Information Sciences 370–371 (2016) 270–287

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Combining neighborhood separable subspaces for classification via sparsity regularized optimization

Pengfei Zhu, Qinghua Hu*, Yahong Han, Changqing Zhang, Yong Du

School of Computer Science and Technology, Tianjin Univerity, Tianjin 150001, China

ARTICLE INFO

Article history:
Received 1 December 2015
Revised 27 July 2016
Accepted 2 August 2016
Available online 2 August 2016

Keywords:
Ensemble learning
Attribute reduction
Neighborhood rough sets
Joint representation
Group sparsity

ABSTRACT

The neighborhood rough set theory has been successfully applied to various classification tasks. The key concept of this theory is to find a sufficient and necessary neighborhood separable subspace for building a compact model. Given a classification learning task, there usually exist numerous neighborhood separable subspaces that maintain the discriminative ability of the original space with respect to a given granularity. These subspaces contain complementary information for classification. However, it is a challenging task to compute these subspaces efficiently. In this paper, we develop a fast neighborhood attribute reduction algorithm based on sample pair selection to find all reducts. Nevertheless, it cannot deal with large-scale data. Then we propose a randomized attribute reduction algorithm based on neighborhood dependency. The randomized algorithm can find a part of all reducts and is very efficient. A classification framework of joint subspace representation is proposed to fully exploit the complementary information in different subspaces. In addition, a weight matrix is learned to combine the representation residuals in the different subspaces via group sparsity regularization. The performances of the proposed attribute reduction algorithms are compared, and the influence of granularity on attribute reduction is discussed. Finally, the proposed technique is compared with other ensemble learning algorithms. Experimental results show that the proposed framework is superior to state-of-the-art classifiers.

1. Introduction

The rapid development of sensors and digital devices has contributed to the emergence of a wide variety of high-dimensional data, e.g., gene microarray data [36] and fault signals [55], which in turn has led to the so-called curse of dimensionality, including extremely high time complexity and storage requirements, as well as the failure of classification models [23]. Thus, to eliminate irrelevancy and redundancy in the feature space, it is necessary to establish low-dimensional structures in high-dimensional data. For this purpose, it is important to determine the characteristics of the raw feature space that should be preserved as well as those that should be removed. Moreover, although the diversity of feature subspaces has been widely exploited, e.g., random subspace method [47] and compressive sensing [52], it has not been investigated extensively from the viewpoint of classification tasks.

Neighborhood rough sets have been employed in vibration diagnosis [55], cancer recognition [20], and tumor classification[44]. The neighborhood rough set theory involves granulation of the feature space into a family of neighborhood

* Corresponding author. Fax.:+862227401839.
E-mail address: huqinghua@tju.edu.cn (Q. Hu).

http://dx.doi.org/10.1016/j.ins.2016.08.004
0020-0255/

information granules and approximation of classification with these granules. If objects with similar feature values are grouped into the same class, the classification is said to be consistent; otherwise, it is inconsistent [20]. Consistency of the neighborhood information granules reflects the separability of the classes. If all the neighborhood information granules are consistent, the task is separable. Neighborhood dependency is essentially the percentage of consistent neighborhood granules [21,22]. It reflects the separation level of classification. Consequently, the neighborhood structure of the feature space can be introduced to evaluate the classification separability [22].

Redundant and irrelevant attributes in the feature space increase the computational complexity, and degrade classification performance. Various attribute reduction methods, including those based on the discernibility matrix [15,35,42,49], [54] and the attribute significance index [22], have been proposed for eliminating such attributes without reducing the approximation ability of the original feature space. In the former category of methods, the discernibility matrix is employed to construct a Boolean discernibility function, and all reducts can be obtained through the reduced disjunction of the discernibility function [29]. The disadvantage of such methods is the computation complexity of the discernibility matrix. It has been shown [10] that only the minimal elements in the discernibility matrix are useful for attribute reduction. Chen [9,10] proposed a fast attribute reduction method using rough sets and fuzzy rough sets to find sample pairs corresponding to minimal elements in the matrix. In the latter category of methods, different attribute evaluation indices, e.g., dependency, Shannon entropy, and mutual information, are used to develop heuristic algorithms with different search strategies [45]. For attribute reduction based on neighborhood rough sets, Hu [22] proposed a greedy algorithm based on neighborhood attribute significance. However, this algorithm can only find one reduct, and usually, it cannot find the optimal solution. According to the rough set theory, given a classification dataset, there are multiple attribute reducts that maintain the approximation ability of the raw feature space.

Obviously, it is useful to find the optimal reduct (i.e., the minimal reduct or some other reduct) for classification. Rough sets emphasize the approximation ability rather than the generalization ability of classification. In fact, different reducts provide diverse information, which describe the original task from different perspectives. Actually, different reducts can be considered as multi-view observations from different sources or sensors. Researches on cognitive psychology and neuroscience show that human brains can automatically store and combine multi-modal information [43]. Different areas of human brain cortex correspond to different modalities, e.g., text, audio, image,etc [39]. Additionally, compared to uni-modal case, multi-modal information can lead to shorter reaction time and higher recognition accuracy [16]. Hence, motivated by the multi-modal organization of the human brains, we can combine the information in different reducts and develop a robust classification model with high accuracy and efficiency.

In fact, subspace ensemble learning has attracted considerable attention, and it has contributed to significant improvements in classification performance. In general, an ensemble is built in three steps, i.e., generating subspaces, learning multiple base learners and combining their predictions [56]. First, subspaces, including sample subspaces and feature subspaces, are generated [28]. Sample subspace techniques perturb the training data with resampling methods, such as bootstrap sampling used in bagging [8]. Feature subspace methods introduce randomness in the feature space, e.g., random forests [6], random subspaces [19], and rotation forests [38]. In general, sample subspaces and feature subspaces are used together to increase the diversity of base learners [26,46,50,56]. As compared to random subspaces, neighborhood separable subspaces can guarantee diversity while maintaining the discrimination ability of the original feature space. Thus, they facilitate the implementation of effective subspace ensembles. Second, base learners are trained in these feature subspaces or sample subspaces. In the third step, majority voting is used in most methods for decision combination. In addition, the following ensemble pruning methods based on ordered aggregation have been proposed for selecting a subensemble: reduce-error pruning [31], orientation ordering [32], margin distance minimization [33], and boosting-based ordering [34]. All these methods attempt to treat base classifiers individually and ignore the class weights of the base classifiers.

In classification tasks, a basic problem is the development of a statistical, generative, or discriminative model to use the training samples to correctly label a test sample. The nearest neighbor classifier (NNC) [11] searches for the nearest sample, and the nearest feature subspace (NFS) [18] searches for the nearest linear subspace in which a class lies. Recently, representation-based classifiers (SRC[47]/CRC[53]) have been developed as a generalization of NNC and NFS; they treat the query sample as a linear combination of all the training samples and use the reconstruction residual of each class for classification. In essence, SRC/CRC search for the most similar samples, and the representation coefficients can be considered as the sample's importance. Representation-based classifiers exhibit excellent performance [17,30,48,60,62], and they can be applied to multi-task learning as well as tasks with nonlinear data distributions. Furthermore, joint representation has been proposed [51] to extend a single modal representation to multi-task presentation. Face blocks and different features are used to construct different dictionaries for recognition. If each neighborhood separable subspace is used as a dictionary, then a joint subspace representation-based classifier can be constructed.

In the present study, as shown in Fig. 1, first, efficient neighborhood attribute reduction algorithms were developed on the basis of the discernibility matrix, neighborhood dependency, and sample pair selection. Then, by considering each neighborhood separable subspace as a dictionary, a joint neighborhood separable subspace representation-based classifier (JNSSRC) was proposed for using the information in different subspaces. To consider the distinctiveness of different subspaces, a weight matrix was learned for combining the representation residuals of the different subspaces via group sparsity imposed on the weights. Finally, experiments were conducted to analyze the performances of the proposed attribute reduction methods, and JNSSRC was compared with some state-of-the-art classifiers.

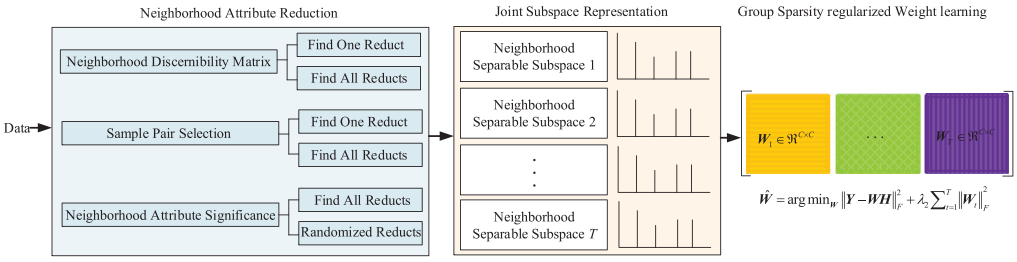

Fig. 1. Flowchart of the proposed method.

The contributions of this paper can be summarized as follows:

- An attribute reduction method based on neighborhood rough sets and the discernibility matrix was designed. This algorithm can find all the reducts of the original data;
- A fast attribute reduction algorithm based on sample pair selection was designed. This algorithm reduces the computational complexity significantly;
- It was verified that, in general, the minimal reduct is not necessarily optimal in terms of generalization ability. Therefore, it is necessary to combine a set of reducts for good generalization performance;
- A new framework of joint neighborhood separable subspace representation-based classification was developed. And, a novel decision combination method was developed for combining the outputs from multiple neighborhood separable subspaces.

The remainder of this paper is organized as follows. Section 2 reviews the concepts of neighborhood rough sets. Section 3 introduces the proposed attribute reduction methods. Section 4 describes the joint subspace representation-based classifier. Section 5 describes some experiments to analyze the generalization ability of the minimal reduct, the impact of neighborhood size, and the performance of the proposed attribute reduction methods and classifier. Finally, Section 6 concludes the paper.

2. Basic concepts

As the classical rough set theory proposed by Pawlak cannot deal with numeric data, Hu et al. constructed a rough set model based on neighborhood granulation [22].

$< U, A, D >$ is called a decision system, where $U = \{x_1, \cdots, x_n\}$ is a non-empty finite set of objects, $A = \{a_1, \cdots, a_m\}$ is a set of conditional attributes that describe the objects, and D is a decision attribute that records the decision labels of the objects. For a binary classification task, D takes the values $\{+1, -1\}$.

