Available online at www.sciencedirect.com
ScienceDirect
Knowledge-Based Systems 21 (2008) 294–304
Knowledge-Based SYSTEMS
www.elsevier.com/locate/knosys

Mixed feature selection based on granulation and approximation

Qinghua Hu *, Jinfu Liu, Daren Yu

Harbin Institute of Technology, Harbin 150001, PR China

Received 22 December 2006; received in revised form 28 May 2007; accepted 28 July 2007
Available online 3 August 2007

Abstract

Feature subset selection presents a common challenge for the applications where data with tens or hundreds of features are available. Existing feature selection algorithms are mainly designed for dealing with numerical or categorical attributes. However, data usually comes with a mixed format in real-world applications. In this paper, we generalize Pawlak's rough set model into δ neighborhood rough set model and k-nearest-neighbor rough set model, where the objects with numerical attributes are granulated with δ neighborhood relations or k-nearest-neighbor relations, while objects with categorical features are granulated with equivalence relations. Then the induced information granules are used to approximate the decision with lower and upper approximations. We compute the lower approximations of decision to measure the significance of attributes. Based on the proposed models, we give the definition of significance of mixed features and construct a greedy attribute reduction algorithm. We compare the proposed algorithm with others in terms of the number of selected features and classification performance. Experiments show the proposed technique is effective.

Keywords: Feature selection; Numerical feature; Categorical feature; δ neighborhood; k-nearest-neighbor; Rough sets

1. Introduction

As the capability of acquiring and storing information increases, more and more candidate features and patterns are gathered in pattern recognition, machine learning and data mining. Generally speaking, the information is usually gathered for multiple learning and mining tasks. Thus, there are a lot of irrelevant or redundant features for a given learning problem. It is observed that irrelevant features will confuse the learning algorithm and deteriorate learning and mining performance [1–3]. Hence it is useful to select parts of features to the learning algorithm in practical applications. Moreover, learning with a subset of features, rather than the whole features, will reduce the cost of acquiring and storing features; speed up learning and recognition.

A great number of feature selection algorithms have been developed in recent years. There are some ways to group these algorithms. There are two key issues in constructing a feature selection algorithm: search strategies and evaluating measures. Accordingly, these algorithms can be grouped in terms of these two dimensions. With respect to search strategies, complete [4], heuristic [5], random [6,7] strategies were introduced in the literatures. Dash and Liu presented an overall review on this issue [8]. Based on evaluating measures, these algorithms can be roughly divided into two classes: classifiers-specific [9,10] and classifier independent. The former employs a learning algorithm to evaluate the goodness of selected features based on the classification accuracies or contribution to the classification boundary, such as the so-called wrapper method [11] and weight based algorithms [12,13]. While the latter constructs a classifier independent measure to evaluate the significance of features, such as inter-class distance [14] mutual information [15,16], dependence measure [17] and consistency measure [5].

In another view of point, one can also group these algorithms into symbolic and numerical methods. Symbolic methods consider that all features take values in a finite set of symbols. The classical rough set model presents a sys-

* Corresponding author. Tel.: +86 451 86413241 252; fax: +86 451 86413241 221.
E-mail address: huqinghua@hcms.hit.edu.cn (Q. Hu).

0950-7051/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2007.07.001

temic theoretic framework for symbolic feature selection [18]. On the contrary, numerical methods assume that the samples are characterized with a set of real-valued variables. In fact, data usually comes with mixed formats in real-world applications, such as medical, marketing, economical analysis. As to numerical algorithms, they recode the symbol attributes with a set of integral numbers, and then treat them as numerical variables, explicitly or implicitly compute the distance between the attribute values [10]; whereas, symbolic methods introduce some discretizing algorithm to partition the value domain of a real-valued variable into several intervals, and the objects in the same interval are assigned with the same symbol, and then the algorithms regard them as symbolic features [19–21]. On one hand, it is sometimes unreasonable to compute similarity or dissimilarity with Euclidian distance as to symbolic attributes. For example, as to attribute *outlook*, it takes values in set {sunny, rainy, overcast}. We can code the value set as 1, 2 and 3, respectively. However, we can also code them with 3, 2 and 1. It is of nonsense to compute distances between the coded values. On the other side, discretizing numeric attributes may bring information loss because the degrees of membership of numerical values to discretized values are not considered [22]. Furthermore, the effectiveness of feature selection significantly depends on the employed discretizing method. A feature selection algorithm for hybrid attributes is thus desirable. However, few researches are focused on this problem in the past. Hall proposed a correlation based feature selection algorithm for discrete and numerical data, where numerical features are discretized [20]. Tang and Mao presented an error probability based measure for mixed feature evaluation [23], they first divided the entire feature space into a set of homogeneous subspaces based on nominal features, then calculated the merit of the mixed feature subset based on sample distributions in the homogeneous subspaces spanned by continuous features. In this algorithm categorical and numerical are differently treated in essence. Moreover Jensen and Shen [24,25], Bhatt and Gopal [26,27], Hu, Yu and Xie [28,29] discussed the fuzzy attribute reduction problems based on fuzzy rough set models, where fuzzy equivalent relations are constructed from fuzzy or numerical attributes.

Pawlak's rough sets are originally proposed to deal with categorical data. In this paper, we will show two generalized rough set model, called k-nearest-neighbor rough sets and δ neighborhood rough sets, for mixed numerical and symbolic feature selection. Symbolic features generate crisp equivalence relations and equivalence classes on the sample spaces, while numerical variables induce a set of so-called k-nearest-neighbor information granules or δ neighborhood information granules in this model, then these granules are used to approximate the decision. We calculate the lower approximation of decision and approximate quality as the significance of features and discuss properties of feature spaces and reduction algorithms based on the proposed models. The contributions of this work are 2-fold. First, we present novel rough set models for mixed feature analysis. Second, we construct a greedy algorithm for mixed numerical and categorical feature selection based on the models. Some experimental results are performed to test the proposed algorithm.

The rest of the paper is organized as follows. Definitions of k-nearest-neighbor and δ neighborhood rough sets are presented in Section 2; analysis on properties of mixed feature spaces is given in Section 3; the feature significance measures and search strategies are shown in Section 4; experiments are described in Section 5. Finally, conclusion comes in Section 6.

