Information Sciences 178 (2008) 3577–3594

Contents lists available at ScienceDirect

Information Sciences

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

Neighborhood rough set based heterogeneous feature subset selection

Qinghua Hu *, Daren Yu, Jinfu Liu, Congxin Wu

Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Article history:
Received 22 November 2007
Received in revised form 11 May 2008
Accepted 19 May 2008

Keywords:
Categorical feature
Numerical feature
Heterogeneous feature
Feature selection
Neighborhood
Rough sets

ABSTRACT

Feature subset selection is viewed as an important preprocessing step for pattern recognition, machine learning and data mining. Most of researches are focused on dealing with homogeneous feature selection, namely, numerical or categorical features. In this paper, we introduce a neighborhood rough set model to deal with the problem of heterogeneous feature subset selection. As the classical rough set model can just be used to evaluate categorical features, we generalize this model with neighborhood relations and introduce a neighborhood rough set model. The proposed model will degrade to the classical one if we specify the size of neighborhood zero. The neighborhood model is used to reduce numerical and categorical features by assigning different thresholds for different kinds of attributes. In this model the sizes of the neighborhood lower and upper approximations of decisions reflect the discriminating capability of feature subsets. The size of lower approximation is computed as the dependency between decision and condition attributes. We use the neighborhood dependency to evaluate the significance of a subset of heterogeneous features and construct forward feature subset selection algorithms. The proposed algorithms are compared with some classical techniques. Experimental results show that the neighborhood model based method is more flexible to deal with heterogeneous data.

1. Introduction

Feature subset selection as a common technique used in data preprocessing for pattern recognition, machine learning and data mining, has attracted much attention in recent years [4,6,7,40,41,44]. Due to the development of information acquirement and storage, tens, hundreds, or even thousands of features are acquired and stored in databases for some real-world applications. With a limited amount of training data, an excessive amount of features may cause a significant slowdown in the learning process, and may increase the risk of the learned classifier to over-fit the training data because irrelevant or redundant features confuse learning algorithms [39,40]. It is desirable to reduce data to get a smaller set of informative features for decreasing the cost in measuring, storing and transmitting data, shortening the process time and leading to more compact classification models with better generalization.

Much progress has been made on feature subset selection these years. There are several viewpoints to categorize such techniques: filter, wrapper and embedded [21,41], unsupervised [20] and supervised [1,11–14,22], etc. Here we roughly divide feature selection algorithms into two categories: symbolic method and numerical method. The former considers all features as categorical variables. The representative is attribute reduction based on rough set theory [24,44], while the latter views all attributes as real-valued variables, which take values in the real-number spaces [21,22]. If there coexist some heterogeneous features, such as categorical and real-valued, symbolic methods introduce a discretizing algorithm to partition the value domains of real-valued variables into several intervals, and then regard them as symbolic features [3,38,39]. By

* Corresponding author.
E-mail address: huqinghua@hcms.hit.edu.cn (Q. Hu).

0020-0255/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2008.05.024

contrast, numerical methods implicitly or explicitly code the categorical features with a series of integer numbers and treat them as numerical variables [22].

Obviously, discretization of numerical attributes may cause information loss because the degrees of membership of numerical values to discretized values are not considered [3,13]. There are at least two categories of structures lost in discretization: neighborhood structure and order structure in real spaces. For example, we know the distances between samples and we can get how the samples are close to each other in real spaces. This information is lost if the numerical attributes are discretized. On the other hand, it is unreasonable to measure similarity or dissimilarity with Euclidean distance as to categorical attributes in numerical methods. In order to deal with mixed feature sets, some of heterogeneous distance functions were proposed [26,35]. However, approaches to selecting features from heterogeneous data have not been fully studied so far [34].

Rough set theory, proposed by Pawlak [24], has been proven to be an effective tool for feature selection, rule extraction and knowledge discovery from categorical data in recent years [2,8,25,28,29,32,33,44]. To deal with fuzzy information, Shen and Jensen presented a fuzzy-rough QUICKREDUCT algorithm by generalizing the dependency function defined in the classical rough set model into the fuzzy case [13,14,31]. In [1], Bhatt and Gopal showed that QUICKREDUCT algorithm is not convergent on many real datasets and they proposed the concept of fuzzy-rough sets on compact computational domain, which is then utilized to improve the computational efficiency. Hu et al. extended Shannon's entropy to measure the information quantity in a set of fuzzy sets [9] and applied the proposed measure to calculate the uncertainty in fuzzy approximation spaces and used it to reduce heterogeneous data [10], where numerical attributes induce fuzzy relations and symbolic features generate crisp relations, then the generalized information entropy is used to compute the information quantity introduced by the corresponding feature or feature subset. However, it is time-consuming to generate a fuzzy equivalence relation from numerical attributes, which causes additional computation in feature selection. Moreover, how to generate effective fuzzy relations in different classification tasks is also an open problem [37].

In this paper, we introduce a neighborhood rough set model for heterogeneous feature subset selection and attribute reduction. As we know, granulation and approximation are two key issues in rough set methodology. Granulation is to segment the universe into different subsets with a certain criterion. The generated subsets are also called elemental granules or elemental concepts. While approximation refers to approximately describe arbitrary subset of the universe with these elemental concepts. Pawlak's rough set model employs equivalence relations to partition the universe and generate mutually exclusive equivalence classes as elemental concepts. This is just applicable to data with nominal attributes. In numerical spaces, the concept of neighborhood plays an important role [15,27,35,36]. Neighborhood relations can be used to generate a family of neighborhood granules from the universe characterized with numerical features, and then we can use these neighborhood granules to approximate decision classes. Based on this observation, a neighborhood rough set model was constructed [11,12,17–19]. Although the basic definitions of neighborhood rough sets was put forward several years ago, to the best of our knowledge, not much work has been conducted on the applications of this model so far [16]. Besides, this idea is similar to the tolerance rough sets [30]. The neighborhood relation can also be considered as a kind of tolerance relations [45,46]. However, no work has been reported to deal with heterogeneous features with tolerance rough set model yet. In fact the neighborhood model is also a natural generalization of Pawlak's rough set model. However, neighborhood rough sets can be used to deal with mixed numerical and categorical data within a uniform framework. The rough set model based on neighborhood relations can thus be introduced to process information with heterogeneous attributes. In this paper, we first show some metrics to compute neighborhoods of samples in general metric spaces, and then we introduce the neighborhood rough set model and discuss the properties of neighborhood decision tables. Based on the proposed model, the dependency between heterogeneous features and decision is defined for constructing measures of attribute significance for heterogeneous data. We present some attribute reduction algorithms with the proposed measures. Numerical experiments are presented and experimental results show that the proposed techniques have great power in heterogeneous attribute reduction. The main contributions of the work are two-fold. First, we extend the neighborhood rough set model to deal with data with heterogeneous features and discuss two classes of monotonicity in terms of consistency, neighborhood sizes and attributes; second, two efficient algorithms are designed for searching an effective feature subset.

The rest of the paper is organized as follows. Section 2 presents the notions and properties of the neighborhood rough set model. Section 3 shows the attribute reduction algorithm. Experimental analysis is given in Section 4. Conclusions come in Section 5.

2. Neighborhood rough set model

2.1. Neighborhood rough sets

Formally, the structural data used for classification learning can be written as an information system, denoted by $IS = \langle U, A\rangle$, where U is a nonempty and finite set of samples $\{x_1, x_2, \ldots, x_n\}$, called a universe, A is a set of attributes (also called features, inputs or variables) $\{a_1, a_2, \ldots, a_m\}$ to characterize the samples. To be more specific, $\langle U, A\rangle$ is also called a decision table if $A = C \cup D$, where C is the set of condition attributes and D is the decision attribute.

Given arbitrary $x_i \in U$ and $B \subseteq C$, the neighborhood $\delta_B(x_i)$ of x_i in feature space B is defined as

$$\delta_B(x_i) = \{x_j | x_j \in U, \Delta^B(x_i, x_j) \leqslant \delta\},$$

where Δ is a distance function. For $\forall x_1,x_2,x_3 \in U$, it usually satisfies:

(1) $\Delta(x_1,x_2) \geqslant 0$, $\Delta(x_1,x_2) = 0$ if and only if $x_1 = x_2$;
(2) $\Delta(x_1,x_2) = \Delta(x_2,x_1)$;
(3) $\Delta(x_1,x_3) \leqslant \Delta(x_1,x_2) + \Delta(x_2,x_3)$.

There are three metric functions widely used in pattern recognition. Considered that x_1 and x_2 are two objects in N-dimensional space $A = \{a_1,a_2,\ldots,a_N\}$, $f(x,a_i)$ denotes the value of sample x in the ith attribute a_i, then a general metric, named Minkowsky distance, is defined as

$$\Delta_P(x_1,x_2) = \left(\sum_{i=1}^{N} |f(x_1,a_i) - f(x_2,a_i)|^P\right)^{1/P}$$

where (1) it is called Manhattan distance Δ_1 if $P = 1$; (2) Euclidean distance Δ_2, if $P = 2$; (3) Chebychev distance if $P = \infty$. A detailed survey on distance functions can be seen in [26].

$\delta_B(x_i)$ is the neighborhood information granule centered with sample x_i and the size of the neighborhood depends on threshold δ. More samples fall into the neighborhood of x_i if δ takes a great value. The shapes of the neighborhoods depend on the used norm. In a two-dimension real space, neighborhood of x_0 in terms of the above three metrics are as shown in Fig. 1. Manhattan distance based neighborhood is a rhombus region around center sample x_0; Euclidean distance based neighborhood is a ball region; while Chebychev distance based neighborhood is a box region.

With the above discussion, we can see there are two key factors to impact on the neighborhood. One is the used distance, the other is threshold δ. The first one determines the shape of neighborhoods and the latter controls the size of neighborhood granules. Furthermore, we can also see that a neighborhood granule degrades to an equivalent class if we let $\delta = 0$. In this case, the samples in the same neighborhood granule are equivalent to each other and the neighborhood rough set model degenerates to Pawlak's one. Therefore, the neighborhood rough sets are a natural generalization of Pawlak rough sets.

In order to deal with heterogeneous features, we give the following definitions to compute neighborhood of samples with mixed numerical and categorical attributes.

Definition 1. Let $B_1 \subseteq A$ and $B_2 \subseteq A$ be numerical attributes and categorical attributes, respectively. The neighborhood granule of sample x induced by B_1, B_2 and $B_1 \cup B_2$ are defined as

(1) $\delta_{B_1}(x) = \{x_i | \Delta_{B_1}(x,x_i) \leqslant \delta, x_i \in U\}$;
(2) $\delta_{B_2}(x) = \{x_i | \Delta_{B_2}(x,x_i) = 0, x_i \in U\}$;
(3) $\delta_{B_1 \cup B_2}(x) = \{x_i | \Delta_{B_1}(x,x_i) \leqslant \delta \wedge \Delta_{B_2}(x,x_i) = 0, x_i \in U\}$, where $\wedge$ means "and" operator.