Definition 1. [22] Let $< U, \Delta >$ be a non-empty metric space, $x \in U$, $\delta \geq 0$. We refer to

$$\delta(x) = \{y | \Delta(x, y) \leq \delta, y \in U\} \tag{1}$$

as the δ neighborhood of x, where Δ is a metric function.

Definition 2. [22] Given $< U, A, D >$, if A generates a family of neighborhood relations N on the universe, then $NDT =< U, N, D >$ is a neighborhood decision system.

Definition 3. [22] Given $NDT =< U, A, N, D >$, D divides U into N equivalence classes, $X_1, X_2, \cdots, X_N$, $B \subseteq A$ generates a neighborhood relation N_B on U, and $\delta_B(x_i)$ are the neighborhood information granules generated in feature space B. Then, the lower and upper approximations of D with respect to B are defined as

$$\begin{aligned} \underline{\mathrm{N}_B} D &= \bigcup_{i=1}^{N} \underline{\mathrm{N}_B} X_i; \\ \overline{\mathrm{N}_B} D &= \bigcup_{i=1}^{N} \overline{\mathrm{N}_B} X_i. \end{aligned} \tag{2}$$

where

$$\begin{aligned} \underline{\mathrm{N}_B} X &= \{x_i | \delta_B(x_i) \subseteq X, x_i \in U\}; \\ \overline{\mathrm{N}_B} X &= \{x_i | \delta_B(x_i) \cap X \neq \emptyset, x_i \in U\}. \end{aligned} \tag{3}$$

Definition 4. [22] Given $NDT =< U, A, D >$, the dependency of D on B is defined as

$$\gamma_B(D) = Card\left(\underline{\mathrm{N}_B} D\right) / Card(U), \tag{4}$$

where *Card* denotes the size of the set. If $\gamma_B(D) = 1$, we say B is a neighborhood separable subspace.

Definition 5. [22] Given $NDT =< U, A, D >$, $B \subseteq A$, $a \subseteq B$, if

$$\begin{aligned} &\gamma_B(D) = \gamma_A(D); \\ &\forall a \in B : \gamma_{(B-a)}(D) < \gamma_B(D), \end{aligned} \tag{5}$$

then we say that B is an attribute reduct and refer to B as a neighborhood separable subspace.

Definition 6. [22] Given $NDT =< U, A, D >$, $\{B_j | j \le r\}$ is a set of attribute reducts, $Core = \underset{j \le r}{\cap} B_j$.

$$\begin{aligned} &K = \underset{j \le r}{\cup} B_j - Core, \\ &K_j = B_j - Core. \end{aligned} \tag{6}$$

Core is a set of attributes that cannot be removed from any reduct; if these attributes are removed, the discrimination ability would decrease. K is a weakly relevant attribute set. The union of *Core* and K_j forms a reduct of the decision system.

In essence, an attribute reduct is a set of condition attributes that maintain the approximation ability of the original feature space and each attribute in the reduct can not be removed for preserving the approximation ability. According to the rough set theory, there are multiple attribute reducts that maintain the approximation ability of the original feature space. Different subspaces contain different types of information, which describe the original feature space from different perspectives. Furthermore, the different subspaces provide supplementary information for each other.

3. Neighborhood attribute reduction

This section presents three attribute reduction methods for finding the neighborhood reducts. First, attribute reduction based on the neighborhood discernibility matrix is introduced in Section 3.1. Second, a fast algorithm based on sample pair selection is introduced in Section 3.2. Third, considering the computation limitation, a randomized reduction algorithm is introduced in Section 3.3. Finally, in Section 3.4, the impact of neighborhood size on attribute reduction is analyzed.

3.1. Attribute reduction based on neighborhood discernibility matrix

In attribute reduction based on the neighborhood discernibility matrix, first, a Boolean discernibility function is constructed using the neighborhood discernibility matrix. Then, the reduced conjunction of the discernibility function is obtained. Finally, the attribute sets that maintain the discernibility ability of the original feature space can be obtained.

Given a neighborhood decision system, $NDT =< U, A, D >$, the neighborhood relation of $B \subseteq A$ in feature space B can be written as a neighborhood relation matrix $M^B = \left(r_{ij}\right)_{n \times n}$, where

$$r_{ij}^B = \begin{cases} 1, & \Delta^B(x_i, x_j) \le \delta \\ 0, & otherwise. \end{cases} \tag{7}$$

Let $M_D(U, A) = \left(n_{ij}\right)_{n \times n}$ denote the neighborhood discernibility matrix of $< U, A, D >$. It satisfies

1. $n_{ij} = \{a : r_{ij}^a = 0, a \in A\}$ if $x_j \notin [x_i]_D$, where $[x_i]_D$ is the equivalence class for x_i;
2. $n_{ij} = \phi$ if $x_j \in [x_i]_D$, where ϕ denotes an empty set.

$M_D(U, A)$ is a symmetrical matrix and $n_{ii} = \phi$.

Let $f_U(A \cup D) = \wedge\{\vee(n_{ij}) : n_{ij} \neq \phi\}$ denote a Boolean function. $f_U(A \cup D)$ can be the discernibility function of $< U, A, D >$. Let $g_D(U, A)$ be the reduced conjunction of $f_D(U, A)$. There exist l and $B_k \subseteq A$ for $k = 1, \ldots, l$, such that $g_D(U, A) = (\wedge B_1) \vee \ldots \vee (\wedge B_l)$, where the elements in B_k appear only once. Let $\mathrm{Red}_D(A)$ be a set of all attribute reducts.

Theorem 1. $\mathrm{Red}_D(A) = \{B_1, \ldots, B_l\}$ *[41].*

From Theorem 1, we can conclude that attribute reduction based on the neighborhood discernibility matrix is essentially the process of using the absorption law to reduce the Boolean discernibility function so as to get its conjunction.

Given a decision table $< U, A, D >$, as shown in Table 1, $A = \{a_1, \ldots, a_6\}$ is the conditional attribute set and U is partitioned by the decision attribute as $U = \{M, N\}$, where $M = \{x_1, x_3, x_7, x_8\}$, $N = \{x_2, x_4, x_5, x_6\}$. If the neighborhood size is set as 0.15 and the Euclidean distance is used, then the neighborhood discernibility matrix $M_D(U, A)$ is

ϕ	{ 1,2,3,5,6}	ϕ	{ 2,4}	{ 1,3,4,5,6}	{ 5,6}	ϕ	ϕ
	ϕ	{ 1,4,5,6}	ϕ	ϕ	ϕ	{ 1,4,5,6}	{ 1,3,4}
		ϕ	{ 3}	{ 1,2,5,6}	{ 2,3,4,5,6}	ϕ	ϕ
			ϕ	ϕ	ϕ	{ 3,5}	{ 5,6}
				ϕ	ϕ	{ 1,2,5,6}	{ 1,2,3}
					ϕ	{ 2,3,4,6}	{ 2,4,5}
						ϕ	ϕ
							ϕ

According to the neighborhood discernibility matrix, we can get the discernibility function $f_U(A \cup D)$. The reduced conjunction can be obtained by the absorption law:

$$g_D(U, A) = n_{43} \wedge n_{14} \wedge n_{16} = a_3 \wedge (a_2, a_4) \wedge (a_5, a_6).$$

Hence, we can get some reducts of the decision table in Table 1: {2, 3, 5} , {3, 4, 5}, {2, 3, 6}, and {3, 5, 6}. It is easy to see that a_3 is the core attribute.

After neighborhood discernibility matrix is calculated, discernibility matrix based attribute reduction algorithms in rough sets can be introduced to find all reducts and one reduct [9,10]. The difference from rough sets is that discernibility matrix is replaced by neighborhood discernibility matrix. The time complexity of finding all the reducts covers two tasks: computing the neighborhood discernibility matrix $M_D(U, A)$ and searching for all the reducts. The time complexity of computing $M_D(U, A)$ is $O(n^2m^2)$, where n and m are the number of samples and number of features, respectively. When $M_D(U, A)$ is calculated, the time complexity of searching for all the reducts is NP-complete. For finding one reduct, the time complexities of computing $M_D(U, A)$ and searching for one reduct are both $O(m^2n^2)$. Hence, the time complexity of finding one reduct based on the neighborhood discernibility matrix is $O(m^2n^2)$.

3.2. Attribute reduction based on sample pair selection

Using the neighborhood discernibility matrix, the computation of all the reducts is NP-hard. Chen [9,10] found that only minimal elements in the discernibility matrix are required for this purpose, whereas the other elements are not. Hence, all elements except the minimal ones are redundant. If the minimal elements in the matrix can be found, the computation load of calculating the discernibility matrix and finding the reducts will be reduced significantly.

For the example presented in Table 1, the minimal elements in the neighborhood discernibility matrix are found to be {3}, {2, 4}, and {5, 6}. In the neighborhood discernibility function $f_U(A \cup D)$, all the other elements in $M_D(U, A)$ are absorbed by the three minimal reducts. Therefore, we can get

$$f_D(U, C) = a_3 \wedge (a_2 \vee a_4) \wedge (a_5 \vee a_6).$$

If we can obtain the minimal elements in $M_D(U, A)$, it is not necessary to compute the entire matrix. Because each minimal element corresponds to a sample pair, we need to find only the corresponding sample pairs.

Definition 7. Given $NDT = < U, A, D >$, the neighborhood relative relationship $DIS(a)$ of conditional attribute a with respect to decision attribute D is defined as

$$DIS(a) = \{(x_i, x_j) \in U \times U : \Delta^a(x_i, x_j) > \delta, x_j \notin [x_i]_D\}.$$

Let $DIS(A)$ be the union of the neighborhood relative relationships of all the conditional attributes, $DIS(A) = \cup_{a \in A} DIS(a)$. When $DIS(a) = DIS(A)$, which implies that $\forall n_{ij} \neq \phi$, $a \in n_{ij}$, $\{a\}$ is the optimal reduct of $< U, A, D >$. For this type of decision system, $\{a\}$ is selected as one reduct and the search for the other reducts is stopped. In the remainder of this paper, we assume that $\forall a \in A$, $DIS(a) \neq DIS(A)$.

For every element n_{ij} ($n_{ij} \neq \phi$) in $M_D(U, C)$, $(x_i, x_j) \in DIS(a) \Leftrightarrow a \in n_{ij}$. Let S_{ij} be the elements absorbed by n_{ij}, $S_{ij} = \cap\{DIS(a) : (x_i, x_j) \in DIS(a)\}$. N_{ij} is the number of conditional attributes that satisfy $(x_i, x_j) \in DIS(a)$. Obviously, $N_{ij} = \left|n_{ij}\right|$. For every $(x_i, x_j) \in DIS(A)$, $(x_i, x_j) \in S_{ij}$; hence, $\cup S_{ij} = DIS(A)$. We can use the following theorem to describe $n_{ij} \in M_D(U, A)$.

