Knowledge-Based Systems 220 (2021) 106908

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A novel approach to attribute reduction based on weighted neighborhood rough sets

Meng Hu [a], Eric C.C. Tsang [a,*], Yanting Guo [a], Degang Chen [b], Weihua Xu [c]

[a] *Faculty of Information Technology, Macau University of Science and Technology, Taipa, Macau*
[b] *Department of Mathematics and Physics, North China Electric Power University, Beijing 102206, China*
[c] *College of Artificial Intelligence, Southwest University, Chongqing 400715, China*

ARTICLE INFO

Article history:
Received 14 October 2020
Received in revised form 23 February 2021
Accepted 25 February 2021
Available online 27 February 2021

Keywords:
Weighted neighborhood rough sets
Attribute reduction
Neighborhood rough sets
Information tables

ABSTRACT

Neighborhood rough sets based attribute reduction, as a common dimension reduction method, has been widely used in machine learning and data mining. Each attribute has the same weight (the degree of importance) in the existing neighborhood rough set models. In this work, we introduce different weights into neighborhood relations and propose a novel approach for attribute reduction. The main motivation is to fully mine the correlation between attributes and decisions before calculating neighborhood relations, and the attributes with high correlation are assigned higher weights. We first construct a Weighted Neighborhood Rough Set (WNRS) model based on weighted neighborhood relations and discuss its properties. Then WNRS based dependency is defined to evaluate the significance of attribute subsets. We design a greedy search algorithm based on WNRS to select an attribute subset which has both strong correlation and high dependency. Furthermore, we use isometric search to find the optimal neighborhood threshold. Finally, ten datasets from UCI machine learning repository and ELVIRA Biomedical data set repository are used to compare the performance of WNRS with those of other state-of-the-art reduction algorithms. The experimental results show that WNRS is feasible and effective, which has higher classification accuracy and compression ratio.

1. Introduction

Rough set theory (RST) is viewed as a powerful mathematical analysis tool in machine learning, pattern recognition, knowledge discovery, etc., which was proposed by Pawlak in 1982 [1]. The classical Pawlak rough set theory needs strict equivalence relations, so it can only mine knowledge in information system with categorical attributes. In order to mine knowledge in information system with real-valued attributes, some researchers have introduced neighborhood relations, fuzzy equivalence relations, dominance relations and similarity relations into Pawlak rough sets to form neighborhood rough sets [2,3], fuzzy rough sets [4–7], dominance-based rough sets [8,9] and similarity relation-rough sets [10,11], respectively. These generalized rough set models have been widely used in attribute reduction [12–16], rule extraction [17,18], decision theory [19,20], incremental learning [21,22] and so on.

The similarity between samples can be well described by neighborhood relations, and the neighborhood relations are easy to calculate in information systems with real-valued attributes, so some scholars use neighborhood rough sets (NRS) to perform attribute reduction in the real-valued information systems. Wu and Zhang [23] proposed generalized rough set approximation operators in neighborhood systems and discussed the relationship between neighborhood operator systems and rough set operator systems. Hu et al. [24] used neighborhood rough sets for feature selection in hybrid systems with categorical and real-valued attributes. Furthermore, a feature evaluation function based on neighborhood decision error minimization is defined to select discrete and continuous features [25]. Chen et al. [26] defined the lower and upper boundary regions based on neighborhood rough sets for imbalanced data and established discernibility matrix and discernibility function to find all reducts of hybrid decision systems. Liang et al. [27] considered the stability of the selected attributes to attribute reduction in neighborhood rough sets and selected a stable attribute subset when the data is disturbed. Chen et al. [28] designed a parallel attribute reduction algorithm in dominance-based neighborhood rough sets, which combines the advantages of dominance relations and neighborhood relations. From the above neighborhood-based rough set models, we know that these models do not consider the weight of attributes. However, in practice, the contribution of each attribute to learning tasks may not be equally important. Sometimes, we need to

[*] Corresponding author.
E-mail addresses: humeng24@sina.com (M. Hu), cctsang@must.edu.mo (E.C.C. Tsang), ytguosx@sina.com (Y. Guo), chengdegang@263.net (D. Chen), chxuwh@gmail.com (W. Xu).

https://doi.org/10.1016/j.knosys.2021.106908
0950-7051/© 2021 Published by Elsevier B.V.

treat attributes differently, that is, assigning different weights to different attributes. If we consider internal relevance between conditional attributes and decisions in advance, then we can highlight the attributes that are highly related to the decisions in neighborhood relations. In this way, the attributes with high correlation and dependency are more easily selected.

The weight of an attribute is an embodiment of its importance. Some researchers have studied how to assign weights of attributes in fuzzy rough sets, decision-theoretic rough sets, k-nearest neighbor rules, fuzzy c-means and so on. Guo et al. [29] used decision tree learning to propose a kind of granulation weighted model for multi-granulation interval-valued decision. Tsang et al. [30] defined a weighted Parzen window function by using fuzzy rough sets based on kernel function and proposed a novel k-nearest-neighbor classification algorithm based on the weighted Parzen window function to improve the classification accuracy. To reduce the loss of hesitant fuzzy multi-attribute decision making, Xu and Zhang [31] calculated the optimal attribute weights of incomplete systems by the maximizing deviation method to reduce the loss of incomplete information. Vluymansa et al. [32] proposed a novel weight selection strategy based on ordered weighted average to increase the antinoise ability of fuzzy rough sets. Other methods to determine attribute weights can be found in [33–36].

Attribute reduction, which is one of the most effective attribute subset selection techniques, is to select informative and compact attributes and to eliminate redundant and inconsistent attributes for learning tasks. There are many attribute reduction algorithms, such as attribute reduction based on classical rough sets, fuzzy rough sets, neighborhood rough sets, entropy and mutual information. Attribute reduction based on classical rough sets cannot deal with continuous attributes. Attribute reduction based on fuzzy rough sets can handle continuous attributes, but cannot handle categories attributes. Neighborhood rough sets can be used to evaluate significance of continuous and categories attributes. The computation time complexity of attribute reduction based on entropy and mutual information is relatively high, it is difficult for big data processing. There are many attribute reduction approaches based on neighborhood knowledge [6,12,13]. Hu et al. [37] used neighborhood mutual information to select features in information systems with discrete and continuous features. Wang et al. [38] constructed a new rough set model, fuzzy neighborhood rough sets, to select feature subsets, and it can reduce the possibility that objects are misclassified. A measure, called the neighborhood discrimination index [39], is defined to evaluate the significance of features, and it has been used to select features with good performance. Patil and Atique [40] proposed a neighborhood positive region (NPR) model based on rough set theory for attribute reduction to decrease running time. In order to solve the problem of low efficiency and over-fitting of limited label data in attribute reduction, Wang et al. [41] combined neighborhood rough sets and local rough sets to define local neighborhood rough set (LNRS) for attribute reduction. Mariello and Battiti [42] combined the advantages of locality sensitive hashing and approximated nearest-neighbors techniques for feature selection. Wang et al. [43] defined four measures to evaluate the significance of features and use neighborhood self-information to remove redundant features. In order to determine the optimal (suboptimal) neighborhood radius of neighborhood rough sets, Yang et al. [44] proposed a pseudo-label neighborhood relation by the distance and pseudo-label of samples, and the neighborhood rough sets and corresponding measures are re-defined by the pseudo-label neighborhood relation to measure the significance of attribute subsets. Considering label information and intra-class and inter-class radii in neighborhood relations simultaneously, Jiang et al. [45] proposed supervised neighborhood relations and studied supervised neighborhood based attribute reduction in depth. Sang et al. [46,47] studied incremental attribute reduction algorithms based on dominance conditional entropy and neighborhood dominance conditional entropy in ordered data and heterogeneous ordered data with the variation of objects, respectively. The methods of attribute reduction based on other types of rough set models can be found in [27,28].

Attributes in the existing neighborhood rough set models have the same weights. If the weights of attributes are the same in neighborhood relations, the attributes that are highly related to the decisions will not be represented to have more degree of importance, and these attributes may not be preferred in attribute reduction. To put more weights on the attributes that are highly relevant to decisions in attribute reduction, we first use the correlation coefficients of attributes with respect to decisions to re-assign weights of attributes, then define weighted neighborhood rough sets (WNRS) and some measures. As with neighborhood rough sets (NRS), the dependency of WNRS can be used to characterize the ability of attribute subsets to distinguish samples with different decisions. We use a greedy search strategy to select an optimal attribute subset which has both strong correlation and high dependency. Furthermore, isometric search strategy is used to find the optimal neighborhood threshold of different datasets. Finally, we use ten datasets from UCI Machine Learning Repository and ELVIRA Biomedical data set repository to verify the validity and stability of attribute reduction algorithm based on weighted neighborhood rough sets (WNRS) and compare it with other typical reduction algorithms. The experimental results show that WNRS is feasible and effective for attribute reduction.