The first item is designed for numerical attributes; the second one is for categorical attributes, and the last one is for mixed numerical and categorical attributes. Therefore Definition 1 is applicable to numerical, categorical data and their mixture. According to this definition, the samples in a neighborhood granule have the same values in terms of categorical features and the distance in term of numerical features is less than threshold δ.

In fact, besides the above definition there are a number of distance functions for mixed numerical and categorical data [26,35], such as Heterogeneous Euclidean-Overlap Metric function (HEOM), Value Difference Metric (VDM), Heterogeneous Value Difference Metric (HVDM) and Interpolated Value Difference Metric (IVDM). HEOM is defined as

$$\mathrm{HEOM}(x,y) = \sqrt{\sum_{i=1}^{m} w_{a_i} \times d_{a_i}^2(x_{a_i},y_{a_i})},$$

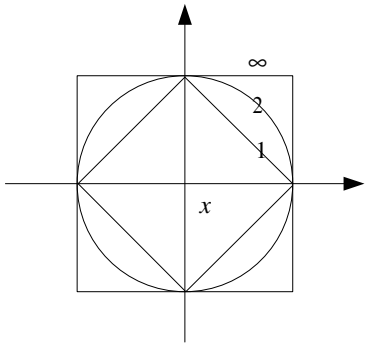


Fig. 1. Neighborhoods of x_0 in terms of different distances.

where m is the number of attributes, w_{a_i} is the weight of attribute a_i, $d_{a_i}(x,y)$ is the distance between samples x and y with respect to attribute a_i, defined as

$$d_{a_i}(x,y) = \begin{cases} 1, & \text{if the attribute value of } x \text{ or } y \text{ is unknown,} \\ overlap_{a_i}(x,y), & \text{if } a_i \text{ is a nominal attribute,} \\ rn\_diff_{a_i}(x,y), & \text{if } a_i \text{is a numerical attribute.} \end{cases}$$

where $overlap_{a_i}(x,y) = \begin{cases} 0, & \text{if } x = y \\ 1, & \text{otherwise} \end{cases}$ and $rn\_diff_{a_i}(x,y) = \frac{|x-y|}{\max_{a_i} - \min_{a_i}}$.

Given a metric space $\langle U, \Delta \rangle$, the family of neighborhood granules $\{\delta(x_i) | x_i \in U\}$ forms an elemental granule system, which covers the universe, rather than partitions it. We have

(1) $\forall x_i \in U$: $\delta(x_i) \neq \emptyset$;
(2) $\cup_{x \in U} \delta(x) = U$.

A neighborhood relation N on the universe can be written as a relation matrix $M(N) = (r_{ij})_{n \times n}$, where

$$r_{ij} = \begin{cases} 1, & \Delta(x_i, x_j) \leqslant \delta, \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to show that N satisfies the properties of reflexivity: $r_{ii} = 1$ and symmetry: $r_{ij} = r_{ji}$.

Obviously, neighborhood relations are a kind of similarity relations, which satisfy the properties of reflexivity and symmetry. Neighborhood relations draw the objects together for similarity or indistinguishability in terms of distances and the samples in the same neighborhood granule are close to each other.

Theorem 1. *Given an information system $\langle U, A \rangle$ and $C_1 \subseteq A$ and $C_2 \subseteq A$, δ is a nonnegative number. N_δ^C is a neighborhood relation induced in feature subspace C with Chebychev distance and δ, we have*

$$N_\delta^{C_1 \cup C_2} = N_\delta^{C_1} \cap N_\delta^{C_2}.$$

Assume $x_i, x_j \in U$, the distance between these samples is $\Delta^C(x_i, x_j) = \max_{a \in C} |f(x_i, a) - f(x_j, a)|$ as to Chebychev distance and features C. We have if and only if $\Delta^{C_1}(x_i, x_j) \leqslant \delta$ and $\Delta^{C_2}(x_i, x_j) \leqslant \delta$ in feature space $C_1 \cup C_2$. Therefore, $N^{C_1 \cup C_2}(x_i, x_j) = 1$ if and only if $N^{C_1}(x_i, x_j) = 1$ and $N^{C_2}(x_i, x_j) = 1$. Based on Theorem 1, we can compute the neighborhood relation over the universe with each attribute independently and the intersection of neighborhood relations is the relation induced with the union of two subsets of features. This property is useful in constructing forward feature selection algorithms, where the first round should compute the neighborhood relations induced by each feature and the rest rounds require computing the relations induced by the feature combinations. With this theorem, we need just compute the neighborhood relations induced by each features and then compute their intersections.

Definition 2. Given a set of objects U and a neighborhood relation N over U, we call $\langle U, N \rangle$ a neighborhood approximation space. For any $X \subseteq U$, two subsets of objects, called lower and upper approximations of X in $\langle U, N \rangle$, are defined as

$$\underline{N}X = \{x_i | \delta(x_i) \subseteq X, x_i \in U\},$$
$$\overline{N}X = \{x_i | \delta(x_i) \cap X \neq \emptyset, x_i \in U\}.$$

Obviously, $\underline{N}X \subseteq X \subseteq \overline{N}X$. The boundary region of X in the approximation space is defined as

$$BNX = \overline{N}X - \underline{N}X.$$

The size of boundary region reflects the degree of roughness of set X in the approximation space $\langle U, N \rangle$. Assuming X is the sample subset with a decision label, generally speaking, we hope the boundary region of the decision should be as small as possible for decreasing uncertainty in decision. The size of boundary region depends on X, attributes to describe U and threshold δ. Delta here can be considered as a parameter to control the granularity level at which we analyze the classification task.

Theorem 2. *Given $\langle U, \Delta, N \rangle$ and two nonnegative δ_1 and δ_2, if $\delta_1 \geqslant \delta_2$, we have*

(1) $\forall x_i \in U$: $N_1 \supseteq N_2$, $\delta_1(x_i) \supseteq \delta_2(x_i)$;
(2) $\forall X \subseteq U$: $\underline{N_1}X \subseteq \underline{N_2}X$; $\overline{N_2}X \supseteq \overline{N_1}X$,

where N_1 and N_2 are the neighborhood relations induced with δ_1 and δ_2, respectively.

Proof. If $\delta_1 \geqslant \delta_2$, obviously, we have $\delta_1(x_i) \supseteq \delta_2(x_i)$. Assuming $\delta_1(x_i) \subseteq X$, we have $\delta_2(x_i) \subseteq X$. Therefore we have $x_i \in \underline{N_2}X$ if $x_i \in \underline{N_1}X$. However, x_i is not necessarily in $\underline{N_1}X$ if we have $x_i \in \underline{N_2}X$. Hence $\underline{N_1}X \subseteq \underline{N_2}X$. Similarly, we can get $\overline{N_2}X \supseteq \overline{N_1}X$. □

Theorem 2 shows that a finer neighborhood relation is produced with a smaller delta; accordingly, the lower approximation is larger than that with a great delta.

Example 1. A dataset, consisting of numerical and nominal attributes, is given in Table 1, where a is a numerical attribute, b is a nominal feature and D is the decision.

We here compute the neighborhood of samples with $\delta = 0.1$. In this case, as to attribute a, $\delta(x_1) = \{x_1\}$; $\delta(x_2) = \{x_2,x_5\}$, $\delta(x_3) = \{x_3\}$; $\delta(x_4) = \{x_4,x_5\}$; $\delta(x_5) = \{x_2,x_4,x_5,x_6\}$; $\delta(x_6) = \{x_4,x_5,x_6\}$. In the same time, we can also divide the samples into a set of equivalence classes according to the feature values of attribute b. $U/b = \{\{x_1,x_2,x_5\},\{x_3,x_4\},\{x_6\}\}$. Based on the decision attribute, the samples are grouped into two subsets: $X_1 = \{x_1,x_3,x_6\}$, $X_2 = \{x_2,x_4,x_5\}$. First we approximate X_1 with the granules induced by attribute a, we get $\underline{N}X_1 = \{x_1,x_3\}$; $\overline{N}X_1 = \{x_1,x_3,x_5,x_6\}$. Analogically, $\underline{N}X_2 = \{x_2,x_4\}$; $\underline{N}X_2 = \{x_2,x_4,x_5,x_6\}$.

According to Definition 1, the information granules induced by a and b are listed as follows:

$\delta(x_1) = \{x_1\} \cap \{x_1,x_2,x_5\} = \{x_1\}$, $\delta(x_2) = \{x_2,x_5\}$, $\delta(x_3) = \{x_3\}$, $\delta(x_4) = \{x_4\}$, $\delta(x_5) = \{x_2,x_5\}$, $\delta(x_6) = \{x_6\}$. With the information provided by attribute a and b, the low and upper approximations of X_1 and X_2 are shown as follows. $\underline{N}X_1 = \{x_1,x_3,x_6\}$, $\overline{N}X_1 = \{x_1,x_3,x_6\}$, $\underline{N}X_2 = \{x_2,x_4,x_5\}$, $\overline{N}X_2 = \{x_2,x_4,x_5\}$.

2.2. Neighborhood decision systems

An information system is called a neighborhood system, denoted by $NIS = \langle U,A,N\rangle$ if there is an attribute in the system generating a neighborhood relation on the universe. To be more specific, a neighborhood information system is also called a neighborhood decision system, denoted by $NDT = \langle U,C \cup D,N\rangle$ if there are two kinds of attributes in the system: condition and decision, and there at least exists a condition attribute which induces a neighborhood relation over the universe.

Definition 3. Given a neighborhood decision system $NDT = \langle U,C \cup D,N\rangle$, $X_1,X_2,\ldots,X_N$ are the object subsets with decisions 1 to N, $\delta_B(x_i)$ is the neighborhood information granule generated by attributes $B \subseteq C$, the lower and upper approximations of decision D with respect to attributes B are defined as

$$\underline{N_B}D = \cup_{i=1}^{N} \underline{N_B}X_i, \quad \overline{N_B}D = \cup_{i=1}^{N} \overline{N_B}X_i,$$

where

$$\underline{N_B}X = \{x_i | \delta_B(x_i) \subseteq X, x_i \in U\}, \quad \overline{N_B}X = \{x_i | \delta_B(x_i) \cap X \neq \emptyset, x_i \in U\}.$$

The decision boundary region of D with respect to attributes B is defined as

$$BN(D) = \overline{N_B}D - \underline{N_B}D.$$

The lower approximation of the decision is defined as the union of the lower approximation of each decision class. The lower approximation of the decision is also called the positive region of the decision, denoted by $POS_B(D)$. $POS_B(D)$ is the subset of objects whose neighborhood granules consistently belong to one of the decision classes. By contraries, the samples in the neighborhood subsets of the boundary region come from more than one decision class. As to classification learning, boundary samples are one class of the sources causing classification complexity because the boundary samples take the similar or the same feature values but belong to different decision classes. This maybe confuses the employed learning algorithm and leads to bad classification performance.

