



# A multiple criteria ensemble pruning method for binary classification based on D-S theory of evidence

Jing Qiu<sup>1</sup> · Zhi Xiao<sup>1</sup> · Bo Zhong<sup>2</sup>

Received: 25 April 2022 / Accepted: 11 October 2022 / Published online: 2 November 2022  
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

## Abstract

Ensemble pruning becomes an important stage in multiple classifier systems, and it has been widely applied to solve binary classification problems. Diversity and performance measures are two widely used evaluation methods to build the selection criterion for ensemble pruning. However, few works consider both of them simultaneously, and they usually use one algorithm to measure the diversity or performance, which may not be enough to capture all the relevant diversities and performance of the base classifiers. To solve this problem, we propose a multiple criteria ensemble pruning method by employing multiple diversity and performance measures to capture the base classifiers' diversity and evaluate their classification ability respectively. Moreover, a multi-criteria decision making method, based on fuzzy soft set and Dempster-Shafer theory of evidence, is used to build the final selection criterion, which can make a good trade-off between the diversity and performance measures. With sixteen binary data sets, the experimental studies show its effectivity and superiority for ensemble pruning over six state-of-the-art benchmark methods.

**Keywords** Ensemble pruning · Multi-criteria decision making · D-S theory of evidence · Binary classification · Diversity and performance measures

## 1 Introduction

Ensemble learning has attracted more and more attention from researchers recently, because it trains multiple weak learners to solve the same problem and achieves stability and accuracy [1]. With the development of Big Data, machine learning methods with increasing data size are usually employed to build the ensemble models, which would bring heavy computational burdens for the ensemble learning [2]. To decrease the computational overheads and improve the accuracy of ensemble methods, ensemble pruning (also known as ensemble selection or selective ensemble) provides a new perspective [3]. And it has been widely used to solve the binary classification problems [4–6]. The goal of ensemble pruning is to select a subset of base classifiers,

which have been shown to perform better in both complexity and accuracy than the full ensemble [7]. Searching for the best subset of an ensemble using an exhaustive method is an NP-complete problem, which is not suitable for ensemble pruning [8]. Therefore, a lot of studies have been done to solve this problem. Among them, we find that the selection criterion and selection or searching method are the two key points for ensemble pruning. Tsoumakas et al. presented a taxonomy of ensemble pruning methods, and concluded that ensemble pruning evaluation measures consists of two major categories: diversity based and performance based [9].

The basic idea of ensemble learning is to combine various classifiers, which may offer complementary information about the patterns to be classified [10]. Thus, the diversity among the base classifiers has been widely applied to build the selection criterion, and it has been proved to have a positive correlation to the ensemble's accuracy [11]. Many diversity measures have been investigated to build the selection criterion, such as Q statistic, the correlation, the disagreement, the double fault, etc. The choice of diversity measures will directly influence the final ensemble [12]. However, Kuncheva et al. found that diversity is not always beneficial, and sometimes it may work toward deterioration of the

✉ Zhi Xiao  
xiaozhicqu@163.com

<sup>1</sup> School of Economics and Business Administration, Chongqing University, Chongqing 400044, People's Republic of China

<sup>2</sup> College of Mathematics and Statistics, Chongqing University, Chongqing 400044, People's Republic of China

ensemble's performance [13]. In their another work, they compared ten diversity measures, and found that the motivation for designing diverse classifiers is correct, but how to choose diversity measures and use them effectively for ensemble pruning is still an open problem [12]. Motivated by that, Cavalcanti et al. found that one diversity measure is not enough to capture all the diversities of ensemble, and combined five diversity measures for ensemble pruning [14].

Besides, accuracy is another important factor in building the selection criterion, because it is the ultimate pursuit of ensemble methods [15]. Santos et al. made a further investigation of GA-based selection method to get the subset of the ensemble with the highest accuracy, and implied that controlling overfitting is an important task [16]. Zhang et al. designed an accuracy guided ensemble pruning strategy for pruning the component classifiers with low complementarity [17]. However, accuracy is unable to capture all the different factors that characterize the performance of a classifier. To evaluate the competence of base classifiers, some other performance measures are also considered, such as Brier score (BS), the area under the receiver operating characteristic (ROC) curve (AUC), etc. The BS measures the classifier's degree of deviation from the true label, which is more accurate to measure the classification ability of the classifiers [18]. The AUC is commonly regarded as a more suitable measurement for evaluating the classifier's overall performances than accuracy [19, 20].

All of the above methods only use diversity or accuracy to build the selection criteria for ensemble pruning. Even though Dai et al. considered both of them for ensemble pruning, they employed only one algorithm to measure the diversity [21], which may be insufficient to capture all the relevant diversities of the base classifiers [14]. Moreover, Bian et al. only employed the mutual information to measure the accuracy, and the normalized variation to measure the diversity to balance the accuracy and diversity [22]. Using a single algorithm is insufficient to measure the classification ability of the base classifiers. Therefore, it is necessary to adopt multiple diversity and performance measures to build the selection criterion simultaneously.

Moreover, Hashemi et al. considered the ensemble feature selection as a Multi-Criteria Decision-Making (MCDM) process [23]. Kou et al. employed MCDM methods for evaluating classification algorithms and suggested that MCDM methods are feasible tools for ensemble pruning [24]. However, they pay less attention to the diversity and performance measures, and do not compare the MCDM methods to other ranking-based selection methods. Motivated by its preferable performance on decision making, we carry out research on considering the ensemble pruning as a MCDM problem based on multiple diversity and performance measures. In addition, Dempster-Shafer (D-S) theory of evidence is a useful tool for combining accumulative evidence of changing

prior opinions [25], and solving the MCDM problems effectively [26]. Meanwhile, fuzzy soft set (FSS), proposed by Maji et al. has also been proved to handle MCDM problems effectively [27]. It provides a theoretic framework to arrange multiple diversity and performance measures of the base classifiers [28]. Therefore, we build a multiple criteria ensemble pruning method by using a MCDM method with FSS and D-S theory of evidence in this paper.

In this study, we use multiple diversity and performance measures to evaluate the competence level of the base classifiers, and then use the fuzzy soft set to organize them. Secondly, we use the D-S theory of evidence to obtain a final score for each base classifier, and then rank the classifiers according to their scores. Finally, we use the greedy ensemble pruning strategy with forward expansion to select the final ensemble from the ordered classifiers. The main contributions of this work are as follows:

- The ensemble pruning method is proposed by considering multiple diversity and performance measures simultaneously, which can evaluate the base classifiers sufficiently.
- The use of a MCDM method for ensemble pruning can make a good trade-off between the diversity and performance measures.
- The D-S theory of evidence is applied to build the combinational rule of the MCDM method. The fuzzy soft set is used to arrange the multiple diversity and performance measures, which help to build the decision matrix.

The rest of this paper is organized as follows. Section 2 gives a preview of the multi-criteria decision making method based on D-S theory of evidence and fuzzy soft set. Section 3 reviews some definitions of diversity and performance measures. Section 4 introduces the research process of the proposed multiple criteria ensemble pruning method. Section 5 displays a detailed description of the experimental setup and Sect. 6 gives a full discussion of the experimental results. Finally, conclusions and avenues for future research are presented in Sect. 7.

## 2 Multi-criteria decision making based on D-S theory of evidence and fuzzy soft set

### 2.1 Fuzzy soft set

Suppose that  $U$  is a universe set, and  $E$  is the parameters set. Then, let  $I^U$  denote the power set of all fuzzy sets of  $U$ . Let  $A$  be a subset of  $E$ , i.e.,  $A \subset E$ . A pair  $(F, A)$  is regarded as a fuzzy soft set (FSS in short) over  $U$ , where  $F$  is a mapping showed by:

$$F : A \rightarrow I^U.$$

The  $F$  is the fuzzy approximate function of the  $(F, A)$ , and it is usually defined according to the characteristics of the problem [28]. For  $e \in A$ ,  $F(e)$  is a fuzzy subset of  $U$ , which is a fuzzy number set regarding to parameter- $e$ . We can show  $F(e)$  with the form of fuzzy set,  $(F, A) = \{(e, F(e)) : e \in A, F(e) \in I^U\}$ .