This paper is organized as follows. In Section 2, we briefly review the basic concept of neighborhood rough sets and point out the shortcomings of neighborhood rough sets. In Section 3, we present the definition of weighted neighborhood rough sets and some measures of evaluation attributes and discuss its properties. In Section 4, we design a heuristic algorithm to find a reduct of a decision information table. In Section 5, we use ten datasets to compare the proposed method with 3 classical attribute reduction algorithms from three aspects. In Section 6, we summarize the paper and propose the future work.

2. Related work

In this section, we review the related knowledge of classical neighborhood rough sets and point out the shortcomings of neighborhood rough sets. Detailed information can be found in [23,24].

2.1. Neighborhood rough sets

Given a decision information table $IS = (U, C, D)$, where $U = \{x_1, x_2, \dots, x_n\}$ is a sample set, $C = \{a_1, a_2, \dots, a_m\}$ is a conditional attribute set to characterize the samples and $D = \{d_1, d_2, \dots, d_r\}$ is a decision attribute set to mark the category of samples. $U/D = \{D_1, D_2, \dots, D_k\}$ is a decision partition on U to D.

In a given decision information table $IS = (U, C, D)$, $\forall x \in U$ and $B \subseteq C$, the neighborhood similarity class of sample x under attribute subset B is defined as

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

where d_B is a distance function with attribute subset B and $\delta(\delta > 0)$ is a neighborhood threshold. The neighborhood similarity class is also called the neighborhood information granule. In this paper, the Euclidean distance is used to measure the distance between two samples. The formula of Euclidean distance is given as follows

$$d_B(x, y) = \sqrt{\sum_{a_i \in B} (f(x, a_i) - f(y, a_i))^2}, \tag{2}$$

where $f(x, a)$ is the value of sample x under attribute a.

In $IS = (U, C, D)$, $\forall X \subseteq U$, $B \subseteq C$ and a given neighborhood threshold δ, two subsets of U, the upper approximation and lower approximation of X with respect to δ in $IS = (U, C, D)$, are defined as

$$\begin{aligned} &\overline{N}_B^\delta(X) = \{x | N_B^\delta(x) \cap X \neq \emptyset\}; \\ &\underline{N}_B^\delta(X) = \{x | N_B^\delta(x) \subseteq X\}, \end{aligned} \tag{3}$$

where $\overline{N}_B^\delta$ and $\underline{N}_B^\delta$ are a pair of approximation operators. X with respect to $\overline{N}_B^\delta$ and $\underline{N}_B^\delta$ is accurate, if $\overline{N}_B^\delta(X) = \underline{N}_B^\delta(X)$; otherwise X is rough. From the definitions of the above two approximation operators, we can get $\underline{N}_B^\delta(X) \subseteq X \subseteq \overline{N}_B^\delta(X)$.

Let $IS = (U, C, D)$ be a decision information table, for a given attribute subset $B \subseteq C$ and a threshold δ, the upper and lower approximations of D with respect to B are defined as

$$\begin{aligned} &\overline{N}_B^\delta(D) = \bigcup_{i=1}^{k} \overline{N}_B^\delta(D_i); \\ &\underline{N}_B^\delta(D) = \bigcup_{i=1}^{k} \underline{N}_B^\delta(D_i), \end{aligned} \tag{4}$$

where $U/D = \{D_1, D_2, \ldots, D_k\}$. The boundary and positive regions of D with respect to B are defined as

$$\begin{aligned} &BN_B^\delta(D) = \overline{N}_B^\delta(D) - \underline{N}_B^\delta(D); \\ &POS_B^\delta(D) = \bigcup_{D_i \in U/D} \underline{N}_B^\delta(D_i). \end{aligned} \tag{5}$$

The size of $BN_B^\delta(D)$ reflects the ability of attribute subset B to approximate D. The smaller the size of $BN_B^\delta(D)$ is, the stronger the ability of B to approximate D is. The size of $POS_B^\delta(D)$ reflects the number of samples which can be classified correctly under B.

In $IS = (U, C, D)$, for a given attribute subset $B \subseteq C$ and a threshold δ, the dependency degree of D with respect to B in $IS = (U, C, D)$ is defined as

$$\gamma_B^\delta(D) = \frac{|POS_B^\delta(D)|}{|U|}, \tag{6}$$

where $|\cdot|$ represents the cardinality of a set. $\gamma_B^\delta(D)$ is used to measure the ability of attribute subset B to approximate D. The larger the $\gamma_B^\delta(D)$ is, the stronger the approximation ability of the attribute subset B is. According to the definition of the dependency degree, $0 \leq \gamma_B^\delta(D) \leq 1$. There are two factors that affect the value of the dependency degree. One is the neighborhood threshold δ to control the size of neighborhood classes. The larger the δ is, the smaller the dependency degree is. Other is the attribute subset B to characterize the samples. As the attributes increase gradually, the value of dependency degree increases

2.2. Shortcomings of neighborhood rough sets

When we calculate the neighborhood classes of neighborhood rough sets, we use the same weight for each attribute. That is to say, the degree of importance of each attribute is the same in attribute reduction. The internal relationship between attributes and decisions is not fully explored, and the calculation of neighborhood classes using same weights may lead to the attributes with large values more easily selected. In order to illustrate the necessity of weighting each attribute before attribute reduction, we use the following example to illustrate the shortcomings of classical neighborhood rough sets.

Table 1
A decision information table.

U	a_1	a_2	a_3	a_4	d
x_1	0.28	0.89	0.21	0.29	1
x_2	0.44	0.87	0.19	0.26	1
x_3	0.48	0.51	0.20	0.39	1
x_4	0.50	0.50	0.26	0.38	1
x_5	0.61	0.69	0.29	0.35	1
x_6	0.39	0.71	0.27	0.18	2
x_7	0.37	0.35	0.33	0.24	2
x_8	0.62	0.90	0.38	0.20	2
x_9	0.76	0.86	0.35	0.27	2
x_{10}	0.89	0.50	0.39	0.28	2

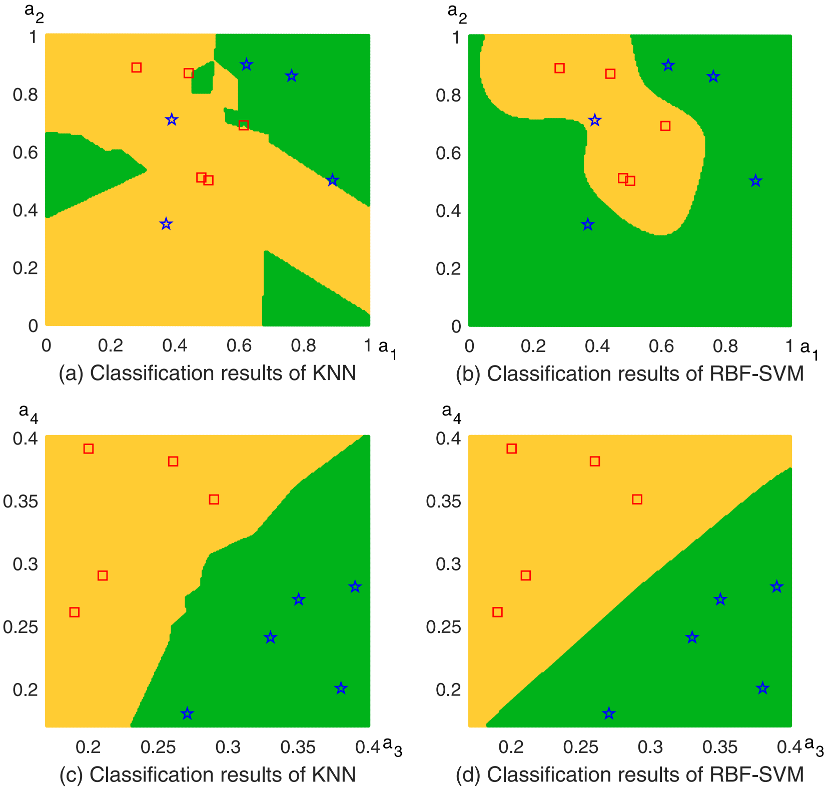
Table 2
Neighborhood information granules are formed by B_1 and B_2.