It is easy to show that

(1) $\overline{N_B}D = U$;
(2) $POS_B(D) \cap BN(D) = \emptyset$;
(3) $POS_B(D) \cup BN(D) = U$.

A sample in the decision system belongs to either the positive region or the boundary region of decision. Therefore, the neighborhood model divides the samples into two subsets: positive region and boundary region. Positive region is the set of

Table 1
Example of heterogeneous data

Object	A	b	D
1	0.20	1	n
2	0.85	1	y
3	0.31	2	n
4	0.74	2	y
5	0.82	1	y
6	0.72	3	n

samples which can be classified into one of the decision classes without uncertainty, while boundary region is the set of samples which can not be determinately classified. Intuitively, the samples in boundary region are easy to be misclassified. In data acquirement and preprocessing, one usually tries to find a feature space in which the classification task has the least boundary region.

Example 2. Approximations are demonstrated as shown in Figs. 2 and 3. In the discrete case, the samples are granulated into a number of mutually exclusive equivalence information granules with their feature values, shown as the lattices in Fig. 2. Assuming we want to describe a subset $X \subseteq U$ with these granules, then we will find two subsets of granules: a maximal subset of granules which are included in X and a minimal subset of granules which includes X. Fig. 4 shows an example of binary classification in a 2-D numerical space, where d_1 is labeled with "*" and d_2 is labeled with "+". Taking samples x_1, x_2 and x_3 as examples, we assign spherical neighborhoods to these samples. We can find $\delta(x_1) \subseteq d_1$ and

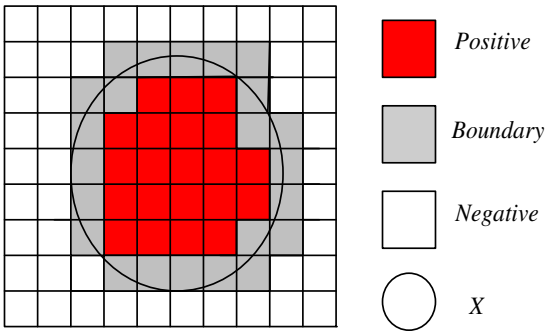


Fig. 2. Rough set in discrete feature space.

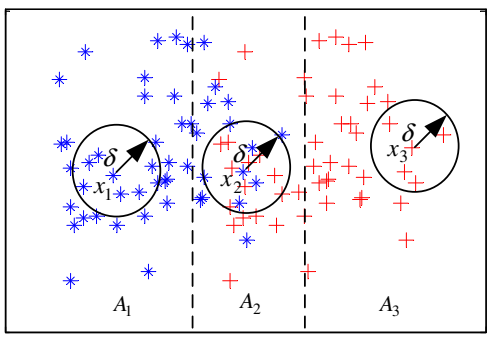
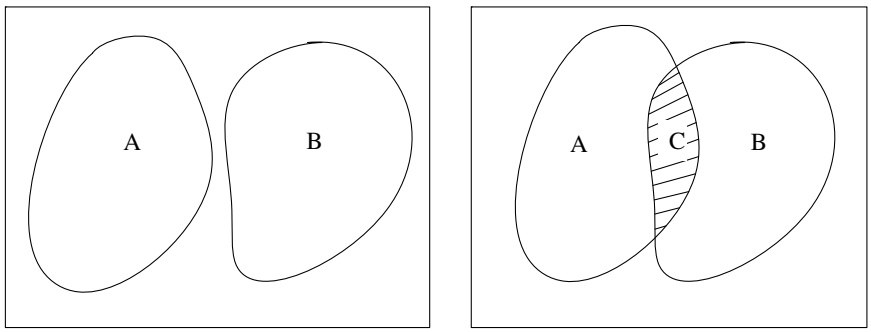


Fig. 3. Rough set in numerical feature space.



(1) Two mutually exclusive classes (2) Two intersection classes

Fig. 4. Geometrical interpretation of dependency.

$\delta(x_3) \subseteq d_2$, while $\delta(x_2) \cap d_1 \neq \emptyset$ and $\delta(x_2) \cap d_2 \neq \emptyset$. According to the above definitions, $x_1 \in \underline{N}d_1$, $x_3 \in \underline{N}d_2$ and $x_2 \in BN(D)$. As a whole, regions A_1 and A_3 are decision positive regions of d_1 and d_2, respectively, while A_2 is the boundary region of decisions.

In practice, the above definitions of lower and upper approximations are sometimes not enough robust for tolerating the noisy samples in the data. For example, assume the sample x_1 in Fig. 3 is mislabeled with d_2. According to the above definitions, all the samples in the neighborhood of x_1 should belong to classification boundary because their neighborhoods are not pure. Obviously, this is not reasonable. Following the idea of variable precision rough sets [47], the neighborhood rough sets can also be generalized by introducing a measure of inclusion degree. Given two sets A and B in universe U, we define A's inclusion degree in B as

$$I(A,B) = \frac{\mathrm{Card}(A \cap B)}{\mathrm{Card}(A)}, \text{ where } A \neq \emptyset.$$

Definition 4. Given any subset $X \subseteq U$ in neighborhood approximation space (U,A,N), we define variable precision lower and upper approximations of X as

$$\underline{N^k}X = \{x_i | I(\delta(x_i), X) \geqslant k, x_i \in U\},$$
$$\overline{N^k}X = \{x_i | I(\delta(x_i), X) \geqslant 1-k, x_i \in U\},$$

where $1 \geqslant k \geqslant 0.5$.

The model degrades to the classical case if $k = 1$. The variable precision neighborhood rough model allows partial inclusion, partial precision and partial certainty which are the coral advantage of granular computing, which simulates the remarkable human ability to make rational decisions in an environment of imprecision [42,43].

2.3. Attribute significance and reduction with neighborhood model

Classification tasks characterized in different feature subspaces have different boundary regions. The size of the boundary region reflects the discernibility of the classification task in the corresponding subspace. It also reflects the recognition power or characterizing power of the corresponding condition attributes. The greater the boundary region is, the weaker the characterizing power of the condition attributes is. Formally, the significance of features can be defined as follows.

Definition 5. Given a neighborhood decision table $\langle U, C \cup D, N\rangle$, distance function Δ and neighborhood size δ, the dependency degree of D to B is defined as

$$\gamma_B(D) = \frac{|POS_B(D)|}{|U|}.$$

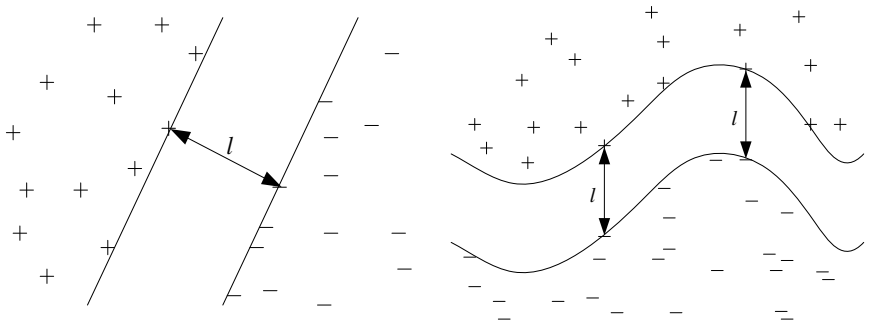
where $|\bullet|$ is the cardinality of a set. $\gamma_B(D)$ reflects the ability of B to approximate D. As $POS_B(D) \subseteq U$, we have $0 \leqslant \gamma_B(D) \leqslant 1$. We say D completely depends on B and the decision system is consistent in terms of Δ and δ if $\gamma_B(D) = 1$; otherwise, we say D depends on B in the degree of γ.

Example 3 (*Geometrical interpretation of dependency*). Fig. 4(1) presents a binary classification task. The patterns in different classes are completely classifiable, and the boundary sample set is empty. In this case, $\gamma_B(D) = |POS_B(D)|/|U| = |A \cup B|/|A \cup B| = 1$. Here the decision is completely dependent on attribute subset B. However, if there is an overlapped region between two classes, as shown in Fig. 4(2), i.e. there are some inconsistent samples, the dependency is computed as

$$\gamma_B(D) = |POS_B(D)|/|U| = |A \cup B|/|A \cup B \cup C| < 1.$$

We can see that the dependency function depends on the size of the overlapped region between classes. Intuitionally, we hope to find a feature subspace, where the classification problem has the least overlapped region because the samples in this region are easy to be misclassified, which will confuse the learning algorithm in training. If the samples are completely separable, the dependency is 1, we say the classification is consistent; otherwise we say it is inconsistent. With an inconsistent classification problem, we try to find the feature subset which gets the greatest dependency.

It is remarkable that there are two kinds of consistent classification tasks: linear and nonlinear, as shown in Fig. 5. Given a classification task in attribute space B, we say the task is consistent if the minimal inter-class distance l between samples is greater than δ because the neighborhood of each sample is pure in this case. We can see that the dependency function can not reflect whether the classification task is linear or nonlinear. On one side, this property lets the proposed measure can be used to deal with linear and nonlinear tasks. On the other side, it shows that the proposed algorithm can not distinguish whether the learning task is linear or nonlinear. The selected features should be validated if a linear learning algorithm is employed after reduction.



(1) linear consistent classification (2) nonlinear consistent classification

Fig. 5. linear and nonlinear neighborhood consistent binary classification. (1) Linear consistent classification and (2) nonlinear consistent classification.

Theorem 3 (Type-1 monotonicity). *Given a neighborhood decision system* $\langle U,C\cup D,N\rangle$, $B_1,B_2\subseteq C$, $B_1\subseteq B_2$, *with the same metric* Δ *and threshold* δ *in computing neighborhoods, we have*

(1) $N_{B_1}\supseteq N_{B_2}$;
(2) $\forall X\subseteq U$, $\underline{N_{B_1}}X\subseteq\underline{N_{B_2}}X$;
(3) $POS_{B_1}(D)\subseteq POS_{B_2}(D)$, $\gamma_{B_1}(D)\leqslant\gamma_{B_2}(D)$.