### 2.2 D-S theory of evidence

Dempster-Shafer (D-S) theory of evidence is originated by the idea of upper probability and lower probability introduced by Dempster [29] and developed by Shafer [30]. It makes the measurement by using belief function and plausible reasoning, which is distinct from other aggregators. The D-S theory of evidence can express the random uncertainty information, the incomplete information and the subjective uncertainty information directly without any prior probability [4], which is similar to the way human collect evidences. Therefore, it has been regard as an increasingly important tool for decision making, because of its advantage in aggregating the information. Some basic concepts of D-S theory of evidence will be recalled in this subsection.

Supposed that  $\Theta$  is a nonempty set of alternatives, which is regarded as a frame of discernment. Let  $A$  be one proposition over the power set  $2^\Theta$ , the basic belief assignment (BBA) is a mapping [30], presented by  $m : 2^\Theta \rightarrow [0, 1]$ . It is a mass function, and  $m$  satisfies:

$$m(\emptyset) = 0, \sum_{A \subseteq \Theta} m(A) = 1.$$

The  $m(A)$  means the evidence which is in support of the proposition  $A$ .

On the frame of discernment  $\Theta$ , the belief function ( $Bel$ ) and plausibility function ( $Pls$ ) are respectively [30] defined as:

$$Bel(A) = \sum_{B \subseteq A} m(B), A \subseteq \Theta,$$

$$Pls : 2^\Theta \rightarrow [0, 1], Pls(A) = \sum_{B \cap A \neq \emptyset} m(B).$$

The  $Bel$  means decision maker’s the full confidence in  $A$ , and  $Pls(A)$  means the extent of no objection to  $A$ .

Besides, the D-S theory of evidence provides a way to combine bodies of evidence. Supposed that  $A_1, A_2, \dots, A_m$  and  $B_1, B_2, \dots, B_n$  are evidences on the frame of discernment  $\Theta$ . The corresponding basic probability assignment functions are  $m_1$  and  $m_2$  respectively. If  $\sum_{A_i \cap B_j = \emptyset} m_1(A_i)m_2(A_j) < 1$ , the evidence combinational rule is given by:

$$m(A) = m_1 \oplus m_2(A) = \frac{1}{1 - K} \sum_{A_i \cap B_j = A} m_1(A_i)m_2(B_j), \forall A \subseteq \Theta, A \neq \emptyset, \tag{1}$$

$$m(A) = m_1 \oplus m_2(A) = 0, A = \emptyset.$$

The  $K = \sum_{A_i \cap B_j} m_1(A_i)m_2(B_j)$ .  $K$  is regarded as the conflict probability, and it indicates the degree of the conflict between the evidences. Coefficient  $\frac{1}{1-K}$  is regarded as normalized factor, and its role is to avoid the probability of assigning non-0 to empty set  $\emptyset$  in the combination [30].

### 2.3 Multi-criteria decision making method

In general, a MCDM method usually consists two stages [31]: (1) information input and construction; (2) aggregation and exploitation. In the first stage, a decision matrix is usually built as the mathematical expression of the MCDM problem, which is defined as follows [32].

Given a MCDM problem with  $n$  distinct alternatives  $A_1, \dots, A_n$  and  $m$  criteria  $Cr_1, \dots, Cr_m$ , where the level of achievement of  $A_i (i = 1, \dots, n)$  with regard to  $Cr_k (k = 1, \dots, m)$  is denoted by  $Cr_k(A_i)$ :

$$\begin{bmatrix} Cr_1(A_1) & Cr_2(A_1) & \dots & Cr_m(A_1) \\ Cr_1(A_2) & Cr_2(A_2) & \dots & Cr_m(A_2) \\ \vdots & \vdots & \ddots & \vdots \\ Cr_1(A_n) & Cr_2(A_n) & \dots & Cr_m(A_n) \end{bmatrix},$$

which is called the decision matrix of the problem.

In the second stage, the aggregation and exploitation are carried out based on the decision matrix, which are used for combining the criteria, such as an utility function  $u(A_i) = \sum_{k=1}^m w_k Cr_k(A_i)$ . From above, we can find that the decision matrix is similar to FSS and D-S theory of evidence provides a novel aggregation rule. Thus, we use the FSS to build the decision matrix and D-S theory of evidence to build the combine rule, which is inspired by Xiao’s work [33].

## 3 Diversity and performance measures

### 3.1 Diversity measures

Even though diversity is widely applied for ensemble pruning, it does not have a generally accepted formal definition. Many diversity measures are presented in previous literatures, and summarized into pairwise measures and non-pairwise measures [12]. Meanwhile, the definition of a “criterion” is “a means or standard of judging” by which one particular choice or course of action could be judged to be more desirable than another in the area of MCDM

**Table 1** Contingency table for the relationship between a pair of classifiers

	$c_i = 1$	$c_i = 0$
$c_j = 1$	$N^{11}$	$N^{10}$
$c_j = 0$	$N^{01}$	$N^{00}$

[34]. Thus, we employ multiple well-known pairwise diversity measures to calculate the classifiers’ diversity.

Suppose a set of  $s$  trained classifiers  $C = \{c_1, \dots, c_s\}$  and  $N$  samples, we let  $c_i = 1$ , if  $c_i$  predicts the sample correctly, and 0, otherwise,  $i = 1, \dots, s$ . Then, we can get a contingency table (Table 1).  $N^{11}$  indicates the number of samples which are correctly classified by both  $c_i$  and  $c_j$ .  $N^{10}$  indicates the number of samples which are correctly classified by  $c_j$  and incorrectly classified by  $c_i$ .  $N^{01}$  indicates the number of samples which are correctly classified by  $c_i$  and incorrectly classified by  $c_j$ .  $N^{00}$  indicates the number of samples which are incorrectly classified by both  $c_i$  and  $c_j$ ,  $i, j = 1, \dots, s$ . With Table 1, we can get following diversity measures [14]:

*Disagreement measure (Dis)*: Its value is calculated by Eq. (2), and the high value of  $dis_{i,j}$  indicates high diversity between  $c_i$  and  $c_j$ .

$$Dis_{i,j} = \frac{N^{01} + N^{10}}{N^{11} + N^{00} + N^{01} + N^{10}} \tag{2}$$

*The Q statistics (Q)*: It is a statistic to assess the similarity of two classifiers, and it ranges from -1 to 1, where the positive values indicate similar predictions and the negative values indicate different predictions made by them. The *Q statistic* of  $c_i$  and  $c_j$  is calculated as follows:

$$Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{3}$$

*The Kappa-statistic (Kappa)*: It is widely used in statistics and also applied to measure the diversity. The value is 1, if the classifiers completely agree. Moreover, it is 0, if they randomly agree. Less than 0 is a rare case. The *Kappa-statistic* of  $c_i$  and  $c_j$  is calculated as follows:

$$Kappa_{i,j} = \frac{\Theta_1 - \Theta_2}{1 - \Theta_2}, \tag{4}$$

where

$$\Theta_1 = \frac{N^{11} + N^{00}}{N^{11} + N^{00} + N^{01} + N^{10}},$$

$$\Theta_2 = \frac{(N^{11} + N^{10})(N^{11} + N^{01}) + (N^{00} + N^{10})(N^{00} + N^{01})}{(N^{11} + N^{00} + N^{01} + N^{10})^2}$$

*The double-fault measure (DF)*: It indicates the proportion of the samples that have been incorrectly predicted by both classifiers, showed as follows:

$$DF_{i,j} = \frac{N^{00}}{N^{11} + N^{00} + N^{01} + N^{10}} \tag{5}$$

### 3.2 Performance measures

As mentioned in the introduction, there are three types of performance measures: some assess the discriminatory ability of the classifiers, some assess the accuracy of the classifiers’ probability predictions, and some assess the correctness of the classifiers’ categorical predictions [18, 35]. In order to evaluate the classifiers’ performance in all directions, we select at least one measure for each type of them. Meanwhile, to keep a balance between diversity and performance measures, we select four performance measures. Considering the set  $C = \{c_1, \dots, c_s\}$  and  $N$  samples, we can get the performance measures as follows.