U	$N(x)^{0.1}_{B_1}(x)$	$N(x)^{0.1}_{B_2}(x)$
x_1	$\{x_1\}$	$\{x_1, x_2, x_5\}$
x_2	$\{x_2\}$	$\{x_1, x_2\}$
x_3	$\{x_3, x_4\}$	$\{x_3, x_4, x_5\}$
x_4	$\{x_3, x_4\}$	$\{x_3, x_4, x_5\}$
x_5	$\{x_5\}$	$\{x_1, x_3, x_4, x_5, x_9\}$
x_6	$\{x_6\}$	$\{x_6, x_7\}$
x_7	$\{x_7\}$	$\{x_6, x_7, x_8, x_9, x_{10}\}$
x_8	$\{x_8\}$	$\{x_7, x_8, x_9, x_{10}\}$
x_9	$\{x_8\}$	$\{x_5, x_7, x_8, x_9, x_{10}\}$
x_{10}	$\{x_{10}\}$	$\{x_7, x_8, x_9, x_{10}\}$

Example 2.1. A given decision information table $IS = (U, C, D)$ is shown in Table 1, where $U = \{x_1, x_2, \ldots, x_{10}\}$ is a sample space, the conditional and decision attribute sets are $C = \{a_1, a_2, a_3, a_4\}$ and $D = \{d\}$, respectively. These samples are divided into two decision classes $D_1 = \{x_1, x_2, \ldots, x_5\}$ and $D_2 = \{x_6, x_7, \ldots, x_{10}\}$ by d. Given two attribute subsets $B_1 = \{a_1, a_2\}$, $B_2 = \{a_3, a_4\}$ and a neighborhood threshold $\delta = 0.1$, the generated neighborhood information granules by B_1 and B_2 under $\delta = 0.1$ are shown in Table 2. From Table 2, we know that the granularity of neighborhood information granules induced by B_1 is finer than that induced by B_2. According to the definition of $POS_B^\delta(D)$ and Table 2, we know that $POS_{B_1}^{0.1}(D) = U$ and $POS_{B_2}^{0.1}(D) = \{x_1, x_2, x_3, x_4, x_6, x_7, x_8, x_{10}\}$. Therefore, we can get $\gamma_{B_1}^{0.1}(D) = 1$ and $\gamma_{B_2}^{0.1}(D) = 0.8$. So from the above point of view, it is obvious that the ability of B_1 to approximate D is better than that of B_2. We use k-nearest neighbor (KNN) and radial basis function support vector machine (RBF-SVM) classifiers to classify the unknown area, and the classification results are shown in Fig. 1, where k of KNN rules is 3 and the control term C of RBF-SVM is set to 1000, and the Gaussian kernel parameter σ of RBF-SVM is automatically optimized by MATLAB. As can be seen from Fig. 1, the separability of attribute subset B_2 is significantly better than that of attribute subset B_1. Under attribute subset B_1, some samples are misclassified by KNN and RBF-SVM and over-fitting has occurred (see Fig. 1(a) and (b)); under attribute subset B_2, all samples are classified correctly by KNN and RBF-SVM and over-fitting has not occurred. The ability of B_1 to approximate D is better than that of B_2, but the separability of B_1 with respect to D is worse than that of B_2. Therefore, it is limited to use the dependency degree of neighborhood rough sets to measure the significance of attribute subsets. Next, we will introduce an effective model to measure the significance of attribute subsets.

3. Weighted neighborhood rough sets

One of the shortcomings of neighborhood rough sets is that it uses the same weight for each attribute to perform attribute reduction. It is important to know that different attributes have



(a) Classification results of KNN
(b) Classification results of RBF-SVM
(c) Classification results of KNN
(d) Classification results of RBF-SVM

Fig. 1. Classification results of two classifiers under $\{a_1, a_2\}$ and $\{a_3, a_4\}$.

different degree of importance for decision-making. So we should use different weights to calculate neighborhood similarity classes for different attributes. A novel rough set model is proposed here, named weighted neighborhood rough sets, to evaluate the significance of attribute subsets.

Given a decision information table $IS = (U, C, D)$, $\forall x \in U$, $\forall a \in C$, $f(x, a)$ is the value of sample x with respect to attribute a. Let the coefficient matrix be

$$A = \begin{pmatrix} f(x_1,a_1) & f(x_1,a_2) & \cdots & f(x_1,a_m) \\ f(x_2,a_1) & f(x_2,a_2) & \cdots & f(x_2,a_m) \\ \vdots & \vdots & \vdots & \vdots \\ f(x_n,a_1) & f(x_n,a_2) & \cdots & f(x_n,a_m) \end{pmatrix},$$

the decision vector be $Y = (f(x_1, d), f(x_2, d), \dots, f(x_m, d))^T$ and the partition coefficients of attributes be $\upsilon = (\upsilon(a_1), \upsilon(a_2), \dots, \upsilon(a_m))^T$. In order to determine the optimal partition coefficients of attributes, we transform the problem of seeking the optimal coefficients into an optimization problem as follows

$$\upsilon^* = \arg\min \|A\upsilon - Y\|^2, \tag{7}$$

where $\|\cdot\|^2$ represents 2-norm of a vector. In order to solve (7), first assuming $A\upsilon = Y$, then both sides of $A\upsilon = Y$ are multiplied by A^T to get $A^TA\upsilon = A^TY$, finally, by solving $A^TA\upsilon = A^TY$, we get

$$\upsilon = (A^TA)^{-1}A^TY. \tag{8}$$

Especially, when matrix (A^TA) is not invertible or the objective function (7) needs a penalty term, we convert (7) to $J(\upsilon) = \|A\upsilon - Y\|^2 + \|\upsilon\|^2$. Since $J(\upsilon)$ is a convex function, the minimum of $J(\upsilon)$ is obtained when $J'(\upsilon) = 0$. $J'(\upsilon) = 2A^T(A\upsilon - Y) + 2\upsilon = 0$. Therefore, $(A^TA + E)\upsilon = A^TY$, where E is an identity matrix. So $\upsilon = (A^TA + E)^{-1}A^TY$. When matrix A^TA or $A^TA + E$ is high dimensional or close to ill-conditioned, we use "\" in MATLAB or subfunction "np.linalg.solve" in Numpy to solve $A^TA\upsilon = A^TY$ or $(A^TA + E)\upsilon = A^TY$, instead of solving the inverse matrix.

$|\upsilon(a)|$ is the absolute value of $\upsilon(a)$, which reflects the relation between attribute a and decision D. The larger the $|\upsilon(a)|$ is, the stronger the internal relevance of attributes and decisions is.

Definition 1. Given a decision information table $IS = (U, C, D)$, $\forall a \in C$, the weight of a is defined as

$$\omega(a) = \frac{|C|\,|\upsilon(a)|}{\sum_{a_i \in C} |\upsilon(a_i)|} \tag{9}$$

Property 1. *Given a decision information table $IS = (U, C, D)$, $\forall a \in C$, the weight vector with attributes $\omega = (\omega(a_1), \omega(a_2), \dots, \omega(a_m))^T$, we have*

$$(1)\ \omega(a) \geq 0;$$

$$(2) \sum_{a_i \in C} \omega(a_i) = |C|. \tag{10}$$

Proof. (1)–(2) can be proved by Definition 1 directly.

From Property 1, we can see that the weights of attributes are assigned by using the partition coefficients between the conditions and decisions. The higher the correlation between the conditional attribute and the decision is, the higher the assigned weight of the conditional attribute is.

Definition 2. Given a decision information table $IS = (U, C, D)$, $\omega = (\omega(a_1), \omega(a_2), \dots, \omega(a_m))^T$ is a weight vector with attributes, for attribute subset $B(B \subseteq C)$ and neighborhood threshold δ, the weighted neighborhood similarity relation is defined as

$$\begin{aligned} W_B^\delta &= \{(x,y)| \sqrt{\sum_{a\in B}(\omega(a)(f(x,a)-f(y,a)))^2} \le \delta\} \\ &= \{(x,y)| \sqrt{\sum_{a\in B}\omega^2(a)(f(x,a)-f(y,a))^2} \le \delta\} \end{aligned} \tag{11}$$

where $\omega(a) \ge 0$, $\sum_{a\in C}|\omega(a)| = |C|$ and $\omega(a)$ is the weight of attribute a. When $\omega(a) > 1$, the degree of importance of attribute a will be increased in the calculation of relations; when $0 < \omega(a) < 1$, the degree of importance of attribute a will be decreased; when $\omega(a) = 1$, the degree of importance of attribute a will remain unchanged; when $\omega(a) = 0$, attribute a has been removed before attribute reduction. $\forall a \in C$, W_B^δ is a classical neighborhood similarity relation, if $\omega(a) = 1$. Therefore, weighted neighborhood relation is a generalization of neighborhood relation, and the neighborhood relation is a special case of weighted neighborhood relation. Obviously, relation matrix W_B^δ is a symmetric matrix.

Property 2. *Given a decision information table $IS = (U, C, D)$, W_B^δ is a weighted neighborhood similarity relation, $\forall x, y \in U$, we have*

$$\begin{aligned} &(1)\ \textit{Reflexivity}: (x,x) \in W_B^\delta; \\ &(2)\ \textit{Symmetry}: (x,y) \in W_B^\delta \Longleftrightarrow (y,x) \in W_B^\delta. \end{aligned} \tag{12}$$

Proof. It is immediate from formula (11).

