since the criterion 3 is better than criterion 2, then criterion 2 is better than criterion 1 for the three datasets, the criterion 3 is used to be first evaluation criterion, then criterion 1 is last one. Seen from the Table 9, the SMCFS has smaller features than f1, f2 and f3 averagely. Besides, the SMCFS has similar time cost as the f1, f2 and f3. Most importantly, the SMCFS has

Table 10 Comparison of SMCFS and wrapper methods (three criteria)

DS	CP	SMCFS-SVM	Wrapper-SVM	SMCFS-BP	Wrapper-BP	SMCFS-RBF	Wrapper-RBF
Letter	ANF	9.6	10	9.5	9.8	9.5	9.8
	ART	5.9341	1845.3	5.9341	2224.7	5.9341	2576.8
	ACA	0.9812	0.9841	0.9799	0.9675	0.9654	0.9709
Wave	ANF	13	12.9	13	13.5	13	13.7
	ART	20.1118	5675.5	20.1118	6348.5	20.1118	6789.6
	ACA	0.9123	0.9467	0.9098	0.9343	0.9019	0.9291
Sonar	ANF	24	23.7	24	24	24	24.5
	ART	27.7675	9987.9	27.7675	10 234.8	27.7675	9878.8
	ACA	0.9821	0.9897	0.9741	0.9798	0.9703	0.9789

* SMCFS-SVM means SMCFS is used to do feature selection, and SVM is used to evaluate the classification accuracy of feature subset from SMCFS; SMCFS-BP and SMCFS-RBF just change the different classifiers. Wrapper-SVM means SVM is used to do feature selection with wrapper method and evaluate the classification accuracy of feature subset; Wrapper-BP and Wrapper-RBF just change the different classifiers

highest classification accuracy which is essential for feature selection.

3.3 Comparison of SMCFS and wrapper methods

In order to show the advantage of SMCFS over wrapper methods, this group of experiments adopts three kinds of classifiers. They are SVM, BP and RBF respectively. Therefore, there are six combinations; they are SMCFS-SVM, SMCFS-BP, SMCFS-RBF, Wrapper-SVM, Wrapper-BP. Wrapper-RBF. The former three combinations mean that firstly SMCFS is used to do feature selection, SVM or BP or RBF is used to evaluate the feature subset obtained to output the corresponding classification accuracy. The latter three combinations mean that SVM or BP or RBF is used to do feature selection and output the classification accuracy of feature subset. The latter three combinations belong to wrapper method. For each method, the same experiments are done for ten times, the final results are statistical average results and are involved in the Table 10. Here, SMCFS uses three evaluation criteria (i.e. f3-f2-f1).

Seen from the Table 10, the average number of features from SMCFS is similar as that from wrapper method regardless of different classifiers. Most apparently, the average time cost of SMCFS is shorter than that of wrapper method greatly. The wrapper method is so slow that it can not be acceptable, especially for some applications with high requirement for time cost. In terms of classification accuracy, the SMCFS is very close to the wrapper method. Considering the low time cost of SMCFS method, the SMCFS method is more attractive.

3.4 Comparison of SMCFS and filter-wrapper methods

Table 11 Comparison of SMCFS and filter-wrapper methods (three criteria)

DS	CP	SMCFS-BP	WFFSA-BP	GFSIC-SBFCV-BP
Letter	ANF	9.5	9.7	10.2
	ART	5.9341	78.9	89.4
	ACA	0.9799	0.9664	0.9781
Wave	ANF	13	12.1	13.5
	ART	20.1118	104.2	123.2
	ACA	0.9098	0.9091	0.9031
Sonar	ANF	24	22.4	22.4
	ART	27.7675	235.5	235.5
	ACA	0.9741	0.9776	0.9643

In order to show the advantage of SMCFS over filter-wrapper method, this group of experiments adopts two filter-wrapper methods for comparison. They are WFFSA [11] and GFSIC-SBFCV [18]. In order to compare the classification accuracy fairly, the three feature selection methods adopt same classifier—BP network. For each method, the same experiments are done for ten times, the final results are statistical average results and are involved in the Table 11. Here, SMCFS uses three evaluation criteria (i.e. f3-f2-f1).