Table 1
Decision table.

sample	a_1	a_2	a_3	a_4	a_5	a_6	D
x_1	0.9	0.11	0.86	0.86	0.11	0.9	1
x_2	0.1	0.95	0.06	0.9	0.88	0.12	2
x_3	0.86	0.95	0.13	0.12	0.1	0.9	1
x_4	0.95	1	0.92	0.1	0.02	0.87	2
x_5	0.06	0.1	0.06	0.11	0.95	0.1	2
x_6	0.87	0.13	0.87	0.86	0.52	0.06	2
x_7	0.9	0.95	0.06	0.06	0.38	0.95	1
x_8	0.95	0.9	0.9	0.11	0.89	0.11	1

Table 2
Pseudo-code for finding all the reducts based on sample pair selection.

Input:	$< U, A, D >, \delta$
Output:	RED
3	Initialize$M_D(U, A) = \phi, RED = \phi$
4	Compute every $DIS(a)$ and $DIS(A)$;
5	Rank $(x_i, x_j) \in DIS(A)$ by N_{ij};
6	Do while $(DIS(A) \neq \phi)$
7	select one sample pair $(x_{i_0}, x_{j_0}) \in DIS(A)$
8	Compute $S_{i_0 j_0}$ and $n_{i_0 j_0} = \{a \in A : (x_{i_0}, x_{j_0}) \in DIS(a)\}$
9	$M_D(U, A) = M_D(U, A) \cup n_{i_0 j_0}$
10	$DIS(A) = DIS(A) - S_{i_0 j_0}$
11	End while
12	Compute $Core_D(U, A) = \{n_{ij} : Card(n_{ij}) = 1\}$, $H = M_D(U, A) - Core_D(U, A)$
13	Compute $Candidate = \cup\{n_{ij} : n_{ij} \in H\}$ and power set PS of $Candidate$
14	Do while $(PS \neq \phi)$
15	Choose a $Cand$ from the power set PS
16	If $Cand \cap n_{ij} \neq \phi, \forall n_{ij} \in H$, then $Red = Core_D(U, A) +$ $Cand$. $RED = [RED, Red]$, delete $Cand$ and the absorbed elements
17	If $\exists n_{ij} \in H, Cand \cap n_{ij} = \phi$, delete $Cand$ from PS
18	End while
19	Output RED

Theorem 2. *For every* $(x_s, x_t) \in S_{ij}$, $(x_s, x_t) \neq (x_i, x_j)$, $S_{ij} \supseteq S_{st}$ *and* $N_{st} \geq N_{ij}$ *holds.*

Proof. If $(x_s, x_t) \in S_{ij}$, then $n_{ij} \subseteq n_{st}$ implies that $S_{ij} \supseteq S_{st}$ and $N_{st} \geq N_{ij}$. □

Theorem 3. *For two minimal elements* $n_{ij} \neq n_{st}$, $(x_i, x_j) \notin S_{st}$ *and* $(x_s, x_t) \notin S_{ij}$ *hold.*

Proof. For two minimal elements $n_{ij} \neq n_{st}$, there exist $P \in n_{ij}$ and $Q \in n_{st}$ such that $P \notin n_{st}$ and $Q \notin n_{ij}$; this implies that $(x_i, x_j) \notin S_{st}$ and $(x_s, x_t) \notin S_{ij}$ hold. □

Theorem 4. *If* S_{ij} *is the maximum,* n_{ij} *is the minimal element in* $M_D(U, C)$.

Proof. S_{ij} is the maximum implies that the number of elements absorbed by n_{ij} is the largest. If S_{ij} is the maximum, there exists no $n_{st} \neq n_{ij}$, such that $S_{ij} \subseteq S_{st}$. Hence, n_{ij} is the minimal element in $M_D(U, C)$. □

Theorem 5. $\cup\{S_{ij}: n_{ij}$ *is the minimal element*$\} = DIS(A)$.

Proof. $\forall(x_s, x_t) \in DIS(A)$, there is always a minimal element n_{ij} that satisfies $(x_s, x_t) \in S_{ij}$, which implies that $\cup\{S_{ij}: n_{ij}$ is the minimal element$\} = DIS(A)$. □

According to Theorem 2, $(x_i, x_j) \in DIS(A)$ can be ranked by N_{ij}.

According to Theorem 2 and Theorem 4, the sample pair (x_i, x_j) with the minimal N_{ij} is bound to correspond to the minimal element n_{ij}.

According to Theorem 2 and Theorem 3, if the S_{ij} to which the minimal element n_{ij} corresponds is deleted from $DIS(A)$, then all $S_{st} \subseteq S_{ij}$ can be removed. However, S_{st} to which the other minimal elements correspond cannot be removed from $DIS(A)$. Among the remaining sample pairs of $DIS(A)$, the sample pair with the minimal N_{ij} is sure to correspond to the minimal element n_{ij}.

According to Theorem 5, only when $DIS(A) = \phi$, the S_{ij} to which all the minimal elements n_{ij} correspond are removed from $DIS(A)$. This implies that the sample pairs $(x_i, x_j) \in DIS(A)$ can be ranked by N_{ij}, and S_{ij} with minimal N_{ij} can be deleted until $DIS(A) = \phi$.

Theorems 2–5 lay the theoretical foundation for the following sample pair selection algorithm. The key to finding all the reducts based on sample pair selection is to find the minimal elements in the neighborhood discernibility matrix. First, the minimal elements cannot be absorbed by other elements. Second, the element with the least N_{ij} is surely the minimal element. Finally, all the minimal elements are found when $DIS(A) = \phi$. After the minimal elements are found, the method for finding all the reducts is the same as the method based on the neighborhood discernibility matrix. The only difference is that the original matrix is substituted by the matrix that contains only the minimal elements.

The detailed algorithm is presented in Table 2. The time complexity of the algorithm covers two tasks: sample pair selection and finding all the reducts. The time complexity of sample pair selection is $O(n^2 m)$, and finding all the reducts is

Table 3
Pseudo-code for finding one reduct based on sample pair selection.

Input:	$< U, A, D >, \delta$
Output:	*REDUCT*
1	Initialize$REDUCT = \phi$
2	Compute every $DIS(a)$ and $DIS(A)$
3	Rank $(x_i, x_j) \in DIS(A)$ by N_{ij};
4	Do while ($DIS(A) \neq \phi$)
5	choose the first sample pair $(x_{i_0}, x_{j_0}) \in DIS(A)$
6	select a where $(x_{i_0}, x_{j_0}) \in DIS(a)$ and put a in *REDUCT*
7	$DIS(A) = DIS(A) - DIS(a)$, where $(x_{i_0}, x_{j_0}) \in DIS(a)$
8	End while
9	Output *REDUCT*

NP-complete. This method can reduce the computation load considerably because it does not need to compute the neighborhood discernibility matrix. In addition, H obtained from the matrix that contains only the minimal elements is far smaller than that obtained from the original matrix. Therefore, the searching speed increases significantly.

To find one reduct, first, we compute the neighborhood relative relationship and rank (x_i, x_j) by N_{ij}. Then, we put the attribute a that can discriminate the first sample pair (x_{i_0}, x_{j_0}) into *REDUCT* and delete $DIS(a)$ from $DIS(A)$. The algorithm is presented in Table 3. The time complexity of the algorithm is $O(n^2 m)$.

3.3. Attribute reduction based on neighborhood attribute significance

As attribute reduction based on the neighborhood discernibility matrix or sample pair selection entails high computation complexity or considerable hardware requirements, we propose a randomized attribute reduction technique based on neighborhood rough sets in this subsection. Unlike the method based on the discernibility matrix, this method ranks features by attribute significance. Hu [22] proposed a heuristic feature selection algorithm with a forward or backward greedy search strategy based on neighborhood attribute significance.

Definition 8. Given $< U, A, D >$,the attribute significance $SIG(a, B, D)$ of conditional attribute $a \subseteq B \subseteq A$ can be defined as

$$SIG(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D), \tag{8}$$

where $\gamma_B(D)$ is the attribute dependency of $B \subseteq A$ with respect to D.

If we use a greedy search strategy to find the reducts based on neighborhood rough sets, the attribute that produces the largest increasement in the discrimination ability is selected. Using neighborhood rough sets, given $NDT =< U, A, D >$, if $B_1 \subset B_2 \subset \ldots \subset B_m \subseteq A$ is a nested sequence of attribute subsets, we have $\gamma_{B_1}(D) \leq \gamma_{B_2}(D) \leq \ldots \leq \gamma_{B_m}(D) = \gamma_B(D)$. For the nested sequence generated by the forward greedy search, $\gamma_{B_1}(D) \prec \gamma_{B_2}(D) \prec \ldots \prec \gamma_{B_m}(D) = \gamma_B(D)$ holds. Here, B_m maintains the approximation ability of the original feature space. The generation of the nested sequence is up to the selection of the newly added feature in each interaction.

To obtain multiple reducts, we can remove the requirement for the newly added feature. We can use the K features with the highest discrimination ability. Thus, we can obtain multiple reducts that maintain the discriminative ability. The algorithm is presented in Table 4.

After we select a random number, we can get a randomized attribute reduct if we run the code once. The computational complexity of this method is $(2n - k)(k + 1) \times m \log m/2$, where n and m are the number of samples and number of features, respectively, and k is the number of attributes in the reduct. When $N = 1$, randomness is eliminated and we can get Red_s. In addition, we should judge whether the present reduct is duplicated when we want to obtain multiple reducts.

Unlike other methods, randomized reduction does not require computation of the power set. Hence, its hardware requirements and computation complexity are significantly lower. Thus, randomized reduction based on neighborhood rough sets is a simple and effective attribute reduction method for generation multiple reducts.

3.4. Multiple granularity neighborhood separable subspace

So far, we have presented three attribute reduction algorithms. In this subsection, we compare the three proposed algorithms. The methods based on the discernibility matrix and sample selection use the discriminative ability of attributes. In contrast, the method based on attribute significance relies on the approximation ability of attributes. The consistency of these two types of methods with respect to each other is related to the definition of lower approximation.

Neighborhood rough sets exploit the neighborhood relation. The lower approximation relies on the selected metric. If we use the infinite norm distance, the two types of methods are consistent. However, if we use the Euclidean distance, the two types of methods cannot be equivalent to each other.