2. Generalized rough set model

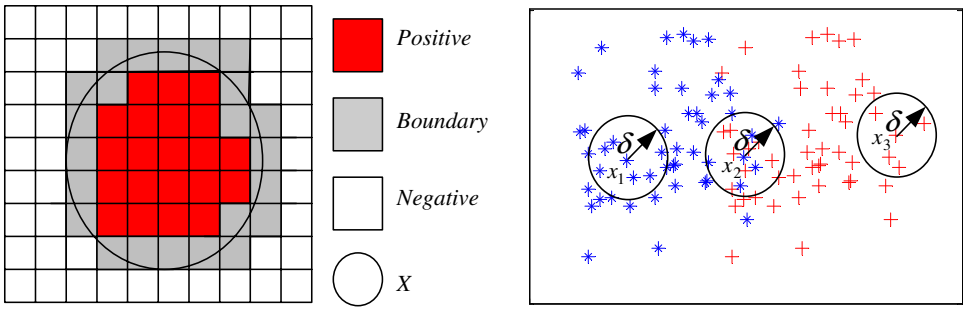
2.1. Basic idea

Granulation and approximation are two fundamental ideas of rough set theory. Information granulation involves partitioning a set of objects into granules, where a granule is a clump of objects which are drawn together by indistinguishability, similarity or functionality [30]. In Pawlak's rough set model, the objects with the same feature values in terms of attributes B are drawn together and form an equivalence class, denoted by $[x]_B$. Equivalence classes are also called elemental information granules or elemental concepts. The family of elemental granules $\{[x_i]_B, x_i \in U\}$ builds a concept system to describe arbitrary subset of the sample space. Due to granularity level and inconsistency in data, subset X may not be precisely described by the elemental information granules. Then two unions of elemental granules are associated with X: lower approximation and upper approximation

$$\underline{B}X = \{[x_i]_B | [x_i]_B \subseteq X\};$$
$$\overline{B}X = \{[x_i]_B | [x_i]_B \cap X \neq \varnothing\}.$$

The lower approximation is the maximal union of elemental granules consistently contained in X, while the upper approximation is the minimal union of elemental granules containing X. The difference between lower approximation and upper approximation is called approximation boundary of X: $BN(X) = \overline{B}X - \underline{B}X$. Elemental granules in boundary region are inconsistent because only parts of their samples belong to X. Rough set based feature selection is to find a minimal subset of feature and the decision has maximal consistent elemental granules in terms of the selected features.

Pawlak's rough set model is built on equivalence relations and equivalence classes. Equivalence relations can be directly induced from categorical attributes based on the attribute values. The samples are said to be equivalent or indiscernible if their attribute values are identical to each other. However, some attributes in data are numerical in real-world applications. Let's consider a two-class problem, as Fig. 1. In the left plot, the sample space is divided into a set of information granules induced with some cate-



(1) Rough sets (2) neighborhood rough sets

Fig. 1. Pawlak's rough sets and neighborhood rough sets.

gorical attributes, where each box denotes an information granule of objects with the same feature values. The granules in the boundary are inconsistent because some of objects in these granules belong to X and the others do not belong to it.

A similar case can also be found in numerical feature spaces, as plot 2. We associate a neighborhood to each object in the sample space, as x_1, x_2 and x_3. It is easy to find that the neighborhood of x_1 are completely contained in class 1, marked with "*", and the neighborhood of x_3 are completely contained in class 2, marked with "+", we say that x_1 and x_3 are the objects in lower approximations of classes 1 and 2, respectively. In the same time, the objects in the neighborhood of x_2 come from classes 1 and 2. Then we define that the samples as x_2 are the boundary objects of the classification. Generally speaking, we hope to find a feature subspace where the boundary region is as little as possible because the samples in boundary region are inconsistent and are easily misclassified. Here we can find that numerical and categorical features can be unified into a framework. In this framework, categorical features generate equivalence information granules of the samples, and numerical features forms neighborhood information granules, and then they are both used to approximate the decision class in the framework of rough sets.

2.2. Neighborhood rough sets

Neighborhood of x_i is a subset of samples close to x_i. There are some ways to define the neighborhoods of samples. One can define it with the fixed radius from the prototype sample. One also can define the neighborhood with fixed k samples in the neighborhood, like k-nearest-neighbor. Whatever, the first issue is to give a metric to compute the distance between objects.

Definition 1. A metric Δ is a function from $R^N \times R^N \to R$ which satisfies the following properties:

P(1) $\Delta(x_1,x_2) \geqslant 0, \quad \forall x_1,x_2 \in R^N; \quad \Delta(x_1,x_2) = 0,$ if and only if $x_1 = x_2$;

P(2) $\Delta(x_1,x_2) = \Delta(x_2,x_1), \quad \forall x_1,x_2 \in R^N;$

P(3) $\Delta(x_1,x_3) \leqslant \Delta(x_1,x_2) + \Delta(x_2,x_3), \quad \forall x_1,x_2,x_3 \in R^N.$

Given a nonempty set X and a metric function Δ, we say X is a metric space, denoted by $\langle X, \Delta \rangle$.

As to real-valued variables, the most frequently used metric is Euclidean distance.

As to categorical attributes, a special metric can be defined as

$$\Delta_C(x,y) = \begin{cases} 1, & \text{if } x \neq y \\ 0, & \text{if } x = y \end{cases}.$$

It is easy to show that Δ_C satisfies the properties of a general metric function. Therefore, categorical spaces are a class of special metric spaces. With the presented metric function, we define the neighborhood of objects.

Definition 2. Given a set of finite and nonempty objects $U = \{x_1, x_2, \cdots, x_n\}$ and a numerical attribute a to described the objects, the δ neighborhood of arbitrary object $x_i \in U$ is defined as

$$\delta_a(x_i) = \{x_j | \Delta(x_i,x_j) \leqslant \delta, x_j \in U\}, \quad \text{where } \delta \geqslant 0.$$

We also called $\delta_a(x_i)$ a neighborhood information granule induced by attribute a and object x_i. The family of neighborhood information granules $\{\delta_a(x) | x \in U\}$ forms a set of elemental concepts in the universe.

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

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

N satisfies the following properties:

P(1) reflectivity : $r_{ii} = 1$;

P(2) symmetry : $r_{ij} = r_{ji}$.

The first property guarantees $\delta_a(x_i) \neq \varnothing$ for $x_i \in \delta_a(x_i)$. As symmetry of distance: $\Delta(x_1,x_2) = \Delta(x_2,x_1)$, we have $\Delta(x_j,x_i) \leqslant \delta$ if $\Delta(x_i,x_j) \leqslant \delta$.

Definition 3. Considering object x and numerical attribute a to describe the object, we call the k-nearest-neighbors of x in terms of a k-nearest-neighbor information granule, denote as $\kappa_a(x)$.

The family of k-nearest-neighbor granules $\{\kappa_a(x) | x \in U\}$ covers the sample space, and we have

P(1) $\forall x_i \in U : \kappa_a(x_i) \neq \varnothing$;

P(2) $\bigcup_{i=1}^{n} \kappa_a(x_i) = U$.

Comparing Definitions 2 and 3, we see that the radium of $\delta_a(x)$ is constant, whereas, the number of objects contained in $\kappa_a(x)$ is fixed. Therefore, the radium of $\kappa_a(x)$ varies with different prototype samples if the samples are not uniformly distributed. In the region of high density, the radium of $\kappa_a(x)$ will become little, and it increases in sparse regions.

$\kappa_a(x)$ operator generates a binary relation over the universe, denoted by κ, where

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

Given relation κ, we have reflexivity: $r_{ii} = 1$. However, the symmetry and transitivity do not hold in this case, namely, it does not necessarily hold that $r_{ij} = r_{ji}$; $r_{ik} = 1$ if $r_{ij} = 1$ and $r_{jk} = 1$.