Definition 3. Given a decision information table $IS = (U, C, D)$ and a weighted neighborhood similarity relation W_B^δ, the weighted neighborhood similarity class is defined as

$$WN_B^\delta(x) = \{y|(y,x) \in W_B^\delta, y \in U\}. \tag{13}$$

The weighted neighborhood similarity class is also called the weighted neighborhood information granule. Threshold δ controls the size of the information granule, the larger the δ is, the larger the size of the information granule is. All weighted neighborhood information granules form a cover on U.

Definition 4. Given a decision information table $IS = (U, C, D)$ and a weighted neighborhood similarity relation W_B^δ, $\forall X \subseteq U$, the upper and lower approximations of X with respect to W_B^δ are defined as

$$\begin{aligned} &\overline{W}_B^\delta(X) = \{x|WN_B^\delta(x) \cap X \ne \emptyset\}; \\ &\underline{W}_B^\delta(X) = \{x|WN_B^\delta(x) \subseteq X\}. \end{aligned} \tag{14}$$

X with respect to the relation W_B^δ is accurate, if $\overline{W}_B^\delta(X) = \underline{W}_B^\delta(X)$; otherwise, X with W_B^δ is rough. Obviously, $\overline{W}_B^\delta(X) \supseteq X \supseteq \underline{W}_B^\delta(X)$. The boundary of X with respect to relation W_B^δ is defined as

$$WBN_B^\delta(X) = \overline{W}_B^\delta(X) - \underline{W}_B^\delta(X). \tag{15}$$

The size of $WBN_B^\delta(X)$ reflects the roughness of X with respect to the relation W_B^δ. The smaller the size of $WBN_B^\delta(X)$ is, the finer the relation W_B^δ is; otherwise, the rougher the relation W_B^δ is.

Definition 5. Given a decision information table $IS = (U, C, D)$ and a weighted neighborhood similarity relation W_B^δ, for $U/D = \{D_1, D_2, \ldots, D_k\}$, the upper and lower approximations of D with respect to the relation W_B^δ are defined as

$$\begin{aligned} &\overline{W}_B^\delta(D) = \bigcup_{i=1}^{k} \overline{W}_B^\delta(D_i); \\ &\underline{W}_B^\delta(D) = \bigcup_{i=1}^{k} \underline{W}_B^\delta(D_i). \end{aligned} \tag{16}$$

The decision boundary and decision positive regions of D with respect to the relation W_B^δ are defined as

$$\begin{aligned} &WBN_B^\delta(D) = \overline{W}_B^\delta(D) - \underline{W}_B^\delta(D); \\ &WPOS_B^\delta(D) = \bigcup_{D_i \in U/D} \underline{W}_B^\delta(D_i). \end{aligned} \tag{17}$$

The size of decision boundary region and the size of decision positive region reflect the roughness of decision D with respect to the relation W_B^δ. $WBN_B^\delta(D)$ is to measure the roughness of W_B^δ from the two aspects of upper and lower approximations. $WPOS_B^\delta(D)$ is to measure the roughness of W_B^δ from the aspects of lower approximation. Generally speaking, the samples of the lower approximation can be classified correctly by W_B^δ, some samples of the upper approximation may be classified correctly and some samples may be misclassified. So the measure ability of the lower approximation is better than that of the upper approximation.

Property 3. *Given a decision information table $IS = (U, C, D)$ and a weighted neighborhood similarity relation W_B^δ, where $U/D = \{D_1, D_2, \ldots, D_r\}$, we have*

$$\begin{aligned} &(1)\ \overline{W}_B^\delta(D) = U; \\ &(2)\ WPOS_B^\delta(D) \cap WBN_B^\delta(D) = \emptyset; \\ &(3)\ WPOS_B^\delta(D) \cup WBN_B^\delta(D) = \overline{W}_B^\delta(D). \end{aligned} \tag{18}$$

Proof. (1) There are $D_i \subseteq \overline{W}_B^\delta(D_i)$ and $\cup_i^r D_i = U$, according to Definition 4, so we have $U \subseteq \overline{W}_B^\delta(D_i)$, obviously $\overline{W}_B^\delta(D) \subseteq U$, so $\overline{W}_B^\delta(D) = U$.

(2) There are $WBN_B^\delta(D) = \overline{W}_B^\delta(D) - \underline{W}_B^\delta(D)$ and $WPOS_B^\delta(D) = \cup_{D_i \in U/D}\underline{W}_B^\delta(D_i)$, so $WPOS_B^\delta(D) \cap WBN_B^\delta(D) = \emptyset$.

(3) From $WBN_B^\delta(D) = \overline{W}_B^\delta(D) - \underline{W}_B^\delta(D)$ and $WPOS_B^\delta(D) = \cup_{D_i \in U/D}\underline{W}_B^\delta(D_i) = \underline{W}_B^\delta(D)$, so $WPOS_B^\delta(D) \cup WBN_B^\delta(D) = \overline{W}_B^\delta(D)$.

Definition 6. Given a decision information table $IS = (U, C, D)$ and W_B^δ is a weighted neighborhood similarity relation, the dependency degree of D with respect to W_B^δ is defined as

$$\gamma_B^\delta(D) = \frac{|WPOS_B^\delta(D)|}{|U|}. \tag{19}$$

$\gamma_B^\delta(D)$ is used to measure the ability of attribute subset B to approximate D, where the attributes of C have different weights. The larger the $\gamma_B^\delta(D)$ is, the stronger the approximation ability of attribute subset B is. From the above definition, we know that $0 \le \gamma_B^\delta(D) \le 1$. There are three factors that affect the value of γ, the first is the neighborhood threshold δ to control the size of neighborhood granules, the second is the attribute subset B to characterize the samples, and the third is the weights of attributes. When the weights of the attributes are given, γ increases with the decrease of δ or the increase of attributes. Next, we will discuss how to determine the weights of attributes.

It can be seen from Definition 2 that when $\omega(a_1) = \omega(a_2) = \cdots = \omega(a_m) = 1$, weighted neighborhood rough sets degenerate into classical neighborhood rough sets. That is to say, weighted neighborhood rough set is a generalized model of neighborhood rough sets. The weights of attributes in weighted neighborhood rough sets are very important for attribute reduction. It is necessary to mine the internal relevance between attributes and decisions before performing attribute reduction. If the internal relevance between attributes and decisions are high, the weights of attributes should be higher; otherwise the weights should be lower.

In order to understand the calculation process of the weighted neighborhood rough set and the difference between it and the classical neighborhood rough set, we continue to calculate the dependency degree of weighted neighborhood rough sets under different attribute subsets in Example 2.1. Firstly, from formulas (8) and (9), we can get $\upsilon = (-0.6601, 0.1850, 7.3070, -1.3485)$ and $\omega = (0.2779, 0.0779, 3.0764, 0.5677)$. Since $\omega(a_1)$, $\omega(a_2)$ and $\omega(a_4)$ are less than 1, the effects of a_1, a_2 and a_4 are reduced in weighted neighborhood rough sets based attribute reduction; $\omega(a_3)$ is more than 1, the effect of a_3 is raised in weighted neighborhood rough sets based attribute reduction. Under attribute subsets $B_1 = \{a_1, a_2\}$, $B_2 = \{a_3, a_4\}$ and neighborhood threshold $\delta = 0.1$, weighted neighborhood similarity relations generated by B_1 and B_2 are represented as follows

$$W_{B_1}^{0.1} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

and

$$W_{B_2}^{0.1} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Weighted neighborhood information granules are induced by $W_{B_1}^{0.1}$ and $W_{B_2}^{0.1}$ in Table 3. From Table 3, we can see that the granularity of the weighted neighborhood information granules induced by B_2 is finer than that induced by B_1. From Definition 5 and weighted neighborhood information granules induced by B_1 and B_2, we can get $WPOS_{B_1}^{0.1}(D) = \emptyset$ and $WPOS_{B_2}^{0.1}(D) = U$. Therefore, we can get $\gamma_{B_1}^{0.1}(D) = 0$ and $\gamma_{B_2}^{0.1}(D) = 1$. That is to say, the ability of attribute subset B_2 to approximate D is better than that of attribute subset B_1 in weighted neighborhood rough sets. However, the ability of attribute subset B_2 to approximate D is worse than that of attribute subset B_1 in classical neighborhood rough sets. From Fig. 1, we can see that the separability of samples under B_2 is better than that of B_1. Therefore, the weighted neighborhood rough set model can repair the shortcomings of neighborhood rough set model. It is more reasonable to measure the degree of importance of attribute subset by the dependency degree of weighted neighborhood rough sets than by the dependency degree of neighborhood rough sets.