Table 4
Pseudo-code for neighborhood randomized attribute reduction (NRAR).

Input:	$< U, A, D >$, δ, random number N
Output:	Reduct *red*
1	Initialize $red = \phi$
2	For $a_i \in A - red$
3	compute attribute dependency $\gamma_{red \cup a_i}(D)$
4	compute $SIG(a_i, red, D) = \gamma_{red \cup a_i}(D) - \gamma_{red}(D)$
5	End
6	select a_k, where a_k is one of the first N features with the largest $SIG(a_k, red, D)$ in $\{A - red\}$
7	If $SIG(a_k, red, D) > 0$
8	$red = [red, a_k]$
9	return to step 2
10	else
11	If *red* does not contain redundant attributes
12	return *red*
13	else
14	delete redundant attributes and return *red*
15	End if
16	End if

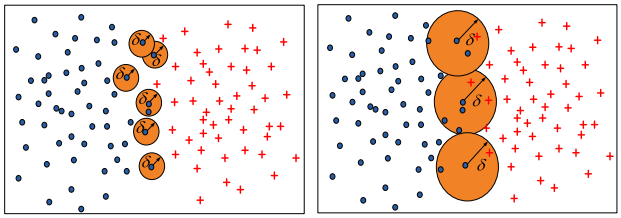

Fig. 2. Impact of neighborhood size on neighborhood information granules.

Proof. Given a neighborhood decision system $NDT =< U, A, D >$ with neighborhood size δ, $B \subseteq A$, Δ is the infinite norm-based distance, i.e., $\Delta^B(x_i, x_j) = \max_{1 \le k \le m} \Delta^{a_k}(x_i, x_j)$. Given two samples x_i and x_j, $x_i \notin [x_j]_D$,

- $\Delta^B(x_i, x_j) < \delta \Leftrightarrow \forall a_k \in B\ \Delta^{a_k}(x_i, x_j) < \delta$ holds. Thus, if x_i and x_j cannot be discerned by an arbitrary conditional attribute, then the δ neighborhood of x_i is inconsistent;
- $\Delta^B(x_i, x_j) > \delta \Leftrightarrow \exists a_k \in B\ \Delta^{a_k}(x_i, x_j) > \delta$ holds. Thus, if there always exists an attribute that can discern x_i and x_j, then x_j is outside the δ neighborhood of x_i. □

Hence, for the infinite norm distance, as long as the discernibility matrix is certain, the attribute dependency can be identified. The former is the sufficient condition for the latter, rather than necessary condition. In the case of the Euclidean distance, there is no inevitable link between $\Delta^B(x_i, x_j) > \delta$ and $\Delta^{a_k}(x_i, x_j) > \delta, k = 1, 2, \ldots, m$.

In neighborhood granulation, the discrimination ability and consistency are influenced by the neighborhood size [59]. As shown in Fig. 2, it is a binary problem. The left and right subfigures show the neighborhood of the boundary samples when the neighborhood is very small or very large, respectively. From the figure, we can see that the smaller the neighborhood is, the more consistent and discriminative the neighborhood information granules are. Further, when the neighborhood is very small, both dimensions can distinguish all the samples. In this case, we cannot select the more discriminative feature. When the neighborhood is very large such that every sample is in the neighborhood of other samples, it is still not possible to select the better feature. Therefore, it is crucial to select a proper neighborhood size for attribute reduction.

Theorem 6. *Given $NDT =< U, A, D >$ and neighborhood size δ ,there always exist α and β such that*

(1) *If $\delta > \alpha$, then every element $n_{ij} = \phi$ in the neighborhood discernibility matrix $M_D(U, A)$;*
(2) *If $\delta < \beta$, then every element $n_{ij} = A - AI$, where $AI = \{a \in A : x_i^a = x_j^a, x_j \notin [x_i]_D\}$.*

Proof.

(1) $R = \left\{d^a(x_i, x_j) : a \in A, x_j \notin [x_i]_D\right\}$ is the distance between any two samples in feature a, and $R^{\max}$ is the maximum value of R. Given $\alpha \ge R^{\max}$, $\delta > \alpha$, $r_{ij} = 1$ in the neighborhood relation matrix M^a, $n_{ij} = \{a : r_{ij}^a = 0, a \in A\}$; therefore, $n_{ij} = \phi$.

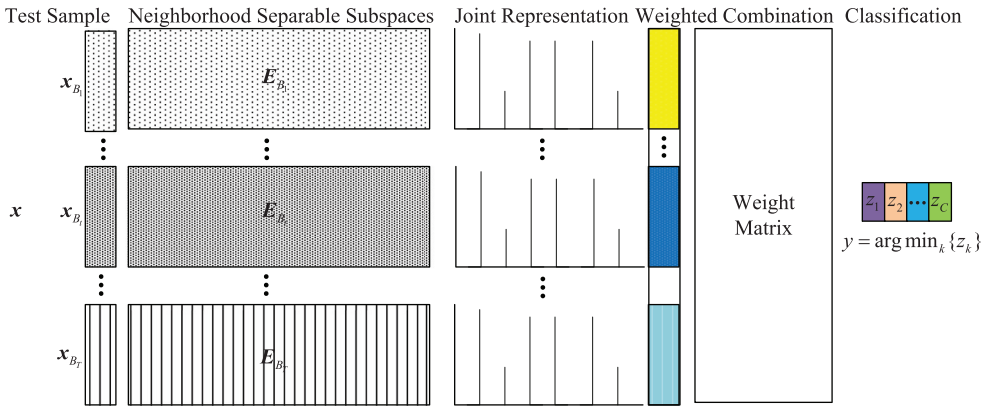

Fig. 3. Multiple neighborhood separable subspace representation-based classifier.

(2) β is a constant close to zero, $\delta < \beta$, attributes $\left\{a \in A : x_i^a \neq x_j^a, x_j \notin [x_i]_D\right\}$ satisfy $r_{ij} = 0$ in M^a. Therefore, we can get $n_{ij} = A - AI$.

If $\{a \in A : \forall i, j, x_i^a = x_j^a, x_j \notin [x_i]_D\}$ is empty, then there is always a constant β such that when $\delta < \beta$, $n_{ij} = A$ in the neighborhood discernibility matrix $M_D(U, A)$.

For example, consider the decision table in Table 1. Given $\alpha = 1,\ \beta = 0.001$

(1) If $\delta > \alpha$, $n_{ij} = \phi$;
(2) If $\delta < \beta$, $n_{ij} = \{1, 2, 3, 4, 5, 6\}$.

From the above example, we can see that δ has a considerable impact on attribute reduction. When $\beta = 0.001$, $n_{ij} = \{1, 2, 3, 4, 5, 6\}$. Any conditional attribute can become a reduct. When $\delta > \alpha$, $n_{ij} = \phi$, the reduct is A. The neighborhood size affects the neighborhood discernibility matrix. Let Nr^δ denote the number of neighborhood separable subspaces, i.e., the neighborhood attribute reducts. In addition, let NR^δ denote the number of attributes in a neighborhood attribute reduct. □

Theorem 7. *Given* $NDT =< U, A, D >$, *if* $\{a \in A : \forall i, j, x_i^a = x_j^a, x_j \notin [x_i]_D\}$ *is empty, then there is always a constant* β, $\delta < \beta$, *such that* NR^δ *is 1 and* Nr^δ *is m, where m is the number of conditional attributes. Moreover,*$\forall a \in A, DIS(a) = DIS(A)$.

Proof. According to Theorem 6, if $\{a \in A : \forall i, j, x_i^a = x_j^a, x_j \notin [x_i]_D\}$ is empty, then there is always a constant β such that when $\delta < \beta$, every element $n_{ij} = A$ in the neighborhood discernibility matrix $M_D(U, A)$. Obviously, $\forall a \in A, DIS(a) = DIS(A)$. In this case, each attribute is one reduct.

Such δ can be considered as a case of overfitting. Although all the samples can be distinguished by any conditional attribute for a given δ, the obtained attribute reduct cannot work well on unseen test data in terms of classification. Theorem 6 and Theorem 7 show that the neighborhood size has a great impact on the discrimination ability and consistency of neighborhood granules. Therefore, it affects the searched neighborhood attribute reducts. Selecting an optimal neighborhood granularity depends on the task at hand [57]. □

4. Joint subspace representation-based classifier

The generalization performance of the minimal reduct is not necessarily optimal. Moreover, the combination of information in different reducts can guarantee information integrity. In this section, we propose a joint neighborhood separable subspace representation-based classifier. The framework of the proposed classifier is shown in Fig. 3. When we obtain multiple neighborhood separable subspaces, the samples in each feature subspace B_t form a dictionary $\boldsymbol{E}_{B_t}$. The sample $\boldsymbol{x}_{B_t}$ is represented in the corresponding dictionary $\boldsymbol{E}_{B_t}$. Then, the representation residuals of different classes for each subspace are obtained. A weight matrix $\boldsymbol{W}$ is learned to combine the representation residuals for the final decision.

4.1. Joint subspace representation

Given a neighborhood decision system $NDT =< U, A, D >$, a set of neighborhood separable subspaces $\{B_1, \ldots, B_t, \ldots, B_T\}$ can be obtained via neighborhood randomized attribute reduction (NRAR).

For a test sample $\boldsymbol{x}$, in each feature subspace B_t, $\boldsymbol{x}_{B_t}$ is represented in the dictionary $\boldsymbol{E}_{B_t}$:

$$\boldsymbol{x}_{\boldsymbol{B_t}} = \boldsymbol{E}_{B_t}\boldsymbol{a}_t + \boldsymbol{e}_t, \tag{9}$$

Fig. 4. Weight matrix $\boldsymbol{W}$.

where the dictionary $\boldsymbol{E}_{B_t}$ consists of the training samples in the neighborhood separable subspace B_t.

To best represent the test sample $\boldsymbol{x}$, the representation residual (i.e., $\left\|\boldsymbol{x}_{B_t} - \boldsymbol{E}_{B_t}\boldsymbol{a}_t\right\|_2^2$) should be minimized. For T neighborhood separable subspaces, we should minimize the joint representation residual:

$$\min \sum_{t=1}^{T} \|\boldsymbol{x}_{B_t} - \boldsymbol{E}_{B_t}\boldsymbol{a}_t\|_2^2. \tag{10}$$

To enforce the class sparsity, a mixed-form regularization is imposed on the coefficients. Then, the optimization objective is

$$\hat{\boldsymbol{A}} = \arg\min \left\{ \begin{matrix} \sum_{t=1}^{T} \|\boldsymbol{x}_{B_t} - \boldsymbol{E}_{B_t}\boldsymbol{a}_t\|_2^2 + \\ \lambda_1 \sum_{j=1}^{C} \left\|\boldsymbol{A}_j\right\|_2 \end{matrix} \right\} \tag{11}$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_t, \ldots, \boldsymbol{a}_T]$ and $\boldsymbol{A}_j$ is the coefficient matrix of the ith class.