The family of $\delta_a(x_i)$ or $\kappa_a(x_i)$, $i = 1, 2, \cdots, n$ forms the elemental information granules in numerical spaces to approximate arbitrary subsets of the samples space.

Definition 4. Given arbitrary subset X of the sample space and a family of neighborhood information granules $\delta_a(x_i)$, $i = 1, 2, \cdots, n$, we define the lower and upper approximations of X with respect to neighborhood relation N_a as

$$\underline{\mathrm{N}_a} X = \{x_i | \delta_a(x_i) \subseteq X,\ x_i \in U\},$$
$$\overline{\mathrm{N}_a} X = \{x_i | \delta_a(x_i) \cap X \neq \varnothing,\ x_i \in U\}.$$

If $\underline{\mathrm{N}_a} X = \overline{\mathrm{N}_a} X$, we say X is N_a-definable; otherwise, X is N_a-rough. As to a N_a-rough set, the difference of lower and upper approximations is called the boundary of X: $BN(X) = \overline{\mathrm{N}_a} - \underline{\mathrm{N}_a} X$.

Similarly, the approximations can also be defined in terms of $\kappa_a(x)$.

Definition 5. Given arbitrary subset X of the sample space and a family of k-nearest-neighbor information granules $\kappa_a(x_i)$, $i = 1, 2, \cdots, n$ we define the lower and upper approximations in terms of relation κ as

$$\underline{\kappa_a} X = \{x_i | \kappa_a(x_i) \subseteq X,\ x_i \in U\},$$
$$\overline{\kappa_a} X = \{x_i | \kappa_a(x_i) \cap X \neq \varnothing,\ x_i \in U\}.$$

Similarly, we say X is κ_a-definable if $\underline{\kappa_a} X = \overline{\kappa_a} X$; otherwise, X is κ_a-rough. As to a κ_a-rough set, the difference of lower and upper approximations is called the boundary region: $BN(X) = \overline{\kappa_a} X - \underline{\kappa_a} X$.

In fact, the sole difference between two definitions is the relations defined over the universe. In order to unify the representation of them, we can denote these two kinds of neighborhood relations as R. Accordingly, the lower and upper approximations are denoted by $\underline{R} X$ and $\overline{R} X$, respectively.

In practice, the above definitions of lower and upper approximations are too strict to tolerate noise in the data. Ziarko introduced variable precision rough set model to deal with this problem [31]. Similarly, the neighborhood rough sets can also be generalized into variable precision neighborhood rough sets by introducing inclusion degree.

Definition 6. Given two crisp sets A and B in the universe U, the inclusion degree of A in B is defined as

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

Definition 7. Given any subset $X \subseteq U$, then we define β lower and upper approximations of X as

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

where $1 \geqslant \beta \geqslant 0.5$. The variable precision neighborhood rough model allows partial inclusion, partial precision, partial certainty, which is the coral advantage of granular computing [30], which simulates the remarkable human ability to make rational decisions in an environment of imprecision.

2.3. Neighborhood information systems for mixed features

Classification problems are usually given a set of samples described with some features. These samples form a tabular pattern set. The table is called a neighborhood information system if the features induce a family of neighborhood relations on the universe.

A neighborhood information system is denoted by $NIS = \langle U, A, V, f \rangle$, where U is the sample set, called the universe, A is the attribute set, V is the domain of attribute values. f is an information function f: $U \times A \to V$. More specifically, a neighborhood information system is also called a neighborhood decision table if there are two kinds of attributes in the system: condition and decision, which is denoted by $NDT = \langle U, A \cup D, V, f \rangle$.

Definition 8. Given $NIS = \langle U, A, V, f \rangle$, B is a subset of numerical attributes, the neighborhood of x in terms of attributes B as

$$R_B(x) = \{x_i | x_i \in R_a(x),\ \forall a \in B\}.$$

Definition 9. Given $NIS = \langle U, A, V, f \rangle$, $B = B^n \cup B^c$, where B^n and B^c are subsets of numerical attributes and categorial attributes, respectively, B^n generates neighborhood relation R_{B^n} and B^c generates equivalence relation R_{B^c}, we define the neighborhood granule of x in terms of attributes B as

$$R_B(x) = \{x_i | x_i \in R_{B^n}(x) \wedge x_i \in R_{B^c}(x),\ \forall a_i \in B^n, b_j \in B^c\}.$$

Definition 10. Given a neighborhood decision table $NDT = \langle U, A \cup D, V, f \rangle$, $X_1, X_2, \cdots, X_N$ are the subsets of objects with decisions 1 to N, $R_B(x_i)$ is the neighborhood information granules including x_i and generated with mixed attributes $B \subseteq A$, then the lower and upper approximations of decision D with respect to B are defined as

$$\underline{R_B}D = \{\underline{R_B}X_1, \underline{R_B}X_2, \cdots, \underline{R_B}X_N\},$$
$$\overline{R_B}D = \{\overline{R_B}X_1, \overline{R_B}X_2, \cdots, \overline{R_B}X_N\},$$

where

$$\underline{R_B}X = \{x_i | R_B(x_i) \subseteq X,\ x_i \in U\},$$
$$\overline{R_B}X = \{x_i | R_B(x_i) \cap X \neq \varnothing,\ x_i \in U\}.$$

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

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

Decision boundary is the neighborhood information granules whose objects belong to more than one decision class. Therefore, they are inconsistent. On the other hand, the lower approximation of decision, also called positive region of decision, denoted by $POS_B(D)$, is the set of information granules whose objects consistently belong to one of the decision classes.

Definition 11. The dependency degree of D to B is defined as the ratio of consistent objects:

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

Dependency function reflects the describing capability of attributes B, which can be considered as the significance of attributes B to approximate decision D.

3. Dependency analysis on feature spaces

As the objective of feature selection is to delete the redundant and irrelevant features, analysis on dependence between condition attributes can discover redundancy of features; whereas analysis on dependence between condition attributes and decision can find which condition attributes are irrelevant with the classification problems. In this section, we will review the existing definitions and present a set of new definitions based on the proposed models.

3.1. Existing definitions

Almuallim and Dietterich [32] defined relevance under the assumption that all features and the label are Boolean and that there is no noise. Here each object X is an element of the set $F_1 \times F_2 \times \cdots \times F_m$, where F_i is the domain of the ith features. Training instances are tuples $\langle X, Y \rangle$, where Y is the label. Given an instance, we denote the value of feature X_i by x_i.

Definition 12. A feature X_i is said to be relevant to a concept C if X_i appears in every Boolean formula that presents C and irrelevant otherwise.

Definition 13. X_i is relevant if there exists some x_i and y for which $p(X_i = x_i) > 0$ such that

$$P(Y = y | X_i = x_i) \neq P(Y = y).$$

Note that the above definition fails to capture the relevance of features in the parity concept, and may be changed as follows. Let S_i be the set of all features except X_i. Denoted by s_i a value assignment to all features in S_i.

