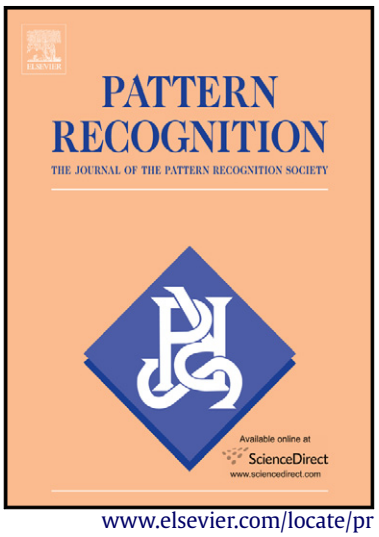
A novel dynamic rough subspace based selective ensemble

Guo, Yuwei; Jiao, Licheng; Wang, Shuang; Wang, Shuo; Liu, Fang; Rong, Kaixuan; Xiong, Tao

DOI:
10.1016/j.patcog.2014.11.001



www.elsevier.com/locate/pr

A novel dynamic rough subspace based selective ensemble

Yuwei Guo, Licheng Jiao, Shuang Wang, Shuo Wang, Fang Liu, Kaixuan Rong, Tao Xiong

PII: S0031-3203(14)00451-8
DOI: http://dx.doi.org/10.1016/j.patcog.2014.11.001
Reference: PR5273

To appear in: *Pattern Recognition*

Cite this article as: Yuwei Guo, Licheng Jiao, Shuang Wang, Shuo Wang, Fang Liu, Kaixuan Rong, Tao Xiong, A novel dynamic rough subspace based selective ensemble, *Pattern Recognition,* http://dx.doi.org/10.1016/j.patcog.2014.11.001

A novel dynamic rough subspace based selective ensemble

Yuwei Guo[1], Licheng Jiao[1], Shuang Wang[1], Shuo Wang[2], Fang Liu[3], Kaixuan Rong[1], Tao Xiong[1]

[1] *Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an, Shaanxi Province 710071, China*

[2] *CERCIA, School of Computer Science, University of Birmingham, UK*

[3] *School of Computer Science and Technology, Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University*

Abstract

Ensemble learning has been a hot topic in machine learning due to its successful utilization in many applications. Rough set theory has been proved to be an excellent mathematical tool for dimension reduction. In this paper, based on rough set, a novel framework for ensemble is proposed. In our proposed framework, the relationship among attributes in rough subspace is first considered, and the maximum dependency degree of attribute is first employed to effectively reduce the searching space of reducts and augment the diversity of selected reducts. In addition, in order to choose an appropriate reduct from the dynamic reduct searching space, an assessment function which can balance the accuracy and diversity is utilized. At last, a new method, i.e., Dynamic Rough Subspace based Selective Ensemble (DRSSE), which is derived from our framework is given. By repeatedly changing the searching space of reducts and selecting the next reduct from the changed searching space, DRSSE finally trains an ensemble system with these selected reducts. Compared with several available ensemble methods, experimental results with several datasets demonstrate that DRSSE can lead to a comparative or even better performance.

Keywords*:* rough set, selective ensemble, dynamic reduct searching space, diversity

1. Introduction

Ensemble learning has been attracted much attention in machine learning for its good generalization ability since 1990s [1]. It is one of the most promising methods for constructing an accurate predictor by combining predictions of a number of base classifiers. It can also reduce overfitting problems and achieve good performance. Typically, building an ensemble system needs two steps. The first step is to train a set of weak classifiers as base learners. Here, "weak" means that the performance of classifier is not particularly good but slightly better than random guess. The second step is to integrate the weak classifiers through various combination strategies. Some typical ensemble learning methods include Boosting [2] and Bagging [3].

It is commonly agreed that the success of ensembles is attributed to the accuracy of all base classifiers and diversity among base classifiers [4,5]. Diversity refers to the degree of disagreement among the base classifiers. Some pairwise and non-pairwise diversity measures have been proposed to estimate the diversity level, such as Q-statistic, double-default measure, entropy, etc. Different from the traditional ensemble methods which integrate the predictions of all available base classifiers, the selective ensemble learning algorithms choose a subset of base classifiers to contribute to the final prediction, based on the understanding that many could be better than all [6]. The advantages of selective ensembles over traditional ensembles lie in a smaller ensemble size and potentially better generalization ability. Therefore, the selective ensemble is believed to be much more effective than a single classifier and the traditional ensemble system.

Rough set theory was first introduced by Polish Mathematician Pawlak in 1980s [7,8]. It is a useful mathematical tool to handle uncertain and vague data [9], and it has been applied successfully in machine learning, pattern recognition, and data mining [10]. The most important feature of rough set theory is its great ability in attribute reduction and feature selection [11,12]. The reduced attribute set based on rough set is named as *reduct* [13,14]. A reduct is a minimal subset of attributes which is sufficient to discern objects with different decision values. Theoretically, it has been proved that a classifier trained based on a reduct

has comparable performance with a classifier trained based on the original information system [22,15]. There is always more than one reduct for any information system, so the classifier can be built in multiple ways. Based on the fact that different data subsets can be obtained by applying rough set theory, and ensemble learning is good at performance generalization. Many recent methods have been proposed to introduce the rough set theory into ensemble learning. These rough set ensembles have been applied to a wide range of practical problems, such as text classification, biomedical classification, tumor classification and web services classification [16-21].

In this paper, a new selective ensemble method, i.e., Dynamic Rough Subspace based Selective Ensemble (DRSSE), which implies a novel rough set ensemble framework is proposed. In this framework, the procedures that using reducts to train base classifiers and selecting base classifiers to construct the ensemble system are no longer mutually independency, which is different from the available methods. The method under this framework takes into consideration the diversity of selected reducts as well as the accuracy and diversity of base classifers. Specifically, the maximum dependency attribute is first used to reduce the searching space of reducts. The aim of reducing the searching space is to increase the difference between alternative reducts and selected reducts as well as improve searching efficiency. A preliminary ensemble system is produced based on the selected reduct. Then, the next selected reduct considers the performance of the preliminary ensemble and the diversity of selected reducts. An assessment function, i.e., Accuracy-Diversity (AD for short) assessment function, which considers the accuracy of each base classifier and the diversity among base classifiers is used to help select reducts. The final ensemble system consists of classifiers trained with the selected reducts.

The rest of this paper is structured as follows. Section 2 introduces the concept of rough set. In Section 3, an algorithm, named DRSSE, is proposed based on a novel framework of selective ensemble using rough set theory. The experiments and the results are presented in Section 4. Finally, Section 5 concludes this paper and points out the future work.

2. Basic Concepts for Rough Set Theory

Before present the details of the proposed DRSSE method, the basic concepts for rough set theory and the methods for generating a set of reducts are given in the following subsection.

2.1 Preliminary

In rough set theory, an information system, which is also called a decision table, is defined as $S=(U,C\cup D,V,f)$, where U is a set of finite and nonempty objects, referred as the universe; C is the condition attribute set characterizing the samples; D is the decision attribute set; V is the attribute values in C and $f:C\to V$ is a description function. Take the constructed data shown in Table 1 as an intuitive example.

Table 1
A constructed dataset.

U	C				D
	c_1	c_2	c_3	c_4	
x_1	1	1	1	1	d_1
x_2	1	3	1	3	d_1
x_3	1	1	2	1	d_1
x_4	2	2	2	1	d_1
x_5	2	1	2	1	d_1
x_6	2	3	2	2	d_2
x_7	3	2	2	2	d_2
x_8	3	3	3	2	d_2
x_9	3	1	3	3	d_2
x_{10}	3	2	1	2	d_2

In the constructed data shown in Table 1, each column represents an attribute (e.g., the condition attribute C which are composed by four attributes c_1,c_2,c_3,c_4 and the decision attribute D) and each row (e.g., $x_1,...,x_{10}$) denotes a sample. The condition attribute describes the property of the sample, and the decision attribute characterizes which category the sample belongs in.

For any $B\subseteq C$, there is an indiscernibility relation $I(B)$, which is defined as follows:

$$I(B)=\{(x,y)\in U\times U \mid \forall a\in B, f(x,a)=f(y,a)\}. \quad (1)$$

That is, for any $a\in B$, if and only if $a(x)=a(y)$ where $a(x)$ denotes the value of attribute a for element x, then $(x,y)\in I(B)$. It is clear that the indiscernibility relation $I(B)$ is an equivalent relation that satisfies reflexivity, symmetry and transitivity. The family of all equivalence classes for $I(B)$, i.e., a partition of an information system determined by B, can be denoted by $U/I(B)$, or simplified by U/B. If (x,y) belongs to $I(B)$, then x and y are B-indiscernible, i.e., there is no difference between x and y with respect to B. Equivalence classes which are generated by B are referred as B-elemental granules or B-information granules, and definition of the equivalence class is as follows:

$$[x]_B=\{x \mid (x_i,x)\in I(B), x\in U\}. \quad (2)$$

That is, x is a sample set and samples in x have the same attribute value in terms of attribute B. As shown in the second column of Table 1, in terms of condition attribute c_1, the samples x_1, x_2 and x_3 belong to an equivalence class, the samples x_4, x_5 and x_6 belong to the other equivalence class, while the samples x_7, x_8, x_9 and x_{10} belong to the third equivalence class.

Given an information system $S=(U,A), X\subseteq U$ and $B\subseteq A$. For any $X\subseteq U$, two unions of elemental granules $\underline{B}(X)$ and $\overline{B}(X)$ can be defined, where $\underline{B}(X)$ and $\overline{B}(X)$ are lower and upper approximations of X regarding B, respectively. Eq. (3) shows the definitions of $\underline{B}(X)$ and $\overline{B}(X)$.

$$\begin{cases}\underline{B}(X)=\{x\in U \mid [x]_B\subseteq X\}\\ \overline{B}(X)=\{x\in U \mid [x]_B\cap X\neq\phi\}\end{cases}. \quad (3)$$

Here, the B-lower approximation is the union of all $[x]_B$ which are included in the set X, whereas the B-upper approximation of a set is the union of all $[x]_B$ which have a

nonempty intersection with the set X. As illustrated in the second column of Table 1, in terms of decision attribute D, $X_1 = \{x_1, x_2, x_3, x_4, x_5\}$ is an equivalence class. $\{x_1, x_2, x_3\}$ and $\{x_1, x_2, x_3, x_4, x_5, x_6\}$ are the lower and upper approximations of X_1 in terms of attribute c_1. Besides, $X_2 = \{x_6, x_7, x_8, x_9, x_{10}\}$ is also an equivalence class regarding decision attribute D. $\{x_7, x_8, x_9, x_{10}\}$ and $\{x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$ are the lower and upper approximations of X_2 regarding attribute c_1, respectively.

The set $BN_B(X) = \overline{B}(X) - \underline{B}(X)$, is defined as the B-boundary region of X. If $\overline{B}(X) = \underline{B}(X)$, it means that X is a definable set, i.e., X can be crisp characterized with respect to B; otherwise, X is referred as rough regarding B. In other words, there are some samples that we are not sure whether they can be accurately classified. In this situation, samples in boundary region cannot be accurately characterized by the condition attribute.

Given a decision table $S = (U, C \bigcup D, V, f)$, where C is a set of condition attribute and D is a set of decision attribute. The positive region regarding D is the union of all low approximations in terms of C. Elements in the union of positive regions can be exactly classified into U / D with respect to C, that is:

$$POS_C(D) = \bigcup_{X \in U/D} \underline{C}(X), \tag{4}$$

where $\underline{C}(X) = \{x \in U \mid [x]_C \subseteq X\}$ is the lower approximation of X regarding C. When it comes to the second column of Table 1, the positive region of D regarding attribute c_1 is $POS_{c_1}(D) = \{x_1, x_2, x_3, x_7, x_8, x_9, x_{10}\}$.

The positive region completely belongs to one of the decision classes and the size of positive region reflects the approximation power of condition attribute C with respect to the decision attribute D. In order to measure the importance of condition attribute, the dependency degree $r_B(D)$ is used as the criterion for attribute selection, and definition of $r_B(D)$ is as follows:

$$r_B(D) = \frac{|POS_B(D)|}{|U|}, \tag{5}$$

where B is a subset of condition attribute C. For each $B \subseteq C$, the decision attribute set D depends on B in terms of $r_B(D)$, and $r_B(D)$ takes values [0, 1]. The larger the dependency degree value is, the more significant of the condition attribute for the approximation of the decision attribute. Take the second column of Table 1 as an example, the dependency degree value $r_{c_1}(D)$ of the decision attribute set D in terms of attribute c_1 is $r_{c_1}(D) = \frac{|POS_{C_1}(D)|}{|U|} = \frac{7}{10} = 0.7$.