The accelerated proximal gradient (APG) method [51] is used to efficiently solve Eq. (11). After the coefficient matrix $\boldsymbol{A}$ is calculated, the reconstruction residual r_{tj} of the tth feature subspace and the jth class is $r_{tj} = \left\|\boldsymbol{x}_{B_t} - \boldsymbol{E}_{B_t,j}\boldsymbol{a}_{tj}\right\|_2^2$, where $\boldsymbol{E}_{B_t} = [\boldsymbol{E}_{B_t,1}, \ldots, \boldsymbol{E}_{B_t,j}, \ldots, \boldsymbol{E}_{B_t,C}]$ and $\boldsymbol{a}_t = [\boldsymbol{a}_{t1}, \ldots, \boldsymbol{a}_{tj}, \ldots, \boldsymbol{a}_{tC}]$. $\boldsymbol{E}_{B_t,j}$ and $\boldsymbol{a}_{tj}$ are the dictionary and coefficient vector of the jth class and the tth neighborhood separable subspace, respectively.

The reconstruction residual r_{tj} is then transformed to the probability $h_{(t\times C+j)}$, i.e., $h_{(t\times C+j)} = \exp(-r_{tj}/\epsilon)$, where ϵ is a constant. A linear projection matrix $\boldsymbol{W} \in R^{C\times TC}$ is introduced for classification. $\boldsymbol{z} = f(h) = \boldsymbol{W}\boldsymbol{h}$, $\boldsymbol{z} \in R^{C\times 1}$. $f(h)$ is a multi-class linear classifier, which has been widely used in multi-class classification and regression tasks [24]. Then, the label of the test sample is $\hat{y} = \arg\max_k\{z_k\}$.

4.2. Group sparsity regularized weight learning

For $\boldsymbol{x}$, $\boldsymbol{y} = [0; \ldots; 1, \ldots; 0]$ is the label vector. If $\boldsymbol{x}$ belongs to the jth class, then the jth element of $\boldsymbol{y}$ is 1. In classification, the projection matrix $\boldsymbol{W}$ should have the following properties: the predicted value z_j of the jth class should be close to 1, whereas the values of the other classes should be close to zero.

Hence, the error between the predicted value and the real value should be minimized.

$$\min \|\boldsymbol{y} - \boldsymbol{z}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{W}\boldsymbol{h}\|_2^2. \tag{12}$$

For the training set, we can get a probability matrix $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_2, \ldots \boldsymbol{h}_n]$, $\boldsymbol{H} \in R^{TC\times n}$, where T, C, and n are the number of feature subspaces, classes, and training samples, respectively. Then, we get

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{H}\|_F^2. \tag{13}$$

As shown in Fig. 4, the projection matrix $\boldsymbol{W}$ consists of T sub-matrixes $\boldsymbol{W}_t$. $\boldsymbol{W} = [\boldsymbol{W}_1, \ldots, \boldsymbol{W}_t, \ldots, \boldsymbol{W}_T]$ and $\boldsymbol{W}_t \in R^{C\times C}$ is the sub-matrix for the t^{th} feature subspace.

Considering that some feature subspaces may be useless for classification, we impose group sparsity on the weight matrix $\boldsymbol{W}$. Group sparsity is widely used in feature selection, classification and regression models [58,61]. The objective function considering the group sparsity is defined as

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{H}\|_F^2 + \lambda_2 \sum_{t=1}^{T} \|\boldsymbol{W}_t\|_F^2. \tag{14}$$

where $\sum_{t=1}^{T} \|\boldsymbol{W}_t\|_F^2$ is the group sparsity regularization on the weight matrix $\boldsymbol{W}$. In this way, the effect of feature subspace with little $\|\boldsymbol{W}_t\|_F^2$ will be constrained in the classification. The problem in Eq.(14) is a group lasso problem. It can be solved by using some sparse toolboxes. In this study, we used SPAMS [2] to solve Eq.(14). The pseudo-code for JNSSRC is presented in Table 5. Please note that in the second step, the training sample $\boldsymbol{x}_k$ should be removed from the dictionary when it is represented.

5. Experimental analysis

In this section, first, we compare the performance of attribute reduction based on the neighborhood discernibility matrix and that based on sample pair selection. Second, we analyze the effect of neighborhood size on attribute reduction. Third, we confirm that the minimal reduct is the simplest but not necessarily the optimal solution in terms of generalization ability. Finally, we test the performance of the proposed classifier.

Table 5
Pseudo-code for JNSSRC.

Input:	Training set $s = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, test sample $\boldsymbol{x}$ and δ
Output:	The prediction of test sample $\boldsymbol{x}$
1	Find T neighborhood separable subspaces $\{B_1, \ldots, B_T\}$ by NRAR;
2	Get representation residual r_{tj} by Eq. (11) for training set;
3	Get probability matrix $\boldsymbol{H}$ for training set;
4	Get weight matrix by $\boldsymbol{W}$ solving Eq. (14);
5	For test sample $\boldsymbol{x}$, get the representation residual vector $\boldsymbol{h}$ by Eq. (11);
6	The prediction of $\boldsymbol{x}$ is $\hat{y} = \arg\max_k \{z_k\}$, where $\boldsymbol{z} = \boldsymbol{Wh}$.

Table 6
Data description.

Data	Features	Class	Instances
heart	13	2	270
iono	34	2	351
wpbc	34	2	198
wine	13	3	178

Table 7
Results of experiment for finding one reduct based on the neighborhood discernibility matrix.

Data	feature	reduct	time	mark	acc.R	acc.O
heart	13	8	126.4	1	77.8 ± 6.5	76.7 ± 9.4
iono	34	10	654.9	1	87.3 ± 6.5	86.4 ± 4.9
wpbc	33	15	96.0	1	70.7 ± 6.7	70.7 ± 6.7
wine	13	5	59.3	1	93.8 ± 5.0	94.9 ± 5.0

Table 8
Results of experiment for finding all the reducts based on the neighborhood discernibility matrix.

Data	feature	reducts	time	minimal
heart	13	74	266.6	6
iono	20	157	41411.0	7
wpbc	20	10	2935.1	14
wine	13	332	43.5	5

Table 9
Results of experiment for finding one reduct based on sample pair selection.

Data	feature	reduct	time	Mark	acc.R	acc.O
heart	13	8	7.5	1	77.8 ± 6.5	76.7 ± 9.4
iono	20	9	172.3	1	88.8 ± 5.7	86.4 ± 4.9
wpbc	20	12	1.7	1	70.7 ± 6.7	70.7 ± 6.7
wine	13	6	14.3	1	94.9 ± 1.8	94.9 ± 5.0

5.1. Attribute reduction

We employed four datasets from the UCI database [5]; detailed descriptions of the data are provided in Table 6. Table 7 and Table 8 summarize the results of attribute reduction based on the neighborhood discernibility matrix. Similarly, Table 9 and Table 10 summarize the results of attribute reduction based on sample pair selection. The neighborhood size is set to 0.15. In Tables 7 and 9, acc.R and acc.O denote the classification accuracy in the neighborhood separable subspace and the original feature space, respectively. The mark value is used to judge whether a reduct is a real reduct; if the mark value is 1, the reduct is a real reduct. In the experiments for finding all the reducts, owing to hardware limitations, we selected 20 features from wpbc and iono.

From Tables 7 and 9, we can see that it takes 7.5 and 1.7 s to find one reduct based on sample pair selection, whereas it takes 126.4 and 96.0 s to find one reduct based on the neighborhood discernibility matrix, for heart and wpbc, respectively. From Tables 8 and 10, we can see that the number of reducts and the number of attributes in the minimal reduct are the same, while the time taken for attribute reduction based on sample pair selection is far less than the time taken for

Table 10
Results of experiment for finding all the reducts based on sample pair selection.

Data	feature	reducts	time	minimal
heart	13	74	20.4	6
iono	34	157	11904.0	7
wpbc	33	10	57.8	14
wine	13	332	28.3	5

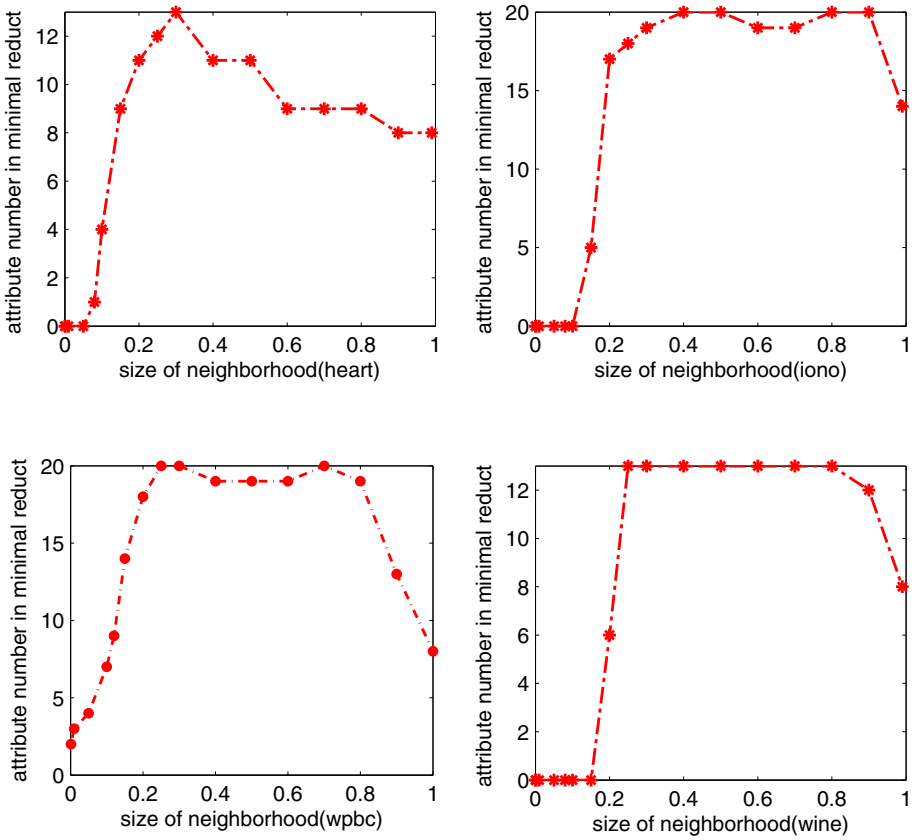

Fig. 5. Relationship between neighborhood size and number of attributes in minimal reduct.

attribute reduction based on the neighborhood discernibility matrix. Hence, we can conclude that attribute reduction based on sample pair selection is more efficient.