Definition 14. X_i is relevant if there exists some x_i, y and s_i which $p(X_i = x_i) > 0$ such that

$$P(Y = y, S_i = s_i | X_i = x_i) \neq P(Y = y, S_i = s_i).$$

Definition 15. X_i is relevant if there exists some x_i, y and s_i which $p(X_i = x_i, S_i = s_i) > 0$ such that

$$P(Y = y | X_i = x_i, S_i = s_i) \neq P(Y = y | S_i = s_i).$$

In [33], John, Kohavi and Pfleger pointed out the deficiency of the above definitions based on XOR problems, and proposed the definitions of weak relevance and strong relevance.

As Definition 16, a feature is strong relevant if it cannot be removed without loss of prediction accuracy. Accordingly, a weak relevance feature is defined as follows:

Definition 16. A feature X_i is weakly relevant if it is not strongly relevant, and there exists a subset of features S_i' of S_i for which there exists some x_i, y and s_i' with $p(X_i = x_i, S_i' = s_i') > 0$.

Weak relevance implies that the feature can sometimes contribute to prediction accuracy. Strong relevant cannot be removed, whereas irrelevant features can never contribute to prediction accuracy.

What' more, Yu and Liu [2] presented a definition of redundancy based on Markov blanket.

Definition 17. Given a feature X_i, let $M_i \subset X (X_i \notin M_i)$, M_i is said to be a Markov blank for X_i if

$$p(X - M_i - \{F_i\}, Y | X_i, M_i) = p(X - M_i - \{F_i\}, Y | M_i).$$

Definition 18. Let G be the current set of features, a feature is redundant and hence should be removed from G if it is weakly and has a Markov blanket M_i within G.

The above definitions give the structure of feature spaces based on prediction accuracy. In next section, we will present the similar definitions in terms of rough sets.

3.2. Dependency analysis with neighborhood rough sets

As mentioned above, $\gamma_B(D)$ reflects the ability of B to approximate D. Obviously, $0 \leqslant \gamma_B(D) \leqslant 1$. We say D completely depends on B if $\gamma_B(D) = 1$, denoted by $B \Rightarrow D$; otherwise, we say $D\gamma$-*depends* on B, denoted by $B \Rightarrow_\gamma D$.

Theorem 1. *$\langle U, A \cup D, V, f \rangle$ is a decision table; A is the set of condition attributes, D is the decision. B_1, $B_2 \subseteq A$, then we have*

P(1) $B_1 \subseteq B_2 : R_{B1} \supseteq R_{B2}$ and $\forall X \subseteq U$, $\underline{R_{B_1}}X \subseteq \underline{R_{B_2}}X$, $\overline{R_{B_1}}X \supseteq \overline{R_{B_2}}X$;

P(2) $B_1 \subseteq B_2 : POS_{B_1}(D) \leqslant POS_{B_2}(D)$, $\gamma_{B_1}(D) \leqslant \gamma_{B_2}(D)$.

Proof. $\forall x \in U$, we $\delta_{B_1}(x) \supseteq \delta_{B_2}(x)$ if $B_1 \subseteq B_2$. Assume $\delta_{B_1}(x) \subseteq \underline{N_{B_1}X}$, where X is one of the decision class, then we have $\overline{\delta_{B_2}(x)} \subseteq \underline{N_{B_2}X}$. In the same time, there may be x_i, $\delta_{B_1}(x_i)n \subset \underline{N_{B_1}X}$ and $\delta_{B_2}(x_i) \subseteq \underline{N_{B_2}X}$. Therefore, $POS_{B_1}(D) \subseteq POS_{B_2}(D)$. Accordingly, we have $\gamma_{B_1}(D) \leqslant \gamma_{B_2}(D)$. □

Theorem 2. *$\langle U, A \cup D, V, f \rangle$ is a decision table, A is the set of condition attributes, D is the decision. $B_1, B_2 \subseteq A$, we have*

P(1) if $B_1 \subseteq B_2$ and $B_1 \Rightarrow D$, then $B_2 \Rightarrow D$;
P(2) if $B_1 \Rightarrow D$, then $B_1 \cup B_2 \Rightarrow D$;
P(3) if $B_1 \Rightarrow B_2$ and $B_2 \Rightarrow D$, then $B_1 \Rightarrow D$.

Definition 19. Given a neighborhood decision table $NDT = \langle U, A \cup D, V, f \rangle$, $B \subseteq A$, $\forall a \in B$, we say a is superfluous in B if $\gamma_{B-a}(D) = \gamma_B(D)$; otherwise, we say a is indispensable. We say attribute B is independent relative to the decision D if $\forall a \in B$ is indispensable.

Definition 20. Given a neighborhood decision table $NDT = \langle U, A \cup D, V, f \rangle$, $B \subseteq A$, we say attribute set B is a relative reduct if

(1) $\gamma_B(D) = \gamma_A(D)$;
(2) $\forall a \in B$, $\gamma_B(D) > \gamma_{B-a}(D)$.

The first condition guarantees that $POS_B(D) = POS_A(D)$. The second condition guarantees there is no superfluous attribute in the reduct. Therefore, a reduct is the minimal subset of attributes which has the same approximating power as the whole attribute set. Theoretically speaking, reducts are the optimal feature subsets for classification.

There are usually multiple reducts in an information system. In other words, we can find more than one subset of features, which has the same prediction capability as the whole features, each reduct presents a point of view to understand the classification problem.

Let $\langle U, A \cup D, V, f \rangle$ be a decision table and $\{B_j | j \leqslant r\}$ is the set of reducts, we denote the following attribute subsets:

$$Core = \bigcap_{j \leqslant r} B_j, \quad K = \bigcup_{j \leqslant r} B_j - Core, \quad K_j = B_j - Core,$$
$$I = A - a \bigcup_{j \leqslant r} B_j.$$

Definition 21. *Core* is the attribute subset of *strong relevance*, which cannot be deleted from any reduct; otherwise the prediction power of the system will decrease. Namely, $\forall a \in Core$, $\gamma_{A-a}(D) < \gamma_A(D)$.

Definition 22. I is the *completely irrelevant* attribute set. The attribute in I will not be included in any reduct, which means I is completely useless in the system.

Definition 23. K_j is a *weak relevant attribute set*. The union of *Core* and K_j forms a reduct of the information system.

Definition 24. Given a feature subset $B = Core \cup k_i$, then $\forall a \in k_j$, $j \neq i$, is said to be redundant.

The structure of attribute and attributes sets are shown in Fig. 2.

4. Feature selection algorithms

The previous definitions discover the structure of feature spaces. Theoretically, there exist $2^N - 1$ combinations of features for a data set with N features and r combinations have the same prediction power as the original data. In most cases, the objective of feature selection is to find one of the r combinations. We cannot try all of the potential combinations to find a reduct in short time if there are many samples and features. Then a fast algorithm is desired.

Definition 25. (*Significance measures*) Consider a decision table $\langle U, A \cup D, V, f \rangle$, where A is the set of condition attributes, D is the decision. $B \subset A$, $a \in A - B$, then the significance of attribute a relative to B and D is defined as

$$SIG_1(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D).$$

$SIG(a, B, D)$ reflects the increment of dependency, which means the positive region increases if we add attribute a in B, the increment of positive region is the significance of attribute a.