Proof. (1) $N_{B_1}\supseteq N_{B_{21}}$ means $r_{ij}=1$ if $\pi_{ij}=1$, where r_{ij} and π_{ij} are the elements in relation matrix N_{B_1} and N_{B_2}. Without loss of generality, assuming $\pi_{ij}=1$, we have $\Delta_{B_2}(x_i,x_j)\leqslant\delta$. $\Delta_{B_2}(x_i,x_j)\geqslant\Delta_{B_1}(x_i,x_j)$ if $B_1\subseteq B_2$. So $\Delta_{B_1}(x_i,x_j)\leqslant\delta$, then we have $r_{ij}=1$.

(2) Given $B_1\subseteq B_2$, x and metric Δ $\forall x_i\in U$, $x_i\in\delta_{B_1}(x)$ if $x_i\in\delta_{B_2}(x)$ because $\Delta_{B_1}(x_i,x)\leqslant\Delta_{B_2}(x_i,x)$. Therefore, we have $\delta_{B_1}(x)\supseteq\delta_{B_2}(x)$ if $B_1\subseteq B_2$. Assume $\delta_{B_1}(x)\subseteq\underline{N_{B_1}}X$, where X is one of the decision classes, then we have $\delta_{B_2}(x)\subseteq\underline{N_{B_2}}X$. In the same time, there may be x_i, $\delta_{B_1}(x_i)\not\subset\underline{N_{B_1}}X$ and $\delta_{B_2}(x_i)\subseteq\underline{N_{B_2}}X$. Therefore, $\underline{N_{B_1}}X\subseteq\underline{N_{B_2}}X$.

(3) Given $B_1\subseteq B_2$, $\underline{N_{B_1}}X\subseteq\underline{N_{B_2}}X$. Assuming $D=\{X_1,X_2,\ldots,X_m\}$, we have $\underline{N_{B_1}}X_1\subseteq\underline{N_{B_2}}X_1,\ldots,\underline{N_{B_1}}X_m\subseteq\underline{N_{B_2}}X_m$. $POS_{B_1}(D)=\cup_{i=1}^m\underline{N_{B_1}}X_i$; $POS_{B_2}(D)=\cup_{i=1}^m\underline{N_{B_2}}X_i$, so $POS_{B_1}(D)\subseteq POS_{B_2}(D)$. Then we have $\gamma_{B_1}(D)\leqslant\gamma_{B_2}(D)$. □

Theorem 3 shows dependency monotonically increases with attributes, which means that adding a new attribute in the attribute subset at least does not decrease the dependency. This property is very important for constructing forward feature selection algorithms. Generally speaking, we hope to find a minimal feature subset which has the same characterizing power as the whole samples. The monotonicity of the dependency function is very important for constructing a greedy forward or backward search algorithm [4]. It guarantees that adding any new feature into the existing subset does not lead a decrease of the significance of the new subset.

Theorem 4 (Type-2 monotonicity). *Given a neighborhood decision system* $\langle U,C\cup D,N\rangle$, $B\subseteq C$ *and metric* Δ *in computing neighborhoods, if* $\delta_1\leqslant\delta_2$ *we have*

(1) $N_{\delta_2}\supseteq N_{\delta_1}$;
(2) $\forall X\subseteq U$, $\underline{N_{\delta_2}}X\subseteq\underline{N_{\delta_1}}X$;
(3) $POS_{\delta_2}(D)\subseteq POS_{\delta_1}(D)$, $\gamma_{\delta_2}(D)\leqslant\gamma_{\delta_1}(D)$.

Proof. Please refer to the proof of Theorem 2. □

Taking parameter δ as the level of granularity where we analyze the classification problem at hand, we can find that the complexity of classification not only depends on the given feature space, but also the granularity level at which the problem is discussed. Granularity, here controlled with parameter δ, can be qualitatively characterized with words fine or coarse. Theorem 3 suggests that the classification complexity is related with the information that hides in the available features, while Theorem 4 shows that complexity is also impacted with the granularity level.

Corollary 1. *Given neighborhood decision system* $\langle U,C\cup D,N\rangle$, $B_1\subseteq B_2\subseteq C$,$\delta_1\leqslant\delta_2$ *are two constants; and* Δ *is a metric function defined in U. We have the following properties:*

(1) *Given granularity* δ, $\langle U,B_2\cup D,N\rangle$ *is consistent if* $\langle U,B_1\cup D,N\rangle$ *is consistent.*
(2) *Given feature space B,* $\langle U,B\cup D\rangle$ *is consistent at the level of granularity* δ_1 *if* $\langle U,B\cup D,N\rangle$ *is consistent at the level of granularity* δ_2.
(3) *Decision system* $\langle U,B_2\cup D,N\rangle$ *is consistent at the level of granularity* δ_1 *if* $\langle U,B_1\cup D,N\rangle$ *is consistent at the level of granularity* δ_2.

Definition 6. Given a neighborhood decision table $NDT = \langle U, C \cup D, N\rangle$, $B \subseteq C$, we say attribute subset B is a relative reduct if the following conditions are satisfied:

(1) sufficient condition: $\gamma_B(D) = \gamma_A(D)$;
(2) necessary condition: $\forall a \in B, \gamma_B(D) \rangle \gamma_{B-a}(D)$.

The first condition guarantees that $POS_B(D) = POS_A(D)$ and the second condition shows there is not any superfluous attribute in a reduct. Therefore, a reduct is a minimal subset of attributes which has the same approximating power as the whole set of attributes.

3. Algorithm design for heterogeneous feature selection

As mentioned above, the dependency function reflects the approximating power of a condition attribute set. It can be used to measure the significance of a subset of attributes. The aim of attribute selection is to search a subset of attributes such that the classification problem has the maximal consistency in the selected feature spaces. In this section, we construct some measures for attribute evaluation, and then present greedy feature selection algorithms.

Given a neighborhood decision system $\langle U, C \cup D, N\rangle$, $B \subseteq C$, $\forall a \in B$, one can define the significance of a in B as

$$Sig_1(a, B, D) = \gamma_B(D) - \gamma_{B-a}(D).$$

Note that the significance of an attribute is related with three variables: a, B and D. An attribute a may be of great significance in B_1 but of little significance in B_2. What's more, the attribute's significance is different for each decision attribute if they are multiple decision attributes in a decision table. The above definition is applicable to backward feature selection, where redundant features are eliminated from the original set of features one by one. Similarly, a measure applicable to forward selection can be written as

$$Sig_2(a, B, D) = \gamma_{B\cup a}(D) - \gamma_B(D) \quad \forall a \in A - B.$$

As $0 \leqslant \gamma_B(D) \leqslant 1$ and $\forall a \in B$: $\gamma_B(D) \geqslant \gamma_{B-a}(D)$, we have

$$0 \leqslant Sig_1(a, B, D) \leqslant 1, \quad 0 \leqslant Sig_2(a, B, D) \leqslant 1.$$

We say attribute a is *superfluous* in B with respect to D if $Sig_1(a, B, D) = 0$; otherwise, a is indispensable in B.

The objective of rough set based attribute reduction is to find a subset of attributes which has the same discriminating power as the original data and has not any redundant attribute. Although there usually are multiple reducts for a given decision table, in the most of applications, it is enough to find one of them. With the proposed measures, a forward greedy search algorithm for attribute reduction can be formulated as follows.

Algorithm 1: Naive forward attribute reduction based on neighborhood rough set model (NFARNRS)

Input: $\langle U, C \cup D, f\rangle$
Delta// Control the size of the neighborhood
Output: reduct *red*.
1: $\emptyset \rightarrow red$;// *red* is the pool to contain the selected attributes
2: For each $a_i \in C - red$
3: Compute $\gamma_{red\cup a_i}(D) = \frac{|POS_{B\cup a_i}(D)|}{|U|}$
4: Compute $SIG(a_i, red, D) = \gamma_{red\cup a}(D) - \gamma_{red}(D)$
5: end
6: select the attribute a_k satisfying $SIG(a_k, red, D) = \max_i(SIG(a_i, red, D))$
7: If $SIG(a_k, red, D) \rangle \varepsilon$,// ε is a little positive real number use to control the convergence
8: $red \cup a_k \rightarrow red$
9: go to step2
10: else
11: return *red*
12: end if

There is a parameter delta, control the size of neighborhoods, to be respecified in the algorithm. We discuss how to specify the value of the parameter in Section 4.

There are four key steps in a feature selection algorithm: subset generation, subset evaluation, stopping criterion and result validation [20]. In algorithm **NFARNRS**, we begin with an empty set *red* of attribute, and we add one feature which makes the increment of dependency maximal into the set *red* in each round. This is the strategy of subset generation. We embed the subset evaluation in this strategy by maximizing the increment of dependency. The algorithm does not stop until the dependency increase less than ε by adding any new feature into the attribute subset *red*.

According to the Geometrical interpretation of dependency function, we can understand that the algorithm tries to find a feature subspace such that there is the least overlapped region between classes for a given classification task.

We obtain the goal by maximizing the positive region, accordingly, maximizing the dependency between the decision and condition attributes, and minimizing the boundary samples. Since the samples in the boundary region are easy to be misclassified.

There are two key steps in Algorithm 1. One is to compute the neighborhood of samples, the other is to analyze whether the neighborhood of a sample is consistent. With sorting technique, we can find the neighborhoods of samples in time complexity $O(n\log n)$[35], while time complexity of the second step is $O(n)$. So the worst case of computational complexity of reduction is $O(N^2 n\log n)$, where N and n are the numbers of features and samples, respectively. Assumed there are k attributes included in the reduct, the total computational times are

$$N\times n\log n+(N-1)\times n\log n+\cdots+(N-k)\times n\log n=(2N-k)(k+1)\times n\log n/2.$$

As Theorem 2 shows that the positive region of decision is monotonous with the attributes, we have the following corollary which can be used to speedup the algorithm.

Corollary 2. *Given neighborhood decision system* $\langle U,C\cup D,N\rangle$ *and a metric* Δ, $M,N\subseteq C$, $M\subseteq N$, *if* $x_i\in POS_M(D)$ *then* $x_i\in POS_N(D)$.

Corollary 2 shows that an object necessary belongs to the positive region with respect to an attribute set if it belongs to the positive region with respect to its subset. In forward attribute selection, the attributes are added into the selected subset one by one according to their significances. That is to say, $\forall M\supseteq B, x_i\in POS_M(D)$ if object $x_i\in POS_B(D)$. Therefore, we need not compute the objects in $POS_B(D)$ when compute $POS_M(D)$ because they are necessary in $POS_M(D)$. In this case, we just need to discuss the objects in $U-POS_B(D)$. The objects in $U-POS_B(D)$ get fewer and fewer as the attribute reduction goes on, and the computation will be reduced in selecting a new feature with this idea. Based on this observation, a fast forward algorithm is formulated as follows.