5.2. Impact of neighborhood size on attribute reduction

Because neighborhood granularity has a significant influence on the discrimination ability, we analyze the impact of neighborhood size on attribute reduction in this subsection. We compute the number of reducts Nr^{δ} in the reduct set *Red*, the number of attributes NR^{δ} in the minimal reduct, and the number of sample pairs while increasing the neighborhood size δ for four datasets: heart, iono, wine, and wpbc. Fig. 5 shows that the number of attributes in the minimal reduct first increases and then decreases when the neighborhood size δ increases.

Fig. 6 shows the variation trend in the number of reducts Nr^{δ} in the reduct set *Red*. As in the above-mentioned case, we can see that it first increases and then decreases as the neighborhood size increases.

The discriminative ability and consistency of neighborhood information granules decrease as δ increases. Fig. 7 shows that the number of selected sample pairs first increases and then decreases. The number of selected sample pairs is the number of minimal elements in the neighborhood discernibility matrix.

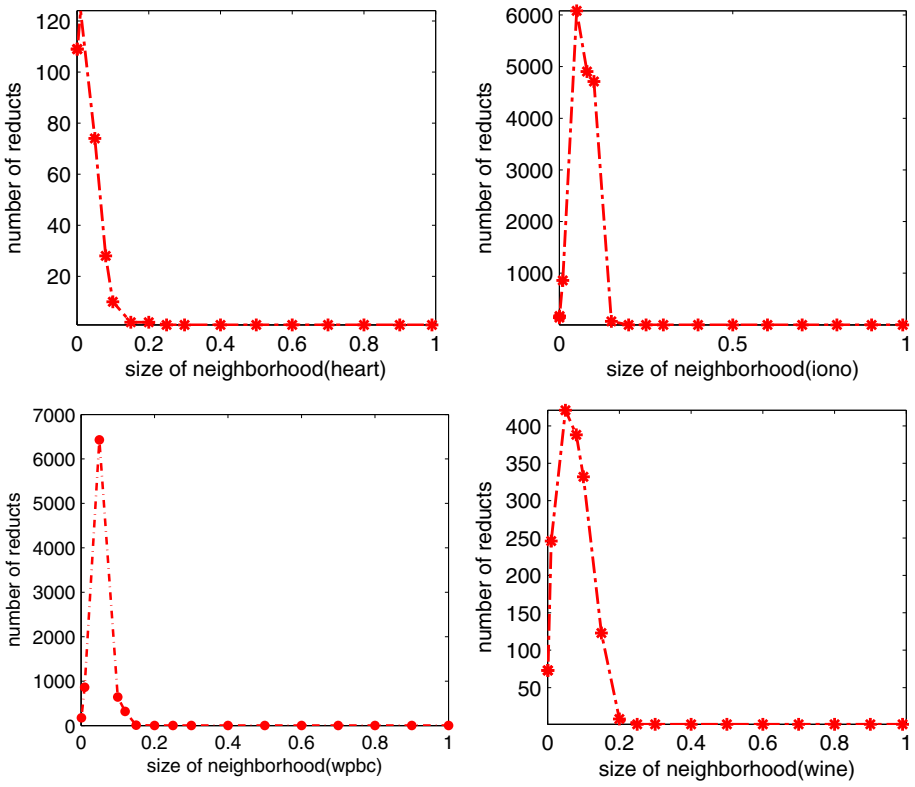

Fig. 6. Relationship between neighborhood size and total number of reducts.

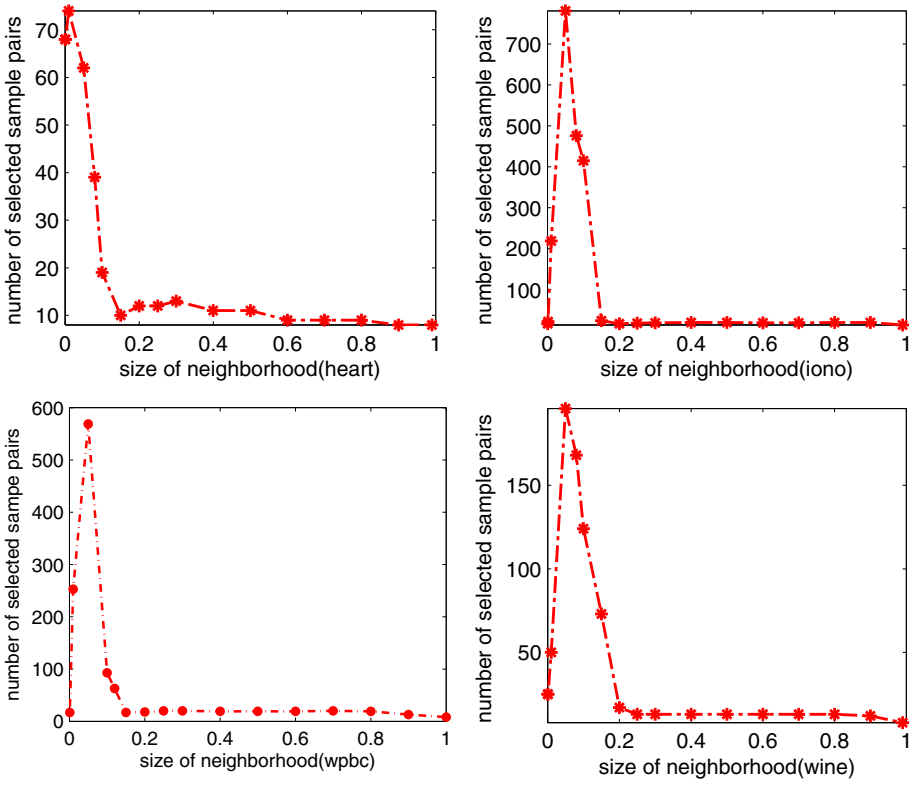

Fig. 7. Relationship between neighborhood size and number of sample pairs.

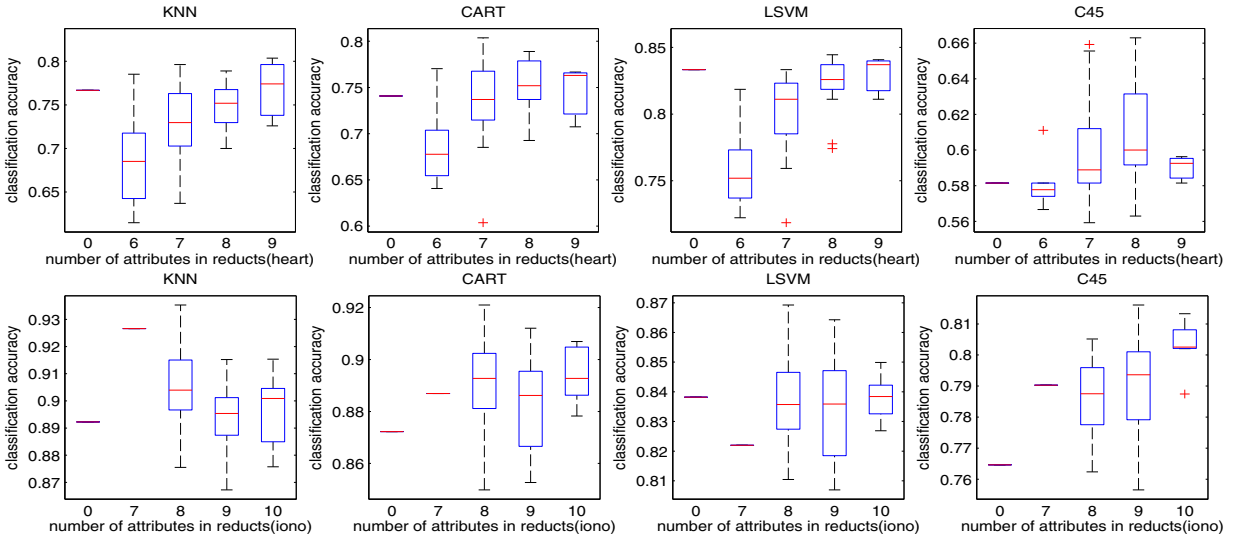

Fig. 8. Box plot of classification performance of the minimal reduct and all the reducts.

5.3. Generalization performance of minimal reduct

The minimal reduct is the simplest description of the original feature space. We want to determine whether the simplest reduct is also the optimal one in terms of generalization performance. According to Theorem 7, we know that there is always a constant β, $\delta < \beta$, such that the number of attributes NR^{δ} in the reducts is 1. In this case, the classification performance of the minimal reduct would be quite poor and attribute reduction would be meaningless.

In Section 5.1, all the reducts are obtained. We compare the performance of the minimal reduct with that of all the reducts using four classifiers (i.e., KNN, CART, LSVM, and C4.5). A box plot of the classification performance of the minimal reduct and all the reducts is shown in Fig. 8. The value corresponding to 0 is the classification accuracy of the minimal reduct. The values corresponding to the other numbers are the classification accuracies of all the other reducts. We can see that the minimal reduct is not necessarily the optimal reduct, and sometimes, its performance is not satisfactory.

5.4. Joint subspace representation-based classifier

Experiments were conducted on seven low-dimensional datasets, and three high-dimensional datasets to analyze the classification performance of the proposed joint neighborhood separable subspaces representation-based classifier (JNSSRC) and to show that it outperforms other state-of-the-art classifiers. 10-fold cross-validation is used for all the classification tasks.

Comparison methods: We compare JNSSRC with the widely used classifiers (1-NN, SVM), neighborhood rough sets based classifier (NEC), and representation based classifiers (NSC, SRC, CRC).

- 1-NN [11]: nearest neighbor classifier;
- NEC [22]: neighborhood classifier;
- NSC [12]: nearest subspace classifier;
- SRC [47]: sparse representation based classifier;
- CRC [53]: collaborative representation based classifier;
- SVM [25]: support vector machines.

JNSSRC is also compared with ensemble learning methods,

- RS [19]: Random Subspace;
- RF [38]: Rotation Forest;
- Bagging [8];
- AdaBoost [37].