Based on Theorem 1, we have $0 \leqslant SIG(a, B, D) \leqslant 1$. If $SIG(a, B, D) = 0$, we say a is superfluous, which means a is useless for B to approximate D.

Similarly, the significance of attribute a can also be written as

$$SIG_2(a, B, D) = \gamma_B(D) - \gamma_{B-a}(D), \quad a \in B.$$

The second important issue in constructing feature selection algorithms is search strategies. In [5], Dash and Liu compared five kinds search strategies: focus [34]: exhaustive search; complete [35]: automatic branch and bound search; SetCover: heuristic search [36]; LVF: probabilistic search [37] and QBB: hybrid search [38]. In this paper, we do not try to compare all kinds of search strategies. We introduce the greedy search strategy for its efficiency [22,27,29,39,42].

Formally, a forward greedy algorithm for mixed feature reduction can be formulated as follows.

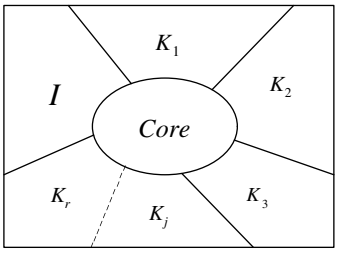


Fig. 2. Structure of an attribute space.

Algorithm: forward attribute reduction based on variable precision neighborhood model (FarVPN)

Input: Hybrid decision table $\langle U, A^c \cup A^n \cup D\rangle$ and β and d *or* k // A^c andA^n are categorical and numerical attributes, respectively.
// β is the threshold for computing variable precision lower approximations; d is the radius of neighborhoods // and k is size of $k-NN$.
Output: One reduct *red*.
Step 1: $\forall a \in A^c$: compute equivalence relation R_a;
Step 2: $\forall a \in A^r$: compute neighborhood relation N_a or κ_a;
Step 3: $\varnothing \to red$; // *red* is the pool to contain the selected attributes
Step 4: For each $a_i \in A - red$

$$\text{Compute } SIG(a_i, red, D) = \gamma^{\beta}_{red\cup a_i}(D) - \gamma^{\beta}_{red}(D),$$
// we define $\gamma^{\beta}_{\varnothing}(D) = 0$

end
Step 5: select the attribute a_k which satisfies: $SIG(a_k, red, D) = \max(SIG(a_i, red, D))$
Step 6: If $SIG(a_k, red, D) > 0$,

$$red \cup a_k \to red$$

go to Step 4
else
return *red*
Step 7: end

If there are N condition attributes and n samples, the time complexity for computing relation between samples is $n \times n$, the worst search time for a reduct is N^2n^2. In real-world applications, only minority of the attributes are included in the reduct. The computing time of forward algorithm will greatly reduce in these cases. Furthermore, fast algorithms for searching k-nearest-neighbors and neighborhood can also be introduced to speed up the procedure [43].

In the following, the algorithm is called **FarVPKNN** if relation matrices are generated with k-nearest-neighbor algorithm as to numerical attributes and **FarVPDN** if relation matrices are computed with δ neighborhood.

5. Experimental analysis

In this section, we empirically evaluate the proposed methods by comparing FarVPN with other attribute reduction algorithms. We compare the numbers of selected features and classification accuracies with the reduced data. Here CART and SVM, two popular learning algorithms, are employed to validate the goodness of selected features based on 10-fold cross validation. What's more, we will show the influence of parameters d, k and β used in FarVPN.

As rough sets based categorical attribute reduction has been reported and compared in the literatures [40], we focus on dealing with numerical or mixed feature reduction in this work. Ten data sets are downloaded from the machine learning data repository, University of California at Irvine. The data sets are outlined in Table 1.

We can find all of the data sets are with numeric attributes; what' more, there are some categorical attributes in data sets *Credit*, *Ecoli* and *Heart*. Before computing reduct, all numerical attributes are normalized into interval $[0,1]$.

The following algorithms are compared:

(1) *Discretization based method:* The traditional method on dealing with numerical features is to discretize the numerical attributes. In order to compare with classical rough sets, here we introduce FCM clustering algorithm divide each numerical attribute into four intervals [41]. Then numerical attributes are recoded and looked as categorical features. We call this method as discretization based method.
(2) *Consistency based method:* Dash and Liu presented a consistency measure for feature selection in [5]. Here we will compare our methods with it. As this method works on discrete domain, we take the same discretization as above.
(3) *Fuzzy entropy based method:* Hu, Yu and Xie presented a fuzzy information entropy based algorithm for hybrid data reduction [29].
(4) *FarVPDN:* δ neighborhood relation is computed with each numerical feature where $\delta = 0.25$, and crisp equivalence relation is generated with each categorical attributes.
(5) *FarVPKNN:* k-nearest-neighbor relations is computed with numerical attributes, where $k = 0.25N$, N is the number of samples. Categorical attributes are computed as FarVPDN.

The experimental results are shown in Tables 2–5. Table 2 shows the comparison of numbers of selected features based on the five feature selection algorithms. Table 3 presents the selected features. Tables 4 and 5 present the comparisons of classification accuracy of selected features based on CART and RBF-SVM learning algorithms, respectively, where the boldface highlights the highest accuracy over different selecting algorithms.

From the tables, we can find all of the feature selection algorithms can remove parts of the candidate features while keep or improve classification accuracies in most of the cases. However, discretization based method cannot choose any feature from the discretized data sets: *diabe* and *heart*. Indeed, this is a limitation of dependence based forward greedy reduction algorithm for $\forall a \in A$, $POS_a(D) = \varnothing$ and $\gamma_a(D) = 0$ in the first cycle of the algorithm. Therefore, no feature will be selected.

Comparing the accuracies in Tables 4 and 5, we can find that the selected features based on fuzzy entropy based method, FarVPDN and FarVPKNN outperform

Table 1
Data description

Data set	Abbreviation	Samples	Numerical features	Categorical features	Classes
Australian credit approval	Crd	690	6	9	2
Pima Indians diabetes	Diab	768	8	0	2
Ecoli	Ecoli	336	5	2	7
Heart disease	Heart	270	7	6	2
Ionosphere	Iono	351	34	0	2
Sonar, mines vs. rocks	Sonar	208	60	0	2
Small soybean	Soy	47	35	0	4
Wisconsin diagnostic breast cancer	WDBC	569	31	0	2
Wisconsin prognostic breast cancer	WPBC	198	33	0	2
Wine recognition	Wine	178	13	0	3

Table 2
Comparison of numbers of selected features based on different selection algorithms

	Original data	Discretization	Consistency	Fuzzy entropy	FarVPDN	FarVPKNN
Crd	15	12	11	13	13	1
Diab	8	0	7	8	7	8
Ecoli	7	1	6	7	6	7
Heart	13	0	8	9	9	7
Iono	34	10	9	13	12	11
Sonar	60	6	6	12	7	6
Soy	35	2	2	2	2	2
WDBC	30	8	11	17	21	7
WPBC	33	7	7	17	11	1
Wine	13	4	4	9	6	6
Average	24.80	5	7.1	10.70	9.40	5.6