Given a subset B for condition attribute set C, for any attribute $c \in B$, the importance of c regarding D is defined as follows:

$$R_B(D) = r_B(D) - r_{B\text{-}c}(D) = \frac{|POS_B(D)|}{|U|} - \frac{|POS_{B-c}(D)|}{|U|}. \tag{6}$$

$R_B(D)$ reflects the importance of c by calculating the difference between $r_B(D)$ and $r_{B\text{-}c}(D)$. The larger the $R_B(D)$ is, the more significant of the attribute c. Take the data shown in Table 1 as an example, considering the subset $B=\{c_1, c_3\}$ and $c = c_1$, then $R_B(D) = r_B(D) - r_{B\text{-}c_1}(D) = 0.7 - 0.7 = 0$. It demonstrates that the attribute c_1 is not essential to decision and classification and c_1 can be removed from the subset B.

2.2 The generation of multiple reducts

In order to produce multiple classifiers, a set of reducts should be utilized. This subsection briefly introduces three methods which can be used for generating multiple reducts. The first method is WADF [22]. In this method, condition attributes are sorted in terms of attribute significance degree firstly, and then the worst condition attribute will be deleted one by one. A good reduct can be found by using the concept of attribute significance. Here, in order to produce a set of reducts that are different from each other, a strategy of changing the order of

condition attributes is applied. This strategy moves away only one dispensable attribute from an obtained good reduct at a time, and finds a new reduct using the remaining condition attribute. Authors in [23] presented another way to find reducts. Their algorithm starts with an indispensable attribute. First of all, the discernibility matrix of the decision system is created, and then according to the significance degree of attributes, the forward stepwise selection and backward elimination strategy are used to generate multiple reducts. The second reducts generating method is "Multiple_Reduct_Generation" [24]. To find multiple reducts, a score function is introduced to assess the importance of each condition attribute. The attribute with lower score means a higher opportunity to become a member of reduct. In terms of obtained scores, these condition attributes are partitioned into several parts. The first part possesses the most important attributes for finding a reduct. Each element in the first part will be deleted from this part and become a member of attributes forming reduct if this element satisfies some constraints. One reduct can be generated according to the previous step, and for finding other reducts, Cartesian product operation is applied iteratively among other parts to modify these parts and finally an approximate set of reducts will be found. As for the third reducts generating method, an existing software ROSETTA [25] can be directly used. ROSETTA has been developed for data processing in the framework of rough set theory. It provides a series of algorithms to generate one reduct or a set of reducts, such as genetic algorithm, Johnson's algorithm and so on. Some rules are used to find multiple reducts, and the importance measure of rule is used to rank these rules. Indispensible attributes of a given dataset need to be computed to rank these rules. Here, the genetic algorithm is implemented for computing minimal hitting sets, as described by Vinterbo and Ohrn [26].

In this paper, ROSETTA software is used to generate a set of reducts as its simplicity. Genetic algorithm with the option of full discernibility is adopted for finding minimal attribute sets which retain the similar classification power as the original dataset.

3. A Novel Framework for Rough Set Theory based Selective Ensemble

In most case, the sample set contains redundant attributes. Removing them may improve the performance of the classifier and save memory consumption. Rough set can be used as a dimension reduction tool to achieve this goal, which makes use of the relationship between the condition attribute and the decision attribute to generate reducts. A reduct is an irreducible subset of features, which keeps the same discernibility as the original set of features. As we know, hundreds of reducts can be generated for a given dataset. That is, for a given dataset, there are not merely one minimal attributes subset A determining specified decision classes in the same degree as the original attribute set, but there exists other minimal attribute subsets with the same properties like A.

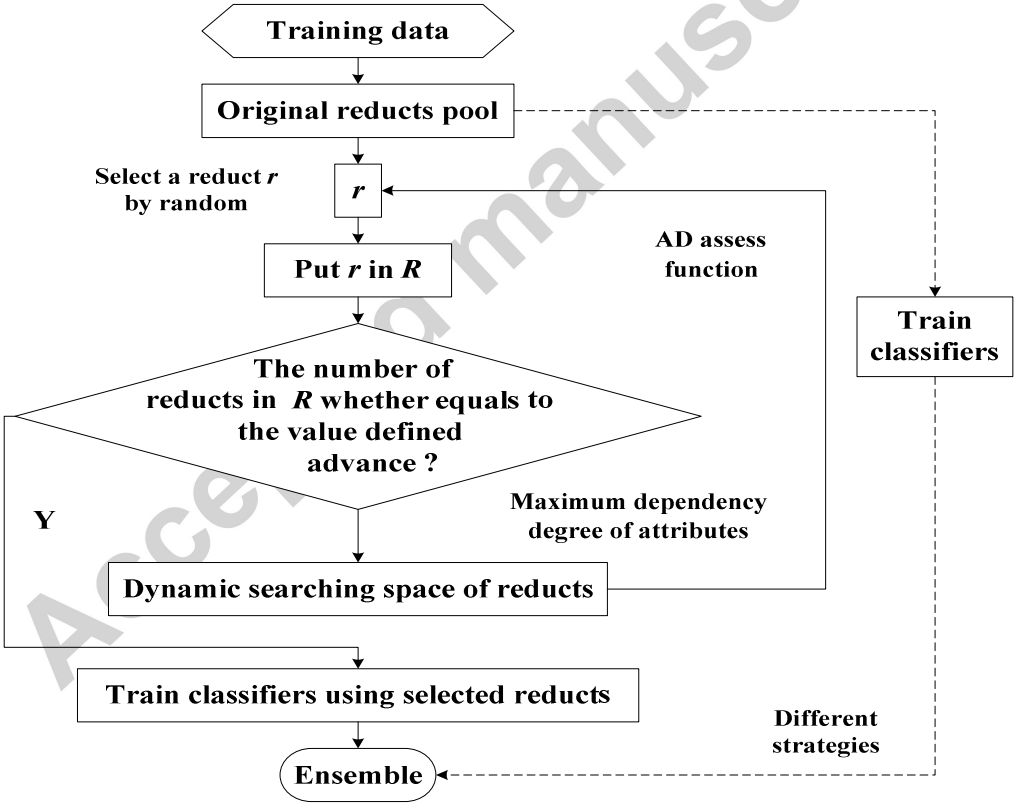


Fig.1. The framework of the proposed algorithm use solid line. Additional procedures are traditional rough set ensemble, use broken line.

Rough set theory provides a new way to produce different feature subsets for training a number of component classifiers. Different from random subspace method and attribute bagging, the selected attributes based on rough set do not lose the discrimination ability of the original information system. Therefore, rough set based ensemble is expected to achieve good classification quality. In the traditional rough set ensemble framework, the first step is to generate a set of reducts. The second step is to build base classifiers with all generated reducts [27,28]. Finally, different strategies are used to integrate these base classifiers. The traditional framework of rough set based ensemble is shown as the dashed line in Fig.1. It can be seen that the traditional framework does not consider the relationship among the generated reducts. As mentioned above, for a given information system, there may exists some redundancy attributes. The performance of classifier may be improved by using the attribute set without redundancy. For an ensemble system, combining some of base classifiers may be better than combining all of base classifiers [6]. This concept can be generalized to reducts based on rough set. That is, a set of reducts also have redundancy, and if we delete the redundant reducts, the efficiency of the built classifiers may be improved.

Diversity among base classifiers is one of the key factors affecting ensemble performance. The classifiers trained with different reducts should have certain diversity inherently. However, this assumption is not always true; that is, different attribute subsets are not directly relevant to diversity among classifiers. For instance, there are two different reducts, but the classifiers trained by the two reducts may have the same classification results. Therefore, there is no diversity between these two classifiers. However, integrating classifiers trained by different attribute subsets should have a higher chance of producing more diversity than integrating classifiers trained by the same attribute subset. In order to obtain good generalization, it is an important task to select appropriate attribute subsets (reducts) for training the base classifiers. So, when we construct ensemble classifiers based on rough set, we not only consider accuracy of base classifiers and diversity among classifiers, but also consider the difference of training set.

With these motivations in mind, we propose a new framework of rough set based selective ensembles, as shown in solid portions of Fig.1. This framework selects reducts to train the base classifiers, which consists of four steps. In the first step, one reduct is randomly chosen from original reducts pool. In the second step, the searching space of reducts is reduced by computing the maximum dependency attribute of selected reduct. In the third step, depending on the accuracy-diversity assessment function, a reduct from the newly searching space is selected. By repeating the second and the third step, a group of proper reducts are selected. Finally, the selected reducts are used to train a set of classifiers, by which the ensemble system is formed. Details of the proposed framework are given next.

3.1 The dependency degree among attributes

The relationship of attributes in reducts is first studied in rough set ensemble. In this subsection, the concept of maximum dependency of condition attributes is introduced [29]. This concept is developed from the criterion of the degree of dependency in rough set theory. The criterion measures the dependency of the decision attribute on the condition attribute, which reflects the approximation power of condition attributes to decision attribute. It is defined in Eq. (5) and considered as an evaluation of attribute important degree. The higher value of $r_B(D)$, the condition attributes B are more important for determining decision attribute D.

In order to choose different reducts from the original reduct pool, the relationship among attributes in a reduct is described. As mentioned earlier, the dependence degree of decision attributes on condition attributes is measured in Eq. (5). This formula can also be used to describe the dependency degree among attributes in a reduct. According to the relationship among attributes in a reduct, we can reduce the original reduct pool, which potential increase the difference between the left reducts and the selected reducts.

Given a new decision system $\langle U, A \rangle$, where $A \subseteq C$ is a minimal attribute subset of the original attribute set C, and A has a similar discrimination ability to the original decision system. For any $c \in A$, the Eq. (5) can be rewritten as:

$$r_P(c) = \frac{|POS_P(c)|}{|U|}, where\ P = A - c, \tag{7}$$

where $POS_P(c) = \bigcup_{X \in U/c} \underline{P}(X)$ is the positive region of c regarding the attribute set P. As for the construction of $\underline{P}(X)$, for any attribute c and attribute set P, two groups of equivalence classes $[X]_c$ and $[X]_P$ can be induced by c and P, respectively. Then, for each equivalence class in $[X]_c$, the size of its lower approximation $\underline{P}(X)$ can be added up by the equivalence classes that induced by P.

When it comes to Eq. (7), the numerator of Eq. (7) indicates the total number of samples which are induced by the attribute set P, and these samples can be positively categorized into the sets induced by attribute c. The dependency degree $r_P(c)$ denotes the proportion of such classifiable samples for which it satisfies the values of attributes in P that determining the values of attribute in c. In other words, $r_P(c)$ expresses the "re-construct" ability of attribute set P to attribute c. If attribute c can be completely "reconstructed" by attribute set P, then c depends totally on P. In this case, c can be removed from A, as the attribute set P can well express c. $r_P(c)$ indicates the dependency degree of attribute c on the attribute set P. With the help of Eq. (7), we expect to find an attribute $c \in A$ which is strongly related to the attribute set P.

3.2 Shrink of the reducts searching space

For base classifiers utilized in ensemble system, a high diversity but with a low generalization error is a dilemma. Rough set theory can be employed to generate multiple attribute subsets, i.e., reducts. A reduct is thought of as a "sufficient set of features", which can fully characterize the knowledge of the original data [30]. In other words, rough set theory can be used to complete attribute reduction with the aim to retain the discriminatory power of original attributes. A variety of works regarding classification based on rough set theory are studied [31-34]，which expect the classifiers trained from reducts can achieve a well accuracy performance. We also hope that the base classifiers trained by reducts can ensure a good accuracy performance than that trained from attribute subsets which are randomly selected from the original data.

In the condition that attribute subsets are produced by rough set theory, we ignore the property of reducts for now. Or suppose the classifiers trained by these reducts have similar classification accuracy. Then we wish to select reducts which are different between each other. Just like the traditional bagging or boosting algorithm generates multiple datasets used for training classifiers by changing the distribution of samples. Classifiers trained by different datasets (based on samples) are thought of having potential for obtaining different result. Likewise, classifiers trained by different reducts (based on attributes) are more likely to produce results in diversity. Different reducts refer to reducts which contain different attributes. Simply to say, our goal is to select reducts with different attributes for the diversity of selected reducts. Eq. (7) measures the approximation ability of attributes, which can be employed to select proper reducts from the reduct pool. If an attribute (*attri*) in a reduct can be well approximation by other attributes, the attribute become relatively less necessary than other attributes. In other word, the influence of deleting this attribute is less than deleting other attributes. Note that, if we produce other reducts under the circumstance – original information system without attribute *attri*, the attribute *attri* will not appear in the newly produced reduct. Actually, a reduct pool can be generated by software. Thus, to ensure the

next selected reduct not containing *attri*, the reducts with *attri* will be deleted from the original reduct pool. To select different reducts, we measure the relationship among attributes in a random selected reduct by Eq. (7), and find the attribute which can be well characterized by other attribute. Then reducts which containing this attribute will be deleted. Thus, this attribute will not appear in the next selected reduct. This process narrows down the searching space of reducts, and the alternative reducts are different from selected reduct. Meanwhile, deleting attribute *attri* causes relatively less impact. The pseudo code of how to shrink the original reduct pool in terms of attribute with maximum dependency degree is shown in Algorithm 1.