Algorithm 2: Fast forward heterogeneous attribute reduction based on neighborhood rough sets (F2HARNRS)

Input: $\langle U,C\cup D\rangle$

delta//the size of the neighborhood

Output: reduct *red*.

1: $\emptyset\rightarrow red$, $U\rightarrow S$;// *red* is used to contain the selected attributes, S is the set of samples out of positive region.
2: while $S\neq\emptyset$
3: for each $a_i\in A-red$
4: generate a temporary decision table $DT_i=\langle U,red\cup a_i,D\rangle$
5: $\emptyset\rightarrow POS_i$
6: for each $O_j\in S$
7: Compute $\delta(O_j)$ in neighborhood decision table DT_i
8: if $\exists X_k\in U/D$, such that $\delta(O_j)\subseteq X_k$
9: $POS_i\cup O_j\rightarrow POS_i$
10: end if
11: end for
12: end for
13: find a_k such that $|POS_k|=\max_i|POS_i|$
14: if $POS_k\neq\emptyset$
15: $red\cup a_k\rightarrow red$
16: $S-POS_k\rightarrow S$
17: else
18: exit while
19: end if
20: end while
21: return *red*, end

Assumed that there are n samples in a decision table and k features in its reduct, and selecting an attribute averagely leads to n/k samples added into the positive region, the computational times of reduction is

$$N\times n\log n+(N-1)\times n\log n\times\frac{k-1}{k}+\cdots+(N-k)\times\frac{1}{k}n\log n\langle\frac{n\log n}{k}(k+k-1+\cdots+1)=N\times n\log n(k+1)/2.$$

In practice, it is usually found that most of samples are grouped into positive regions at the beginning of reduction, therefore, the computation will be greatly reduced at the sequential round and the reduction procedure greatly speedups.

Moreover, the attributes divide the samples into exclusive subsets if they are discrete. In this case, we don't require comparing samples $U-POS_i$ with U and we just need to compute the relation between $U-POS_i$. Therefore, it is enough to generate a temporary decision table $DT_i=\langle U-POS,red\cup a_i,D\rangle$ in step 4. Algorithm 2 can be written as follows.

Algorithm 3: Fast forward discrete attribute reduction based on Pawlak rough sets (F2DARPRS)

Input: $\langle U, C \cup D \rangle$
Output: reduct *red*

$\emptyset \to red$, $U \to S$;// *red* is used to contain the selected attributes, S is the set of samples out of positive region.
while $S \neq \emptyset$
  for each $a_i \in A - red$
    generate a temporary decision table $DT_i = \langle S, red \cup a_i, D \rangle$
    $\emptyset \to POS_i$
    for each $O_j \in S$
      Compute $\delta(O_j)$ in neighborhood decision table DT_i
      if $\exists X_k \in U/D$, such that $\delta(O_j) \subseteq X_k$
        $POS_i \cup O_j \to POS_i$
      end if
    end for
  end for
  find a_k such that $|POS_k| = \max_i |POS_i|$
  if $POS_k \neq \emptyset$
    $red \cup a_k \to red$
    $S - POS_k \to S$
  else
    exit while
  end if
end while
return *red*, end

If there are n samples in the decision table and k features in the reduct, and selecting an attribute averagely leads to n/k samples added into the positive region, the computational times of reduction are

$$N \times n \log n + (N-1) \times \frac{k-1}{k} \times n \times \log\left(\frac{k-1}{k} \times n\right) + \cdots + (N-k) \times \frac{1}{k} \times n \times \log\left(\frac{1}{k} \times n\right).$$

4. Experimental analysis

In this section, we first compare the computational time of these algorithms on different datasets. Then we present the comparative results with some classical feature selection algorithms. Finally, we discuss the influence of parameters in the neighborhood model.

The effectiveness of classical rough set method in categorical attribute reduction has been discussed in some literatures [40]. Here, we mainly show that the neighborhood model is applicable to the data with numerical attributes or heterogeneous attributes in this work. In order to test the proposed feature selection algorithms, we download some data sets from the machine learning data repository, University of California at Irvine [23]. The datasets are outlined in Table 2. There are nine sets including heterogeneous features, and four sets just with numerical features and the other five databases only with categorical attributes. Furthermore, the numbers of samples are between 101 and 20 000. The last two columns marked with CART and SVM show the classification performances of the original data based on 10-fold cross validation, which is a widely used technique to test and evaluate classification performance. We randomly divide the samples into 10 subsets, and use nine of them as training set and the rest one as the test set. After 10 rounds, we compute the average value and variation as the final performance. Here CART and linear SVM are used as learning algorithms, respectively.

We test the computational efficiency of the three algorithms on data sets mushroom and letter, the computations are conducted on a PC with P4 3.0 GHz CPU and 1 GB memory. Figs. 6–9 present the computational time with these algorithms. We increase the number of samples in the reduction and compare the variation of computational time with different reduction algorithms. Figs. 6 and 7 show the results in the case where the categorical attributes are considered as numerical ones. We compute the distance between different samples and find their neighborhood subsets, whereas, Figs. 8 and 9 are the efficiency that we get the relation between samples by comparing their value and whether they are equivalent. We do not compute the distances in this case.

It is easy to see that the computational time with Algorithm 1 significantly increases with the numbers of samples; however, the computation complexity of Algorithms 2 and 3 is not sensitive with the size of sample sets compared with Algorithm 1.

Now, we test the effectiveness of the algorithms to categorical data. Five datasets are used in the experiment. Table 3 presents the comparison of the numbers of the selected features, the classification performance with CART and linear SVM classifiers, respectively. Lymphography, mushroom, soybean and zoo are greatly reduced with the rough set based attribute

Table 2
Data description

	Data	Samples	Numerical	Categorical	Class	CART	SVM
1	Anneal	798	6	32	5	99.89 ± 0.35	99.89 ± 0.35
2	Credit	690	6	9	2	82.73 ± 14.86	81.44 ± 7.18
3	German	1000	7	12	2	69.90 ± 3.54	73.7 ± 4.72
4	Heart1	270	7	6	2	74.07 ± 6.30	81.11 ± 7.50
5	Heart2	303	6	7	5	48.27 ± 8.25	59.83 ± 68.6
6	Hepatitis	155	6	13	2	91.00 ± 5.45	83.50 ± 5.35
7	Horse	368	7	15	2	95.92 ± 2.30	72.30 ± 3.63
8	Sick	2800	6	22	2	98.43 ± 1.16	96.54 ± 0.67
9	Thyroid	9172	7	22	2	62.68 ± 2.88	67.01 ± 3.39
10	Iono	351	34	0	2	87.55 ± 6.93	93.79 ± 5.08
11	Sonar	208	60	0	2	72.07 ± 13.94	85.10 ± 9.49
12	Wdbc	569	31	0	2	90.50 ± 4.55	98.08 ± 2.25
13	Wine	178	13	0	3	89.86 ± 6.35	98.89 ± 2.34
14	Lymphography	148	0	18	4	69.94 ± 21.95	56.23 ± 5.83
15	Letter	20000	0	16	26	86.56 ± 1.05	82.26 ± 1.09
16	Mushroom	8124	0	22	2	96.37 ± 9.90	92.34 ± 12.61
17	Soybean	683	0	35	19	91.84 ± 5.21	52.85 ± 6.35
18	Zoo	101	0	16	7	90.65 ± 9.13	86.15 ± 9.01

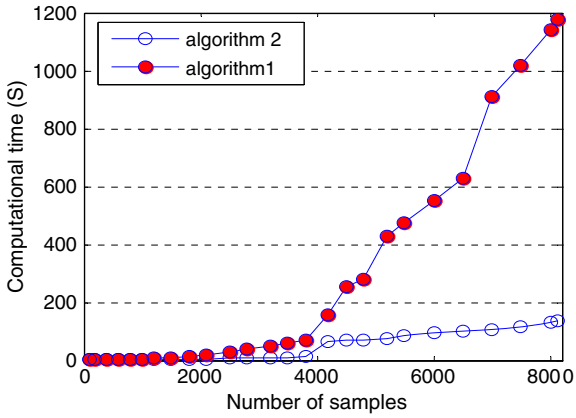


Fig. 6. Comparison of computational efficiency (mushroom).

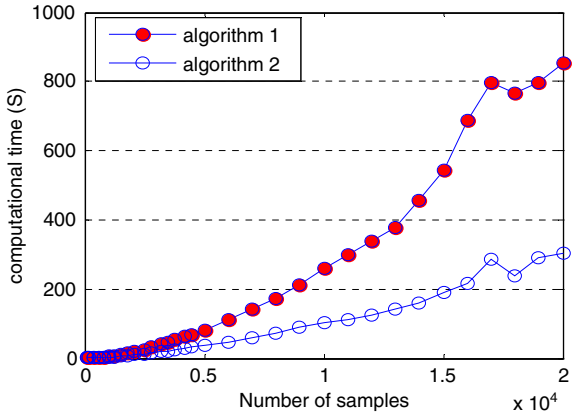


Fig. 7. Comparison of computational efficiency (letter).

reduction algorithm. At the same time, the reduced data of lymphography, soybean and zoo keep or improve the classification performance of the raw datasets.

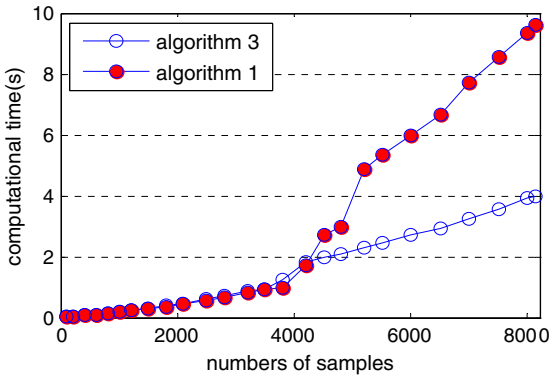


Fig. 8. Comparison of computational efficiency (mushroom).

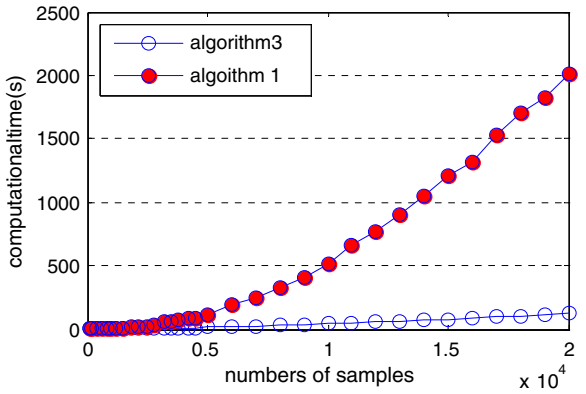


Fig. 9. Comparison of computational efficiency (letter).