Seen from the Table 11, the best advantage of SMCFS over the other two methods is average running time. SMCFS is faster than them very apparently. The two methods are so slow that they can not be acceptable, especially for some applications with high requirement for time cost. Besides, the number of features from SMCFS is similar as that from filter-wrapper methods. In terms of classification accuracy, the SMCFS is better than the GFSIC-SBFCV. The major reason is that first step (filter) maybe loses some important features which can not be taken back by the second step (wrapper).

3.5 Comparison of SMCFS and other multi-criteria methods

In order to show the advantage of SMCFS over other multi-criteria method, this group of experiments adopts one multi-criteria method [26]. Besides, for three criteria, there are six combinations of serial multi-criteria; they are f1-f2-f3, f1-f3-f2, f2-f1-f3, f2-f3-f1, f3-f2-f1, and f3-f1-f2. Considering the feature selection capability of three criteria, the f3-f2-f1 is SMCFS, and maybe is best. Here, f1-f2-f3 and f2-f1-f3 are adopted to compare with f3-f2-f1 (SMCFS). In order to compare the classification accuracy fairly, the three feature selection methods adopt same classifier—BP network. For each method, the same experiments are done for ten times, the final results are statistical average results and are involved in the Table 10. Here, SMCFS uses three evaluation criteria (i.e. f3-f2-f1).

Seen from the Table 12, for the three datasets, the number of features from SMCFS is less than that from EFS. The major reason is that the processing on three feature subsets with different single evaluation criterion is '∪' operation. Therefore, some less important features is possible to be involved into the final feature subset. The time cost of EFS is longer than SMCFS slightly. Theoretically, the EFS is parallel multi-criteria method and should be 3 times than filter method with single criterion, therefore it should be longer than SMCFS apparently. But the search algorithm within SMCFS is more complex than that within EFS, so advantage of the time cost of SMCFS over EFS is not too apparent. Since some less important features are involved wrongly, the average classification accuracy of SMCFS is better than that of EFS. The experimental results reflect the fact. In terms of the multi-criteria method f1-f2-f3 and f2-f1-f3, the SMCFS is better than them averagely. In terms of number of features, they are similar, SMCFS is slightly better than f1-f2-f3 and f2-f1-f3. The average running time of SMCFS is shorter than f1-f2-f3 and f2-f1-f3. The average classification accuracy of SMCFS is better than that of f1-f2-f3 and f2-f1-f3 apparently. The major reason is SMCFS adopts the best criterion as first criterion and the worst criterion as last criterion to make the beginning of searching is better and efficiency of searching for optimal feature subset is better.

3.6 Stability of classification accuracy of SMCFS

In order to show the SMCFS can obtain satisfying classification accuracy stably, same experiments are done for ten times. Figure 4 shows the classification accuracy from five feature selection methods during ten times experiments respectively. The five feature selection methods are wrapper method, SMCFS, f1, f2 and f3 respectively. All of them use BP network as classifier.

Seen from the figure, some things can be found. Firstly, the SMCFS is very stable. The classification accuracy from SMCFS is similar. However, the f1, f2 and f3 are not stable, the classification accuracy from them change a lot with different experiments. Secondly, the classification accuracy from SMCFS is close to wrapper method, it means the classification accuracy from SMCFS is satisfying. But the classification accuracy from f1, f2 and f3 is worse than wrapper method apparently.