Table 3
Features sequentially selected against different significance criterions

Data	FarVPDN	FarVPKNN
Credit	15, 11, 6, 9, 7, 10, 12, 4, 3, 2, 1, 13, 8	3
Heart	5, 10, 12, 13, 3, 1, 7, 11, 2	13, 5, 10, 12, 3, 1, 4
Iono	1, 5, 28, 8, 12, 29, 31, 34, 7, 24, 32, 3	7, 5, 1, 15, 3, 32, 12, 8, 27, 2, 18
Sonar	55, 1, 48, 19, 37, 23, 12	11, 12, 49, 15, 1, 4
WDBC	23, 28, 21, 12, 22, 9, 25, 10, 8, 19, 2, 26, 5, 16, 27, 30, 29, 1, 11, 15, 3	28, 14, 22, 12, 19, 25, 3
Wine	13, 10, 7, 1, 5, 2	10, 13, 7, 6, 2, 1

Table 4
Comparison of classification accuracy based on CART algorithm

	Original data	Discretization	Consistency	Fuzzy entropy	FarVPDN	FarVPKNN
Crd	0.8217 ± 0.0459	0.8274 ± 0.1398	0.8158 ± 0.1446	0.8144 ± 0.1416	0.8288 ± 0.1496	$\mathbf{0.8548 \pm 0.1851}$
Diab	0.7227 ± 0.0512	0.0000 ± 0.0000	0.7253 ± 0.0548	0.7213 ± 0.0404	$\mathbf{0.7253 \pm 0.0493}$	0.7227 ± 0.0512
Ecoli	0.8197 ± 0.0444	0.4262 ± 0.0170	0.8168 ± 0.0429	0.8197 ± 0.0444	0.8168 ± 0.0429	$\mathbf{0.8197 \pm 0.0444}$
Heart	0.7407 ± 0.0630	0.0000 ± 0.0000	0.7815 ± 0.0863	0.7593 ± 0.0766	0.7593 ± 0.0766	$\mathbf{0.7851 \pm 0.0757}$
Iono	0.8755 ± 0.0693	$\mathbf{0.9089 \pm 0.0481}$	0.9062 ± 0.0600	0.9068 ± 0.0564	0.9063 ± 0.0396	0.9034 ± 0.0528
Sonar	0.7207 ± 0.1394	0.6926 ± 0.0863	0.6976 ± 0.0760	0.7160 ± 0.0857	0.7550 ± 0.0683	$\mathbf{0.8074 \pm 0.0986}$
Soy	0.9750 ± 0.0791	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
WDBC	0.9050 ± 0.0455	0.9351 ± 0.0339	0.9069 ± 0.0273	0.9193 ± 0.0318	0.9228 ± 0.0361	$\mathbf{0.9473 \pm 0.0394}$
WPBC	0.6963 ± 0.0826	0.6955 ± 0.1018	0.6924 ± 0.1395	0.7103 ± 0.1092	0.6453 ± 0.1292	$\mathbf{0.7434 \pm 0.0907}$
Wine	0.8986 ± 0.0635	0.8972 ± 0.0741	0.8972 ± 0.0741	0.9097 ± 0.0605	0.9208 ± 0.0481	$\mathbf{0.9382 \pm 0.0409}$
Average	**0.8176**	**0.6383**	**0.8240**	**0.8277**	**0.8280**	**0.8522**

those selected with discretization and consistency based methods. Especially, although FarVPKNN method deletes most of the candidate features, average classification accuracy greatly improve. It shows FarVPKNN is able to find the most informative features for classification.

Table 5
Comparison of classification accuracy based on RBF-SVM algorithm

	Original data	Discretization	Consistency	Fuzzy entropy	FarVPDN	FarVPKNN
Crd	0.8144 ± 0.0718	0.8058 ± 0.0894	0.8058 ± 0.0894	0.8144 ± 0.0718	0.8144 ± 0.0718	**0.8548 ± 0.1851**
Diab	0.7747 ± 0.0430	0.0000 ± 0.0000	**0.7669 ± 0.0377**	0.7747 ± 0.0430	0.7747 ± 0.0430	0.7747 ± 0.0430
Ecoli	0.8512 ± 0.0591	0.4262 ± 0.0170	0.8512 ± 0.0591	0.8512 ± 0.0591	**0.8512 ± 0.0591**	**0.8512 ± 0.0591**
Heart	0.8111 ± 0.0750	0.0000 ± 0.0000	0.8074 ± 0.0488	0.8074 ± 0.0488	0.8074 ± 0.0488	**0.8519 ± 0.0462**
Iono	0.9379 ± 0.0507	0.9348 ± 0.0479	**0.9519 ± 0.0423**	0.9462 ± 0.0365	0.9293 ± 0.0627	0.9404 ± 0.0544
Sonar	0.8510 ± 0.0948	0.7074 ± 0.1004	0.7843 ± 0.0742	0.8271 ± 0.0902	**0.8364 ± 0.0837**	0.8317 ± 0.0728
Soy	0.9300 ± 0.1135	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
WDBC	0.9808 ± 0.0225	0.9649 ± 0.0183	0.9579 ± 0.0238	0.9702 ± 0.0248	**0.9790 ± 0.0161**	0.9684 ± 0.0162
WPBC	0.7779 ± 0.0420	0.7837 ± 0.0506	0.7632 ± 0.0304	**0.8087 ± 0.0601**	0.7842 ± 0.0769	0.7632 ± 0.0304
Wine	0.9889 ± 0.0234	0.9486 ± 0.0507	0.9486 ± 0.0507	**0.9833 ± 0.0268**	**0.9833 ± 0.0268**	0.9778 ± 0.0287
Average	**0.8718**	**0.6571**	**0.8637**	**0.8783**	**0.8760**	**0.8814**

Furthermore, we empirically study the impact of parameters d, k and β on selected features. First we try $d = 0$–1 with step 0.05 and $\beta = 0.5$–1 with step 0.05, and perform the reduction algorithm on wine data. Similarly, we try $k = 0.1N$ to $0.5N$ with step $0.05N$ and $\beta = 0.5$–1 with step 0.05, where N is the number of samples.

From Figs. 3–8, we can see although the numbers of selected features vary from 3 to 20 with parameters D, K and β, most of the regions in Figs. 5–8 get high classification accuracies; they are higher than 90%. This shows that we can assign the parameters with values in wide arranges. They do not influence the classification performance too much. However, a combination of a great D and β or great K and β is not recommended because there will be no or few features selected in this case.