Table 3
Performance comparison on discrete data

Data	Raw data			Reduced data		
	Feature	CART	SVM	Feature	CART	SVM
Lymphography	18	69.94 ± 21.95	56.23 ± 5.83	6	68.25 ± 18.22	79.03 ± 11.49
Letter	16	86.56 ± 1.05	82.26 ± 1.09	12	85.69 ± 1.30	78.05 ± 1.40
Mushroom	22	96.37 ± 9.90	92.34 ± 12.61	5	96.37 ± 9.90	87.36 ± 10.39
Soybean	35	91.84 ± 5.21	52.85 ± 6.35	16	85.60 ± 6.64	89.43 ± 5.22
Zoo	16	90.65 ± 9.13	86.15 ± 9.01	5	92.76 ± 9.87	88.51 ± 12.19
Average	21.4	87.07	73.97	8.8	85.73	84.48

Now, we compare the proposed algorithm with some classical feature selection techniques: information entropy based feature selection widely used [11], consistency based algorithm [4] and rough set based QUICKREDUCT algorithm [13]. As these algorithms can just work on categorical data, we discretize the numerical features with MDL [5] in preprocessing. The selected features for each classification problems are presented in Table 4, where numerical features are standardized to interval [0,1], delta takes values in interval [0.1,0.2] for neighborhood rough set based algorithm and 2-norm distance is used. The orders of the feature subsets in Table 4 reflect the sequence the features are selected in.

As different learning algorithms make use of available features in distinct ways, in other words, different learning algorithms may require different feature subsets to produce the best classification performance, we here change the value of delta in the neighborhood rough set model. Delta takes value from 0.02 to 0.4 with step 0.02 to get different feature subsets, and then evaluate the selected features with CART and linear SVM based on 10-fold cross validation. The feature subsets with the greatest average accuracies and the corresponding values of delta are shown in Table 5. We can find with Table 5 that the best feature subsets for CART and SVM are identical to each other in some of the cases; however, they are usually different. This suggests that no algorithm is consistently better than others for all kinds of learning tasks and classification algorithms. Delta here is a parameter to control the granularity of analysis. We can get different properties at

Table 4
Features sequentially selected against different significance criterions

Data	Discretization + entropy	Discretization + consistency	Discretization + RS	NRS
Anneal	27,13,36	27,1,3	27, 3,1	27, 3, 1
Credit	9,11,15,6,14,4,8,3,1,2,7	9,4,10,15,14,2,6,8,3,1,11	4,7,9,15,1,3,11,6,14,8,2	15,8,6,9,7,10,12,4,2,11,13
German	1,4,7,3,11,8,12,2,17,9,5,14	3,1,4,7,11,8,12,2,6,9,14,15	–	2,13,4,1,5,8,9,3,7
Heart1	13,12,3,11,1,7,2,9,8,6,10	13,12,3,11,7,1,8,10,2,9,6	–	10,12,13,3,1,4,5,8,7
Heart2	12,13,3,11,7,9,10,8,2,6	12,3,13,11,7,9,10,8,2,6	12,7,3,11,13,8,9,2,10,6	4,8,1,5,12,3
Hepatitis	18,14,19,13,3,15,17,11,4	18,6,7,13,14,3,9	2,18,8,10,4,5,17,19,13,15,3,12	2,14,17,18,11,15,9,1
Horse	15, 21, 17, 6	15, 21, 17, 6	15, 21, 17, 6	19, 5, 17, 18, 21
Iono	5,6, 3, 7, 33, 27, 12, 13	34, 5, 17, 4, 13, 8, 14, 7	5,3,6,34,17,14,22,4	1,5,19,32,24,20,7,8,3
Sick	2,5,9,10,15,20,22	22,19,29,20,2,1,24,3,11,23,6,10, 14,17,4,9,16,5,8,13	22,19,29,20,2,1,24,3,11,23,6,10, 14,17,4,9,16,5,8,13	24,22,20,19,29,1,26,2,18,10,6,11
Sonar	11, 4, 45, 36, 52, 47, 21, 13, 35, 51, 28, 5, 9, 49	11, 4, 36, 45, 46, 5, 20, 48, 9, 47, 13, 10, 21, 52	–	58,1,45,35,21,27,12,29,54,16, 55,22,10,33,48,34
Thyroid	3,5,6,8,9,10,11,12,13,14,16,17, 18,20,21,22,23,24,25,26,27	28,9,27,2,26,20,3,17,22,24,11,19, 10,14,5,8,4,6,16,7,18,12,13	28,9,27,2,26,20,22,17,24,3,10,16,18, 11,6,14,19,4,8,7,5,13,12,1	21,24,29,20,26,19,2,1,22,10, 3,6,18,11,16,17,4
Wdbc	23, 28, 22, 14, 25, 9, 17	21, 27, 28, 22, 29, 13, 7	24,8,22,26,13,5,14	23, 8, 22, 25, 29, 19
Wine	7, 1, 10, 2, 4	7, 1, 10, 2, 4	10, 13, 7, 2	13, 10, 7, 6,1

Table 5
Optimal features selected for CART and SVM against different deltas

Data	CART	Delta	SVM	Delta
Anneal	27,3,1	0.10	27,3,1	0.10
Credit	11,2,6,14,3,9	0.02	11,2,6,14,3,9	0.02
German	2,4,6,7,1,8,12,11,9,17,14	0.16	2,4,13,1,7,11,8,3,9	0.06
Heart1	10,12,3,1,4,5,13,7	0.12	5,10,9,12,13,3,7,1,2,11,4,8	0.24
Heart2	12,5,10,1,8,3,4,11,6,7	0.12	12,5,10,1,3,4,8,13,7,9	0.14
Hepatitis	2,1,18,17,11	0.08	2,1,18,17,11	0.08
Horse	5,15,17,20,21,16,4	0.06	5,15,17,20,21,16,4	0.06
Iono	1,5,3,28,19,24,31,8,34,21,25,30,32,4,11,7,23,9,18,22,26,10,12	0.4	1,5,19,4,30,34,25,8,3,7,14,12	0.20
Sick	21,24,29,20,26,19,2,1,22,10,3,6,18,11,16,17,4	0.10	24,20,22,19,1,29,2	0.02
Sonar	55,1,48,12,21,26,42,17,6	0.24	55,1,48,12,21,26,42,17,6	0.24
Thyroid	28,9,27,2,26,20,24,22,3,17,10,16,6,11,18,14,8,4,13,19,5,7,12,1	0.04	28,9,27,2,26,20,24,22,3,16,10,11,18,6,17,14,8,4,13,1,5	0.02
Wdbc	23,22, 28,26,25, 8	0.08	8,21,22,19,28,12,25,2,10,29,27,9,23,7,16,1,11,15,6,26,30,18	0.26
Wine	10,7,1	0.02	13,10,11,1,12,5,2,7,3,4	0.32

each level of granularity; accordingly, get a best subset of features to describe the recognition problem for different learning algorithms.

Table 6 gives the numbers of features in raw data and reduced data with different techniques, where the last two columns are the feature numbers with different deltas when CART or SVM produces the best classification performance on the corresponding feature subsets. Mark "$\surd$" after the numbers means the feature subsets are minimal among these feature selection algorithms. As a whole, entropy based algorithm obtains minimal subsets four times, consistency three, RS 3, NRS 4, CART+NRS 7 and SVM+NRS 6, respectively. Furthermore, there are three recognition tasks getting an empty subset of features as to rough set based greedy algorithm. As to these data, $\forall a \in A, POS_a(D) = \emptyset$, then the dependency of each single feature

Table 6
Number of features selected against different significance criterions

Data	raw	Entropy	Consistency	RS	NRS	CART + NRS	SVM + NRS
Anneal	38	3$\surd$	3$\surd$	3$\surd$	3$\surd$	3$\surd$	3$\surd$
Credit	15	11	11	11	11	6$\surd$	6$\surd$
German	19	12	12	0	9$\surd$	11	9$\surd$
Heart1	13	11	11	0	9	8$\surd$	12
Heart2	13	10	10	10	6$\surd$	10	10
Hepatitis	19	9	8	12	8	5$\surd$	5$\surd$
Horse	22	4$\surd$	4$\surd$	4$\surd$	5	7	7
Iono	34	8$\surd$	8$\surd$	8$\surd$	9	23	12
Sick	29	7$\surd$	20	20	12	17	7$\surd$
Sonar	60	14	14	0	16	9$\surd$	9$\surd$
Thyroid	29	21	23	24	17	24	21
Wdbc	31	7	7	7	6$\surd$	6$\surd$	22
Wine	13	5	5	4	5	3$\surd$	10

in these discretized data is zero, so no feature is selected in the first round according to Algorithm 3. This is a main disadvantage of classical rough set based algorithm.

CART and linear support vector machine (SVM) are introduced to test the quality of the selected subsets of features. The classification performances of the raw data and the reduced data based on 10-fold cross validation are shown in Tables 7 and 8, where ONRS denotes the optimal feature subsets selected with variable delta, marks ↑ and ↓ mean that the performances improve or decline compared with the raw data, and √ highlights the highest accuracies among the reduced datasets. As to CART, ONRS outperforms the raw data 12 times over the 13 classification tasks. In the same time, ONRS outperforms the raw datasets seven times with respect to SVM. Moreover, 10 reduced datasets with ONRS and CART produce the best performances over the 5 feature selection techniques, and ONRS and SVM are better than the other algorithms 12 times. As a whole, the average accuracy of ONRS outperforms any other feature selection algorithm in terms of CART and SVM learning algorithms.

Threshold delta plays an important role in neighborhood rough sets. It can be considered as a parameter to control the granularity of data analysis. The significances of attributes vary with the granularity levels. Accordingly, the neighborhood based algorithm selects different feature subsets. Fig. 10 shows the numbers of the selected features and classification accuracies with these features, where 2-norm distance function is used. The classification performances of the feature subsets are based on the average classification accuracies of CART and SVM algorithms. The classification accuracies vary with the threshold delta, where the little subplot in each plot presents the numbers of the selected features when δ takes values from 0.05 to 1 with step 0.05. All the results of the four datasets show a common rule that the numbers of the selected features increase with the value of delta at first, arrive at a peak value, and then decrease. Correspondingly, classification accuracies based on CART and SVM have the similar variation trend. Roughly speaking, [0.1, 0.3] is a candidate interval for delta, where both CART and SVM get good classification performance. Beyond 0.4, the neighborhood rough set based feature selection algorithm can not get enough features to distinguish the samples. Furthermore, we can also see that the numbers of the selected features are different when delta takes values in interval [0.1, 0.3]; however, these features sometimes produce the same classification performance. We should vary the value of delta in [0.1, 0.3] to find a minimal subset of features with a comparable classification performance.