*Accuracy (ACC)*: It is the most commonly used measure to evaluate the classifier and build the ensemble pruning method. The *ACC* is usually measured using the percentage of correctly classified observations. It is gained upon a confusion matrix for binary classification presented in Table 2. The *ACC* of  $c_i$  is calculated as follows [18]:

$$ACC_i = \frac{TN + TP}{TN + FP + FN + TP} \tag{6}$$

*Brier score (BS)*: It measures the classifier’s degree of deviation from the true probability. For binary classification, if  $p_j$  indicates the estimated probability of  $j$ th sample to the negative class, and  $f_j$  indicates its true probability of negative, then the *BS* for  $c_i$  is calculated as follows [18]:

$$BS_i = \frac{\sum_{i=1}^N (f_i - p_i)^2}{N} \tag{7}$$

*The area under the receiver operating characteristic (ROC) curve (AUC)*: The ROC shows the tradeoff between TP rate and FP rate. The larger the *AUC*, the better the classifier. A simple method to calculate the *AUC* of a classifier is showed as follows [36]:

$$AUC = \frac{\sum_{i=1}^{TP+FN} rank_i - \frac{(TP+FN) \times (TP+FN+1)}{2}}{(TP + FN) \times (TN + FP)}, \tag{8}$$

**Table 2** Confusion matrix

Actual class	Predicted result	
	Negative	Positive
Negative	True negative (TN)	False positive (FP)
Positive	False negative (FN)	True positive (TP)

where  $rank_i$  indicates the rank of  $i^{th}$  positive sample in the ordered list, which is usually built based on the predicted scores or posterior probability of the positive samples.

*H-measure (H)*: It gives a normalized classifier assessment based on expected minimum misclassification loss [36]. Its cost ratio distribution is given based on a beta distribution and equal parameter values, which use different distributions to different classifiers different from AUC. Thus, it improves the AUC.

### 4 The proposed method: DSFSS-P

In this section, a detailed description of the proposed method is presented. The base classifiers are generated based on heterogeneous binary classification algorithms, which are given in Sect. 5.2. To prune the best subset of the ensemble, a multiple criteria ensemble pruning method is proposed by combining multiple diversity and performance measures of the base classifiers. Meanwhile, we solve this problem from the perspective of MCDM with FSS and D-S theory of evidence. In general, with the base classifiers' diversity and performance measures on validation set, the FSS is built, which is a decision matrix of the MCDM problem. Then the D-S theory of evidence is used to fuse the multiple criteria and get the selection criterion for each base classifiers. Finally, we apply the greedy selection strategy with forward expansion to search the best ensemble. We call the proposed method DSFSS-P for short, and the research process of it can be divided into three phases shown in Fig. 1.

- (a) In phase 1, the  $s$  base classifiers are trained based on training set, and the eight diversity and performance measures, denoted by  $E = \{e_1, e_2, \dots, e_8\}$ , are used to evaluate them based on the validation set. The evaluation results are organized by the fuzzy soft set, which is the decision matrix of the MCDM problem, denoted

$$\text{by } (F, E) = \left\{ \begin{matrix} F_{e_1}(c_1) & F_{e_2}(c_1) & \dots & F_{e_8}(c_1) \\ F_{e_1}(c_2) & F_{e_2}(c_2) & \dots & F_{e_8}(c_2) \\ \dots & \dots & \dots & \dots \\ F_{e_1}(c_s) & F_{e_2}(c_s) & \dots & F_{e_8}(c_s) \end{matrix} \right\}.$$

- (b) In phase 2, the basic probability assignment functions are built to compute the basic probability of the classifiers, and the evidences from the eight measures are combined with D-S theory of evidence to get the ordered classifiers  $C' = \{c'_1, c'_2, \dots, c'_s\}$ .
- (c) In phase 3, the greedy selection strategy with forward expansion is used to search the top  $t$  classifiers with the best ensemble classification accuracy.

### 4.1 Build the decision matrix using multiple diversity and performance measures

One diversity measure is not enough to capture all the diversities of ensemble, and accuracy is not the most effective measure to evaluate classifier's overall performances, as introduced in Sect. 1. Thus using one single criterion to select an ensemble classifier is biased [37]. Meanwhile, diversity and performance measures are two key points to build the selection criterion. It would be comprehensive and reasonable to consider multiple diversity and performance measures. In this work, four diversity measures and four performance measures are selected to estimate the base classifiers' performance showed in Sect. 3. To select the best ensemble, we solve the problem from the perspective of the MCDM, and the FSS is used to build the decision matrix based on the classifiers' performance. So the base classifiers are the alternatives of the MCDM problem. Meanwhile, the diversity and performance measures are selection criteria. Then we can get the universe  $C = \{c_1, \dots, c_s\}$ , and the set of parameters  $E = \{e_1, e_2, \dots, e_8\}$ , where  $e_1 = Dis$ ,  $e_2 = Q$ ,  $e_3 = Kappa$ ,  $e_4 = DF$ ,  $e_5 = ACC$ ,  $e_6 = BS$ ,  $e_7 = AUC$ , and  $e_8 = H$ . Then a pair  $(F, E) = \{(e, F(e)) : e \in E, F(e) \in I^C\}$  is a FSS, and for each  $e_j \in E$  ( $j = 1, 2, \dots, 8$ ), we can get  $F(e_j) = \{(c, F_{e_j}(c)) : c \in C, F_{e_j}(c) \in [0, 1]\}$ , which is the mapping under the parameter  $e_j$ .

With the purpose of building the decision matrix based on the FSS, constructing the fuzzy set corresponding to each parameter are required primarily. The crucial step is building the fuzzy approximate function  $F$ . However, the data dimension and meaning of it corresponding to the parameters are diverse, where some drops into the scope of  $[-1, 1]$ , and some drops into the scope of  $[0, 1]$ . Therefore, the functions should be built adaptively.