Algorithm 1. Shrink the searching space of reducts in terms of attribute with maximum dependency degree.

Input: original reducts searching space RED, any reduct A in RED

Output: a changed searching space RED_C

1. for each $c_i \in A$
2. using Eq.(7), compute the dependency degree r_i of attribute c_i upon the left attribute $A - c_i$
3. end for
4. select c_m such that $r_m = \max_{a_i \in A}(d_i)$
5. delete reducts red which contained the attribute c_m from original reducts pool RED, $RED\_C \leftarrow RED - red$

Based on a set of produced reducts shown in Table 2, the procedures of choosing the different reducts are illustrated.

Table 2
A set of reducts.

	Reducts
1	$c_1,\ c_2,\ c_5,\ c_6,\ c_7$
2	$c_1,\ c_2,\ c_8,\ c_9,\ c_{10}$
3	$c_1,\ c_4,\ c_5,\ c_8,\ c_{10}$
4	$c_1,\ c_4,\ c_6,\ c_7,\ c_9$
5	$c_2,\ c_4,\ c_5,\ c_7,\ c_{10}$
6	$c_3,\ c_5,\ c_6,\ c_7,\ c_8$
7	$c_1,\ c_2,\ c_3,\ c_5,\ c_9,\ c_{10}$
8	$c_1,\ c_2,\ c_6,\ c_8,\ c_9,\ c_{10}$
9	$c_1,\ c_3,\ c_4,\ c_5,\ c_6,\ c_7$
10	$c_1,\ c_3,\ c_4,\ c_5,\ c_6,\ c_{10}$
11	$c_1,\ c_3,\ c_4,\ c_6,\ c_9,\ c_{10}$
12	$c_2,\ c_3,\ c_4,\ c_7,\ c_8,\ c_9$
13	$c_2,\ c_3,\ c_7,\ c_8,\ c_9,\ c_{10}$
14	$c_2,\ c_4,\ c_6,\ c_7,\ c_8,\ c_{10}$
15	$c_3,\ c_6,\ c_7,\ c_8,\ c_9,\ c_{10}$
16	$c_1,\ c_2,\ c_3,\ c_5,\ c_6,\ c_9,\ c_{10}$
17	$c_1,\ c_2,\ c_4,\ c_6,\ c_7,\ c_8,\ c_{10}$
18	$c_1,\ c_2,\ c_5,\ c_6,\ c_7,\ c_9,\ c_{10}$
19	$c_2,\ c_3,\ c_4,\ c_5,\ c_6,\ c_7,\ c_8$
20	$c_2,\ c_3,\ c_4,\ c_5,\ c_6,\ c_9,\ c_{10}$

Firstly, we will select a reduct at random and suppose the random selected reduct is $reduct5 = \{c_2, c_4, c_5, c_7, c_{10}\}$. The information system of reduct5 is shown in Table 3.

Table 3
A information system constructed with a reduct.

U	c_2	c_4	c_5	c_7	c_{10}
x_1	1	2	2	2	1
x_2	1	1	1	2	2
x_3	2	2	2	2	1
x_4	1	1	1	1	1
x_5	1	2	1	2	2
x_6	2	2	2	2	2
x_7	1	2	2	1	1
x_8	2	1	2	2	1
x_9	1	2	2	2	1
x_{10}	1	1	1	2	2

Secondly, we calculate the dependence degree of each attribute on the other attributes in the random selected reduct by Eq. (7). The group of equivalence classes induced by singleton attribute and other attributes in reduct5 are given below.

(a) $X(c_2=1)=\{x_1,x_2,x_4,x_5,x_7,x_9,x_{1o}\}$, $X(c_2=2)=\{x_3,x_6,x_8\}$.

For the set of attributes $P_2=\{x_4,x_5,x_7,x_{1o}\}$, we have the following equivalence classes:

$\{x_1,x_3,x_9\},\{x_2,x_{1o}\},\{x_4\},\{x_5\},\{x_6\},\{x_7\},\{x_8\}$.

(b) $X(c_4=1)=\{x_2,x_4,x_8,x_{1o}\}$, $X(c_4=2)=\{x_1,x_3,x_5,x_6,x_7,x_9\}$.

For the set of attributes $P_4=\{x_2,x_5,x_7,x_{1o}\}$, we have the following equivalence classes:

$\{x_1,x_9\},\{x_2,x_5,x_{10}\},\{x_3,x_8\},\{x_4\},\{x_6\},\{x_7\}$.

(c) $X(c_5=1)=\{x_2,x_4,x_5,x_{1o}\}$, $X(c_5=2)=\{x_1,x_3,x_6,x_7,x_8,x_9\}$.

For the set of attributes $P_5=\{x_2,x_4,x_7,x_{1o}\}$, we have the following equivalence classes:

$\{x_1,x_9\},\{x_2,x_{1o}\},\{x_3\},\{x_4\},\{x_5\},\{x_6\},\{x_7\},\{x_8\}$.

(d) $X(c_7=1)=\{x_4,x_7\}$, $X(c_7=2)=\{x_1,x_2,x_3,x_5,x_6,x_8,x_9,x_{1o}\}$.

For the set of attributes $P_7=\{x_2,x_4,x_5,x_{1o}\}$, we have the following equivalence classes:

$\{x_1,x_7,x_9\},\{x_2,x_{1o}\},\{x_3\},\{x_4\},\{x_5\},\{x_6\},\{x_8\}$.

(e) $X(c_{1o}=1)=\{x_1,x_3,x_4,x_7,x_8,x_9\}$, $X(c_{1o}=2)=\{x_2,x_5,x_6,x_{1o}\}$.

For the set of attributes $P_{10}=\{x_2,x_4,x_5,x_7\}$, we have the following equivalence classes:

$\{x_1,x_9\},\{x_2,x_{1o}\},\{x_3,x_6\},\{x_4\},\{x_5\},\{x_7\},\{x_8\}$.

Thirdly, obtain the lower and upper approximation of X based on attribute c_2 ($c_4, c_5, c_7,$ and c_{1o}) with respect to the rest attributes in $reduct5$. The lower approximation and upper approximation can be calculated using the formula in Eq. (4), which are shown as follow:

(a) c_2 with respect to P_2:

$$\underline{P_2}\left(X_{c_2=1}\right)=\{x_2,x_4,x_5,x_7,x_{1o}\},\ \overline{P_2}\left(X_{c_2=1}\right)=\{x_1,x_2,x_3,x_4,x_5,x_7,x_9,x_{1o}\},$$

$$\underline{P_2}\left(X_{c_2=2}\right)=\{x_6,x_8\},\qquad \overline{P_2}\left(X_{c_2=2}\right)=\{x_1,x_3,x_6,x_8,x_9\}.$$

(b) c_4 with respect to P_4:

$$\underline{P_4}\left(X_{c_4=1}\right)=\{x_4\},\qquad \overline{P_4}\left(X_{c_4=1}\right)=\{x_2,x_3,x_4,x_5,x_8,x_{1o}\},$$

$$\underline{P_4}\left(X_{c_4=2}\right)=\{x_1,x_6,x_7,x_9\},\ \overline{P_4}\left(X_{c_4=2}\right)=\{x_1,x_2,x_3,x_5,x_6,x_7,x_8,x_9,x_{1o}\}.$$

(c) c_5 with respect to P_5:

$$\underline{P_5}\left(X_{c_5=1}\right)=\{x_2,x_4,x_5,x_{1o}\},\qquad \overline{P_5}\left(X_{c_5=1}\right)=\{x_2,x_4,x_5,x_{1o}\},$$

$$\underline{P_5}\left(X_{c_5=2}\right)=\{x_1,x_3,x_6,x_7,x_8,x_9\},\ \overline{P_5}\left(X_{c_5=2}\right)=\{x_1,x_3,x_6,x_7,x_8,x_9\}.$$

(d) c_7 with respect to P_7:

$$\underline{P_7}\left(X_{c_7=1}\right)=\{x_4\},\qquad \overline{P_7}\left(X_{c_7=1}\right)=\{x_1,x_4,x_7,x_9\},$$

$$\underline{P_7}\left(X_{c_7=2}\right)=\{x_2,x_3,x_5,x_6,x_8,x_{1o}\},\ \overline{P_7}\left(X_{c_7=2}\right)=\{x_1,x_2,x_3,x_5,x_6,x_7,x_8,x_9,x_{1o}\}.$$

(e) c_{1o} with respect to P_{1o}:

$$\underline{P_{1o}}\left(X_{c_{1o}=1}\right)=\{x_1,x_4,x_7,x_8,x_9\},\ \overline{P_{1o}}\left(X_{c_{1o}=1}\right)=\{x_1,x_3,x_4,x_6,x_7,x_8,x_9\},$$

$$\underline{P_{1o}}\left(X_{c_{1o}=2}\right)=\{x_2,x_5,x_{1o}\},\qquad \overline{P_{1o}}\left(X_{c_{1o}=2}\right)=\{x_2,x_3,x_5,x_6,x_{1o}\}.$$

Fourthly, obtain the positive regions of X with respect to P by Eq. (4), which are given as follow:

(a) The positive region of the partition $X=U/c_2$ with respect to P_2,

$$POS_{p_2}(c_2)=\bigcup\underline{P_2}(X)=\{x_2,x_4,x_5,x_6,x_7,x_8,x_{1o}\}.$$

(b) The positive region of the partition $X=U/c_4$ with respect to P_4,

$$POS_{p_4}(c_4)=\bigcup\underline{P_4}(X)=\{x_1,x_4,x_6,x_7,x_9\}.$$

(c) The positive region of the partition $X = U/c_5$ with respect to P_5,

$$POS_{p_5}(c_5) = \bigcup \underline{P_5}(X) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{1o}\}.$$

(d) The positive region of the partition $X = U/c_7$ with respect to P_7,

$$POS_{p_7}(c_7) = \bigcup \underline{P_7}(X) = \{x_2, x_3, x_4, x_5, x_6, x_8, x_{1o}\}.$$

(e) The positive region of the partition $X = U/c_{1o}$ with respect to P_{1o},

$$POS_{P_{1o}}(c_{1o}) = \bigcup \underline{P_{1o}}(X) = \{x_1, x_2, x_4, x_5, x_7, x_9, x_{1o}\}.$$

Fifthly, the dependency degree of each attribute, which indicates the approximation power of other attributes in a reduct to this attribute, is calculated by Eq. (7) are shown as follow:

(a) The approximation power of P_2 to c_2 is measured by

$$r_2 = \frac{|POS_{P_2}(c_2)|}{|U|} = \frac{7}{10} = 0.7.$$

(b) The approximation power of P_4 to c_4 is measured by

$$r_4 = \frac{|POS_{P_4}(c_4)|}{|U|} = \frac{5}{10} = 0.5.$$

(c) The approximation power of P_5 to c_5 is measured by

$$r_5 = \frac{|POS_{P_5}(c_5)|}{|U|} = \frac{10}{10} = 1.$$

(d) The approximation power of P_7 to c_7 is measured by

$$r_7 = \frac{|POS_{P_7}(c_7)|}{|U|} = \frac{7}{10} = 0.7.$$

(e) The approximation power of P_{1o} to c_{1o} is measured by

$$r_{1o} = \frac{|POS_{P_{1o}}(c_{1o})|}{|U|} = \frac{7}{10} = 0.7.$$

Sixthly, we find out the attribute (i.e. c_5) with the maximal value of dependency degree,

which means other attribute (i.e. P_5) can relatively well express this attribute.

At last, reducts contained attribute c_5 will be deleted for the difference between selected reducts. The remained reducts are shown in Table 4. The next reduct will be selected from this temporary reduct pool.

Table 4
The remained reducts.