4 Conclusions

From the theoretical analysis of SMCFS and the experimental results, it can be seen that the SMCFS can have similar time cost as and higher classification accuracy than those of the filter method with single evaluation criterion. Although the wrapper method has better precision than SMCFS, the high time cost of wrapper method makes it unacceptable for some applications with high requirement for time cost. The filter-wrapper method is one kind of eclectic method between filter method and wrapper method, but as shown in experimental result, it still needs lots of time cost and

Table 12 Comparison of SMCFS and other multi-criteria methods

DS	CP	SMCFS (f3-f2-f1)	EFS	Random Serial multi-criteria method	
				(f1-f2-f3)	(f2-f1-f3)
Letter	ANF	9.5	12.5	10	10.5
	ART	5.9341	6.4523	6.7892	7.6547
	ACA	0.9799	0.9367	0.9515	0.9573
Wave	ANF	13	17	15.6	14.5
	ART	20.1118	22.2167	21.8321	23.6754
	ACA	0.9098	0.8197	0.8314	0.8371
Sonar	ANF	24	31	24	24
	ART	27.7675	22.8432	31.7653	30.9856
	ACA	0.9741	0.9142	0.9271	0.9243

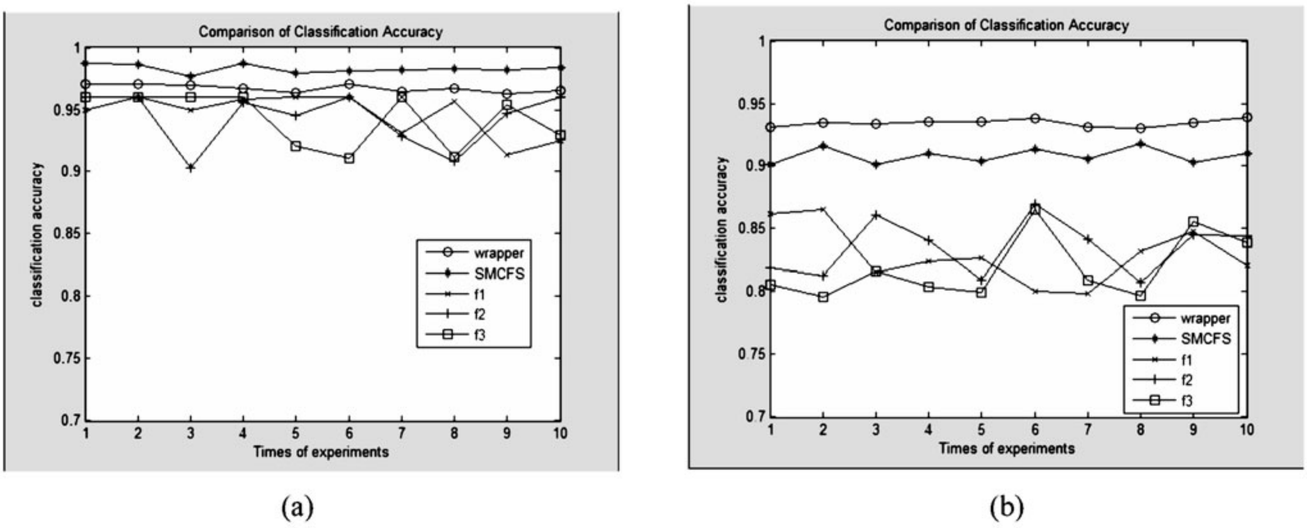

Fig. 4 Stability of classification accuracy of SMCFS: (**a**) for letter dataset; (**b**) for wave dataset

has similar time cost as SMCFS. The parallel multi-criteria method and random multi-criteria method is worse than SMCFS from the experimental data.

Besides, except the three datasets, some other datasets with more than 800 features are considered for comparison. The experimental results show that the precision of SMCFS is similar as filter method with single evaluation criterion. The average running time of SMCFS is longer than filter method with single evaluation criterion. The reason is due to that for the datasets with too much features, the difference between SMCFS and filter method with single evaluation criterion is not apparent. With increasing of features, the time cost for feature selection increases dramatically. SMCFS is not suitable for those datasets.