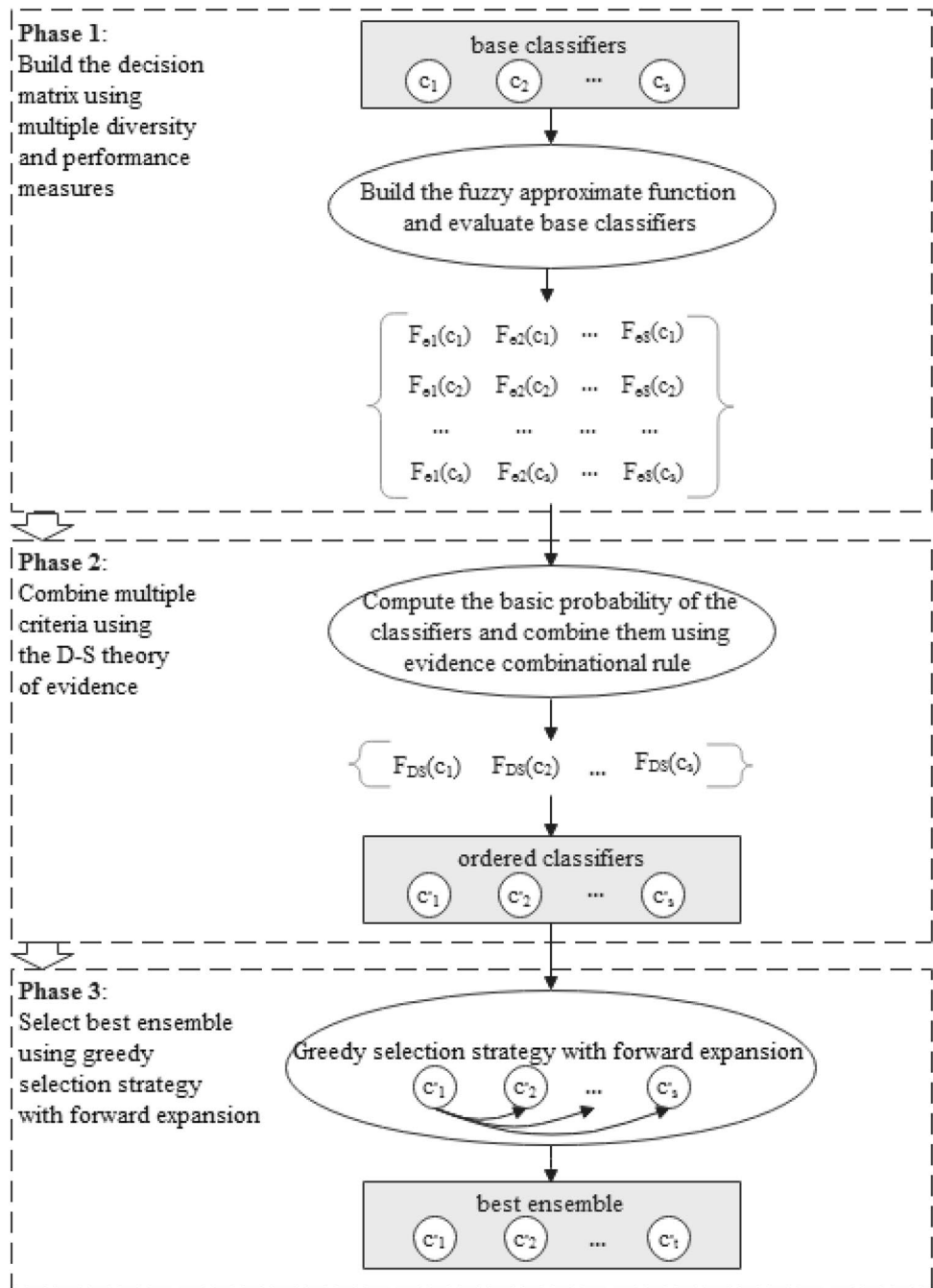
Suppose  $s$  base classifiers  $C = \{c_1, \dots, c_s\}$ ,  $c_i(e_j)$  indicates the value of  $c_i$  corresponding to the parameter  $e_j$  based on validation set,  $i = 1, \dots, s, j = 1, \dots, 8$ . For  $e_5, e_6, e_7, e_8$ , the  $c_i(e_j)$  is computed based on their definitions. For  $e_1, e_2, e_3, e_4$ , the  $c_i(e_j)$  is computed by Eq. (9):

$$c_i(e_j) = \sum_{k=1}^s \text{diversity}_{ik}(e_j) \tag{9}$$

The  $\text{diversity}_{ik}(e_j)$  indicates the pairwise diversity of  $c_i$  and  $c_k$  under parameter  $e_j$ , which is calculated by using Eqs. (2, 3, 4, 5) respectively,  $i, k = 1, \dots, s, j = 1, 2, 3, 4$ .

The high values of *disagreement measure* of them indicate high diversity. Meanwhile, the high values of the *ACC*, *AUC*, *H* indicate good performance of the classifiers. Therefore, for parameters  $e_1, e_5, e_7, e_8$ , the fuzzy approximate function  $F(e)$  is built as follows:

**Fig. 1** Overview of the proposed DSFSS-P method



$$F_{e_j}(c_i) = \frac{c_i(e_j) - c_{\min}(e_j)}{c_{\max}(e_j) - c_{\min}(e_j)} \tag{10}$$

The  $c_{\min}(e_j)$  and  $c_{\max}(e_j)$  indicates the minimal and maximal value of the  $s$  base classifiers corresponding to the parameter  $e_j$  respectively, and  $F_{e_j}(c_i)$  indicates the fuzzy value or membership of  $c_i$  corresponding to the parameter  $e_j$ .

However, the high values of the  $Q$  statistics,  $Kappa$ -statistic and  $double$ -fault measure of them indicate low diversity, and high value of  $Brier$  score indicates high posterior

probability error. Therefore, for parameters  $e_2, e_3, e_4, e_6$ , the fuzzy approximate function  $F(e)$  is built as follows:

$$F_{e_j}(c_i) = \frac{c_{\max}(e_j) - c_i(e_j)}{c_{\max}(e_j) - c_{\min}(e_j)} \tag{11}$$

Therefore, the FSS  $(F, E)$  over  $C$  is presented by Table 3, which is the decision matrix.

**Table 3** Matrix representation of fuzzy soft set  $(F, E)$  over  $C$

$C$	$e_1$	$e_2$	...	$e_8$
$c_1$	$F_{e_1}(c_1)$	$F_{e_2}(c_1)$	...	$F_{e_8}(c_1)$
$c_2$	$F_{e_1}(c_2)$	$F_{e_2}(c_2)$	...	$F_{e_8}(c_2)$
...	...	...	...	...
$c_s$	$F_{e_1}(c_s)$	$F_{e_2}(c_s)$	...	$F_{e_8}(c_s)$

### 4.2 Combine multiple criteria using the D-S theory of evidence

In the decision matrix, the attributes or criteria are sometimes inconsistent because of the fuzzy relations of the diversity and the accuracy [20, 38]. The D-S theory of evidence can deal with imprecise and uncertain information without prior information, and sometimes can handle the conflicts among evidences [26]. Thus the combinational rule of D-S theory of evidence is an effective tool to fuse the multiple criteria based on the diversity and performance measures.

According to the definition of D-S theory of evidence and the decision matrix  $(F, E)$ , the frame of discernment of best classifiers is denoted by  $\Theta=C=\{c_1, \dots, c_s\}$ , and  $E = \{e_1, e_2, \dots, e_8\}$  are the eight performance measures to provide the evidences denoted by  $(F, E) = \{(e, F(e)) : e \in E, F(e) \in I^C\}$ . In order to combine the multiple measures by the D-S theory of evidence based combinational rule, the study defines the basic probability assignment functions based on Gong and Hua’s work [39] of basic utility assignments:

$$m(F_{e_j}(c_i)) = \frac{F_{e_j}(c_i)}{\sum_{i=1}^s F_{e_j}(c_i) + \beta}, \tag{12}$$

where  $\beta$  ranges in  $[0, 1]$  and usually takes 1 [40].  $m(F_{e_j}(c_i))$  indicates basic probability assignment function for degree of membership of  $c_i(i = 1, 2, \dots, s)$  respectively under  $e_j(j = 1, 2, \dots, 8)$ . From the perspective of D-S theory of evidence, it indicates the BPA of the singleton set  $\{c_i\}, i = 1, 2, \dots, s$ , and  $\beta$  is used to represent the comprehensive evidence of the other subsets of the power set of  $\Theta$ . Since searching for the best subset from the  $\Theta$  using an exhaustive method is an NP-complete problem, it is hard to derive the evidences of all the elements of the power set of  $\Theta$ . Besides, the purpose of this section is to select the best classifier under the  $\Theta$ . Thus the BPAs of the singleton sets are mainly built.

In order to fuse the BPAs, the Eq. (8) is usually used to combine  $m(F_{e_j}(c_i))$ . The result can be regarded as a fuzzy set  $C'' = \{c, F_{DS}(c) : c \in C, F_{DS}(c) \in [0, 1]\}$ , where  $F_{DS}(c)$  is a fuzzy value. It is the final score of classifier  $c$ , indicating

the membership of it belonging to best alternative of the ensemble. The algorithm of combing multiple criteria using D-S theory of evidence to get  $F_{DS}(c)$  is showed detailedly in Fig. 2. With the fuzzy set  $C''$ , we rank the classifiers by the descending order of the  $F_{DS}(c)$ , and get the ordered set  $C' = \{c'_1, c'_2, \dots, c'_s\}$ .