	Reducts
2	$c_1, c_2, c_8, c_9, c_{10}$
4	c_1, c_4, c_6, c_7, c_9
8	$c_1, c_2, c_6, c_8, c_9, c_{10}$
11	$c_1, c_3, c_4, c_6, c_9, c_{10}$
12	$c_2, c_3, c_4, c_7, c_8, c_9$
13	$c_2, c_3, c_7, c_8, c_9, c_{10}$
14	$c_2, c_4, c_6, c_7, c_8, c_{10}$
15	$c_3, c_6, c_7, c_8, c_9, c_{10}$
17	$c_1, c_2, c_4, c_6, c_7, c_8, c_{10}$

3.3 The accuracy-diversity assessment function

In this paper, we employ an assessment function [35], called the Accuracy-Diversity assessment function (AD assessment function), to choose appropriate reducts from the dynamic searching space of reducts. Reducts can be picked out as the training dataset of classifiers.

AD assessment function balances accuracy and diversity. Generally speaking, if base classifiers have lower generalization error and higher diversity among each other, the performance of ensemble system is better. However, having a low generalization error and a high diversity may face a dilemma. That is, if there are two classifiers and both of them have perfect classification result, they may make the same predictions. Therefore, there are no differences between these two classifiers. So, in order to get a better ensemble performance, we should balance the trade-off between accuracy and diversity. When we construct an ensemble system based on rough set theory, both accuracy of base classifiers and diversity among classifiers should be explicitly considered.

The basic idea of AD assessment function is to balance the diversity and accuracy of classifiers. The function is defined as follows:

$$AD(f,N)=1-A_{emp}(f,N)+\omega\times D_{div}(f,N). \tag{8}$$

Here, the second term $A_{emp}(f,N)$ corresponds to the empirical loss of base classifiers f; the third term $D_{div}(f,N)$ corresponds to the diversity among base classifiers f. ω is a cost parameter balancing the importance of the second term and third term. In this paper, the term $A_{emp}(f,N)$ in Eq. (8) is evaluated with l_2 loss:

$$A_{emp}(f,N)=\frac{1}{mN}\sum_{k=1}^{m}\sum_{i=1}^{N}[f_k(x_i)-y_i]^2, \tag{9}$$

where, m is the number of base classifiers, N is the number of test objects.

There is no widely accepted definition for diversity, however, some measures of diversity have been proposed. In [36], the relationship of four pairwise (Q statistic, Correlation coefficient, Disagreement measure and Double-fault measure) and six non-pairwise (Kohavi-Wolpert variance, Interrater agreement, Entropy measure, Measure of difficulty, Generalized diversity, and Coincident failure diversity) measures have been studied. The ten measures are deemed to have strong correlations between each other, and the experimental facts show that these measures can be categorized into three groups: the Coincident Failure Diversity (CFD), the Double Fault measure (DF) and all the remaining measures. Three typical measures DF, CFD and ENTropy measure (ENT) are chosen to measure the diversity among base classifiers depending on the analysis above.

The definition of the Coincident Failure Diversity (CFD) is

$$\begin{cases} CFD(f,N)=\begin{cases}0, & p_0=1.0\\ \dfrac{1}{1-p_0}\displaystyle\sum_{i=1}^{m}\frac{m-i}{m-1}p_i, & p_0<1.0\end{cases}\\ p_i=\dfrac{\displaystyle\sum_{j=1}^{N}1_{\left[i=\sum_{k=1}^{m}(1-O_{kj})\right]}}{N}, \quad (0\le i\le m)\end{cases}, \tag{10}$$

where $1_{[\pi]} = \begin{cases} 1, & if\ predicate\ \pi\ holds \\ 0, & otherwise \end{cases}$, the definition of $1_{[\pi]}$ is also used in [37]. Here, $O = [O_{kj}]_{m \times N}$ is the oracle output matrix. That is, if the k-th base classifier correctly classifies the j-th object, where $1 \le k \le m, 1 \le j \le N$, the value of $O_{kj} = 1$, otherwise, $O_{kj} = 0$.

The Double-Fault measure (DF) is determined by

$$\begin{cases} DF(f,N) = \dfrac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} df_{ik} \\ df_{ik} = \dfrac{\sum_{j=1}^{N} (1 - O_{ij}) \cdot (1 - O_{kj})}{N} \end{cases}. \tag{11}$$

We usually use $1 - DF$ to instead of DF, the higher value of $1 - DF$, the more diverse among base classifiers.

The ENTropy measure (ENT) which is defined as follows:

$$ENT = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{m - \lceil m/2 \rceil} \min \left\{ \sum_{i=1}^{m} o_{ij}, m - \sum_{i=1}^{m} o_{ij} \right\}, \tag{12}$$

where $\lceil \bullet \rceil$ means the operation of rounding up to an integer. For example, $\lceil 5/2 \rceil$ returns the value 3 in terms of this operation.

From the above formulas, we can see that base classifier with higher accuracy is represented by the lower value of the second term, and larger value of $D_{div}(f,N)$ implies higher diversity among base classifiers. In order to keep the consistency of the second term that measures the accuracy and the third term which measures the diversity, $A_{emp}(f,N)$ is replaced by $1 - A_{emp}(f,N)$. Then, the classifiers which have the higher AD function value will be selected to construct an ensemble system.

The parameter ω is used to balance the relationship of diversity and accuracy. We adaptively adjust the value of ω in the searching process [35]. Our purpose is to maximize the value of AD function, so the value of ω is not changed if AD is increasing. When

AD is not increasing, we increase ω if A_{emp} is not increasing and D_{div} is decreasing, we decrease ω if A_{emp} is increasing and D_{div} is not decreasing. That is, we adjust the parameter ω in deference to obtain A_{emp} and D_{div}, which determine to pay more attention to accuracy or diversity when we choose the next classifier. We set ω to 1 as the initial value, and set the changing amount of ω to 10% based on its current value.

3.3 DRSSE

Based on the concept of maximum dependency degree of attributes and accuracy-diverse assessment function, a novel algorithm, Dynamic Rough Subspace based Selective Ensemble (DRSSE), is proposed. DRSSE uses a new rough set based selective ensemble framework. In this new framework, the searching space of reducts is continually changing based on the selected reducts. Dynamic searching space formed by attribute with maximum dependency degree ensures that the deleted attribute has limited effect on the ensemble system, and the reducts in this dynamic searching space are more likely to differ from the selected reduct. The role of AD assessment function is to select proper reduct from the dynamic searching space. The DRSSE learning procedure is detailed in Algorithm 2.

Algorithm 2. Dynamic Rough Subspace based Selective Ensemble (DRSSE)

Input: a decision system $\langle U, C \cup D, f \rangle$, the amount of reducts required is M

Output: ensemble system

1. find multiple reducts RED using ROSSETTA software
2. $RED\_temp \leftarrow RED, R \leftarrow \phi, Cls \leftarrow \phi$
3. choose a reduct A_1 randomly from the temporary reducts pool RED_temp, $R \leftarrow \{R, A_1\}$, training a classifier f_1 with reduct A_1, $Cls \leftarrow \{Cls, f_1\}$
4. do
5. $A = A_j$
6. the same with the process 1-4 of Algorithm 1
7. produce a new temporary reducts pool RED_temp through deleting reducts red which contained the attribute a_m from original reducts pool RED

   $RED\_temp \leftarrow RED - red$
8. choose a new reduct A_j from RED_temp, $Cls \leftarrow \{Cls, f_j\}$, $R \leftarrow \{R, A_j\}$

   A_j satisfies: $AD(Cls, f_j) = \max_{A_k \in RED} (AD(Cls, f_k))$
9. until the number of reducts in R is equal to M
10. ensemble system is formed by classifiers trained with R

In summary, the framework of DRSSE considers not only the diversity of training subspace but also the accuracy and diversity of classifiers in an ensemble system. It is worthwhile to highlight some outstanding characteristics of DRSSE:

(1) A new rough set ensemble framework is proposed. The relationship among attributes in rough subspace is first considered in the framework of DRSSE. A number of studies show that the performance improvement of ensemble systems benefits from the difference of input attributes. Classifiers constructed with diverse attributes will increase their diversity.

(2) The attribute dimension is reduced. In most case, there are usually some redundant attributes. Training classifier without redundant and irrelevant attributes can significantly increase the generalization power of base classifiers and improve the performance of ensemble.

(3) High quality attribute subsets are used. According to the theory of rough set, a reduct is the essential part of the original dataset and has the same discernible power. DRSSE trains classifiers by using reducts that should have better performance than the algorithms using randomly selected attributes.

(4) The concept of attribute with maximum dependency degree is first used to change the searching space of reducts. An assessment function, called accuracy-diversity, is also used to select suitable reducts based on the balance of accuracy and diversity.

4. Experimental results and analysis

In order to evaluate the performance of the proposed algorithm, we validate the proposed DRSSE method on ten datasets from UCI machine learning data repository [38], three face databases (FERET, ORL and CMU PIE) and a handwritten database (MNIST). The detailed description of these datasets is shown in Table 5. In our experiments, the original dataset is split into two parts, i.e., training dataset and test dataset. The proportion of training set is 50% of all samples, and the rest is the test set.

Table 5
Description of experimental datasets.

Dataset	Sample	Attributes	Classes
air	359	65	3
australian	690	15	2
dermatology	366	35	6
german	1000	25	2
heart	270	14	2
ionosphere	351	35	2
sonar	208	61	2
vowel	528	11	11
wdbc	569	31	2
wine	178	13	3

It is clear that selecting different classifiers may lead to dissimilar effects on the results of the proposed algorithm. Thus, we select two different base classifiers, i.e., CART (Classification and Regression Trees) [39] and SVM (Support Vector Machine) [40]. The ensemble system is composed of base classifiers which are trained with different feature subsets (reducts). The combination strategy for ensemble is majority vote. The training method is run 20 times for each setting and we compare the averages. The size of ensemble should be specified in advance in DRSSE algorithm.

The size of UCI datasets ranges between 178 and 1000, and the number of attributes varies from 11 to 65. The information about the datasets is shown in Table 5. FERET face database [41] consists of 1400 face images of 200 individuals, which were taken under varying pose, illumination and expression. The resolution of each face image is cropped to 32×32. ORL face database [42] have 400 face images of 40 individuals. Each of the face images contains 32×32 pixels. As in [43], for CUM PIE face dataset [44], we choose the frontal pose (C27) with varying lighting, which leaves us 3329 face images of 68 individuals. In our experiments,

the cropped images of size 32×32 are used. MNIST handwritten digits database [45] contains 70,000 images, and each image consists of 28×28 pixels. Fig.2 shows some images from these three face databases and the MNIST database.

Fig.2. (a) all faces of the first person in the FERET database, (b) all faces of the first person in the ORL database, (c) partial faces of the first person in the PIE database, and (d) some images from digits 0 through 9 in the MNIST database.

As introduced in Algorithm 2, the reducts are produced by the ROSSETTA software. The results of UCI dataset are shown in Table 6. Some examples of the obtained reducts with a different number of attributes are given in the second column of Table 6. The third column and the fourth column present the range of the number of attributes in reducts and the total number of generated reducts, respectively. From Table 6 we can see that: (1) There is no clear link between different reducts of the same dataset. (2) The number of attributes of generated reducts is less than half of the number of original attributes in most cases. (3) A large number of reducts exist in one dataset, e.g. air and sonar have more than 200 reducts.

Table 6
The detail of reduction results of UCI dataset.

Dataset	some reducts with different size	the number of attributes in reducts	The number of reducts
air	17, 26, 27, 31, 34, 37, 40, 45, 46 1, 11, 18, 19, 24, 28, 31, 33, 41, 57, 62 1, 11, 19, 23, 24, 28, 30, 31, 33, 34, 35, 41, 49	9-13	210
australian	2, 3, 7 2, 7, 8, 14 1, 2, 5, 6, 10	3-5	28
dermatology	1, 3, 4, 14, 16, 34 1, 3, 4, 11, 25, 28, 32, 34 1, 3, 4, 7, 12, 20, 24, 28, 32, 34	6-10	172
german	2, 4, 7, 10, 16 1, 2, 4, 6, 10, 12, 20 2, 4, 6, 7, 10, 15, 21, 22	5-8	90
heart	1, 8, 10 3, 5, 7, 10 1, 3, 7, 9, 10	3-5	66
ionosphere	4, 6, 10, 11, 13, 14, 18, 20, 23, 24, 27 1, 4, 6, 9, 11, 13, 15, 16, 17, 22, 23, 24, 25, 26, 30 1, 3, 4, 7, 8 , 11, 12, 13, 14, 18, 19, 20, 21, 26, 28, 30, 34	11-17	193
sonar	1, 3, 12, 20, 27, 48, 55 1, 5, 6, 9, 12, 24, 40, 59 9, 12, 21, 25, 26, 38, 56, 58, 60	7-9	245
vowel	7, 9	2	44
wdbc	2, 13, 16, 18, 21, 22, 25, 27 2, 5, 10, 11, 14, 17, 24, 26, 27, 28 1, 5, 7, 9, 10, 11, 14, 17, 19, 20, 24, 25, 26, 30	8-14	197
wine	4, 6, 9, 10, 12 1, 3, 4, 6, 10, 12 1, 3, 4, 6, 7, 8, 9	5-7	94

The produced reducts of these face databases and MNIST database are described in Table 7. Min and Max indicates the minimal and maximum number of attribute that the produced reducts possess. The fourth column shows the number of attributes in original database. The fifth column to the eighth column shows the number of reducts when the range of the number of attributes in reducts is given. The last column presents the total number of the produced reducts. Table 7 shows that: (1) Hundreds of reducts can be produced by these face databases and MNIST database. (2) The number of attributes of original dataset can be down by about a

third to a half.