Property 4 (*Tpye-I Monotonicity*)**.** *Given a decision information table $IS = (U, C, D)$, for $B_1 \subseteq B_2 \subseteq C$ and a neighborhood threshold δ, we have*

$$\begin{aligned} &(1)\ W_{B_1}^{\delta} \supseteq W_{B_2}^{\delta}; \\ &(2)\ \forall X \subseteq U, \overline{W}_{B_1}^{\delta}(X) \supseteq \overline{W}_{B_2}^{\delta}(X), \underline{W}_{B_1}^{\delta}(X) \subseteq \underline{W}_{B_2}^{\delta}(X); \\ &(3)\ WPOS_{B_1}^{\delta}(D) \subseteq WPOS_{B_2}^{\delta}(D), \gamma_{B_1}^{\delta}(D) \leq \gamma_{B_2}^{\delta}(D). \end{aligned} \tag{20}$$

Proof. (1) $\forall x, y \in U$, from $B_1 \subseteq B_2$ and Definition 2, we have $\sum_{a \in B_1} |\omega^2(a)|(f(x,a)-f(y,a))^2 \leq \sum_{a \in B_2} |\omega^2(a)|(f(x,a)-f(y,a))^2$, so $W_{B_1}^{\delta} \supseteq W_{B_2}^{\delta}$.

Table 3
Weighted neighborhood information granules under B_1 and B_2.

U	$WN_{B_1}^{0.1}(x)$	$WN_{B_2}^{0.1}(x)$
x_1	$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$	$\{x_1, x_2, x_3\}$
x_2	$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\}$	$\{x_1, x_2, x_3\}$
x_3	$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\}$	$\{x_1, x_2, x_3\}$
x_4	$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\}$	$\{x_4, x_5\}$
x_5	$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$	$\{x_4, x_5\}$
x_6	$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$	$\{x_6\}$
x_7	$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$	$\{x_7, x_9\}$
x_8	$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$	$\{x_8, x_{10}\}$
x_9	$\{x_2, x_3, x_4, x_5, x_8, x_9, x_{10}\}$	$\{x_7, x_9\}$
x_{10}	$\{x_5, x_8, x_9, x_{10}\}$	$\{x_8, x_{10}\}$

(2) According to $B_1 \subseteq B_2$ and (1) , there is $WN_{B_2}^{\delta}(x) \subseteq WN_{B_1}^{\delta}(x)$, if $x \in \overline{W}_{B_2}^{\delta}(X)$, we have $WN_{B_2}^{\delta}(x) \cap X \neq \emptyset$, and then $WN_{B_1}^{\delta}(x) \cap X \neq \emptyset$, from Definition 4 we can get $x \in \overline{W}_{B_1}^{\delta}(X)$, so $\overline{W}_{B_1}^{\delta}(X) \supseteq \overline{W}_{B_2}^{\delta}(X)$; in the same way, we can get $\underline{W}_{B_1}^{\delta}(X) \subseteq \underline{W}_{B_2}^{\delta}(X)$.

(3) According to (2), $\forall D_i \in U/D$, we have $\underline{W}_{B_1}^{\delta}(D_i) \subseteq \underline{W}_{B_2}^{\delta}(D_i)$, $WPOS_{B_1}^{\delta}(D) = \cup_{D_i \in U/D} \underline{W}_{B_1}^{\delta}(D_i)$ and $WPOS_{B_2}^{k}(D) = \cup_{D_i \in U/D} \underline{W}_{B_2}^{k}(D_i)$, so $WPOS_{B_1}^{\delta}(D) \subseteq WPOS_{B_2}^{\delta}(D)$; then we have $\gamma_{B_1}^{\delta}(D) \leq \gamma_{B_2}^{\delta}(D)$.

Property 4 shows that the dependency degree of weighted neighborhood rough sets increases monotonically with the increase of attributes. We often use greedy search strategy to find a minimum attribute subset which has the same ability to characterize samples as the original attribute set. The monotonicity of the dependency function about attributes just satisfies the demand of designing greedy search algorithm.

Property 5 (*Tpye-II Monotonicity*)**.** *Given a decision information table $IS = (U, C, D)$, for $B \subseteq C$ and two given neighborhood thresholds δ_1 and δ_2, $\delta_1 \leq \delta_2$, we have*

$$\begin{aligned} &(1)\ W_B^{\delta_1} \subseteq W_B^{\delta_2}; \\ &(2)\ \forall X \subseteq U, \overline{W}_B^{\delta_1}(X) \subseteq \overline{W}_B^{\delta_2}(X) \textit{ and } \underline{W}_B^{\delta_1}(X) \supseteq \underline{W}_B^{\delta_2}(X); \\ &(3)\ WPOS_B^{\delta_1}(D) \supseteq WPOS_B^{\delta_2}(D), \gamma_B^{\delta_1}(D) \geq \gamma_B^{\delta_2}(D). \end{aligned} \tag{21}$$

Proof. (1) As $\delta_1 \leq \delta_2$, $\forall x, y \in U$, if $\sqrt{\sum_{a \in B} \omega^2(a)(f(x,a) - f(y,a))^2} \leq \delta_1$, then $\sqrt{\sum_{a \in B} \omega^2(a)(f(x,a) - f(y,a))^2} \leq \delta_2$, so $W_B^{\delta_1} \subseteq W_B^{\delta_2}$.

(2) According to $\delta_1 \leq \delta_2$ and (1), there is $WN_B^{\delta_1}(x) \subseteq WN_B^{\delta_2}(x)$, if $x \in \overline{W}_B^{\delta_1}(X)$, we have $WN_B^{\delta_1}(x) \cap X \neq \emptyset$, and then $WN_B^{\delta_2}(x) \cap X \neq \emptyset$, therefore $x \in \overline{W}_B^{\delta_2}(X)$, so $\overline{W}_B^{\delta_1}(X) \subseteq \overline{W}_B^{\delta_2}(X)$; in the same way, we can get $\underline{W}_B^{\delta_1}(X) \supseteq \underline{W}_B^{\delta_2}(X)$.

(3) According to (2), $\forall D_i \in U/D$, we have $\underline{W}_B^{\delta_1}(D_i) \supseteq \underline{W}_B^{\delta_2}(D_i)$, so $WPOS_B^{\delta_1}(D) = \cup_{D_i \in U/D} \underline{W}_B^{\delta_1}(D_i)$, $WPOS_B^{\delta_2}(D) = \cup_{D_i \in U/D} \underline{W}_B^{\delta_2}(D_i)$, so $WPOS_B^{\delta_1}(D) \supseteq WPOS_B^{\delta_2}(D)$; then we have $\gamma_B^{\delta_1}(D) \geq \gamma_B^{\delta_2}(D)$.

Property 5 shows that the value of dependency degree is related to the weighted neighborhood information granules. However, for a given attribute subset, the size of information granules is controlled by threshold δ of weighted neighborhood rough sets. Therefore, the dependency degree increases monotonically with the decrease of threshold δ. A reasonable threshold δ is very important for attribute reduction in weighted neighborhood rough sets. Later, we will show how to set δ.

From Properties 4 and 5, we know that the ability of attribute subsets to approximate decisions depends not only on the attribute subset to characterize samples of universe, but also on the threshold δ of weighted neighborhood rough sets.

Definition 7. Given a decision information table $IS = (U, C, D)$, an attribute subset $B \subseteq C$, a neighborhood threshold δ and an

Table 4
A new decision information table.

U	a_1	a_2	a_3	a_4	d
x_1	0.28	0.89	0.21	0.29	1
x_2	0.44	0.87	0.19	0.26	1
x_3	0.48	0.51	0.20	0.39	1
x_4	0.50	0.50	0.26	0.38	1
x_5	0.61	0.69	0.29	0.35	1
x_6	0.39	0.71	0.27	0.18	2
x_7	0.37	0.35	0.33	0.24	2
x_8	0.62	0.90	0.38	0.20	2
x_9	0.76	0.86	0.35	0.27	2
x_{10}	0.89	0.50	0.39	0.28	2
x_{11}	1.00	1.00	0	0	1
x_{12}	0	0	1.00	1.00	2

attribute $a \in B$. a is called the redundant attribute of B with respect to D, if $\gamma^{\delta}_{B-\{a\}}(D) = \gamma^{\delta}_{B}(D)$. a is called the necessary attribute of B with respect to D, if $\gamma^{\delta}_{B-\{a\}}(D) < \gamma^{\delta}_{B}(D)$. Attribute subset B is called a reduct of C with respect to D, if the following two conditions are satisfied

$$\begin{aligned}&(1)\ \textit{sufficiency} : \gamma^{\delta}_{B}(D) = \gamma^{\delta}_{C}(D);\\&(2)\ \textit{necessity} : \forall a \in B, \gamma^{\delta}_{B-\{a\}}(D) < \gamma^{\delta}_{B}(D).\end{aligned} \tag{22}$$

From (1), we know that the dependency degree of attribute subset B with respect to D is the same as that of attribute subset C with respect to D. According to (2), we can see that all attributes in B are necessary attributes. Therefore, B is a minimum attribute subset with the same dependency degree as C. There are different reducts using different neighborhood thresholds. For a given threshold, reduct is not unique in decision information table $IS = (U, C, D)$. Further considering Example 2.1, we have $\gamma^{0.1}_{B_2-\{a_3\}}(D) = 0$, $\gamma^{0.1}_{B_2-\{a_4\}}(D) = 0.7$ and $\gamma^{0.1}_{B_2}(D) = \gamma^{0.1}_{C}(D) = 1$. Therefore, $\gamma^{0.1}_{B_2-\{a_3\}}(D) < \gamma^{0.1}_{B_2}(D)$, $\gamma^{0.1}_{B_2-\{a_4\}}(D) < \gamma^{0.1}_{B_2}(D)$ and $\gamma^{0.1}_{B_2}(D) = \gamma^{0.1}_{C}(D)$. So $B_2 = \{a_3, a_4\}$ is a reduct in Example 2.1.