6. Conclusions and future work

We give two rough set models, named δ neighborhood rough sets and k-nearest-neighbor rough sets, for mixed numerical and categorical feature selection and reduction in this work. A forward greedy mixed attribute reduction algorithm is constructed to find minimal subsets of features which can keep classification ability based on the proposed model. In order to test the effectiveness of the proposed method, we compare discretization based method, consistency based method and fuzzy entropy based method with the proposed one on 10 UCI data sets and we introduce

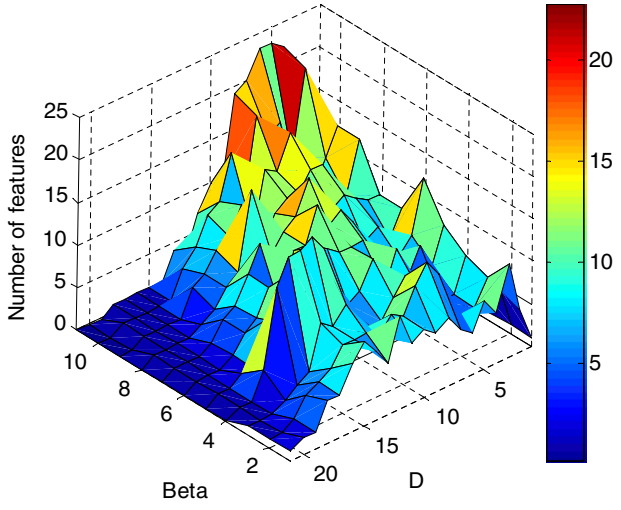


Fig. 3. Number of features varies with D and β.

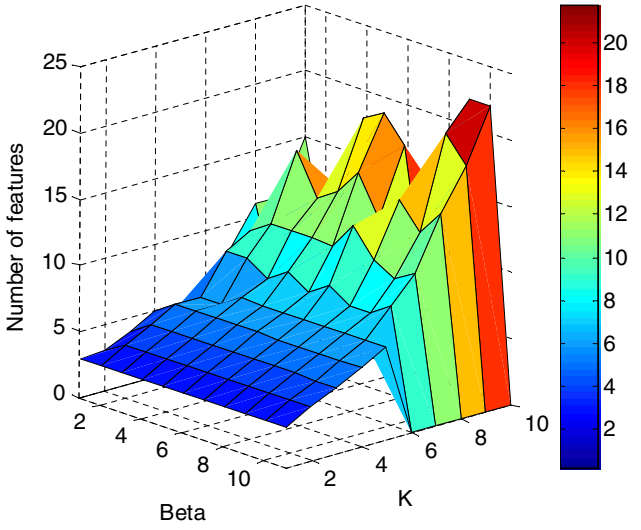


Fig. 4. Number of features varies with D and K.

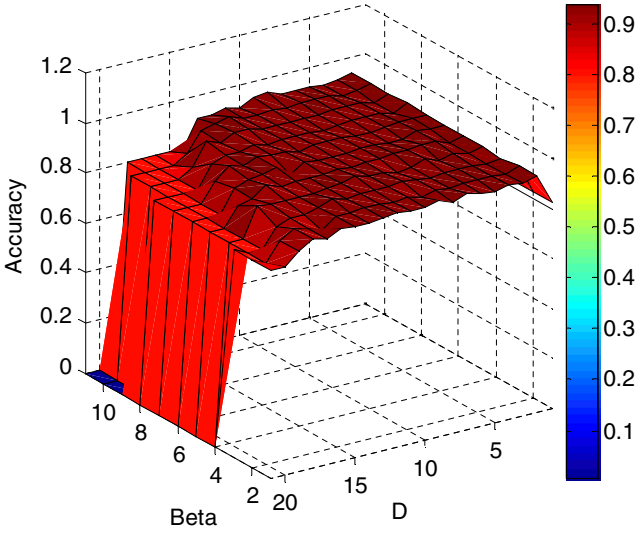


Fig. 5. Accuracy varies with D and β (CART).

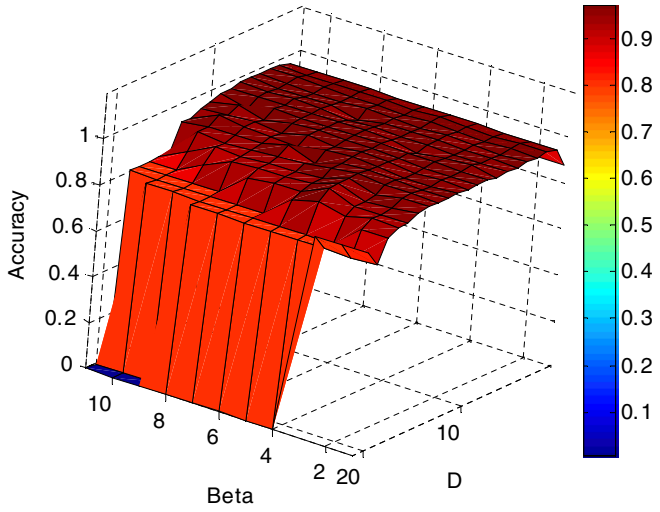


Fig. 6. Accuracy varies with D and β (SVM).

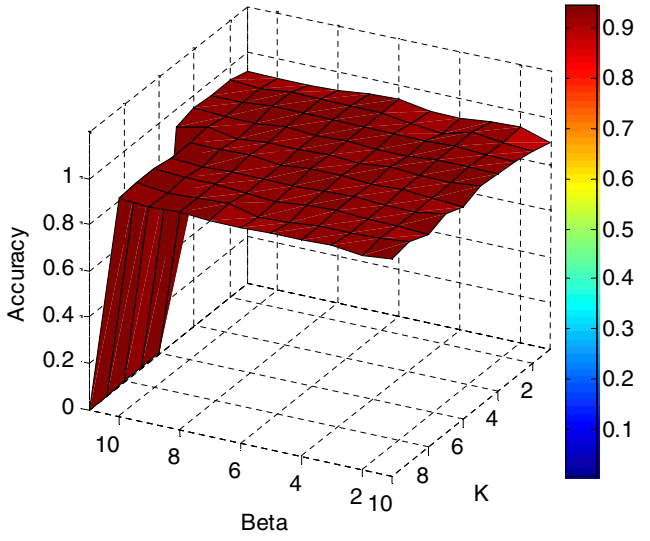


Fig. 7. Accuracy varies with K and β (CART).

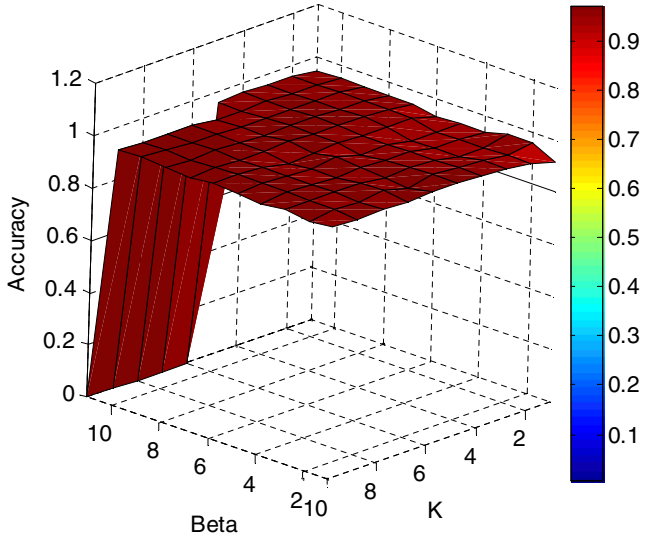


Fig. 8. Accuracy varies with K and β (SVM).

two popular learning algorithms: CART and SVM, to validate the selected features based on 10-fold cross validation. The experimental result shows that the performance of the proposed method outperforms the others with respect to the number of selected features and classification accuracies.