Among the current feature selection problems, some of them need low time cost and high precision and have middle size of features, such as feature selection problems in biomedical image classification, fault detection and so on. Their common characteristics are as follows: firstly, in order to deal with dynamic specimens, feature selection algorithm should be dynamic, and the optimal feature subset changes with change of specimens, so the time cost should be low. Secondly, in order to classify the specimens correctly, the feature selection algorithm needs to be precise. Thirdly, the size of feature selection problems is not too big, the usual domain is [10, 150]. The SMCFS is most suitable for this kind of applications. Fourthly, the relationship between features is very complex, the correlation between features should be considered. According to these applications with high precision and low time cost, this paper proposes one new multi-criteria feature selection method-sequential multi-criteria feature selection algorithm (SMCFS). By combining the consistency and otherness of different evaluation criteria, the SMCFS adopts more than one evaluation criteria sequentially to enhance the efficiency of feature selection. Based on one new agent genetic algorithm (CAGA), the SMCFS can obtain high precision of feature selection and low time cost which is close to filter method with single evaluation criterion. Several groups of experiments are done to demonstrate the performance of the SMCFS. According to three datasets from UCI database, SMCFS is compared with different kinds of feature selection methods. The experimental results show that the SMCFS can have low time cost and high precision of feature selection, and is very suitable for some applications of feature selection with high requirement for time cost and precision.

In the future, more works will be done as following: Firstly, we will consider more than three evaluation criteria. Secondly, the feature selection algorithm will be incorporated into classification system for real applications.

Acknowledgements We would like to express our sincere thanks for the valuable suggestions and comments made by the reviewers and the editor. This research is funded by NSFC (National Nature Science Fund of China, No: 30570473), Chongqing Natural Science Foundation (No. is CSTC, 2008BB2164, CSTC, 2008BB2322) and Innovation Ability Training Foundation of Chongqing University (CDCX018).

References

1. De Stefano C, Fontanella F, Marrocco C (2008) A GA-based feature selection algorithm for remote sensing images. In: EvoWorkshops 2008. LNCS, vol 4974. Springer, Berlin, pp 285–294
2. Suzuki E (2005) Worst case and a distribution-based case analysis of sampling for rule discovery based on generality and accuracy. Appl Intell 22(1):29–36
3. Eriksson T, Kim S, Kang H-G, Lee C (2005) An information-theoretic perspective on feature selection in speaker recognition. IEEE Signal Process Lett 12(7):500–503