Low-dimensional datasets

1. **heart** is a dataset for the prediction of the presence of heart disease in a patient. It consists of 270 samples, 13 attributes, and 2 classes (i.e., presence and absence) [13].
2. **hepatitis** is a medical dataset that consists of 155 samples, 2 classes, and 19 attributes [14].

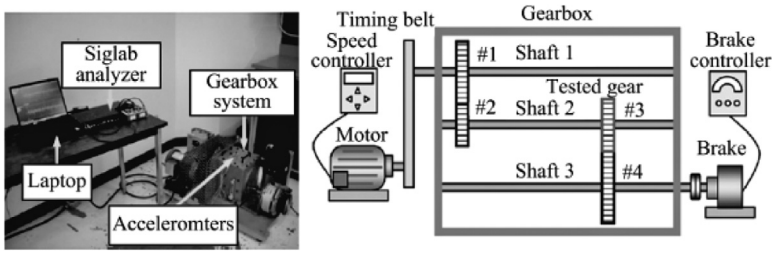

Fig. 9. Experimental setup of vibration dataset[55].

Table 11
Accuracy of JNSSRC and other classifies.

Data	1-NN	NEC	NSC	CRC	SRC	SVM
heart	76.7 ± 9.4	79.3 ± 6.3	80.0 ± 7.2	82.6 ± 5.0	78.9 ± 6.5	82.2 ± 6.0
hepatitis	82.5 ± 7.6	83.5 ± 10.4	82.0 ± 3.2	83.3 ± 3.5	84.7 ± 7.1	86.7 ± 5.4
iono	86.4 ± 4.9	83.5 ± 4.7	87.3 ± 5.8	90.7 ± 5.6	91.5 ± 4.5	87.6 ± 6.5
thyroid	94.4 ± 5.8	93.5 ± 7.5	94.9 ± 4.1	91.1 ± 7.5	93.5 ± 6.3	92.1 ± 8.0
wine	94.9 ± 5.0	96.6 ± 2.9	97.2 ± 3.9	98.9 ± 2.3	98.3 ± 2.8	98.9 ± 2.3
wpbc	70.7 ± 6.7	78.3 ± 7.3	76.3 ± 3.0	76.3 ± 6.1	76.3 ± 7.3	77.4 ± 7.7
vibration	80.9 ± 11.2	76.3 ± 12.0	77.4 ± 11.2	79.6 ± 11.5	80.5 ± 10.8	79.6 ± 10.3
Average	82.0	82.5	82.8	84.1	84.3	84.0
Data	RS	RF	Bagging	AdaBoost	JNSSRC(U)	JNSSRC
heart	80.7 ± 6.3	82.6 ± 7.2	82.2 ± 7.8	80.0 ± 5.8	79.6 ± 6.8	83.3 ± 5.7
hepatitis	88.4 ± 7.7	91.6 ± 5.5	93.0 ± 6.4	89.6 ± 5.4	90.0 ± 5.4	91.0 ± 1.8
iono	93.7 ± 5.5	93.4 ± 4.3	92.1 ± 5.9	93.2 ± 4.6	93.8 ± 4.4	95.5 ± 3.6
thyroid	93.5 ± 5.4	95.8 ± 4.1	94.4 ± 5.1	93.9 ± 6.1	94.9 ± 6.3	96.3 ± 4.3
wine	96.6 ± 4.4	98.3 ± 2.3	95.4 ± 3.8	97.2 ± 3.0	97.7 ± 4.0	99.4 ± 1.8
wpbc	76.3 ± 3.2	78.3 ± 6.8	77.7 ± 6.5	79.8 ± 6.5	78.6 ± 6.7	81.7 ± 6.3
vibration	83.1 ± 16.8	84.8 ± 10.1	80.6 ± 9.4	84.4 ± 15.6	83.2 ± 12.5	83.6 ± 11.2
Average	85.3	87.1	85.7	86.1	86.1	88.0

3. **iono** is a radar dataset collected by a system in Goose Bay, Labrador [40]. The targets were free electrons in the ionosphere. "Good" radar returns are those that show evidence of some type of structure in the ionosphere, whereas "bad" returns are those that do not; these signals pass through the ionosphere. Hence, there are two classes. In all, there are 351 instances and 34 attributes.
4. **thyroid gland** is a dataset for thyroid disease recognition. There are 215 instances, 4 attributes, and 3 classes (i.e., normal, hyperthyroidism, and hypothyroidism) [7].
5. **wine** is a dataset that contains the results of a chemical analysis of wines made in the same region of Italy but derived from three different cultivars [1]. There are 178 samples and 13 attributes, including alcohol, ash, and color intensity.
6. **wpbc** (i.e., Wisconsin-Madison prognostic breast cancer) is a dataset for breast cancer classification. There are 198 individual cancer cases, including only those cases of invasive breast cancer with no evidence of distant metastases at the time of diagnosis. There are 34 attributes, of which the first 30 features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass [4].
7. **vibration** is a dataset used for gear crack recognition [55]. Gear crack level identification is a crucial part of gearbox diagnosis. The experimental setup in shown in Fig. 9. The data were collected under 3 load levels, 12 speeds, and 5 crack levels. Furthermore, each experiment was repeated 3 times. Subsequently, 540 samples were obtained. Then, through time and frequency analysis, 52 features, including mean and variance, were extracted.

The number of features in the seven datasets varies from 4 to 52. For the proposed method, we present the results with and without the learned weight matrix $\boldsymbol{W}$ (JNSSRC and JNSSRC(U), respectively). The parameters λ_1 and λ_2 were set to 0.001 and 0.1, respectively. The classification accuracies are listed in Table 11. From Table 11, we can see that JNSSRC outperformed the other classifiers and ensemble learning methods. As compared to the result without the learned weight matrix $\boldsymbol{W}$, the result with $\boldsymbol{W}$ showed improved classification performance (improvement of 1.9%).

High-dimensional datasets

We use three gene classification datasets and there are 9216, 7129, and 2308 features in the three datasets, respectively.

1. **breast** is a dataset for human breast tumor recognition. There are 84 experimental examples and 9216 gene features. In addition, there are 5 classes for prediction [36].

Table 12
Accuracy of gene classification.

Data	1-NN	NEC	NSC	CRC	SRC	SVM
breast	69.3 ± 15.8	69.3 ± 15.8	67.0 ± 8.2	70.3 ± 16.9	70.6 ± 15.9	67.9 ± 16.1
lung	69.6 ± 14.6	71.8 ± 15.9	68.6 ± 21.4	74.2 ± 22.2	76.7 ± 22.0	68.5 ± 16.0
SRBCT	73.0 ± 16.4	74.3 ± 15.3	81.1 ± 12.7	90.7 ± 11.2	92.0 ± 11.0	92.0 ± 11.0
Average	70.5	71.8	72.2	78.4	79.7	76.1
Data	RS	RF	Bagging	AdaBoost	JNSSRC(U)	JNSSRC
breast	69.8 ± 11.4	72.6 ± 10.3	70.4 ± 12.1	70.8 ± 5.0	71.5 ± 11.1	73.5 ± 9.4
lung	71.6 ± 21.8	72.3 ± 15.8	68.2 ± 22.8	64.8 ± 19.9	78.9 ± 24.0	81.1 ± 22.1
SRBCT	88.9 ± 10.6	90.7 ± 8.8	84.7 ± 12.2	88.9 ± 11.9	95.9 ± 6.0	97.2 ± 3.8
Average	76.7	78.5	74.4	74.8	82.1	83.9

Table 13
Accuracy on MNIST and CIFAR-10 datasets.

Data	1-NN	NEC	NSC	CRC	SRC	SVM
MNIST	98.89	98.82	98.89	98.74	98.78	98.92
CIFAR-10	89.15	89.18	89.66	89.37	89.34	89.14
Data	RS	RF	Bagging	AdaBoost	DL	JNSSRC
MNIST	98.96	98.91	98.88	98.92	99.05	98.95
CIFAR-10	89.22	89.19	89.28	89.32	89.24	89.72

2. **Lung** was first used in [3]. It consists of gene-expression profiles for 86 primary lung adenocarcinomas, including 67 stage I and 19 stage III tumors, and 10 non-neoplastic lung samples are generated for study purposes. Each sample is described with 7129 genes.
3. **SRBCT** is a dataset of small round blue cell tumors (SRBCT), reported in [27]. There are five different childhood tumors, including the Ewing family of tumors, neuroblastoma, non-Hodgkin lymphoma, and rhabdomyosarcoma. This dataset consists of 88 samples, and each sample is described with 2308 genes.

The parameters λ_1 and λ_2 are set the same as that for low-dimensional datasets. Table 12 shows the classification results for the three high-dimensional gene datasets. The experiment results indicate that the performance of the proposed method is superior to that of state-of-the-art methods. As compared to CRC and SRC, the improvement in accuracy is 5.5 % and 4.2%, respectively. Further, as compared to other ensemble learning algorithms, the improvement is more significant than that for the low-dimensional datasets. In addition, as compared to the result without the learned weight matrix $\boldsymbol{W}$, the improvement is 1.8%, which is similar to that for the low-dimensional datasets. Therefore, these experimental results for datasets with different feature dimensions show that JNSSRC can be applied to both low-dimensional and high-dimensional classification tasks.

Large scale datasets

The MNIST database of handwritten digits is composed of a training set with 60,000 samples and a test set with 10,000 samples. The image size of handwritten digits is 28×28 and there are ten classes. The CIFAR-10 dataset consists of 60,000 samples with 6000 images per class. There are 50,000 training samples and 10,000 test samples. There are ten classes and the image size is 32×32.

Deep learning has achieved great success on large-scale image classification and object recognition tasks by learning hierarchical neural networks. In this section, beside other comparison methods, we also compare with deep learning algorithms. For fair comparison, we extract high-level deep feature by convolutional neural networks and use it to evaluate different classifiers. Table 13 shows the accuracy on MNIST and CIFAR-10 datasets. DL represents the accuracy of the Softmax classifier. From the result, we can see that the difference of different classifiers is little. On both datasets, the performance of JNSSRC is superior to other representation based classifiers, including NSC, CRC and SRC. Additionally, the result shows that JNSSRC is comparable to DL with Softmax classifier.

6. Conclusion

In this paper, fast neighborhood attribute reduction algorithms are developed to find all reducts by sample pair selection. As the fast algorithm cannot apply to large-scale problems, a randomized attribute reduction algorithm is developed based on neighborhood dependency. A joint representation-based classifier was proposed to combine the information in different neighborhood separable subspaces. Experiments showed that the proposed classification model was shown to be superior to other state-of-the-art classifiers. We will investigate the extension of the proposed model in classification tasks with specific data distributions in the future work.