Remark. Can WNRS be replaced by neighborhood rough sets after normalizing all attribute values to [0,1]? Further considering Example 2.1, we first normalize all attribute values to [0,1], then compute dependencies of neighborhood rough sets under B_1 and B_2 at $\delta = 0.35$. $\gamma^{0.35}_{B_1} = 0.4$ and $\gamma^{0.35}_{B_2} = 1$. It seems that neighborhood rough sets after normalizing all attribute values to [0,1] can achieve the same effect as WNRS in Example 2.1, (i.e. the ability of B_2 to approximate D is better than that of B_1 under neighborhood rough sets after normalizing all attribute values to [0,1]). In fact, WNRS performs attribute reduction better than neighborhood rough sets after normalizing all attribute values to [0,1] in most cases. Next, we will add two objects to Example 2.1 to illustrate that neighborhood rough sets after normalizing all attribute values to [0,1] cannot replace WNRS.

Example 3.1. There is a decision information table shown in Table 4, where x_1-x_{10} are from Example 2.1, x_{11} and x_{12} are two new objects.

For Example 3.1, we normalize all attribute values to [0,1], and the normalized result is itself. For $B_1 = \{a_1, a_2\}$ and $B_2 = \{a_3, a_4\}$, the dependencies of neighborhood rough sets are $\gamma^{0.1}_{B_1} = 1$ and $\gamma^{0.1}_{B_2} = 0.8333$. That is to say, the ability of B_1 to approximate D is better than that of B_2 under neighborhood rough sets after normalizing all attribute values to [0,1]. However, the dependencies of WNRS are $\gamma^{0.1}_{B_1} = 0.1667$ and $\gamma^{0.1}_{B_2} = 1$. From the above discussion, we know that neighborhood rough sets after normalizing all attribute values to [0,1] cannot replace WNRS. WNRS can mine the correlation of attributes and decisions before computing neighborhood relations. However, neighborhood rough sets after normalizing all attribute values to [0,1] do not consider the correlation of attributes and decisions, only scale the size of values of an attribute by the maximum value and minimum value of the attribute.

4. Attribute reduction algorithm based on weighted neighborhood rough sets

From the above analysis, we can see that the dependency degree of weighted neighborhood rough sets can be used to evaluate the significance of attribute subsets. If the dependency degree of an attribute subset is greater, its ability to distinguish samples from different decisions is stronger. For a decision information table with m conditional attributes, there are $2^m - 1$ candidate attribute subsets. It is unrealistic to calculate the dependency degree of each attribute subset one by one. There are many strategies to find a reduct, such as genetic algorithm, branch and bound, greedy search, etc. In this paper, our main contribution is to evaluate attribute subsets, so we choose a greedy search algorithm to find an optimal attribute subset. Next, we will define two measures to evaluate the significance of an attribute relative to an attribute subset.

Definition 8. Given a decision information table $IS = (U, C, D)$, an attribute subset $B \subseteq C$, a neighborhood threshold δ and an attribute $a \in B$. the internal significance of a relative to B under D is defined as

$$sig_{in}(a, B, D) = \gamma^{\delta}_{B}(D) - \gamma^{\delta}_{B-\{a\}}(D). \tag{23}$$

Obviously, according to Definition 6 and Property 4, we have $0 \le sig_{in}(a, B, D) \le 1$. a is not an internal necessary attribute relative to B, if $sig_{in}(a, B, D) = 0$, and a can be removed from B. a is an internal necessary attribute relative to B, if $sig_{in}(a, B, D) > 0$.

Definition 9. Given a decision information table $IS = (U, C, D)$, an attribute subset $B \subseteq C$, a neighborhood threshold δ and an attribute $a \in C - B$. the external significance of a relative to B under D is defined as

$$sig_{out}(a, B, D) = \gamma^{\delta}_{B\cup\{a\}}(D) - \gamma^{\delta}_{B}(D). \tag{24}$$

Obviously, from Definition 6 and Property 4, we have $0 \le sig_{out}(a, B, D) \le 1$. a is not an external necessary attribute relative to B, if $sig_{out}(a, B, D) = 0$. a is an external necessary attribute relative to B, if $sig_{out}(a, B, D) > 0$. When we are doing forward search, each round will select the attribute with the most value of $sig_{out}(a, B, D)$ and $sig_{out}(a, B, D) > 0$ to the selected attribute subset.

There are two greedy search strategies for attribute reduction, one is sequential forward search, the other is sequential backward elimination. First, we use the sequential forward search to select attributes, then use the backward elimination to eliminate attributes of $sig_{in}(a, B, D) = 0$ in the selected attribute subset. In order to use sequential forward search, when $B = \emptyset$, we rule $\gamma^{\delta}_{B}(D) = 0$ and $\emptyset$ is not a reduct. Weighted neighborhood rough sets based attribute reduction (WNRS) is shown in Algorithm 1. There is a parameter δ in Algorithm 1. δ is the threshold that controls the size of weighted neighborhood granules, and it needs to be set in advance. δ can be set by the prior knowledge of experts, or found by isometric search. In the experimental section we will show how to search for the threshold δ.

In step 1, the initial state of *red* for sequential forward search is an empty set, and the time complexity is $O(1)$. In step 2, we use formula (8) to calculate the weights of all conditional attributes, and the time complexity is $O(|U| \times |C|)$. In steps 3–5, weighted neighborhood similarity relations W^{δ}_{a} of all conditional attributes are calculated by formula (11), and the time complexity

Algorithm 1 Attribute reduction based on weighted neighborhood rough sets (WNRS)

```
Require: A decision information table $IS = (U, C, D)$ and a threshold $\delta$.
Ensure: Attribute subset $red$.
1:  Initialize: $red \leftarrow \emptyset$; // $red$ is initialized to an empty set;
2:  Calculate the weight of each conditional attribute by formula (9);
3:  for each $a \in C$ do
4:      Compute weighted neighborhood similarity relation $W_a^\delta$ by formula (11);
5:  end for
6:  while $C - red \neq \emptyset$ do
7:      for each $a \in C - red$ do
8:          Compute the dependency degree $\gamma^\delta_{red\cup\{a\}}(D)$ by formula (19);
9:          $sig_{out}(a, red, D) = \gamma^\delta_{red\cup\{a\}}(D) - \gamma^\delta_{red}(D)$;
10:     end for
11:     Find $a_k$ with maximum value of $sig_{out}(a_k, red, D)$;
12:     if $sig_{out}(a_k, red, D) = 0$ then
13:         break; // Loop termination.
14:     else
15:         $red \leftarrow red \cup \{a_k\}$; // Put $a_k$ into $red$.
16:     end if
17: end while
18: for each $a \in red$ do
19:     $sig_{in}(a, red, D) = \gamma^\delta_{red}(D) - \gamma^\delta_{red-\{a\}}(D)$;
20:     if $sig_{in}(a_k, red, D) = 0$ then
21:         $red \leftarrow red - \{a_k\}$; // Remove internal unnecessary attributes.
22:     end if
23: end for
24: return $red$;
```

is $O(|U|^2 \times |C|)$. In steps 6–17, According to the idea of greedy search, we find out the attribute with the greatest external significance relative to the current reduction and add it to the current reduction, and the time complexity is $O(|U| \times |C| \times |U/D|)$. The attributes selected in steps 6–9 may have internal unimportant attributes. So we need to remove the attributes with 0 internal significance in steps 18–23, and the time complexity is $O(|U| \times |red| \times |U/D|)$.