Table 7
Classification accuracy of different feature subsets with CART (%)

Data	Raw data	Entropy	Consistency	RS	NRS	ONRS
Anneal	99.89 ± 0.35	100.00 ± 0.0√	100.00 ± 0.0√	100.00 ± 0.00√	100.0 ± 0.0√	100.00 ± 0.0↑√
Credit	82.73 ± 14.86	82.59 ± 13.88	81.86 ± 14.37	82.88 ± 14.34	82.03 ± 13.54	83.03 ± 18.51↑√
German	69.90 ± 3.54	70.20 ± 5.67	72.60 ± 4.81√	–	69.30 ± 5.03	70.60 ± 5.23↑
Heart1	74.07 ± 6.30	76.30 ± 6.34	76.30 ± 6.34	–	75.93 ± 7.66	78.15 ± 6.49↑√
Heart2	48.27 ± 8.25	51.10 ± 10.15	51.10 ± 10.15	51.39 ± 10.64	51.67 ± 4.99	53.79 ± 6.36↑√
Hepatitis	91.00 ± 5.45	91.00 ± 3.16	87.00 ± 7.45	91.00 ± 4.46	90.33 ± 4.57	92.33 ± 4.57↑√
Horse	95.92 ± 2.30	89.11 ± 4.45	89.11 ± 4.45	89.11 ± 4.45	95.13 ± 3.96	96.18 ± 4.72↑√
Iono	87.55 ± 6.93	91.49 ± 5.25	89.22 ± 5.55	93.18 ± 3.61√	90.06 ± 5.19	92.90 ± 6.51↑
Sick	98.43 ± 1.16	96.61 ± 0.99	97.82 ± 0.85	97.82 ± 0.85	98.54 ± 1.11	98.57 ± 1.13↑√
Sonar	72.07 ± 13.94	74.93 ± 12.89	74.48 ± 10.28	–	73.55 ± 7.22	77.83 ± 8.71↑√
Thyroid	62.68 ± 2.88	61.49 ± 2.61	61.24 ± 4.02	62.55 ± 2.99	62.51 ± 3.01	62.62 ± 3.05√
Wdbc	90.50 ± 4.55	94.01 ± 3.94	94.00 ± 2.48	94.20 ± 3.43	93.14 ± 3.08	94.72 ± 2.23↑√
Wine	89.86 ± 6.35	94.37 ± 3.71	94.37 ± 3.71√	92.08 ± 4.81	92.08 ± 4.81	93.26 ± 4.37↑
Average	81.76	82.55	82.24	–	82.64	84.15

Table 8
Classification accuracy of different feature subsets with SVM (%)

Data	Raw data	Entropy	Consistency	RS	NRS	ONRS
Anneal	99.89 ± 0.35	100.00 ± 0.00√	100.00 ± 0.00√	100.00 ± 0.00√	100.00 ± 0.00√	100.00 ± 0.00↑√
Credit	81.44 ± 7.18	85.48 ± 18.51√	85.48 ± 18.51√	85.48 ± 18.51√	85.48 ± 18.51√	85.48 ± 18.51↑√
German	73.70 ± 4.72	73.70 ± 5.08	73.60 ± 4.45	–	74.00 ± 4.94	74.50 ± 5.82↑√
Heart1	81.11 ± 7.50	84.07 ± 4.64	84.07 ± 4.64	–	83.336.59	84.81 ± 3.68↑√
Heart2	59.83 ± 68.6	58.61 ± 6.34	58.61 ± 6.34	58.61 ± 6.34	58.04 ± 4.11	60.64 ± 5.36↑√
Hepatitis	83.50 ± 5.35	85.67 ± 9.17	84.00 ± 10.52	85.00 ± 7.24	89.00 ± 4.46	90.33 ± 4.57↑√
Horse	92.96 ± 4.43	89.11 ± 4.45	89.11 ± 4.45	89.11 ± 4.45	87.24 ± 3.61	91.55 ± 4.21↓√
Iono	93.79 ± 5.08	82.94 ± 6.74	82.66 ± 6.32	83.30 ± 5.97	87.26 ± 6.06	89.82 ± 5.62↓√
Sick	96.54 ± 0.67	93.89 ± 0.11	97.00 ± 1.07√	97.00 ± 1.07	93.89 ± 0.11	93.89 ± 0.11↓
Sonar	85.10 ± 9.49	77.93 ± 7.47	77.00 ± 8.51	–	74.05 ± 7.60	79.33 ± 8.71↓√
Thyroid	67.01 ± 3.39	62.16 ± 4.53	67.34 ± 3.36	67.30 ± 3.40	67.30 ± 3.40	67.39 ± 3.44↑√
Wdbc	98.08 ± 2.25	95.97 ± 2.32	95.61 ± 2.51	95.09 ± 2.83	95.96 ± 2.62	97.73 ± 2.19↓√
Wine	98.89 ± 2.34	93.82 ± 6.11	93.82 ± 6.11	95.00 ± 4.10	97.22 ± 2.93	98.89 ± 2.34√
Average	85.53	83.33	83.72	–	84.06	85.72

Threshold delta plays an important role in neighborhood rough sets. It can be considered as a parameter to.

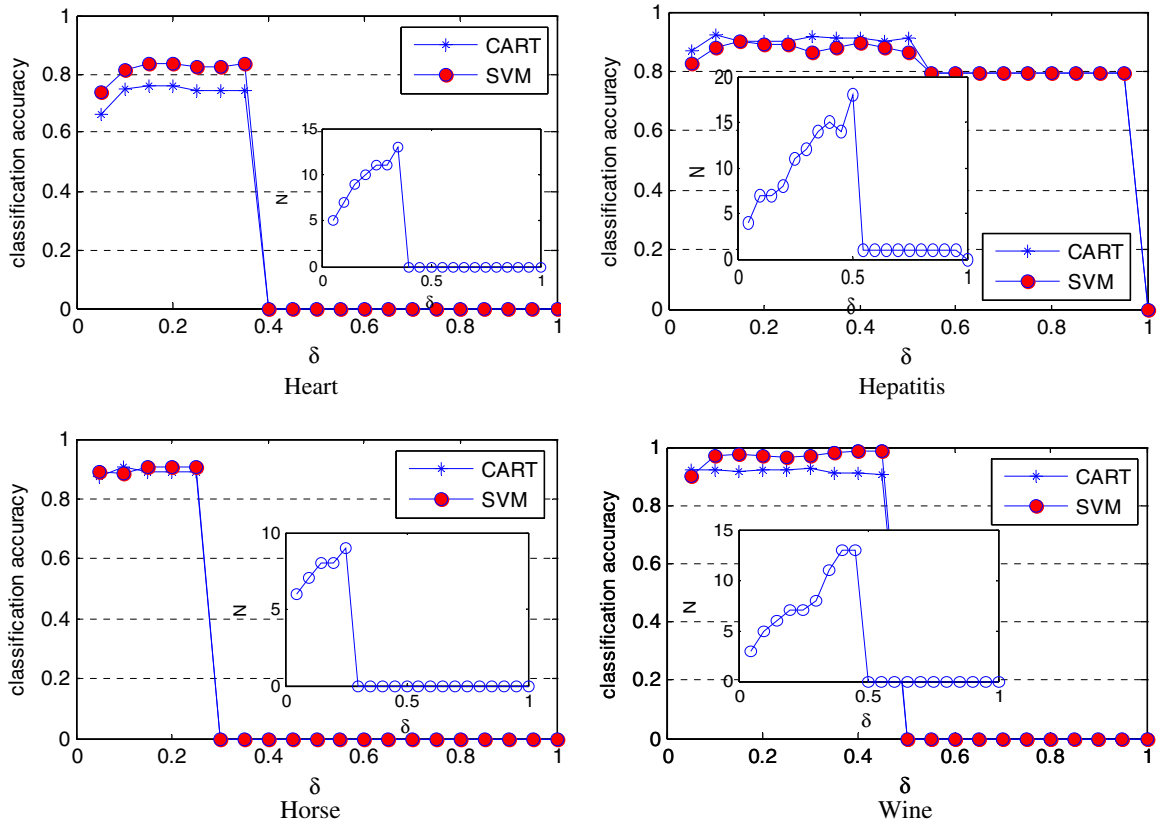


Fig. 10. Variation of feature numbers and classification accuracies with delta (2-norm based distance).

Fig. 11 shows the curves of the numbers of the selected features and the corresponding classification accuracies varying with the size of neighborhood, where infinite norm is introduced to compute the distances between samples. Comparing Fig. 10 with Fig. 11, we can see that the results produced with the Chebychev distance and Euclidean distance based algorithms are much similar with each other. In practical applications, we can select one of the norms in computing the neighborhood and reduction.

5. Conclusion and future work

Reducing redundant or irrelevant features can improve classification performance in most of cases and decrease cost of classification. The classical rough set model is widely discussed in the applications of feature selection and attribute reduction. However, this model can just deal with nominal data. In this work, we propose a technique for heterogeneous feature selection based on neighborhood rough sets. We design a feature evaluating function, called neighborhood dependency, which reflects the percentage of samples in the decision positive region. Theoretical arguments show that the significance of features monotonically increases with the feature subset. This property is important for integrating the evaluating function into some search strategies. Then greedy feature selection algorithms are constructed based on the dependency function.

The experimental results show that the proposed method can be used to deal with both categorical attributes and numerical variables without discretization and it is able to find a small and effective subsets of features in comparison with the classical rough set based reduction algorithm, the consistency based algorithm and the fuzzy entropy method. We also experimentally discuss the influence of the size of neighborhood. It is found that the algorithm gets the good subsets of feature for classification learning if δ takes value in interval [0.1,0.3].

This work shows a novel approach to dealing with heterogeneous feature selection. The future work will be focused on constructing neighborhood classifiers with the proposed model to lay a foundation for neighborhood based learning systems, such as k-nearest neighbor methods and neighborhood counting methods [35]. Indeed, the proposed model takes the similar assumption as KNN methods; they share the idea that classification can be obtained from the information of neighborhoods. Different from KNN, the neighborhood rough set model maybe forms a systematic theoretic framework for heterogeneous data analysis, sample reduction, attribute reduction, dependency analysis and classification learning.