Our experiments consist of four parts. In Section 4.1, we first show the performance of base classifiers trained with different reducts. In Section 4.2, we present the classification accuracies of DRSSE with three different diversity measures and determine which measure should be selected for DRSSE. The impact of the number of classifiers on the result of DRSSE is presented in Section 4.3. Finally, we compare DRSSE with other available ensemble methods in Section 4.4.

Table 7
The detail of reduction results of UCI dataset.

Database	Min	Max	Original	Min-400	400-500	500-600	600-Max	Sum
FERET	312	651	1024	39	41	42	15	137
ORL	309	632	1024	33	40	40	20	133
PIE	332	660	1024	37	39	45	14	135
MNIST	267	477	784	107	21	0	0	128

4.1 The performance of reducts

CART and SVM are introduced to train base classifiers. Fig.3 evaluates the performance of classifiers trained with reducts which are produced by the ten UCI datasets. Box plot is used to show the results, which represents the lower quartile, median, and upper quartile values in lines. The x-axis represents the number of attributes in reducts and the last column is the total number of attributes in dataset. The y-axis is the accuracies of classifiers. In Fig.3, the left column and the right column present the classification results acquired by using CART and SVM as the base classifiers, respectively. It can be seen from Fig.3, firstly, the classifiers trained with reducts have much more opportunity to achieve even better (for CART, ionosphere, sonar, wdbc and wine; for SVM, sonar and wine) or comparative (for CART,

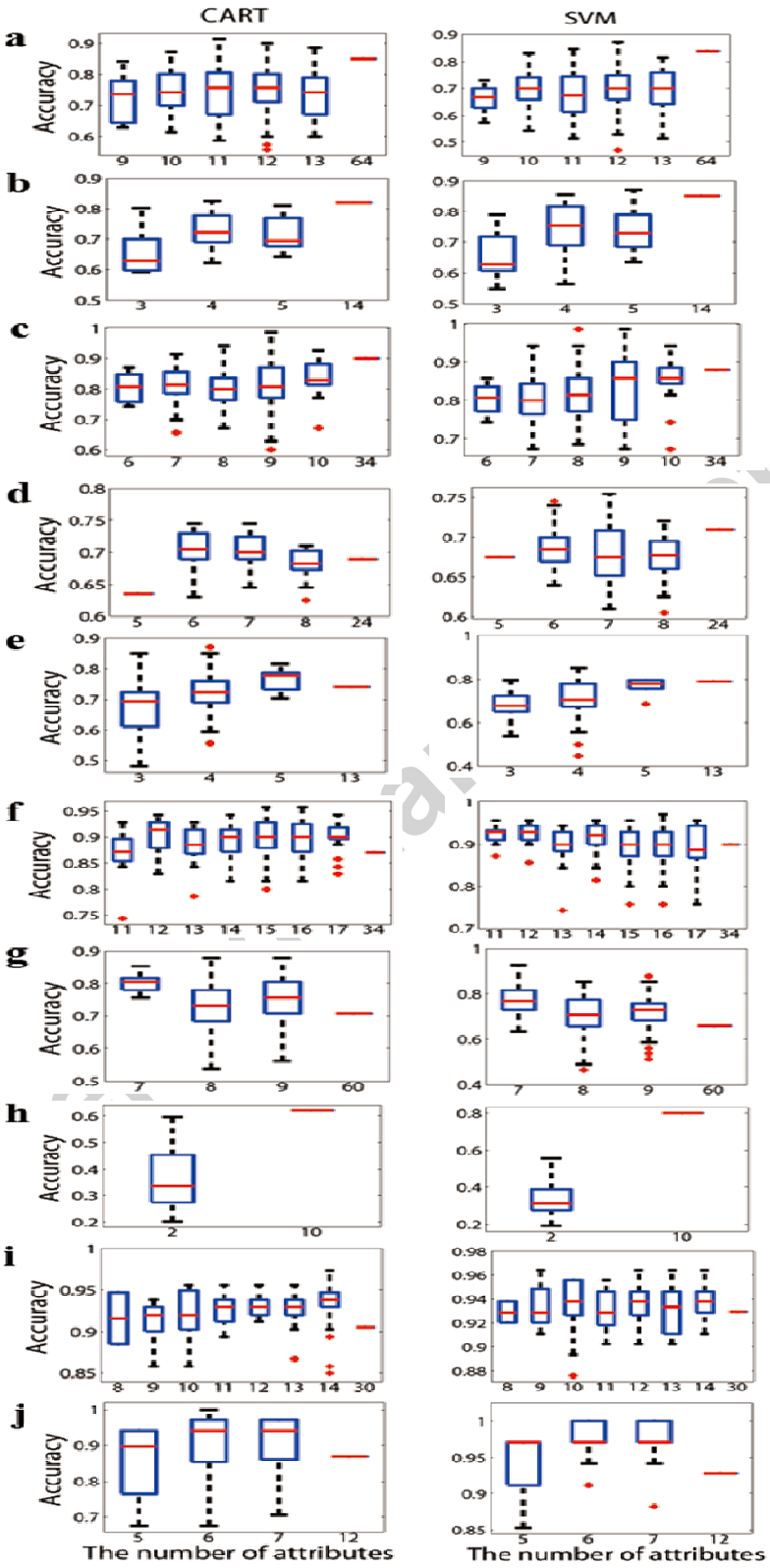


Fig.3. Box plots of the classification accuracies of reducts. (a) air. (b) australian. (c) dermatology. (d) german. (e) heart. (f) ionosphere. (g) sonar. (h) vowel. (i) wdbc. (j) wine.

german and heart; for SVM, dermatology, heart, ionosphere and wdbc) performance than that trained with original dataset. Secondly, as for vowel dataset, performance of the classifiers trained with reducts is poorer than that trained with the original dataset. This is because the number of attributes in vowel's reduct is only two, which is too few to reflect the original dataset's information.

Table 8 shows the classification results of reducts generated by FERET, ORL, PIE and MNIST databases. Because the number of attributes in these reducts has a wide range, we divide the reducts into several groups according to the quantity of attributes in these reducts. Min and Max denote the minimal and the maximum amount of attributes, respectively. The second to the fifth columns show the average accuracy of classifiers trained by the reducts in the corresonding interval. The last column is the results acquired by the original database. CART and SVM are the two base classifiers. Note that, the symbol "--" means the value is non-exist, this is because the maximum attribute number in the reducts generated by MNIST database is 477. It can be concluded from Table 8 that: (1) For PIE and MNIST, with CART as the base classifier, the classification results using reducts are slightly poorer than that using original data. However, for FERET and ORL, the results are not satisfied. (2) When SVM is used as the base classifier, the classifiers trained by the reducts can reach a comparative performance than that trained with the original data. (3) For the three face dataset, the results in the interval [600-Max] are slightly better than that in the interval [Min-400].

Table 8
The accuracies of classifiers trained by reducts.(Mean % ± Std %)

Dataset	Min-400		400-500		500-600		600-Max		Original	
	CART	SVM	CART	SVM	CART	SVM	CART	SVM	CART	SVM
FERET	12.03 ± 2.1	23.54 ± 1.9	13.13 ± 1.6	24.32 ± 1.0	14.99 ± 1.1	24.31 ± 0.7	14.97 ± 1.7	24.45 ± 0.9	18.50 ± 1.5	24.58 ± 1.5
ORL	42.17 ± 3.6	77.10 ± 0.9	41.97 ± 2.8	75.60 ± 1.9	43.75 ± 3.2	77.56 ± 1.2	43.27 ± 2.1	78.42 ± 1.3	52.75 ± 4.5	77.05 ± 2.8
PIE	60.44 ± 1.3	61.07 ± 1.9	61.31 ± 1.7	63.14 ± 1.4	61.17 ± 1.8	62.38 ± 1.0	61.89 ± 1.4	62.51 ± 0.9	62.26 ± 1.4	62.74 ± 2.4
MNIST	84.89 ± 1.1	91.22 ± 2.3	85.05 ± 0.8	92.42 ± 1.9	--	--	--	--	85.51 ± 2.8	93.46 ± 1.5

4.2 DRSSE with different diversity measures

In DRSSE, we use CFD, DF and ENT to measure diversity among base classifiers. DRSSE with the three measurements are simply described as DRSSE_CFD, DRSSE_DF and DRSSE_ENT, respectively. Fig.4 (CART is used as the base classifier) and Fig.5 (SVM is used as the base classifier) show the performance of DRSSE with different diversity measurements on different datasets. Note that, the number 1 to 14 in the x-axis of the two figure denote the datasets air, australian, dermatology, german, heart, ionosphere, sonar, vowel, wdbc, wine, FERET, ORL, PIE and MNIST, respectively. From Fig.4 and Fig.5 we can observe the following: (1) There are differences in the performance of DRSSE with different measurements. (2) The performance of DRSSE_DF is relatively good, which achieves the best classification accuracy in 19 out of 28 cases. (3) For all these dataset, the accuracy of DRSSE_DF is superior to DRSSE_CFD. For DRSSE with CART, the maximum difference of classification accuracy for the two measures ranges from 0.29% (sonar) to 2.76% (australian); for DRSSE with SVM, the accuracy difference of DRSSE_DF and DRSSE_CFD ranges from 0.37% (PIE) to 3.05% (FERET). (4) DRSSE_DF and DRSSE_ENT have merits and demerits on different datasets. However, on the whole, the performance of DRSSE_DF is slightly better than DRSSE_ENT.

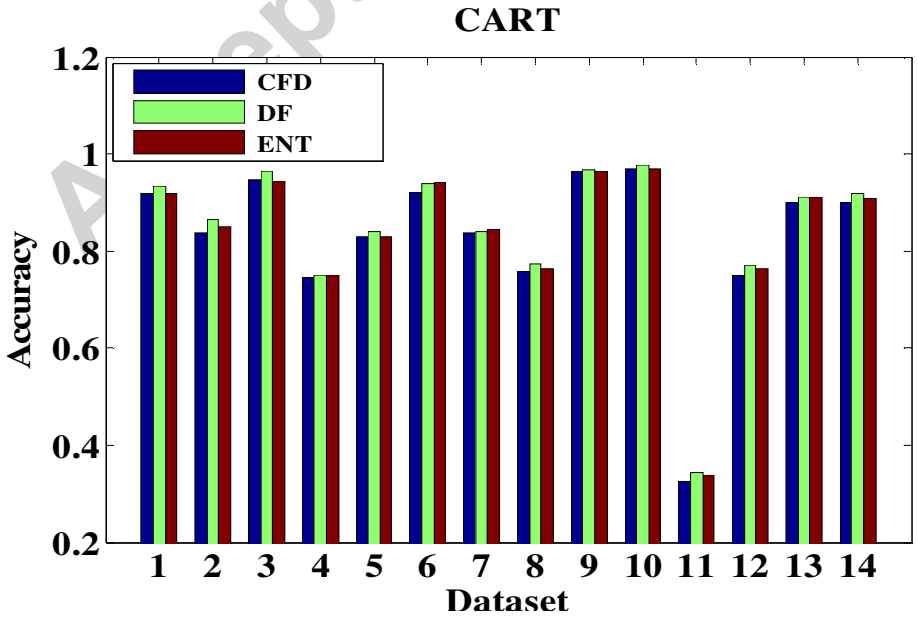


Fig.4. The accuracies of DRSSE with different diversity measurements (base classifier is CART).