### 4.3 Select best ensemble using greedy selection strategy with forward expansion

Greedy selection strategy is a commonly used method to search for the optimal choice of ensemble based on the selection criteria, which can reduce the search space appropriately [21]. For ensemble pruning, this method greedily selects the classifier by searching the neighbors of the current ensemble according to the selection criteria. Therefore, there are usually two steps involved: building the selection criteria and determining the direction of the search [41]. With the FSS  $(F, E)$  and D-S theory of evidence, we build a new selection criteria  $F_{DS}(c)$  and rank the base classifiers to get the ordered set  $C' = \{c'_1, \dots, c'_s\}$ . Meanwhile, forward expansion and backward shrinkage are the common search directions, which have been extensively studied and proved to be an efficient method for finding the optimal solution. In this paper, we employ the forward expansion to select classifiers to form the final ensemble, and the algorithm is showed in Fig. 3.

## 5 Experimental setup

### 5.1 Datasets

For evaluating the performance of the DSFSS-P, sixteen binary datasets are gathered from the UC Irvine Machine Learning Repository [42], which are broadly applied for assessing the ensemble models’ results. A descriptive outline of the databases is presented in Table 4. For the concision of the experiment, normalization is implemented after missing data imputation by mean value of all data sets. For robustness check, a  $2 \times 5$ -fold cross-validation is used to diminish the effect of variability created by random partitioning on forecasting results. In particular, the samples are randomly split into five folds at first. In each time for validation, we use one fold to evaluate the results of the methods, and the last four folds are randomly divided equally for training and pruning respectively. The validation experiments are conducted for five times with different fold for testing. Meanwhile, 5-fold cross-validation are conducted for two times based on different partitions. The final experimental results are averaged over the 10 validations.

**Fig. 2** Algorithm of combining multiple criteria using D-S theory of evidence

---

**Algorithm 1:** Combine multiple attributes using D-S theory of evidence

---

**Input:** The fuzzy soft set  $(F, E)$  over  $C = \{c_1, \dots, c_s\}$  ;

- 1: Initialize  $\beta=1$  ;
- 2: **For**  $i = 1 : s$
- 3:   **For**  $j = 1 : 8$
- 4:     Compute basic probability for each classifier using Eq.(12) and get  $m(F_{e_j}(c_i))$  ;
- 5:   **End For**
- 6: **End For**
- 7: Combine  $m(F_{e_1}(c_i))$  and  $m(F_{e_2}(c_i))$  using Eq.(8) to get  $m(A)$ ;
- 8: **For**  $j = 3 : 8$
- 9:    $m(B) = m(F_{e_j}(c_i))$  ;  
    Combine  $m(A)$  and  $m(B)$  using Eq.(8) to get new  $m(A)$ ;
- 10: **End For**
- 11: Output the  $m(A)$ ;

---

**Fig. 3** Algorithm of selecting the best ensemble using greedy selection strategy with forward expansion

---

**Algorithm 2:** Greedy selection strategy with forward expansion

---

**Input:** The ordered set  $C' = \{c'_1, \dots, c'_s\}$  ;

- 1: Initialize  $BestEnsemble = \emptyset, MaxACC=0$ ;
- 2:  $T = \{c'_1\}$ ;
- 3: **For**  $i = 2 : s$
- 4:    $T = T \cup \{c'_i\}$  ;
- 5:   Compute  $ACC_T$  using Eq. (5);
- 6:   **If**  $ACC_T > MaxACC$
- 7:      $MaxACC = ACC_T$ ;
- 8:      $BestEnsemble = T$ ;
- 9:   **End If**
- 10: **End For**
- 11: Output the  $BestEnsemble$ ;

---



**Table 4** Descriptive outline of the databases

Datasets	Abbreviations	Instance	Features	Number of different classes
Activity recognition system based on multisensory data fusion (standing and walking)	AReM1	13,920	6	6720/7200
Activity recognition system based on multisensory data fusion (bending1 and bending2)	AReM2	6240	6	3360/2880
Activity recognition system based on multisensory data fusion (sitting and lying)	AReM3	14,399	6	7200/7199
Australian credit approval	Australian	690	14	307/383
Blood transfusion service center	Blood	748	5	570/178
Diabetic retinopathy debrecen	Diabetic	1150	19	539/611
Dresses attribute sales	Dresses	500	12	290/210
EEG eye state	EEG	14,980	14	8257/6723
German credit data	German	1000	20	700/300
Indian liver patient dataset	ILPD	583	10	167/416
Japanese credit screening	Japanese	690	15	307/383
Mammographic mass data	MMD	961	5	516/445
Pima indians diabetes	Pima	768	8	500/268
Steel plates faults	Steel	1941	27	1268/673
Vertebral column	Vertebral	310	6	100/210
Wisconsin breast cancer	Wisconsin	699	10	458/241

## 5.2 Base classifiers and benchmark methods development

Base classifiers play an essential role in the ensemble model's performance, as every classifier has its merits and advantages. A trade-off is demanded between synthesizing all the base classifiers' advantages and the complexity of the ensemble model. Fruitful results have been gained for building classifiers with machine learning methods, such as DT, SVM, NN [43], and comparative studies are also implemented to show their better performances [35]. Therefore, they are employed to build the base classifiers. They are generated by using bagging and various parameters, and the ensemble size is 300 at the first place. Table 5 displays the settings of the classification algorithms' parameters used for training base classifiers [35].

In order to investigate the effectivity of the proposed ensemble pruning method based on multiple diversity and performance measures, six benchmark methods are selected. Two common selection methods, single best (SB) and fusing

all base classifiers (ALL), are chosen as benchmarks. Four other popular ensemble pruning methods are selected. Hill-climbing ensemble selection (HCES) and Kappa pruning (Kappa) are two ordered based ensemble pruning methods, which build the selection criteria based on *ACC* and *Kappa* respectively. One ensemble pruning method built based on simultaneous diversity and *ACC* (SDAcc) is selected, which is proposed by [21]. Moreover, the method combining multiple diversity measures for ensemble pruning (DivP) is also selected [14].

The experiments described in this study are conducted on a PC with a 3.2 GHz Intel Core 4 CPU and 8 GB RAM, using the Windows 10 operating system. MATLAB, version 2020b, is used for modelling.

## 5.3 Performance indicators measures and statistical tests of significance

For the credibility and robustness, the *ACC* is selected to measure its classification ability, and the *disagreement*

**Table 5** Settings of the classification algorithms' parameters

Classification algorithm	Number of classifiers	Parameters	Candidate settings
SVM	100	Width of RBF kernel	$2^{(-6, -5, -4, -3, -2, -1)}$
DT	100	Min. leaf size	$n^*[0.01, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175]^*$
NN	100	Number of hidden nodes	[5, 10, 15, 20, 25, 30]

\*n is the number of samples applied for training base classifiers

*measure* is selected to measure diversity of the ensemble classifiers generated by it.

Next, a non-parametric test was used to assess the statistical differences between models [44]. To be specific, Friedman test was carried out to check the differences of models' ranking performance. The null hypothesis is that all classifiers perform identically, which indicates the average ranks of them are same. If the hypothesis is rejected, a post-hoc test called Holm's method is employed to take pairwise comparisons among them, which show significant differences [45].