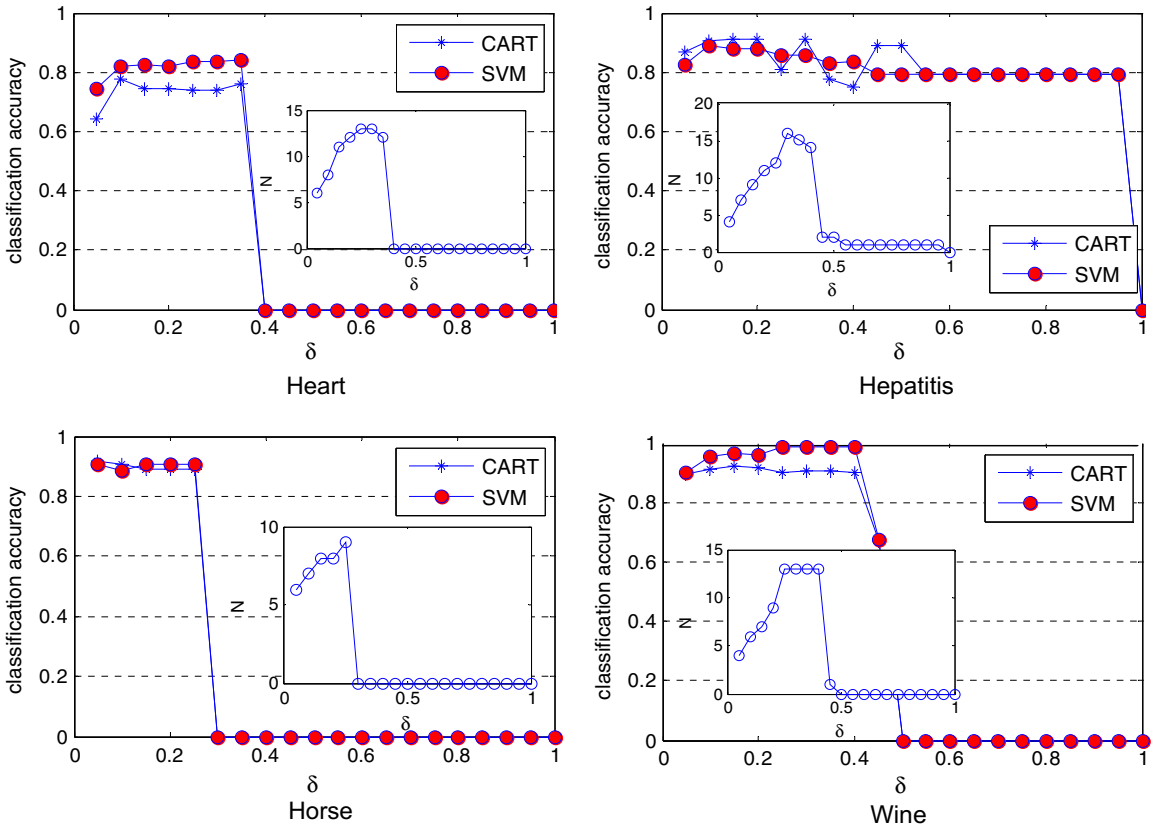


Fig. 11. Variation of feature numbers and classification accuracies with delta (infinite-norm based distance).

Acknowledgement

The authors would like to thank the anonymous reviewers for their constructive comments. This work is partly supported by National Natural Science Foundation of China under Grant 60703013 and Development Program for Outstanding Young Teachers in Harbin Institute of Technology under Grant HITQNJS.2007.017.

References

[1] R.B. Bhatt, M. Gopal, On fuzzy-rough sets approach to feature selection, Pattern Recognition Letters 26 (2005) 965–975.
[2] D.G. Chen, C.Z. Wang, Q.H. Hu, A new approach to attribute reduction of consistent and inconsistent covering decision systems with covering rough sets, Information Sciences 177 (2007) 3500–3518.
[3] J.Y. Ching, A.K.C. Wong, K.C.C. Chan, class-dependent discretization for inductive learning from continuous and mixed-mode data, IEEE Transactions on PAMI 17 (1995) 641–651.
[4] M. Dash, H. Liu, Consistency-based search in feature selection, Artificial Intelligence 151 (2003) 155–176.
[5] U. Fayyad, K. Irani, Discretizing continuous attributes while learning Bayesian networks, in: Proc. 13th International Conference on Machine Learning, Morgan Kaufmann, 1996, pp. 157–165.
[6] M. Hall, Correlation based feature selection for machine learning, Ph.D. Thesis, University of Waikato, Department of Computer Science, 1999.
[7] M. Hall, Correlation-based Feature selection for discrete and numeric class machine learning, in: Proc. 17th ICML. CA, 2000, pp. 359–366.
[8] Q.H. Hu, X.D. Li, D.R. Yu, Analysis on classification performance of rough set based reducts, in: Q. Yang, G. Webb (Eds.), PRICAI 2006, LNAI, vol. 4099, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 423–433.
[9] Q.H. Hu, D.R. Yu, Z.X. Xie, J.F. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Transactions on Fuzzy Systems 14 (2006) 191–201.
[10] Q.H. Hu, D.R. Yu, Z.X. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Letters 27 (2006) 414–423.
[11] Q.H. Hu, D.R. Yu, Z.X. Xie, Neighborhood classifiers, Expert Systems with Applications 34 (2008) 866–876.
[12] Q.H. Hu, J.F. Liu, D.R. Yu, Mixed feature selection based on granulation and approximation, Knowledge-Based Systems 21 (2008) 294–304.
[13] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches, IEEE Transactions of Knowledge and Data Engineering 16 (2004) 1457–1471.
[14] R. Jensen, Q. Shen, Fuzzy-rough sets assisted attribute selection, IEEE Transactions on Fuzzy Systems 15 (1) (2007) 73–89.
[15] W. Jin, Anthony K.H. Tung, J. Han, W. Wang, Ranking outliers using symmetric neighborhood relationship, PAKDD (2006) 577–593.
[16] Y. Li, S.C.K. Shiu, S.K. Pal, Combining feature reduction and case selection in building CBR classifiers, IEEE Transactions on Knowledge and Data Engineering 18 (3) (2006) 415–429.
[17] T.Y. Lin, Neighborhood systems and approximation in database and knowledge base systems, in: Proceedings of the Fourth International Symposium on Methodologies of Intelligent Systems, Poster Session, October 12–15, 1989, pp. 75-86.

[18] T.Y. Lin, Granulation and Nearest Neighborhoods: Rough Set Approach, Granular Computing: An Emerging Paradigm, Physica-Verlag, Heidelberg, Germany, 2001. pp. 125–142.
[19] T.Y. Lin, Neighborhood systems: mathematical models of information granulations, in: 2003 IEEE International Conference on Systems, Man & Cybernetics, Washington, DC, USA, October 5–8, 2003.
[20] H. Liu, L. Yu, Towards integrating feature selection algorithms for classification and clustering, IEEE Transactions on Knowledge and Data Engineering 17 (2005) 491–502.
[21] D.P. Muni, N.R. Pal, J. Das, Genetic programming for simultaneous feature selection and classifier design, IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics 36 (1) (2006) 106–117.
[22] J. Neumann, C. Schnorr, G. Steidl, Combined SVM-based feature selection and classification, Machine Learning 61 (2005) 129–150.
[23] D.J. Newman, S. Hettich, C.L. Blake, C.J. Merz, UCI Repository of Machine Learning Databases, University of California, Department of Information and Computer Science, Irvine, CA, 1998. <http://www.ics.uci.edu/~mlearn/MLRepository.html>.
[24] Z. Pawlak, Rough Sets, Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Dordrecht, 1991.
[25] Z. Pawlak, A. Skowron, Rough Sets: Some Extensions, Information Sciences 177 (2007) 28–40.
[26] D. Randall Wilson, Tony R. Martinez, Improved heterogeneous distance functions, Journal of Artificial Intelligence Research 6 (1997) 1–34.
[27] H. Shin, S. Cho, Invariance of neighborhood relation under input space to feature space mapping, Pattern Recognition Letters 26 (2005) 707–718.
[28] D. Slezak, Approximate reducts in decision tables, in: Proceedings of IPMU' 96, Granada, Spain, 1996, pp. 1159–1164.
[29] R. Slowinski, D. Vanderpooten, A generalized definition of rough approximations based on similarity, IEEE Transactions on Knowledge and Data Engineering 12 (2000) 331–336.
[30] A. Skowron, J. Stepaniuk, Tolerance approximation spaces, Fundamenta Informaticae 27 (1996) 245–253.
[31] Q. Shen, R. Jensen, Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring, Pattern recognition 37 (2004) 1351–1363.
[32] J. Stefanowski, On rough set based approaches to induction of decision rules, in: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery, Physica-Verlag, Heidelberg, Germany, 1998, pp. 501–529.
[33] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Letters 24 (2003) 833–849.
[34] W.Y. Tang, K.Z. Mao, Feature selection algorithm for mixed data with both nominal and continuous features, Pattern Recognition Letters 28 (5) (2007) 563–571.
[35] H. Wang, Nearest neighbors by neighborhood counting, IEEE Transactions on PAMI 28 (2006) 942–953.
[36] Y.Y. Yao, Relational interpretations of neighborhood operators and rough set approximation operators, Information Sciences 111 (1998) 239–259.
[37] D.S. Yeung, D.G. Chen, E.C.C. Tsang, J.W.T. Lee, X.Z. Wang, On the generalization of fuzzy rough sets, IEEE Transactions on Fuzzy Systems 13 (3) (2005) 343–361.
[38] D.R. Yu, Q.H. Hu, W. Bao, Combining rough set methodology and fuzzy clustering for knowledge discovery from quantitative data, Proceedings of the CSEE 24 (6) (2004) 205–210.
[39] L. Yu, H. Liu, Efficiently handling feature redundancy in high-dimensional data, in: Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-03), Washington, DC, August, 2003, pp. 685–690.
[40] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, Journal of Machine Learning Research 5 (2004) 1205–1224.
[41] Z.X. Zhu, Y.S. Ong, M. Dash, Wrapper-filter feature selection algorithm using a memetic framework, IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics 37 (2007) 70–76.
[42] L. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems 19 (1997) 111–127.
[43] L. Zadeh, Fuzzy logic equals computing with words, IEEE Transactions on Fuzzy Systems 4 (2) (1996) 103–111.
[44] N. Zhong, J. Dong, S. Ohsuga, Using rough sets with heuristics for feature selection, Journal of Intelligent Information Systems 16 (3) (2001) 199–214.
[45] W. Zhu, Generalized rough sets based on relations, Information Sciences 177 (2007) 4997–5011.
[46] W. Zhu, F.-Y. Wang, Reduction and axiomization of covering generalized rough sets, Information Sciences 152 (2003) 217–230.
[47] W. Ziarko, Variable precision rough sets model, Journal of Computer and System Sciences 46 (1) (1993) 39–59.