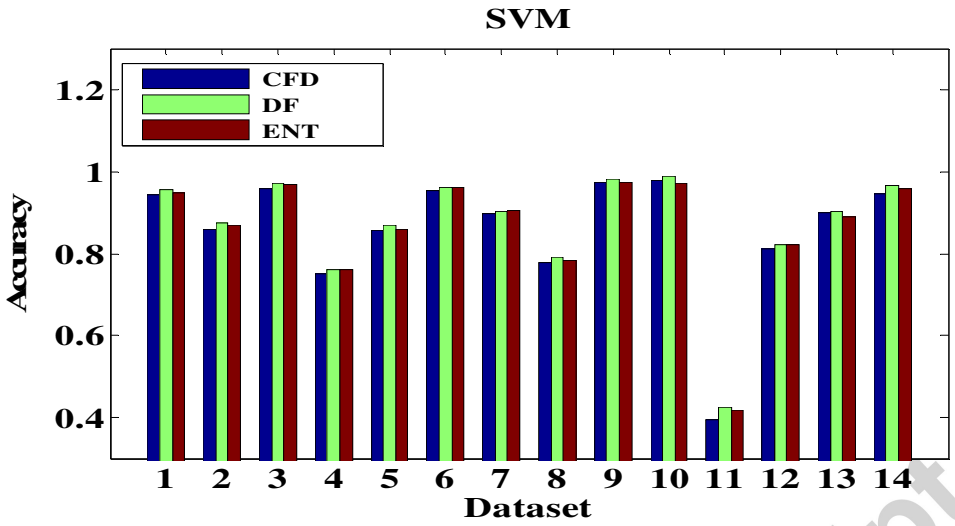


Fig.5. The accuracies of DRSSE with different diversity measurements (base classifier is SVM).

4.3 The impact of the size of ensemble on the performance of DRSSE

In the above experiments, for the UCI dataset, the number of classifiers in DRSSE is set to 20, and the number of classifiers is set to 25 for the three face databases and MNIST. Here, we investigate the influence of the number of base classifiers on the performance of DRSSE. We investigate the classification accuracy of DRSSE with DF measurement when the size of ensemble is set to 5, 10, 15, 20, 25, and 30 respectively. Note that, for australian, because the total number of reducts is 28, so the maximum size of ensemble is set to 25. The experimental results are shown in Fig.6 and Fig.7. The left column in both figures shows the results of DRSSE with CART as the base classifier, and the right column represents the results of DRSSE with SVM as the base classifier. From Fig. 6 and Fig.7, the following conclusions can be summarized.

(1) For all datasets, the performance of DRSSE with CART and SVM are influenced by the number of base classifiers.

(2) In most case, as the number of base classifiers increases, the accuracy of ensemble system goes up at first, and reaches its culmination, then decreases or retains. However, the dataset vowel is an exception. The accuracy curves of vowel in Fig.6 (h) are continually rising when the number of classifiers is increasing. This is because the attribute amount in

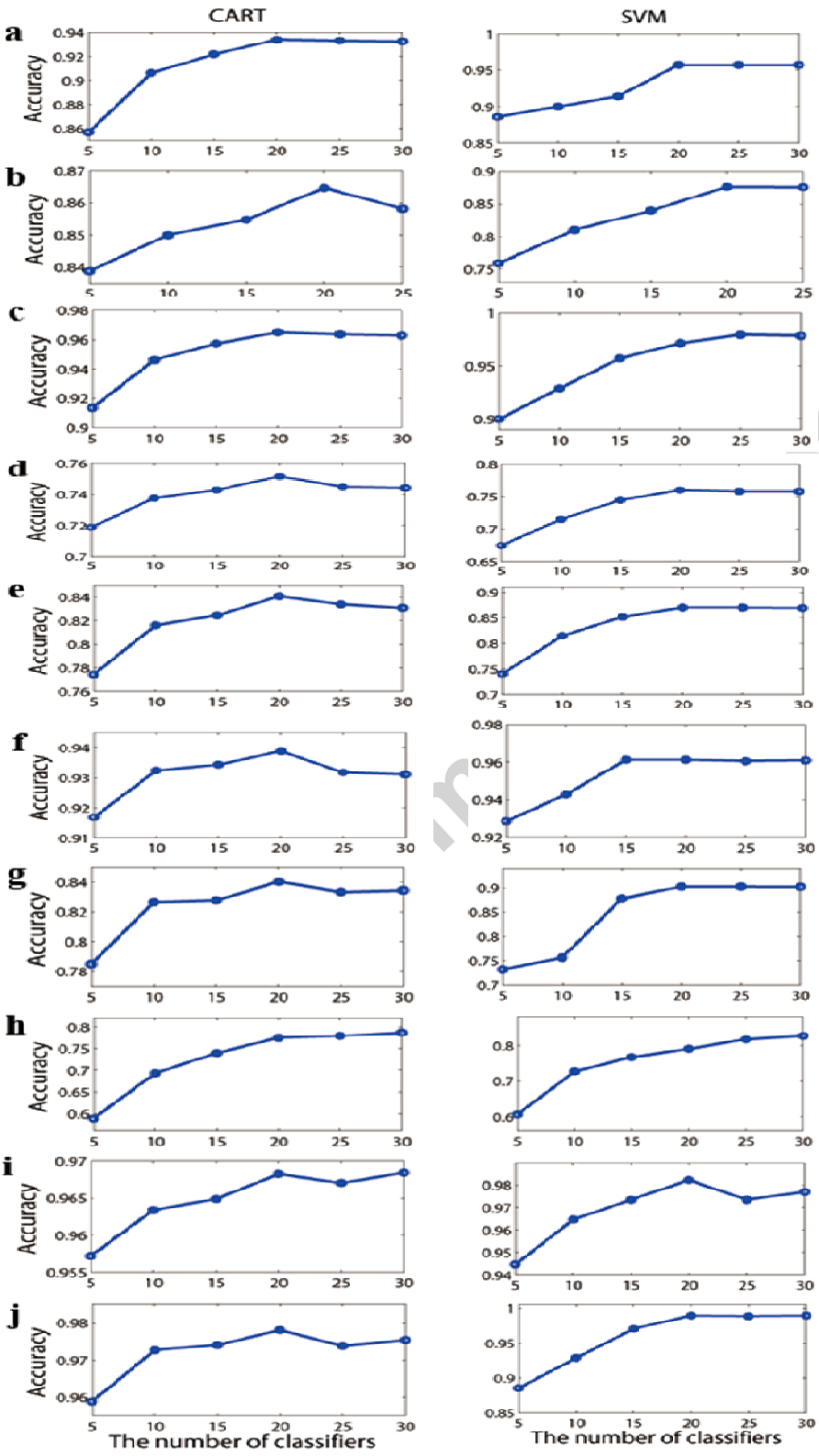


Fig.6. The classification accuracy of DRSSE versus the number of classifiers on different datasets. (a) air. (b) australian. (c) dermatology. (d) german. (e) heart. (f) ionosphere. (g) sonar. (h) vowel. (i) wdbc. (j) wine.

each single vowel's reduct is small, so we have to use more reducts to capture the original dataset's information.

(3) For the ten datasets from UCI, except for the dataset vowel and dermatology (DRSSE with SVM), DRSSE can obtain a well performance when the size of ensemble is 20. Therefore, in order to improve the efficiency of DRSSE, we set the ensemble size as 20 when deal with the UCI datasets.

(4) For ORL dataset, DRSSE with CART can reach a better performance when the size of the ensemble system is 20. However, for other seven results shown in Fig.7, when the number of classifiers in the ensemble system is set to 25, the classification performance is much better. Therefore, it is appropriate to choose the size of the ensemble system as 25.

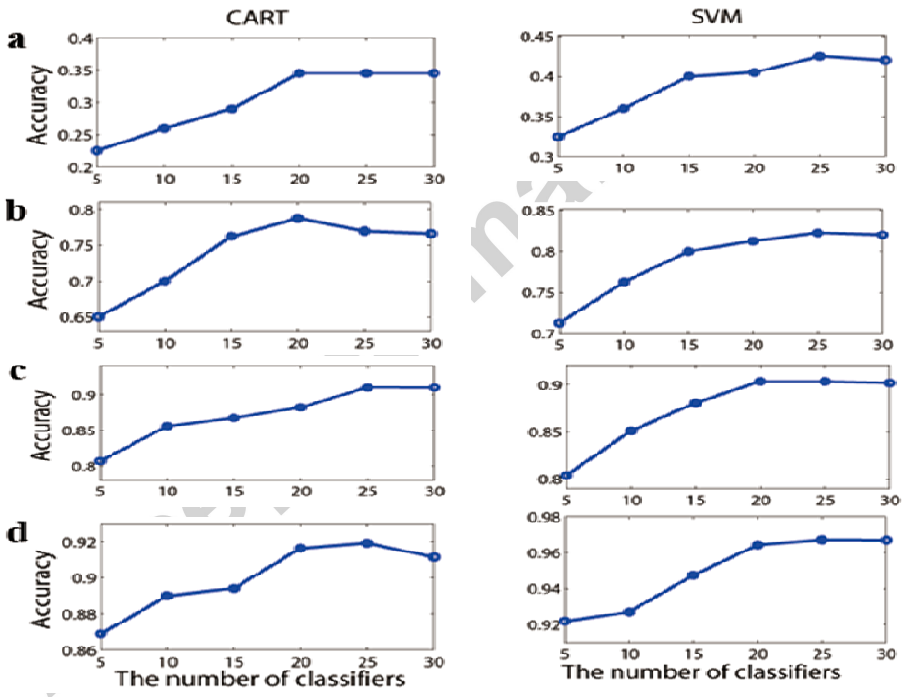


Fig.7. The classification accuracy of DRSSE versus the number of classifiers on different datasets. (a) FERET. (b) ORL. (c) PIE. (d) MNIST.

The average accuracy of DRSSE on all UCI datasets (without dataset australian) with the size variation of ensemble is shown in Fig.8. For australian, the maximum number of reducts is not come up to 30, so when we compute the average value, the dataset is not considered. The average performance of DRSSE on FERET, ORL, PIE and MNIST with the changed ensemble size is presented in Fig.9.

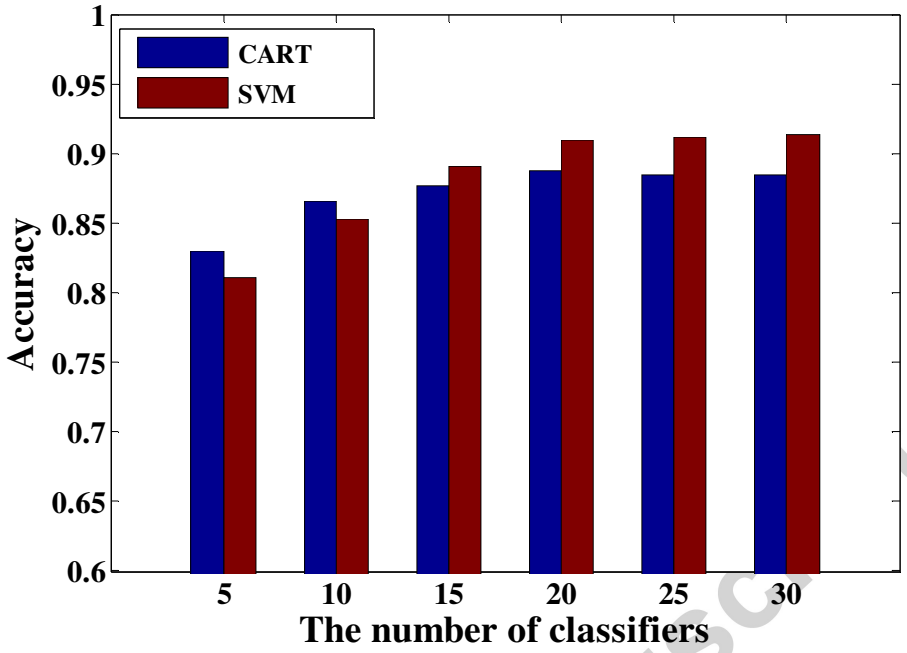


Fig.8. The average performance of DRSSE on the UCI datasets versus the number of classifiers.