Acknowledgement

This work was supported by the National Program on Key Basic Research Project under Grant 2013CB329304, the National Natural Science Foundation of China under Grants 61502332, 61432011, 61222210.

References

[1] S. Aeberhard, D. Coomans, O. De Vel, The classification performance of rda, Tech. Rep, Dept. Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland„ 1992. 92–01
[2] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, Convex optimization with sparsity-inducing norms, Optim. Mach. Learn. (2011) 19–53.
[3] D.G. Beer, S.L. Kardia, C.-C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, et al., Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nat. Med. 8 (8) (2002) 816–824.
[4] K.P. Bennett, Decision Tree Construction via Linear Programming, Center for Parallel Optimization, Computer Sciences Department, University of Wisconsin, 1992.
[5] M. Lichman, UCI machine learning repository, 2013.
[6] A. Bosch, A. Zisserman, X. Munoz, Image classification using random forests and ferns, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE, 2007, pp. 1–8.
[7] R.W. Brause, Medical analysis and diagnosis by neural networks, in: Medical data analysis, Springer, 2001, pp. 1–13.
[8] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[9] D. Chen, L. Zhang, S. Zhao, Q. Hu, P. Zhu, A novel algorithm for finding reducts with fuzzy rough sets, Fuzzy Syst. IEEE Trans. 20 (2) (2012) 385–389.
[10] D.G. Chen, S.Y. Zhao, L. Zhang, Y.P. Yang, X. Zhang, Sample pair selection for attribute reduction with rough set, IEEE Trans. Knowl. Data Eng. 24 (2012) 2080–2093.
[11] T. Cover, P. Hart, Nearest neighbor pattern classification, Inf. Theory IEEE Trans. 13 (1) (1967) 21–27.
[12] K. Crammer, R. Gilad-Bachrach, A. Navot, N. Tishby, Margin analysis of the lvq algorithm, Adv. Neural Inf. Process. Syst. (2003) 479–486.
[13] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K.H. Guppy, S. Lee, V. Froelicher, International application of a new probability algorithm for the diagnosis of coronary artery disease, Am. J. Cardiol. 64 (5) (1989) 304–310.
[14] P. Diaconis, B. Efron, Computer-intensive methods in statistics, Sci. Am. 248 (5) (1983) 116–130.
[15] L. Deer, M. Restrepo, C. Cornelis, J. Gómez, Neighborhood operators for covering-based rough sets, Inf. Sci. 336 (2016) 21–44.
[16] M.H. Giard, F. Peronnet, Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study, J. Cognit. Neurosci. 11 (5) (1999) 473–490.
[17] M. Harandi, M. Salzmann, Riemannian coding and dictionary learning: Kernels to the rescue, CVPR, 2015.
[18] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, D. Kriegman, Clustering appearances of objects under varying illumination conditions, in: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, IEEE, vol. 1, 2003, pp. I–11.
[19] T.K. Ho, The random subspace method for constructing decision forests, Patt. Anal. Mach. Intell. IEEE Trans. 20 (8) (1998) 832–844.
[20] Q. Hu, W. Pan, S. An, P. Ma, J. Wei, An efficient gene selection technique for cancer recognition based on neighborhood mutual information, Int. J. Mach. Learn. Cybern. (2010) 1–12.
[21] Q. Hu, W. Pedrycz, D. Yu, J. Lang, Selecting discrete and continuous features based on neighborhood decision error minimization, Syst. Man Cybern. Part B 40 (1) (2010) 137–150.
[22] Q. Hu, D. Yu, Z. Xie, Neighborhood classifiers, Expert Syst. Appl. 34 (2) (2008) 866–876.
[23] P. Indyk, R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality, in: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, ACM, 1998, pp. 604–613.
[24] Z. Jiang, Z. Lin, L.S. Davis, Learning a discriminative dictionary for sparse coding via label consistent k-svd, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1697–1704.
[25] T. Joachims, Making large-scale svm learning practical, Tech. Rep. 8 (1998) 499–526.
[26] I. Khan, J.Z. Huang, K. Ivanov, Incremental density-based ensemble clustering over evolving data streams, Neurocomputing 191 (2016) 34–43.
[27] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Med. 7 (6) (2001) 673–679.
[28] Z. Lai, W.K. Wong, Y. Xu, J. Yang, D. Zhang, Approximate orthogonal sparse embedding for dimensionality reduction, IEEE Trans. Neural Networks Learn. Syst. 27 (4) (2016) 723–735.
[29] M.S. Lazo-Cortés, J.F. Martínez-Trinidad, J.A. Carrasco-Ochoa, G. Sanchez-Diaz, On the relation between rough set reducts and typical testors, Inf. Sci. 294 (2015) 152–163.
[30] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, C. Yuan, Low-rank preserving projections, IEEE Trans. Cybern. 46 (8) (2016) 1900–1913, doi:10.1109/TCYB.2015.2457611.
[31] D.D. Margineantu, T.G. Dietterich, Pruning adaptive boosting, in: MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-, Citeseer, 1997, pp. 211–218.
[32] G. Martınez-Munoz, A. Suárez, Aggregation ordering in bagging, in: Proc. of the IASTED International Conference on Artificial Intelligence and Applications, Citeseer, 2004, pp. 258–263.
[33] G. Martínez-Muñoz, A. Suárez, Pruning in ordered bagging ensembles, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 609–616.
[34] G. Martınez-Munoz, A. Suárez, Using boosting to prune bagging ensembles, Patt. Recogn. Lett. 28 (1) (2007) 156–165.
[35] W. Pedrycz, Granular Computing: Analysis and Design of Intelligent Systems, CRC press, 2013.
[36] C.M. Perou, T. Sørlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, et al., Molecular portraits of human breast tumours, Nature 406 (6797) (2000) 747–752.
[37] G. Rätsch, T. Onoda, K.-R. Müller, Soft margins for adaboost, Mach. Learn. 42 (3) (2001) 287–320.
[38] J.J. Rodriguez, L.I. Kuncheva, C.J. Alonso, Rotation forest: A new classifier ensemble method, Patt. Anal. Mach. Intell. IEEE Trans. 28 (10) (2006) 1619–1630.
[39] J. Sepulcre, M.R. Sabuncu, T.B. Yeo, H. Liu, K.A. Johnson, Stepwise connectivity of the modal cortex reveals the multimodal organization of the human brain, J. Neurosci. 32 (31) (2012) 10649–10661.
[40] V.G. Sigillito, S.P. Wing, L.V. Hutton, K.B. Baker, Classification of radar returns from the ionosphere using neural networks, Johns Hopkins APL Tech. Digest 10 (1989) 262–266.
[41] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, Theor. Decis. Lib. 11 (1992) 331–362.
[42] E.C. Tsang, Q. Hu, D. Chen, Feature and instance reduction for pnn classifiers based on fuzzy rough sets, Int. J. Mach. Learn. Cybern. 7 (1) (2016) 1–11.
[43] M.T. Wallace, L.K. Wilkinson, B.E. Stein, Representation and integration of multiple sensory inputs in primate superior colliculus, J. Neurophysiol. 76 (2) (1996) 1246–1266.
[44] S. Wang, X. Li, S. Zhang, J. Gui, D. Huang, Tumor classification by combining pnn classifier ensemble with neighborhood rough set based gene reduction, Comput. Biol. Med. 40 (2) (2010) 179–189.
[45] S. Wang, W. Pedrycz, Q. Zhu, W. Zhu, Subspace learning for unsupervised feature selection via matrix factorization, Patt. Recogn. 48 (1) (2015) 10–19.

[46] X.-Z. Wang, H.-J. Xing, Y. Li, Q. Hua, C.-R. Dong, W. Pedrycz, A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning, IEEE Trans. Fuzzy Syst. 23 (5) (2015) 1638–1654.
[47] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, Patt. Anal. Mach. Intell. IEEE Trans. 31 (2) (2009) 210–227.
[48] M. Yang, P. Zhu, F. Liu, L. Shen, Joint representation and pattern learning for robust face recognition, Neurocomputing 168 (2015) 70–80.
[49] J.T. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: perspectives and challenges, IEEE Trans. Cybern. 43 (6) (2013) 1977–1989.
[50] Z.-H. You, Y.-K. Lei, L. Zhu, J. Xia, B. Wang, Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis, BMC Bioinformatics 14 (8) (2013) 1.
[51] X.-T. Yuan, S. Yan, Visual classification with multi-task joint sparse representation, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 3493–3500.
[52] K. Zhang, L. Zhang, M.-H. Yang, Real-time compressive tracking, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 864–877.
[53] L. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: Which helps face recognition? in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 471–478.
[54] X. Zhang, J. Dai, Y. Yu, On the union and intersection operations of rough sets based on various approximation spaces, Inf. Sci. 292 (2015) 214–229.
[55] X. Zhao, Q. Hu, Y. Lei, M. Zuo, Vibration-based fault diagnosis of slurry pump impellers using neighbourhood rough set models, Proc. Inst. Mech. Eng. Part C 224 (4) (2010) 995–1006.
[56] Z.-H. Zhou, Y. Yu, Ensembling local learners throughmultimodal perturbation, Syst. Man Cybern. Part B 35 (4) (2005) 725–735.
[57] P. Zhu, Q. Hu, Adaptive neighborhood granularity selection and combination based on margin distribution optimization, Inf. Sci. 249 (2013) 1–12.
[58] P. Zhu, Q. Hu, C. Zhang, W. Zuo, Coupled dictionary learning for unsupervised feature selection, AAAI 2016, 2016.
[59] P. Zhu, Q. Hu, W. Zuo, M. Yang, Multi-granularity distance metric learning via neighborhood granule margin maximization, Inf. Sci. 282 (282) (2014) 321–331.
[60] P. Zhu, L. Zhang, W. Zuo, X. Feng, Q. Hu, A self-representation induced classifier, IJCAI 2016, 2016.
[61] P. Zhu, W. Zuo, L. Zhang, Q. Hu, S.C. Shiu, Unsupervised feature selection by regularized self-representation, Patt. Recogn. 48 (2) (2015) 438–446.
[62] L. Zhuang, T.-H. Chan, A.Y. Yang, S.S. Sastry, Y. Ma, Sparse illumination learning and transfer for single-sample face recognition with image corruption and misalignment, IJCV 114 (2–3) (2015) 272–287.