## 6 Experimental results

### 6.1 Ensemble performance and diversity evaluation

Comparison of the average accuracies of the DSFSS-P method and 6 state-of-the-art benchmark methods on the 16 datasets over 10 validations are showed in Table 6. As we can see from it, the DSFSS-P achieves the highest classification accuracy in most cases by 13 times, which is much higher than SB, ALL, HCES, Kappa, SDAcc and DivP with

**Table 6** Comparison of the models' average accuracy on the 16 datasets

	SB	ALL	HCES	DSFSS-P	Kappa	SDAcc	DivP
AReM1	0.9879	0.9885	0.9885	<b>0.9886</b>	0.9879	0.9881	0.9877
AReM2	0.9612	0.9583	0.9636	<b>0.9644</b>	0.9612	0.9620	0.9617
AReM3	0.9740	0.9505	<b>0.9781</b>	0.9759	0.9740	0.9751	0.9753
Australian	0.8551	0.8536	0.8667	<b>0.8696</b>	0.8522	0.8594	0.8609
Blood	0.7699	0.7673	0.7780	<b>0.7874</b>	0.7766	0.7833	0.7686
Diabetic	0.7087	0.7061	0.7096	0.7122	0.7078	<b>0.7209</b>	0.6887
Dresses	0.5800	0.5760	0.5520	<b>0.5820</b>	0.5800	0.5760	0.5760
EyeState	0.8347	0.6986	<b>0.8452</b>	0.8426	0.8348	0.8385	0.8378
German	0.7240	0.7520	0.7510	<b>0.7530</b>	0.7260	0.7380	0.7220
ILPD	0.6775	0.7102	0.7050	<b>0.7153</b>	0.6809	0.7051	0.7067
Japanese	0.8522	0.8609	0.8580	<b>0.8667</b>	0.8565	0.8609	0.8522
MMD	0.8096	0.8304	0.8283	<b>0.8315</b>	0.8159	0.8273	0.8273
Pima	0.7773	0.7760	0.7760	<b>0.7812</b>	0.7773	0.7773	0.7812
Steel	0.7481	0.7676	0.7764	<b>0.7862</b>	0.7661	0.7785	0.7852
Vertebral	0.8000	0.8355	0.8258	<b>0.8548</b>	0.8323	0.8226	0.8323
Wisconsin	0.9642	<b>0.9671</b>	0.9628	<b>0.9671</b>	0.9642	0.9642	0.9656

**Table 7** Comparison of the ensemble models' average diversity on the 16 datasets

	ALL	HCES	DSFSS-P	Kappa	SDAcc	DivP
AReM1	0.0165	<b>0.0178</b>	0.0168	0.0171	0.0111	0.0167
AReM2	0.0312	0.0350	0.0372	<b>0.0477</b>	0.0103	0.0304
AReM3	0.0715	0.0499	0.0648	<b>0.1503</b>	0.0220	0.0332
Australian	0.1876	0.3510	0.3232	<b>0.4154</b>	0.0612	0.1200
Blood	0.1419	0.2269	0.1833	<b>0.3240</b>	0.0964	0.1685
Diabetic	0.3637	0.4140	0.3595	<b>0.5073</b>	0.1927	0.3415
Dresses	0.4241	<b>0.4443</b>	0.4067	0.4364	0.3588	0.2583
EyeState	0.3101	0.3337	0.1975	<b>0.5036</b>	0.1309	0.2121
German	0.2674	0.2819	0.2714	<b>0.3447</b>	0.1510	0.2171
ILPD	0.1977	0.2966	0.2424	0.3167	0.1302	<b>0.3330</b>
Japanese	0.1250	<b>0.1838</b>	0.1673	0.1733	0.0789	0.1340
MMD	0.1496	0.2045	0.2346	<b>0.3828</b>	0.0531	0.1992
Pima	0.2156	0.2481	0.2168	<b>0.2900</b>	0.1193	0.1903
Steel	0.2614	0.2934	0.2781	<b>0.3258</b>	0.1762	0.2684
Vertebral	0.2146	0.2682	0.2327	<b>0.2965</b>	0.0893	0.1989
Wisconsin	0.0879	0.1956	0.0879	<b>0.3796</b>	0.0119	0.0286

0, 1, 2, 0, 1, 0 times. On datasets AReM3, Diabetic and EyeState, though DSFSS-P is not the best one, it is second only to HCES, SDAcc and HCES respectively in terms of the average accuracy. On dataset Wisconsin, the DSFSS-P and ALL get the same accuracy, which is highest among the seven methods. Distinct differences are hardly found between seven methods.

In order to compare the diversities of the six ensemble methods in a clearer way, we summarize the results in Table 7. Bold values indicate the highest diversity among the six ensemble models. We can find that diversity of the ensemble pruned by DSFSS-P is at a medium level among the six methods on all datasets, but it obtains the best accuracy in most cases. The Kappa achieves the most diversified ensembles on most of the datasets, since it prunes the ensemble based on the *Kappa*. However, the accuracy of their measurements is very unstable and unsatisfactory, that may be due to the fact that only Kappa is insufficient to capture all the diversity of the ensemble. Divp, pruning the ensemble based on multiple diversity measures, achieves the lower diversity than SDAcc, which prunes the ensemble based on accuracy and diversity. They both obtain the lower diversity than the HCES on all datasets except for ILPD,

which selects the ensemble only based on accuracy. It's an interesting and surprising discovery. Furthermore, it is found that there is no obvious correlation between the accuracy and the diversity of the ensemble.

Meanwhile, the averaged run-time of ensemble pruning modes is recorded in Table 8. From the results, we can see that the DSFSS-P is much faster than the other ensemble pruning methods in all datasets. It is attribute to that our proposed method does not use iterative methods for the ensemble pruning, and the D-S combine rule is also a fast algorithm. The overall time complexity of the proposed method is  $O(MS)$ , where  $M$  is the number of criteria, and  $S$  is the number of the base classifiers.

### 6.2 Statistical tests of significance

Friedman test was carried out to check the significant differences of the methods' results, which is shown in Table 9. With the level of significance  $\alpha = 0.05$ , i.e., 95% confidence, the chi-square statistic is 40.781, and the degree of freedom is 6. The results show that the null hypotheses are rejected with the statistical results  $p$  value = 0.000, and it can be concluded that all the models play significantly different role in solving the binary classification problems. Moreover, it can be seen that DSFSS-P obtains the first rank whereas SB obtains the seventh rank, which again proves the superiority of the ensemble methods. SDAcc and HCES obtains the second and third rank respectively. They beat the other two pruning methods, i.e., Kappa and Divp, which select the ensembles only based on diversity measures. Kappa is worse than DivP and ALL, which indicates that one diversity measure is

**Table 8** Comparison of the ensemble pruning models' average runtime(s) on the 16 datasets

	HCES	DSFSS-P	Kappa	SDAcc	DivP
AReM1	644.19	<b>57.20</b>	529.09	640.20	917.13
AReM2	267.81	<b>9.03</b>	232.05	289.58	391.86
AReM3	639.41	<b>57.11</b>	549.42	676.33	391.47
Australian	32.08	<b>1.45</b>	27.53	43.10	50.28
Blood	30.30	<b>1.19</b>	29.02	44.49	51.55
Diabetic	46.05	<b>2.49</b>	43.18	62.55	76.53
Dresses	21.50	<b>1.35</b>	20.02	32.80	32.25
EyeState	679.96	<b>56.90</b>	578.90	691.17	814.81
German	44.59	<b>1.96</b>	39.89	59.31	37.11
ILPD	25.60	<b>1.15</b>	24.19	40.91	38.20
Japanese	30.48	<b>1.24</b>	28.34	46.14	42.72
MMD	41.01	<b>1.60</b>	38.56	58.01	75.76
Pima	32.90	<b>1.41</b>	30.87	49.15	46.63
Steel	85.16	<b>2.62</b>	74.62	102.51	66.34
Vertebral	13.84	<b>0.93</b>	12.91	26.81	30.17
Wisconsin	28.61	<b>1.13</b>	27.05	42.61	49.33

**Table 10** Results of post-hoc test after rejection of null hypothesis based on accuracy values (significance level of 0.05)

Comparison	p values	Sidak alpha	Result
DSFSS-P vs SB	0.00000	0.0024	H0 is rejected
DSFSS-P vs ALL	0.00000	0.0026	H0 is rejected
DSFSS-P vs HCES	0.00000	0.0027	H0 is rejected
DSFSS-P vs Kappa	0.00000	0.0028	H0 is rejected
DSFSS-P vs SDAcc	0.00002	0.0030	H0 is rejected
DSFSS-P vs DivP	0.00005	0.0032	H0 is rejected

**Table 9** Friedman test based on accuracy values (significance level of 0.05)

Friedman test							
Statistic			p value			Result	
40.781			0.000			H0 is rejected	
Algorithm	SB	ALL	HCES	DSFSS-P	Kappa	SDAcc	DivP
Ranking	5.31	4.31	3.69	1.19	4.81	3.56	4.25

unstable for ensemble pruning, and it is unable to capture all diversities of the ensemble.