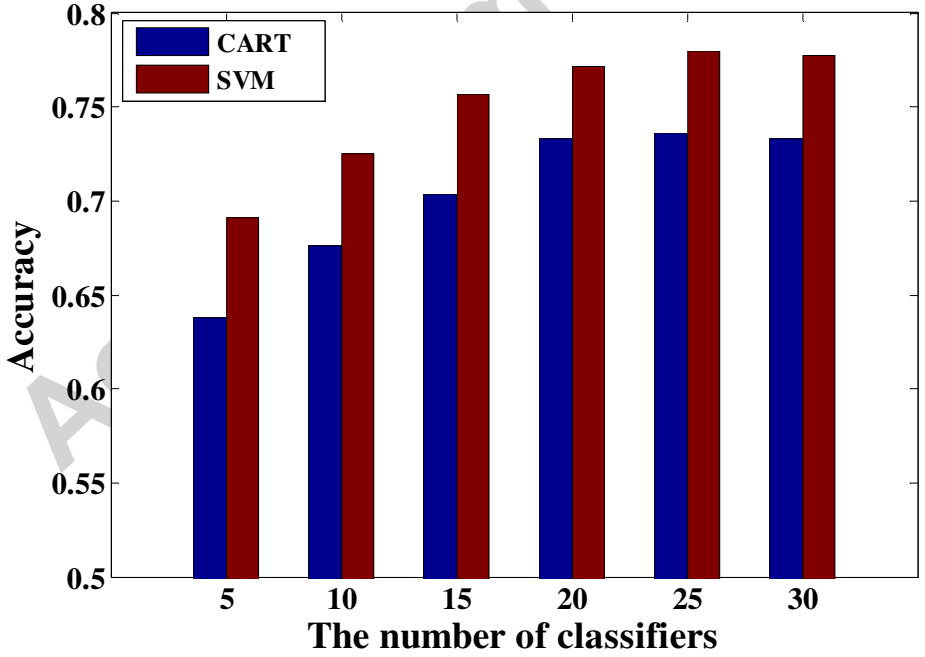


Fig.9. The average performance of DRSSE on FERET, ORL, PIE and MNIST versus the number of classifiers.

From Fig.8 and Fig.9, it is clear that as the ensemble size increases, the accuracy is increasing at first, then drops down or retains. Fig.8 denotes that when the ensemble size is small (eg. 5 and 10), DRSSE with CART has a better performance than DRSSE with SVM on the UCI datasets. In addition, the best results are acquired when the size of ensemble is 20. For the face datasets and MNIST, the number of the classifiers can be set to 25 according to the results shown in Fig.9.

4.4 Comparison with other ensemble methods

Adopted by UCI datasets, with CART as the base classifier, we first validate the effectiveness of both using the reduced searching space and AD assessment function. The experimental results are shown in Table 9.

Table 9
Comparison of proposed algorithm with random select method and SREMD. The best result for each dataset is in bold.

Dataset	RS	SREMD	DRSSE
air	84.57 ± 4.4	87.16 ± 4.3	**93.39 ± 3.6**
australian	82.41 ± 3.2	82.28 ± 1.9	**86.46 ± 2.9**
dermatology	91.71 ± 3.4	91.37 ± 1.9	**96.51 ± 1.7**
german	71.50 ± 3.4	71.83 ± 2.0	**75.15 ± 2.3**
heart	75.74 ± 6.0	79.03 ± 3.9	**84.07 ± 3.8**
ionosphere	90.14 ± 3.4	92.10 ± 2.7	**93.89 ± 3.5**
sonar	73.90 ± 8.2	78.36 ± 5.2	**84.01 ± 4.5**
vowel	70.40 ± 4.5	66.78 ± 3.3	**77.50 ± 3.4**
wdbc	93.10 ± 2.3	94.90 ± 2.1	**96.82 ± 1.0**
wine	91.18 ± 5.6	94.62 ± 4.6	**97.81 ± 2.4**
Average	82.47 ± 4.4	83.84 ± 3.2	**88.56 ± 2.9**

RS denotes the Random Select reducts from the original reducts pool that are used to train base classifiers. SREMD denotes the Select Reducts Ensemble based on attribute with Maximum Dependency degree. The ensemble size in RS and SREMD is also set to 20. The results shown in Table 9 imply the following two points. Firstly, both RS and SREMD select reducts randomly. The performance of selecting reducts from a reduced searching space is better than that choosing reducts from the original searching space. This can be attributed to

the fact that SREMD deletes part of reducts from the original reducts pool based on the selected reduct. This procudure not only shrinks the reduct pool but also potentially increases the diversity between the next selecting and the selected reducts. In other words, the SREMD algorithm deletes reducts associated with the selected reducts and retains reducts that are different from the selected reducts. Therefore, the final selected reducts are different. Secondly, choosing reducts from a reduced searching space based on AD assessment function can receive a better performance. Classifiers trained by the selected reducts in DRSSE have a good accuracy and diversity with each other.

Table 10
Comparison of classification accuracies of different algorithms with CART. The best result for each dataset is in bold. (Mean % ± Std %)

Dataset	TCV	ALL	Bagging	AdaBoost	RS	EROS	DESCD	DRSSE
air	85.81 ± 6.2	87.29 ± 4.9	82.68 ± 6.6	88.83 ± 6.5	86.03 ± 5.9	89.14 ± 4.1	87.91 ± 3.5	**93.39 ± 3.6**
australian	82.45 ± 4.6	84.45 ± 4.0	85.22 ± 4.5	84.93 ± 5.6	79.42 ± 4.1	85.11 ± 3.8	84.38 ± 5.9	**86.46 ± 2.9**
dermatology	90.04 ± 6.8	92.86 ± 3.2	94.97 ± 3.5	90.50 ± 5.0	91.06 ± 6.7	95.86 ± 2.7	91.43 ± 2.8	**96.51 ± 1.7**
german	69.30 ± 3.7	72.50 ± 1.5	73.00 ± 3.5	72.40 ± 3.3	71.40 ± 3.3	72.65 ± 3.6	73.14 ± 4.3	**75.15 ± 2.3**
heart	74.44 ± 8.9	77.96 ± 2.4	82.22 ± 5.3	81.48 ± 8.1	79.60 ± 7.4	78.33 ± 6.8	83.00 ± 3.1	**84.07 ± 3.8**
ionosphere	87.26 ± 6.9	91.29 ± 3.5	85.57 ± 6.4	88.57 ± 4.7	89.14 ± 7.0	92.86 ± 3.5	89.95 ± 4.2	**93.89 ± 3.5**
sonar	70.74±11.5	74.15 ± 5.9	74.04 ± 8.5	78.85 ± 7.7	75.96 ± 9.4	77.07 ± 6.0	80.44 ± 5.6	**84.01 ± 4.5**
vowel	62.40 ± 4.6	78.08 ± 4.5	72.73 ± 6.1	65.45 ± 3.4	64.24 ± 3.7	75.86 ± 4.1	**79.02 ± 3.9**	77.50 ± 3.4
wdbc	90.50 ± 4.6	94.96 ± 2.1	94.72 ± 4.4	95.07 ± 3.9	95.07 ± 4.3	95.96 ± 1.8	94.20 ± 3.7	**96.82 ± 1.0**
wine	86.94 ± 7.9	92.35 ± 6.1	88.76 ± 7.1	94.38 ± 3.7	89.89 ± 8.2	93.47 ± 5.9	95.35 ± 4.4	**97.81 ± 2.4**
FERET	18.50 ± 1.5	33.91 ± 2.5	31.02 ± 2.2	34.10 ± 2.9	30.77 ± 3.2	33.30 ± 1.8	32.15 ± 3.9	**34.50 ± 2.3**
ORL	52.75 ± 4.5	67.63 ± 5.3	69.38 ± 4.7	71.33 ± 4.3	68.91 ± 4.9	72.75 ± 5.3	74.68 ± 4.1	**78.75 ± 3.9**
PIE	62.26 ± 1.4	86.01±1.9	87.32 ± 1.2	88.64 ± 1.7	86.99 ± 2.3	89.93 ± 3.2	87.26 ± 2.8	**91.01 ± 2.8**
MNIST	85.51 ± 3.2	86.14 ± 2.8	85.47 ± 2.9	87.32 ± 3.3	84.28 ± 3.8	87.46 ± 4.6	86.59 ± 2.7	**91.93 ± 3.2**

The comparison results among our proposed DRSSE method, some other classical ensemble learning methods and the rough set ensemble methods are shown in Table 10 and Table 11. The two tables give the classification accuracy comparisons with CART and SVM, respectively. The size of the ensemble system is set to 20 for UCI datasets and the size of the ensemble system is set to 25 for the other four datasets. The compared methods include the Ten-fold Cross Validation (TCV), integrating ALL the base classifiers (ALL), Bagging,

AdaBoost, Random Subspace (RS), EROS proposed by Hu [27], and DES_CD proposed in [46].

Table 11
Comparison of classification accuracies of different algorithms with SVM. The best result for each dataset is in bold. (Mean % ± Std %)

Dataset	TCV	ALL	Bagging	AdaBoost	RS	EROS	DESCD	DRSSE
air	83.49 ± 4.3	86.23 ± 3.2	82.99 ± 4.5	87.64 ± 4.8	83.47 ± 3.6	90.51 ± 3.6	91.03 ± 4.2	**95.71 ± 3.9**
australian	85.53 ± 3.9	85.90 ± 4.6	86.32 ± 3.2	85.81 ± 5.3	84.66 ± 3.7	86.26 ± 4.2	85.41 ± 3.8	**87.59 ±4.1**
dermatology	88.96 ± 5.9	84.62 ± 3.7	90.12 ± 4.5	88.74 ± 5.4	91.43 ± 6.7	92.17 ± 6.3	90.23 ± 5.4	**97.14 ± 6.1**
german	70.33 ± 3.2	70.05 ± 1.7	73.76 ± 3.6	72.54 ± 2.9	71.45 ± 3.4	73.00 ± 2.5	74.69 ± 3.9	**76.15 ± 2.8**
heart	79.43 ± 4.8	80.99 ± 2.4	84.23 ± 4.9	84.36 ± 4.2	82.68 ± 3.6	82.14 ± 4.0	84.13 ± 2.9	**87.04 ± 3.8**
ionosphere	90.11 ± 6.2	91.54 ± 3.8	90.46 ± 4.8	91.29 ± 6.7	90.24 ± 5.0	92.12 ± 4.8	90.41 ± 4.3	**96.14 ± 5.3**
sonar	66.29 ±7.4	76.10 ± 5.3	76.43 ± 6.9	82.13 ± 7.3	79.41 ± 5.8	80.03 ± 6.0	85.24 ± 4.6	**90.24 ± 6.7**
vowel	76.06 ± 3.5	78.61 ± 4.7	77.41 ± 4.1	74.34 ± 3.9	73.56 ± 4.2	77.62 ± 4.1	**84.85 ± 3.7**	81.13 ± 3.2
wdbc	93.00 ± 3.1	94.06 ± 2.9	93.87 ± 4.3	96.96 ± 3.8	95.47 ± 3.7	95.76 ± 4.7	94.66 ± 3.7	**98.23 ± 3.8**
wine	92.88 ± 4.9	95.77 ± 4.4	93.67 ± 3.5	95.28 ± 4.7	94.99 ± 3.2	96.19 ± 4.2	95.23 ± 5.0	**98.97 ± 3.2**
FERET	24.58 ± 2.1	25.75 ± 1.6	32.74 ± 2.6	33.42 ± 2.4	32.64 ± 3.2	31.95 ± 2.3	37.62 ± 3.4	**42.50 ± 1.9**
ORL	77.00 ± 4.5	79.33 ± 3.3	76.43 ± 4.1	77.02 ± 3.2	77.23 ± 4.9	80.21 ± 5.3	79.64 ± 3.8	**82.25 ± 4.0**
PIE	80.69 ± 2.9	84.97 ± 3.0	83.14 ± 3.7	85.07 ± 1.7	84.11 ± 3.2	85.25 ± 3.3	84.13 ± 2.7	**90.36 ± 2.4**
MNIST	93.46 ± 2.6	94.41 ± 1.3	93.27 ± 2.5	94.61 ± 3.3	93.61 ± 3.7	95.57 ± 3.5	93.21 ± 2.9	**96.71 ± 1.9**

The best result for each dataset is bolded. Both Table 10 and Table 11 demonstrate that the performance of our method is better than other methods on the same dataset in most of case. For vowel dataset, because each reduct contains only two attributes, so the ensemble performance by selecting part of reducts to train classifiers is poorer than DESCD which use all attributes to train classifiers. As presented in Table 10, compared with other algorithms (TCV, ALL, Bagging, AdaBoost, RS, EROS, DESCD), the maximum difference of accuracy is 28.75% (PIE), 11.12% (ORL), 10.71% (air), 12.05% (vowel), 13.26% (vowel), 6.94% (sonar), 5.48% (air), respectively. The corresponding minimum difference of classification accuracy is 4.01% (australian), -0.58% (vowel), 1.24% (australian), 0.4% (FERET), 1.75% (wdbc), 0.65% (dermatology), -1.52% (vowel), respectively. Note that, the minus sign denotes the performance of DRSSE is lower than the compared algorithm. From Table 10 and Table

11, we can see that DRSSE with SVM have a better performance than DRSSE with CART.