4. Cho BH, Yu H, Kim K-W, Kim TH, Kim IY, Kim SI (2008) Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artif Intell Med 42(1):37–53
5. Yan J, Zhang B, Liu N, Yan S, Cheng Q, Fan W, Yang Q, Xi W, Chen Z (2006) Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing. IEEE Trans Knowl Data Eng 18(3):320–333
6. Su C-T, Yang C-H (2008) Feature selection for the SVM: an application to hypertension diagnosis. Expert Syst Appl 34(1):754–763
7. Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17(4):491–502
8. Oh I-S, Lee J-S, Moon B-R (2004) Hybrid genetic algorithms for feature selection. IEEE Trans Pattern Anal Mach Intell 26(11):1424–1437
9. Huang C-L, Wang C-J (2006) A GA-based feature selection and parameters optimization for support vector machines. Expert Syst Appl 31:231–240
10. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1):273–324
11. Zhu Z, Ong Y-S, Dash M (2007) Wrapper–filter feature selection algorithm using a memetic framework. IEEE Trans Syst Man Cybern—Part B: Cybern 37(1):70–76
12. Wang L (2008) Feature selection with kernel class separability. IEEE Trans Pattern Anal Mach Intell 30(9):1534–1546
13. Gunal S, Edizkan R (2008) Subspace based feature selection for pattern recognition. Inf Sci 178(19):3716–3726
14. Chen L-H, Hsiao H-D (2008) Feature selection to diagnose a business crisis by using a real GA-based support vector machine: an empirical study. Expert Syst Appl 35(3):1145–1155
15. Jiang Z, Yamauchi K, Yoshioka K et al (2006) Support vector machine-based feature selection for classification of liver fibrosis grade in chronic hepatitis C. J Med Syst 30(5):34–45
16. Pizzi NJ, Pedrycz W (2008) Effective classification using feature selection and fuzzy integration. Fuzzy Sets Syst 159(21):2859–2872
17. Yang Y, He GWK (2007) An approach for selective ensemble feature selection based on rough set theory. In: RSKT2007. LNAI, vol 4481. Springer, Berlin, pp 518–525
18. Yuan H, Tseng S-S, Gangshan W, Fuyan Z (1999) A two-phase feature selection method using both filter and wrapper. In: IEEE SMC'99 conference. IEEE Press, New York, pp 132–136
19. Li Y, Lu B-L, Wu Z-F (2006) A hybrid method of unsupervised feature selection based on ranking. In: The 18th international conference on pattern recognition (ICPR'06), vol 2. IEEE Press, New York, pp 687–690
20. Huang J, Cai Y, Xu X (2007) A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern Recognit Lett 28(13):1825–1844
21. Banerjee M, Mitra S, Banka H (2007) Evolutionary-rough feature selection in gene expression data. IEEE Trans Syst Man Cybern Part C: Appl Rev 37(4):622–632
22. Mitra S, Banka H (2006) Multi-objective evolutionary biclustering of gene expression data. Pattern Recognit 39(12):2464–2477
23. Wing WY, Yeung DS, Firth M et al (2008) Feature selection using localized generalization error for supervised classification problems using RBFNN. Pattern Recognit 41(12):3706–3719
24. Huang C-L, Tsai C-Y (2007) A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. Expert Syst Appl, available online 15 December 2007
25. Blazadonakis ME, Zervakis M (2008) Wrapper filtering criteria via linear neuron and kernel approaches. Comput Biol Med, available online 24 July 2008
26. Doan S, Horiguchi S (2004) An efficient feature selection using multi-criteria in text categorization. In: Proceedings of the fourth international conference on hybrid intelligent systems (HIS'04)
27. 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(8):1226–1238
28. Emmanouilidis C, Hunter A, Machtyre J, Cox C (1999) Multiple-criteria genetic algorithms for feature selection in neuro-fuzzy modeling. In: Proceeding of international joint conference on neural networks, IJCNN '99, 10–16 July 1999, vol 6, pp 4387–4392
29. Auarth B, L'opez M, Cerquides J (2008) Hopfield networks in relevance and redundancy feature selection applied to classification of biomedical high-resolution micro-CT images. In: ICDM 2008. LNAI, vol 5077. Springer, Berlin, pp 16–31
30. Zeng X-P, Li Y-M, Qin J (2008) A dynamic chain-like agent genetic algorithm for global numerical optimization and feature selection. Neurocomputing. doi:10.1016/j.neucom.2008.02.010. Available online at www.sciencedirect.com
31. Srinivas M, Patnaik LM (1994) Adaptive probabilities of crossover and mutation in genetic algorithms. IEEE Trans Syst Man Cybern 24(4):656–659
32. Zhong W, Liu J, Xue M, Jiao L (2004) A multiagent genetic algorithm for global numerical optimization. IEEE Trans Syst Man Cybern—Part B: Cybern 34(2):1128–1141
33. Gong D-W, Sun X-Y, Guo X-J (2002) Novel survival of the fittest genetic algorithm. Control Dec 17(6):908–911
34. Michalewicz Z, Fogel DB (2000) How to solve it: modern heuristics. Springer, Berlin, pp 83–234

Yongming Li received the bachelor's degree from University of electronic science and technology of china in 1999, the M.S. degrees from the Chongqing University in 2003, and Ph.D. doctor degree in Chongqing University in 2007 respectively. Li works in the college of Communication engineering at Chongqing University (CQU) where he researches intelligent computing, pattern recognition, image processing, data analysis and their applications to real-world problems. Until now, he has published over thirty technical papers on relevant field. He is member of several academic leagues and reviewer over ten journals and international conferences.

Xiaoping Zeng received the bachelor's degree, the M.S. degrees and Ph.D. from the Chongqing University. Professor Zeng works in the college of Communication engineering at Chongqing University (CQU) where he researches signal processing and their applications to real-world problems. Until now, he has published more than seventy technical papers on relevant field.