Then, a post-hoc test called Holm’s method was carried out for a pairwise comparison between the rankings achieved by each model. Table 10 shows the results of post-hoc test on the confidence level  $\alpha = 0.05$ , and all the hypotheses are rejected. We can state that DSFSS-P

performance significantly better than any other benchmark model in terms of accuracy values.

### 6.3 Analysis of the selection criteria

In order to analyze the influence of the selection criteria on the performance the ensemble pruning method, we conduct the experiment by combining different number of the criteria.

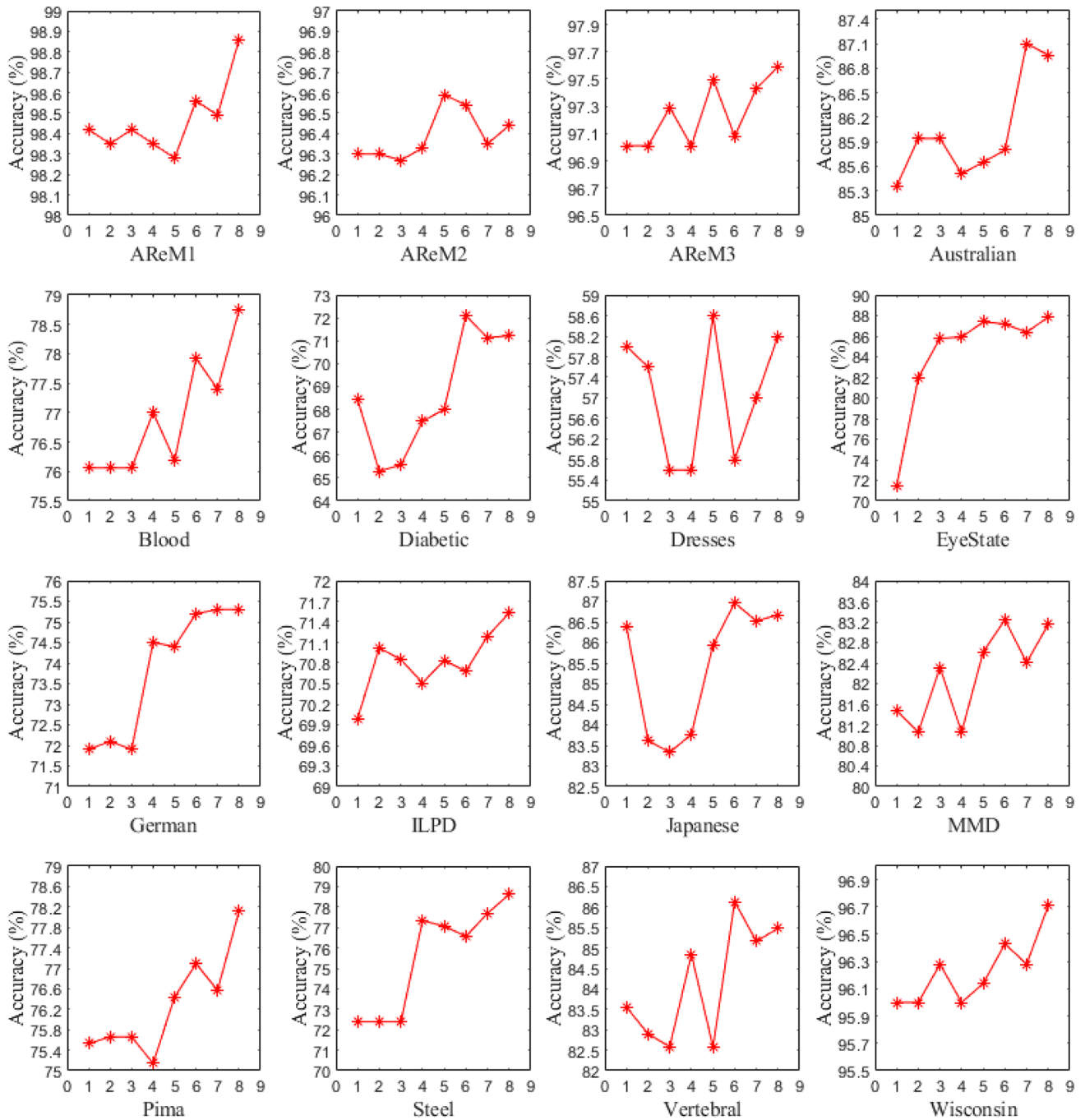


Fig. 4 The influence of the number of selection criteria on the performance the ensemble pruning method

The result is shown in Fig. 4, where the horizontal axis shows the number of selection criteria. The sequence of the criteria is according to  $\{DF, Kappa, Q, Dis, H, AUC, BS, ACC\}$ . Specifically speaking, 1 means use  $DF$  as the criteria to build the ensemble pruning method, 2 means combine  $DF$  and  $Kappa$  as the criteria to build the method, and so on in a similar fashion. The vertical axis represents the result of methods based on ACC.

From the results, we can see that combining all the eight criteria obtains the highest ACC 9 times and second high ACC 6 times among the 16 datasets. Meanwhile, the 16 lines of the experiment show an upward trend, which means the more criteria, the better performance the ensemble pruning method. It shows the multiple criteria ensemble pruning method's effectivity and superiority. Moreover, only combining the criteria with diversity measures for ensemble pruning doesn't obtain the highest accuracy for any dataset. Combining multiple diversity and performance measures for ensemble pruning (DSFSS-P) can generate more accurate ensemble than that only considering one performance or diversity measure and that combining multiple diversity measures.

## 7 Conclusions and future work

In this study, we propose a multiple criteria ensemble pruning method for binary classification by combining multiple diversity and performance measures simultaneously with a MCDM method. The  $Q$  statistics,  $Kappa$ -statistic, *double-fault* and *disagreement measure* are used to capture the base classifiers' diversity. The *accuracy*, *Brier score*, *ROC* and *H-measure* are used to measure their classification ability. They are the criteria giving a full evaluation to the base classifier and the relation of them. In order to achieve a trade-off between the diversity and performance measures, the FSS is applied to arrange the base classifiers' levels of achievements with regard to the criteria, which provides a mathematical theory support to build the decision matrix. The D-S theory of evidence is applied to aggregate the uncertainty and conflicting information of the criteria, which help to give a comprehensive evaluation to the base classifiers.

To validate the performance of the proposed ensemble pruning method, a comprehensive empirical evaluation is carried out on 16 binary datasets from UCI. Compared to 6 state-of-the-art benchmark methods, the proposed method DSFSS-P achieves better performance in 13 out of the 16 data sets than other methods do in terms of classification accuracy, which show its effectivity and superiority for ensemble pruning. Thus, ensemble pruning by considering multiple diversity and performance measures can bring

more accurate and generalized ensemble. In our further research, there are two tasks for investigation to make the proposed ensemble pruning method more promising. On one hand, the choice of diversity and performance measures is still an open problem. On the other hand, the MCDM based ensemble pruning method's classification performance can be further improved with more customized design.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China [grant number 72071021, 71671019].