5. Experimental analysis

In this section, we will design a series of experiments to verify the effectiveness and robustness of the proposed WNRS algorithm. Three excellent attribute reduction algorithms, neighborhood rough sets based attribute reduction (NRS) [24], neighborhood discrimination index based attribute reduction (NDI) [39] and neighborhood self-information based attribute reduction (NSI) [43], are selected to compare with WNRS. We will compare these four algorithms from three aspects: (1) the classification accuracies under different classifiers, (2) the running time of reduction algorithms and (3) the number of selected attributes. All algorithms are executed in MATLAB 2015b, and run in hardware environment with Inter(R) Core(TM) i7-4790 CPU @3.60 GHz 3.60 GHz, with 16 GB RAM.

We use KNN and RBF-SVM classifiers to evaluate the performance of these algorithms. The ten datasets downloaded from UCI machine learning repository [48] and ELVIRA Biomedical data set repository [49] are described in Table 5. The first eight datasets are from UCI, and the last two datasets are from ELVIRA. Values of all attributes are first normalized into the interval [0, 1].

Table 5
Data description.

No.	Datasets	Attributes	Samples	Classes
1	Seeds	8	210	3
2	Wine	14	178	3
3	vowel-context	14	990	11
4	Wdbc	31	569	2
5	Wpbc	34	198	2
6	sat-tst	37	2000	6
7	sat-trn	37	4435	6
8	sonar	61	208	2
9	Lung-Cancer	181	12 534	2
10	Prostate-Cancer	136	12 601	2

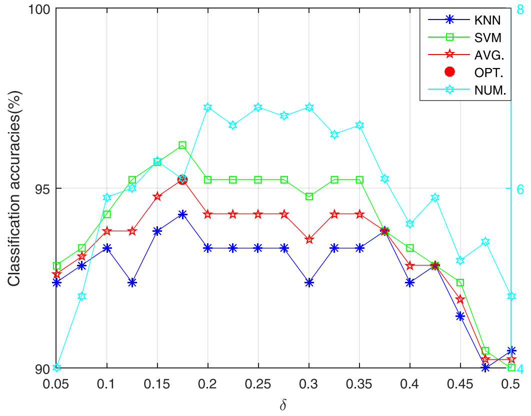


Fig. 2. Number of selected attribute and accuracy varying with neighborhood δ (Seeds).

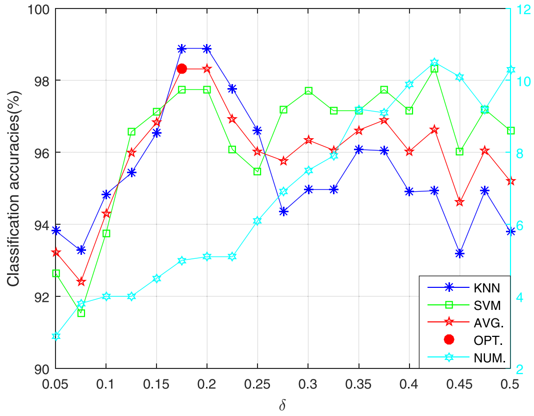


Fig. 3. Number of selected attribute and accuracy varying with neighborhood δ (Wine).

We employ 10-fold cross validation to evaluate the performances of these algorithms. In the training stage, we use the training set to reduce attributes and select an optimal attribute subset. In the validation stage, the sub data extracted from the validation set by the selected attribute subset is used to calculate the classification accuracies of KNN and RBF-SVM. After ten cycles, the average performance of the ten cycles is regarded as the final performance. In the experiments, we search δ from 0.05 to 0.5 with step 0.05 and find the optimal δ for each dataset. The neighborhood thresholds of the other three algorithms are set in the same way. For WNRS, we use $\upsilon = (A^TA + E)^{-1}A^TY$ to compute υ of the high dimensional datasets (Lung-Cancer and Prostate-Cancer) and use $\upsilon = (A^TA)^{-1}A^TY$ to compute υ of the low dimensional datasets.

For WNRS algorithm, the number of selected attributes and classification accuracies are shown in Figs. 2–11 under different

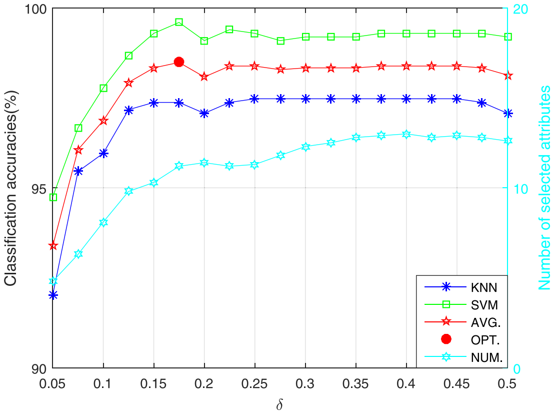


Fig. 4. Number of selected attribute and accuracy varying with neighborhood δ (vowel-context).

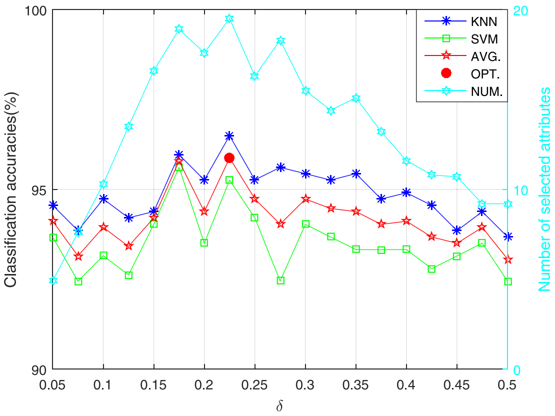


Fig. 5. Number of selected attribute and accuracy varying with neighborhood δ (Wdbc).

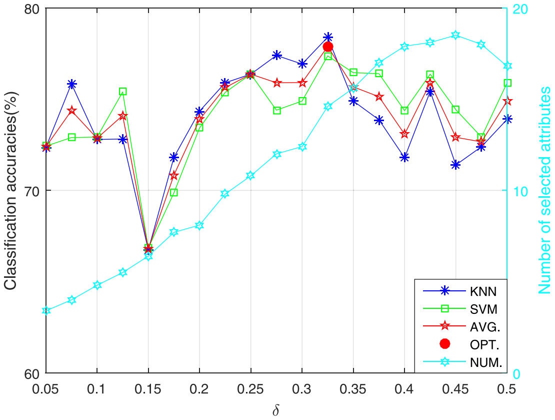


Fig. 6. Number of selected attribute and accuracy varying with neighborhood δ (Wpbc).

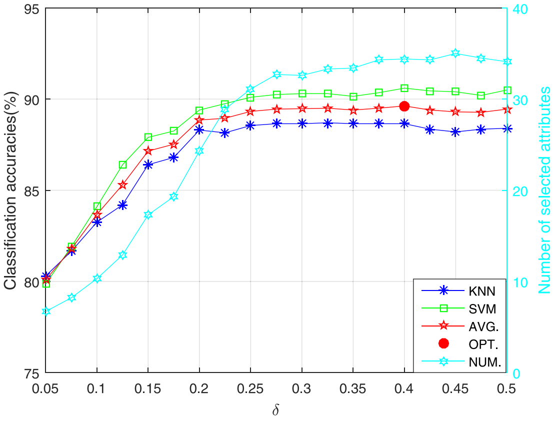


Fig. 7. Number of selected attribute and accuracy varying with neighborhood δ (sat-tst).

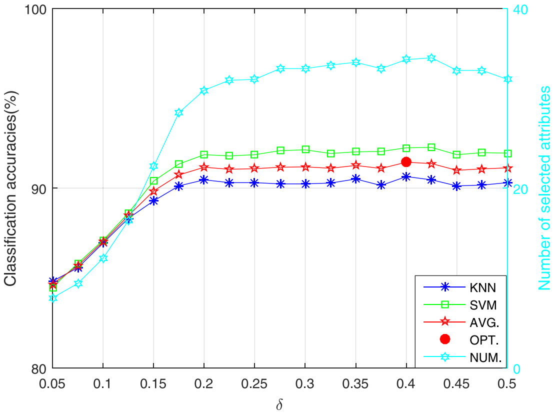


Fig. 8. Number of selected attribute and accuracy varying with neighborhood δ (sat-trn).

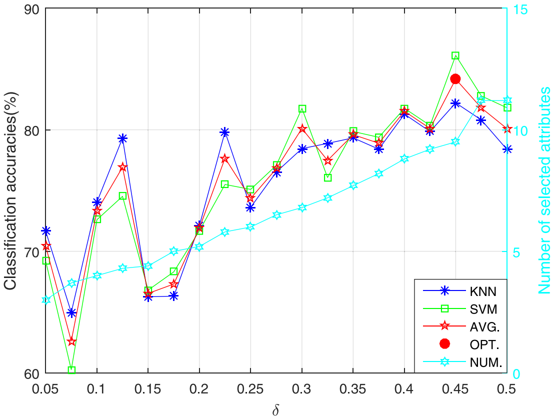


Fig. 9. Number of selected attribute and accuracy varying with neighborhood δ (sonar).