Further work would include both theoretical and experimental comparison of different attribute reduction algorithms. As most of feature selection algorithms are designed either for numerical attributes [10,13] or for categorical features [2,5,15], whereas the proposed algorithm can directly deal with mixed numerical and categorical features, more experimental comparison are desired for other typical algorithms. Moreover, in this paper the proposed model computes the joint relation of two features as the intersection operation of two relations induced by the attributes. In fact, there is more than one method to compute the relation between numerical and discrete samples. The difference in computing relations would lead to different reducts. We can develop new techniques to calculate the relation induced by mixed multiple features and compare the yielded reducts.

Acknowledgement

This work is supported by Natural Science of 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] E. Guyon, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.
[2] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy? Journal of Machine Learning Research 5 (2004) 1205–1224.
[3] T. Dietterich, Machine-learning research: four current directions, AI Magazine 18 (4) (1997) 97–136.
[4] P. Somol, P. Pudil, J. Kittler, Fast branch & bound algorithms for optimal feature selection, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (7) (2004) 900–912.
[5] M. Dash, H. Liu, Consistency-based search in feature selection, Artificial Intelligence 151 (2003) 155–176.
[6] I.S. Oh, J.S. Lee, B.R. Moon, Hybrid genetic algorithms for feature selection, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (11) (2004) 1424–1437.
[7] M.L. Raymer, W.E. Punch, E.D. Goodman, et al., Dimensionality reduction using genetic algorithms, IEEE Transactions on Evolutionary Computation 4 (2) (2000) 164–171.
[8] H. Dash, H. Liu, Feature selection for classification, Intelligent Data Analysis 1 (1997) 131–156.
[9] E. Gasca, J.S. Sanchez, R. Alonso, Eliminating redundancy and irrelevance using a new MLP-based feature selection method, Pattern Recognition 39 (2) (2006) 313–315.
[10] J. Neumann, C. Schnorr, G. Steidl, Combined SVM-based feature selection and classification, Machine Learning 61 (2005) 129–150.
[11] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1–2) (1997) 273–324.
[12] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (2002) 389–422.
[13] Z.X. Xie, Q.H. Hu, D.R. Yu, Improved feature selection algorithm based on SVM and correlation, ISNN 1 (2006) 1373–1380.

[14] K. Kira, L.A. Rendell, The feature selection problem: traditional methods and a new algorithm, in: Proceedings of AAAI-92, San Jose, CA, 1992, pp. 129–134.
[15] F. Fleuret, Fast binary feature selection with conditional mutual information, Journal of Machine Learning Research 5 (2004) 1531–1555.
[16] H.C. Peng, F.H. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1226–1238.
[17] M. Modrzejewski, Feature selection using rough sets theory, in: P.B. Brazdil (Ed.), Proceedings of the European Conference on Machine Learning, Vienna, Austria, 1993, pp. 213–226.
[18] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Letters 24 (2003) 833–849.
[19] H. Liu, F. Hussian, C.L. Tan, M. Dash, Discretization: an enabling technique, Journal of Data Mining and Knowledge Discovery 6 (4) (2002) 393–423.
[20] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann Publishers, Stanford University, CA, 2000.
[21] L. Yu, H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, in: Proceedings of the Twentieth International Conference on Machine Leaning (ICML-03), Washington, DC, August 2003, pp. 856–863.
[22] R. Jenson, Q. Shen, Fuzzy-rough sets for descriptive dimensionality reductions, in: Proceedings of IEEE International Conference on Fuzzy Systems, 2002, pp. 29–34.
[23] W.Y. Tang, K.Z. Mao, Feature selection algorithm for data with both nominal and continuous features, in: T.B. Ho, D. Cheung, H. Liu (Eds.), PAKDD 2005, LNAI 3518, Springer-Verlag, Berlin, Heidelberg, 2005, pp. 683–688.
[24] Q. Shen, R. Jensen, Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring, Pattern Recognition 37 (7) (2004) 1351–1363.
[25] R. Jensen, Q. Shen, Fuzzy-rough attribute reduction with application to web categorization, Fuzzy Sets and Systems 141 (3) (2004) 469–485.
[26] R.B. Bhatt, M. Gopal, On fuzzy-rough sets approach to feature selection, Pattern Recognition Letters 26 (2005) 965–975.
[27] R.B. Bhatt, M. Gopal, On the compact computational domain of fuzzy-rough sets, Pattern Recognition Letters 26 (2005) 1632–1640.
[28] 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 (2) (2006) 191–201.
[29] Q.H. Hu, D.R. Yu, Z.X. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Letters 27 (5) (2006) 414–423.
[30] L.A. 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.
[31] W. Ziarko, Variable precision rough sets model, Journal of Computer and System Sciences 46 (1) (1993) 39–59.
[32] H. Almuallim, T.G. Dietterich, Learning with many irrelevant features, in: Ninth National Conference on Artificial Intelligence, 1991, pp. 547–552.
[33] G.H. John, R. Kohavi, K. Pfleger, Irrelevant features and the subset selection problem, in: Proceeding of the 11th International Conference on Machine Learning, 1994, pp. 121–129.
[34] H. Almuallim, T.G. Dietterich, Learning Boolean concepts in the presence of many irrelevant features, Artificial Intelligence 69 (1–2) (1994) 279–305.
[35] H. Liu, H. Motoda, M. Dash, A monotonic measure for optimal feature selection, in: Proceedings of European Conference on Machine Learning, Chemnitz, Germany, 1998, pp. 101–106.
[36] M. Dash, Feature selection via set cover, in: Proceedings of IEEE Knowledge and Data Engineering Exchange Workshop, Newport, CA, IEEE Computer Society, 1997, pp. 165–171.
[37] G. Brassard, P. Bratley, Fundamentals of Algorithms, Prentice Hall, Englewood Cliffs, NJ, 1996.
[38] M. Dash, H. Liu, Hybrid search of feature subsets, in: Proceedings of Pacific Rim International Conference on Artificial Intelligence (PRICAI-98), Singapore, 1998, pp. 238–249.
[39] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches, IEEE Transactions on Knowledge and Data Engineering 16 (2004) 547–1471.
[40] Q.H. Hu, X.D. Li, D.R. Yu, Analysis on classification performance of rough set based reducts, in: Proceeding of 9th Pacific Rim International Conference on Artificial Intelligence, 2006.
[41] 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.
[42] Q.H. Hu, Z.X. Xie, D.R. Yu, Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognition (2007), doi:10.1016/j.patcog.2007.03.017.
[43] J.Z.C. Lai, Y.C. Liaw, J. Liu, Fast k-nearest-neighbor search based on projection and triangular inequality, Pattern Recognition 40 (2007) 351–359.