## References

1. Nguyen TT, Luong AV, Dang MT, Liew AWC, McCall J (2020) Ensemble selection based on classifier prediction confidence. *Pattern Recogn* 100:107104
2. Zhou HF, Zhao XH, Wang X (2014) An effective ensemble pruning algorithm based on frequent patterns. *Knowl-Based Syst* 56:79–85
3. Zhang Y, Burer S, Street WN (2006) Ensemble pruning via semi-definite programming. *J Mach Learn Res* 7:1315–1338
4. Wang Z, Wang RX, Gao JM, Gao ZY, Liang YJ (2020) Fault recognition using an ensemble classifier based on Dempster-Shafer theory. *Pattern Recogn* 99:107079
5. Zhang C-X, Kim S-W, Zhang J-S (2020) On selective learning in stochastic stepwise ensembles. *Int J Mach Learn Cybern* 11(1):217–230. <https://doi.org/10.1007/s13042-019-00968-9>
6. Mohammed AM, Onieva E, Wozniak M, Martinez-Munoz G (2022) An analysis of heuristic metrics for classifier ensemble pruning based on ordered aggregation. *Pattern Recogn* 124:108493
7. Ykhlef H, Bouchaffra D (2017) An efficient ensemble pruning approach based on simple coalitional games. *Information Fusion* 34:28–42
8. Wozniak M, Grana M, Corchado E (2014) A survey of multiple classifier systems as hybrid systems. *Information Fusion* 16:3–17. <https://doi.org/10.1016/j.inffus.2013.04.006>
9. Tsoumakas G, Partalas I, Vlahavas I (2009) An ensemble pruning primer. In: Okun O, Valentini G (eds) *Applications of supervised and unsupervised ensemble methods*. Springer, Berlin, Heidelberg, pp 1–13. [https://doi.org/10.1007/978-3-642-03999-7\\_1](https://doi.org/10.1007/978-3-642-03999-7_1)
10. Yu ZW, Zhang YD, Chen CLP, You J, Wong HS, Dai D, Wu S, Zhang J (2019) Multiobjective semisupervised classifier ensemble. *IEEE T Cybernetics* 49(6):2280–2293
11. Jackowski K (2018) New diversity measure for data stream classification ensembles. *Eng Appl Artif Intel* 74:23–34
12. Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach Learn* 51(2):181–207
13. Kuncheva LI, Whitaker CJ, Shipp CA, Duin RPW (2003) Limits on the majority vote accuracy in classifier fusion. *Pattern Anal Appl* 6(1):22–31. <https://doi.org/10.1007/s10044-002-0173-7>
14. Cavalcanti GDC, Oliveira LS, Moura TJM, Carvalho GV (2016) Combining diversity measures for ensemble pruning. *Pattern Recogn Lett* 74:38–45
15. Zhang ZL, Chen YY, Li J, Luo XG (2019) A distance-based weighting framework for boosting the performance of dynamic ensemble selection. *Inform Process Manag* 56(4):1300–1316
16. Dos Santos EM, Sabourin R, Maupin P (2009) Overfitting cautious selection of classifier ensembles with genetic algorithms.

- Information Fusion 10(2):150–162. <https://doi.org/10.1016/j.inffus.2008.11.003>
17. Zhang YQ, Cao G, Li XS (2021) Multiview-based random rotation ensemble pruning for hyperspectral image classification. *IEEE T Instrum Meas* 70:1–14
  18. Ferri C, Hernandez-Orallo J, Modroui R (2009) An experimental comparison of performance measures for classification. *Pattern Recogn Lett* 30(1):27–38
  19. Huang J, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. *Ieee T Knowl Data En* 17(3):299–310
  20. Sabzevari M, Martinez-Munoz G, Suarez A (2022) Building heterogeneous ensembles by pooling homogeneous ensembles. *Int J Mach Learn Cybern* 13(2):551–558. <https://doi.org/10.1007/s13042-021-01442-1>
  21. Dai Q, Ye R, Liu ZA (2017) Considering diversity and accuracy simultaneously for ensemble pruning. *Appl Soft Comput* 58:75–91
  22. Bian YJ, Wang YJ, Yao YQ, Chen HH (2020) Ensemble pruning based on objection maximization with a general distributed framework. *IEEE T Neur Net Lear* 31(9):3766–3774
  23. Hashemi A, Dowlatshahi MB, Nezamabadi-pour H (2022) Ensemble of feature selection algorithms: a multi-criteria decision-making approach. *Int J Mach Learn Cybern* 13(1):49–69. <https://doi.org/10.1007/s13042-021-01347-z>
  24. Peng Y, Kou G, Wang GX, Shi Y (2011) FAMCDM: A fusion approach of MCDM methods to rank multiclass classification algorithms. *Omega-Int J Manage Sci* 39(6):677–689
  25. Shafer G. (1976) A mathematical theory of evidence, vol 29. vol 4. Princeton University Press
  26. Yuan KJ, Deng Y (2019) Conflict evidence management in fault diagnosis. *Int J Mach Learn Cybern* 10(1):121–130. <https://doi.org/10.1007/s13042-017-0704-6>
  27. Roy AR, Maji PK (2007) A fuzzy soft set theoretic approach to decision making problems. *J Comput Appl Math* 203(2):412–418
  28. Cagman N, Enginoglu S, Citak F (2011) Fuzzy soft set theory and its applications. *Iranian J Fuzzy Syst* 8(3):137–147
  29. Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. *Ann Math Stat* 38(2):325–339
  30. Shafer G (1976) A mathematical theory of evidence, vol 1. Princeton University Press, Princeton
  31. Xu XZ, Martel JM, Lamond BF (2001) A multiple criteria ranking procedure based on distance between partial preorders. *Eur J Oper Res* 133(1):69–80
  32. Wang R, Kwong S (2014) Active learning with multi-criteria decision making systems. *Pattern Recogn* 47(9):3106–3119
  33. Xiao FY (2018) A hybrid fuzzy soft sets decision making method in medical diagnosis. *IEEE Access* 6:25300–25312
  34. Zavadskas EK, Turskis Z, Kildiene S (2014) State of art surveys of overviews of Mcdm/Madm methods. *Technol Econ Dev Eco* 20(1):165–179
  35. Lessmann S, Baesens B, Seow HV, Thomas LC (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. *Eur J Oper Res* 247(1):124–136
  36. Hand DJ (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach Learn* 77(1):103–123
  37. Zhang J, Dai Q, Yao CS (2021) DEP-TSPmeta: a multiple criteria dynamic ensemble pruning technique ad-hoc for time series prediction. *Int J Mach Learn Cybern* 12(8):2213–2236. <https://doi.org/10.1007/s13042-021-01302-y>
  38. Taghavi ZS, Niaki STA, Niknamfar AH (2019) Stochastic ensemble pruning method via simulated quenching walking. *Int J Mach Learn Cybern* 10(7):1875–1892. <https://doi.org/10.1007/s13042-018-00912-3>
  39. Gong B, Hua Z. (2007) The evidential reasoning approach for multi-attribute decision making problem with incomplete decision matrix. In: Fourth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE, pp 416–421
  40. Xiao Z, Yang XL, Niu Q, Dong YX, Gong K, Xia SS, Pang Y (2012) A new evaluation method based on D-S generalized fuzzy soft sets and its application in medical diagnosis problem. *Appl Math Model* 36(10):4592–4604
  41. Partalas I, Tsoumakas G, Vlahavas I (2010) An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. *Mach Learn* 81(3):257–282. <https://doi.org/10.1007/s10994-010-5172-0>
  42. UCI Machine Learning Repository (2013) University of California, School of Information and Computer Science. <http://archive.ics.uci.edu/ml>
  43. Xu LX, Wang XF, Bai L, Xiao J, Liu Q, Chen EH, Jiang XY, Luo B (2020) Probabilistic SVM classifier ensemble selection based on GMDH-type neural network. *Pattern Recogn* 106:107373
  44. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30
  45. Sidhu P, Bhatia MPS (2019) A two ensemble system to handle concept drifting data streams: recurring dynamic weighted majority. *Int J Mach Learn Cybern* 10(3):563–578. <https://doi.org/10.1007/s13042-017-0738-9>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.