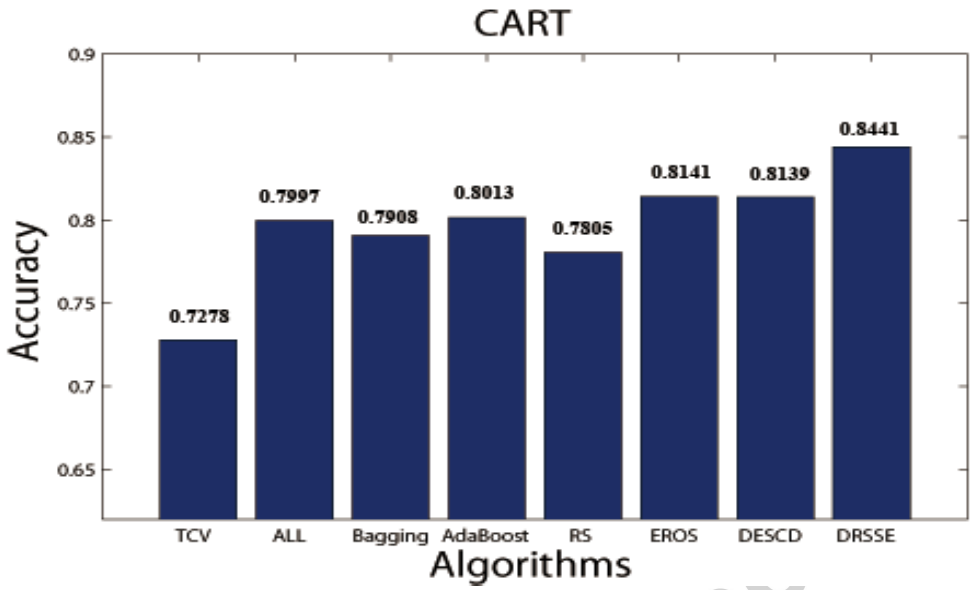


Fig.10. The average classification accuracy on all datasets (base classifier is CART).

The average classification accuracies on all datasets with CART and SVM are plotted in Fig.10 and Fig.11, respectively. Both of the two figures show that DRSSE can obtain the best average performance. From Fig.10, we can see that DRSSE outperforms TCV, ALL, Bagging, AdaBoost, RS, EROS, DESCD algorithm by 11.64%, 4.44%, 5.34%, 4.28%, 6.36%, 3.00%, 3.02%, respectively. As shown in Fig.11, the base classifier is SVM, compared with other algorithms, the maximum difference of average accuracy is 8.45% (DRSSE versus TCV), and the corresponding minimum difference is 3.55%.

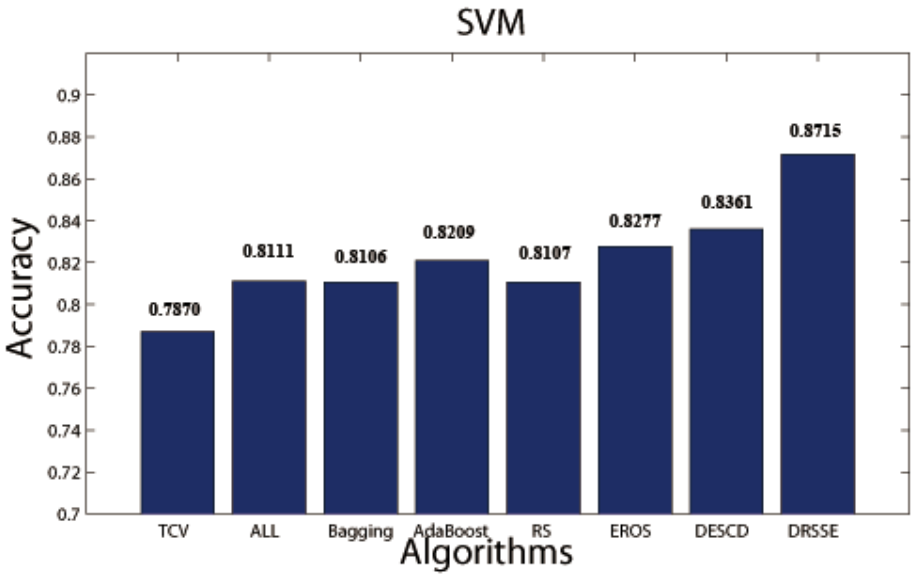


Fig.11. The average classification accuracy on all datasets (base classifier is SVM).

5. Conclusions and future work

In this paper, a new rough set based ensemble method is proposed. Instead of using all reducts to train base classifiers, we select diverse reducts from the reduct pool to train classifiers. First, a reduct is randomly selected from original reduct pool, and we narrow down the reduct searching space by deleting reducts containing attributes with the maximum dependency degree from the original reduct pool. Then, a new reduct is selected from the new searching space according to the AD assessment function. The AD function estimates both diversity between classifiers and accuracy of base classifiers. Finally, in each iteration, only one reduct is picked out, and our algorithm will not be terminated until the pre-defined size of ensemble is reached. Experimental results with a series of datasets demonstrate that DRSSE can lead to a better performance.

In the future, the rough set ensemble will be investigated further. The traditional rough set based ensemble methods are limited to supervised learning. However, most practical problems involve unlabeled data. Therefore, it is worth to explore semi-supervised rough set to solve ensemble learning problem.

Acknowledgements

This work was supported by the National Basic Research Program (973 Program) of China, No. 2013CB329402, the National Natural Science Foundation of China, No. 61003198, the Fund for Foreign Scholars in University Research and Teaching Programs(the 111 Project), No. B07048, and the Program for Cheung Kong Scholars and Innovative Research Team in University, No. IRT1170.

References

[1] Dietterich T G. Machine-learning research [J]. AI magazine, 1997, 18(4): 97.

[2] Schapire R E. The boosting approach to machine learning: An overview [J]. LECTURE NOTES IN STATISTICS-NEW YORK-SPRINGER VERLAG-, 2003: 149-172.

[3] Breiman L. Bagging predictors [J]. Machine learning, 1996, 24(2): 123-140.

[4] Kuncheva L I, Skurichina M, Duin R P W. An experimental study on diversity for bagging and boosting with linear classifiers [J]. Information fusion, 2002, 3(4): 245-258.

[5] Brown G, Wyatt J, Harris R, et al. Diversity creation methods: a survey and categorization [J]. Information Fusion, 2005, 6(1): 5-20.

[6] Zhou Z H, Wu J, Tang W. Ensembling neural networks: many could be better than all [J]. Artificial intelligence, 2002, 137(1): 239-263.

[7] Pawlak Z. Rough sets [J]. International Journal of Parallel Programming, 1982, 11(5): 341-356.

[8] Pawlak Z, Wong S K, Ziarko W. Rough sets: Probabilistic versus deterministic approach [J]. International Journal of Man-Machine Studies, 1988.

[9] Pawlak Z. Rough set theory and its applications[J]. Journal of telecommunications and information technology, 2002: 7-10.

[10] Kaneiwa K, Kudo Y. A sequential pattern mining algorithm using rough set theory [J]. International Journal of Approximate Reasoning, 2011, 52(6): 881-893.

[11] Banka H, Mitra S. Feature selection, classification and rule generation using rough sets[M]//Rough Sets: Selected Methods and Applications in Management and Engineering. Springer London, 2012: 51-76.

[12] Hu Q, Pan W, Zhang L, et al. Feature selection for monotonic classification[J]. Fuzzy Systems, IEEE Transactions on, 2012, 20(1): 69-81.

[13] Pawlak Z.: Rough sets: Theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht 1991.

[14] Rauszer C, Skowron A. The discernibility matrices and functions in information systems [J].Intelligent Decision Support-Handbook of Applications and Advances of the Rough Sets Theory, Knowledge Engineering and Problem Solving, 1992, 11: 331-362.

[15] Shen Q, Jensen R. Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring[J]. Pattern recognition, 2004, 37(7): 1351-1363.

[16] Saha S, Murthy C, Pal S. Classification of web services using tensor space model and rough ensemble classifier [J]. Foundations of Intelligent Systems, 2008: 508-513.

[17] Shi, Lei, et al. "Rough set and ensemble learning based semi-supervised algorithm for text classification." Expert Systems with Applications 38.5 (2011): 6300-6306.

[18] Lei, S. H. I., et al. "Rough Set Based Decision Tree Ensemble Algorithm for Text Classification." Journal of Computational Information Systems6 1 (2010): 89-95.

[19] Zhu, Haiping, et al. "Integration of rough set and neural network ensemble to predict the configuration performance of a modular product family." International Journal of Production Research 48.24 (2010): 7371-7393.

[20] Shi L, Xi L, Ma X, et al. A novel ensemble algorithm for biomedical classification based on Ant Colony Optimization [J]. Applied Soft Computing, 2011, 11(8): 5674-5683.

[21] Wang S L, Li X, Zhang S, et al. Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction [J]. Computers in biology and medicine, 2010, 40(2): 179-189.

[22] Wu Q X, Bell D, McGinnity M. Multiknowledge for decision making [J]. Knowledge and information systems, 2005, 7(2): 246-266.

[23] Ishii N, Morioka Y, Bao Y, et al. Control of variables in reducts-kNN classification with confidence [M] //Knowledge-Based and Intelligent Information and Engineering Systems. Springer Berlin Heidelberg, 2011: 98-107.

[24] Kumar Das A, Sil J. An efficient classifier design integrating rough set and set oriented database operations [J]. Applied Soft Computing, 2011, 11(2): 2279-2285.

[25] Aleksander Ohrn: Discernibility and Rough Sets in Medicine: Tools and Applications. PhD Thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway, NTNU report 1999:133, IDI report 1999:14, ISBN 82-7984-014-1, 239 pages. 1999.

[26] Vinterbo S, Øhrn A. Minimal approximate hitting sets and rule templates [J]. International Journal of approximate reasoning, 2000, 25(2): 123-143.

[27] Hu Q, Yu D, Xie Z, et al. EROS: Ensemble rough subspaces [J]. Pattern recognition, 2007, 40(12):

3728-3739.
[28] Zhao H. Intrusion Detection Ensemble Algorithm based on Bagging and Neighborhood Rough Set [J]. International Journal of Security and Its Applications, 2013, 7(5): 193-204.
[29] Herawan T, Deris M M, Abawajy J H. A rough set approach for selecting clustering attribute [J]. Knowledge-Based Systems, 2010, 23(3): 220-231
[30] Thangavel K, Pethalakshmi A. Dimensionality reduction based on rough set theory: A review[J]. Applied Soft Computing, 2009, 9(1): 1-12.
[31] Lingras P. Unsupervised rough set classification using gas[J]. Journal of Intelligent Information Systems, 2001, 16(3): 215-228.
[32] Bazan J G, Nguyen H S, Nguyen S H, et al. Rough set algorithms in classification problem[M]//Rough set methods and applications. Physica-Verlag HD, 2000: 49-88.
[33] Dai J, Xu Q. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification[J]. Applied Soft Computing, 2013, 13(1): 211-221.
[34] Liu D, Li T, Liang D. Incorporating logistic regression to decision-theoretic rough sets for classifications[J]. International Journal of Approximate Reasoning, 2014, 55(1): 197-210.
[35] Opitz D W. Feature selection for ensembles[C] //AAAI/IAAI. 1999: 379-384.
[36] Kuncheva L I, Whitaker C J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy[J]. Machine learning, 2003, 51(2): 181-207.
[37] Zhang M L, Zhou Z H. Exploiting unlabeled data to enhance ensemble diversity [J]. Data Mining and Knowledge Discovery, 2013, 26(1): 98-129.
[38] C.L. Blake, C.J. Merz, UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/mlearn/MLSummary.html/, 1998.
[39] Breiman L, Friedman J, Stone C J, et al. Classification and regression trees [M]. CRC press, 1984.
[40] Christopher J.C. Burges, A tutorial on support vector machine for pattern recognition, Data Mining and Knowledge Discovery 2 (2) (1998) 121–167.
[41] Phillips P J, Moon H, Rizvi S A, et al. The FERET evaluation methodology for face-recognition algorithms[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2000, 22(10): 1090-1104.
[42] ORL Database of Faces, (1994). http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.
[43] D. Cai, X. He, J. Han, Semi-supervised discriminant analysis, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2007, pp. 1–7.
[44] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression database, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (12) (2003) 1615–1618.
[45] Tenenbaum J B, De Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.
[46] Lysiak R, Kurzynski M, Woloszynski T. Optimal selection of ensemble classifiers using measures of competence and diversity of base classifiers [J]. Neurocomputing, 2014, 126: 29-35.