δ. In Figs. 2–11 , the δ corresponding to the red dot indicates the best performing δ on average classification accuracy under KNN and RBF-SVM. From Figs. 2, 3, 5 and 6, we can see that as the δ increases, the number of selected attributes first increases and then decreases (Seeds, Wine, Wdbc and Wpbc). From Figs. 4, 7, 8, 9, 10 and 11, as the δ increases, the number of selected attributes increases (vowel-context, sat-tst, sat-trn, sonar, Lung-Cancer and Prostate-Cancer). With the increase of δ, the classification accuracy first increases, then remains unchanged or decreases. The position of the red dot in each figure is the optimal average classification accuracy and the selected attribute subset with the smaller size. We choose the corresponding value of δ as the neighborhood threshold. From Figs. 2–11, we find when δ is in the interval [0.15, 0.45], classification accuracies of KNN and SVM are relatively high in most cases, so we recommend the optimal interval of neighborhood threshold as [0.15, 0.45].

Table 6
Classification accuracies of reduced data and raw data with KNN (%).

Dataset	Raw data	NRS	NSI	NDI	WNRS
Seeds	92.38 ± 5.59	93.81 ± 3.92	93.33 ± 3.33	91.9 ± 3.92	**94.29 ± 6.66**
Wine	94.90 ± 7.33	94.93 ± 5.53	96.63 ± 2.90	96.63 ± 3.91	**98.89 ± 2.34**
vowel-context	95.76 ± 2.77	96.97 ± 1.72	97.17 ± 1.24	96.57 ± 2.74	**97.37 ± 1.28**
Wdbc	96.31 ± 0.99	**96.84 ± 2.15**	**96.84 ± 2.45**	96.13 ± 2.59	96.32 ± 2.4
Wpbc	71.16 ± 12.99	73.16 ± 8.57	72.26 ± 10.82	74.74 ± 10.02	**78.39 ± 9.33**
sat-tst	88.15 ± 1.40	88.40 ± 1.91	87.95 ± 3.14	88.00 ± 1.86	**88.65 ± 2.68**
sat-trn	90.33 ± 1.11	90.46 ± 1.40	90.48 ± 1.44	90.51 ± 1.34	**90.58 ± 1.17**
sonar	81.69 ± 6.80	78.81 ± 8.88	81.74 ± 6.25	77.93 ± 14.1	**82.21 ± 12.5**
Lung-Cancer	94.44 ± 6.93	**98.89 ± 2.34**	98.33 ± 3.75	**98.89 ± 2.34**	98.33 ± 2.68
Prostate-Cancer	76.37 ± 10.35	81.65 ± 7.66	77.97 ± 6.65	80.27 ± 8.79	**88.85 ± 6.52**
Average	88.1490 ± 5.6260	89.3920 ± 4.4080	89.2700 ± 4.1970	89.1570 ± 5.1610	**91.3880 ± 4.7560**

Table 7
Classification accuracies of reduced data and raw data with RBF-SVM (%).

Dataset	Raw data	NRS	NSI	NDI	WNRS
Seeds	91.90 ± 5.96	94.29 ± 3.76	94.76 ± 3.51	94.29 ± 7.38	**96.19 ± 4.38**
Wine	96.67 ± 4.68	96.60 ± 3.93	96.63 ± 3.91	97.19 ± 5.42	**97.75 ± 2.91**
vowel-context	99.29 ± 1.07	**99.70 ± 0.49**	**99.70 ± 0.49**	99.39 ± 0.85	99.60 ± 0.71
Wdbc	94.91 ± 3.25	94.37 ± 3.40	94.56 ± 3.55	**95.61 ± 2.65**	94.91 ± 2.54
Wpbc	74.74 ± 11.07	75.74 ± 8.48	75.74 ± 8.52	74.74 ± 7.51	**77.37 ± 9.71**
sat-tst	90.25 ± 2.00	90.50 ± 2.35	90.50 ± 2.08	90.55 ± 2.15	**90.60 ± 2.08**
sat-trn	91.75 ± 0.85	92.00 ± 1.06	91.91 ± 1.19	92.04 ± 1.06	**92.18 ± 0.93**
sonar	85.57 ± 7.11	81.29 ± 9.60	83.21 ± 7.11	77.9 ± 11.25	**86.12 ± 8.77**
Lung-Cancer	97.78 ± 2.87	**99.44 ± 1.76**	**99.44 ± 1.76**	98.89 ± 2.34	98.33 ± 2.68
Prostate-Cancer	73.63 ± 17.23	75.99 ± 14.85	80.82 ± 14.19	81.1 ± 12.65	**86.76 ± 6.75**
Average	89.6490 ± 5.6090	89.9920 ± 4.9680	90.7270 ± 4.6310	90.1700 ± 5.3260	**91.9810 ± 4.1460**

Table 8
Running time of four reduction algorithms (s).

Dataset	NRS	NSI	NDI	WNRS
Seeds	0.0158 ± 0.0034	0.0999 ± 0.0131	0.0131 ± 0.0015	0.0143 ± 0.0004
Wine	0.0309 ± 0.0020	0.2343 ± 0.0155	0.0223 ± 0.0024	0.0233 ± 0.0004
vowel-context	3.2924 ± 0.1540	7.1796 ± 0.1633	1.3814 ± 0.0801	2.3048 ± 0.0472
Wdbc	4.0310 ± 0.1452	7.367 ± 0.1147	3.3560 ± 0.4355	2.7692 ± 0.4430
Wpbc	0.2944 ± 0.0408	1.2595 ± 0.07	0.3792 ± 0.0358	0.2687 ± 0.0537
sat-tst	123.2618 ± 0.8427	205.8333 ± 12.9959	96.0991 ± 7.8208	103.2826 ± 7.2974
sat-trn	677.5580 ± 20.6859	978.1363 ± 148.2848	447.4005 ± 34.4245	537.2471 ± 18.5589
sonar	1.0150 ± 0.1288	3.1969 ± 0.2286	0.4160 ± 0.0344	0.3330 ± 0.0545
Lung-Cancer	34.6689 ± 2.8309	186.8655 ± 15.9527	10.7487 ± 1.3415	16.7987 ± 0.8505
Prostate-Cancer	24.1142 ± 3.3243	1152.6995 ± 215.4227	28.9416 ± 2.7797	17.5283 ± 1.0708
Average	86.8282 ± 2.8158	254.2872 ± 39.3261	58.8758 ± 4.6956	68.0570 ± 2.8377

Table 9
Number of selected attributes with four reduction algorithms.

Dataset	Raw data	NRS	NSI	NDI	WNRS
Seeds	7	5.6 ± 0.52	6.9 ± 0.32	6.5 ± 0.53	6.1 ± 0.32
Wine	13	6.9 ± 0.32	7.4 ± 0.52	6.8 ± 0.42	5.0 ± 0.00
vowel-context	13	11.0 ± 0.00	10.6 ± 0.52	10.0 ± 0.00	11.2 ± 0.79
Wdbc	30	26.7 ± 0.82	28.1 ± 0.32	21.1 ± 2.92	19.1 ± 2.85
Wpbc	33	16.0 ± 1.05	19.2 ± 1.03	19.4 ± 1.65	14.6 ± 2.27
sat-tst	36	34.6 ± 1.17	34.6 ± 1.07	35.6 ± 0.97	34.4 ± 1.35
sat-trn	36	34.4 ± 0.52	33.1 ± 0.99	34.0 ± 0.47	34.3 ± 0.95
sonar	60	20.2 ± 1.23	19.6 ± 1.35	11.2 ± 0.63	9.5 ± 0.97
Lung-Cancer	12 533	7.5 ± 0.53	6.1 ± 0.57	3.8 ± 0.42	2.9 ± 0.32
Prostate-Cancer	12 600	9.7 ± 1.06	46.5 ± 6.52	13.4 ± 0.7	5.3 ± 0.48
Average	2536.1	17.26 ± 0.72	21.21 ± 1.32	16.18 ± 0.87	14.24 ± 1.03

The classification accuracies of the raw data and the reduced data by using the four reduction algorithms under KNN and RBF-SVM are shown in Tables 6 and 7, where the boldface highlights the best performance over different reduction algorithms. From Tables 6 and 7, we can see that the average accuracies of NRS, NSI, NDI and WNRS are better than those of raw data with KNN and RBF-SVM. Compared with the average accuracy of raw data, we can find that the accuracies of NRS, NSI, NDI and WNRS with KNN improved by 1.2430%, 1.1210%, 1.0080% and 3.2390%, respectively. The performances of NRS, NSI, NDI and WNRS with RBF-SVM improved by 0.3430%, 1.0780%, 0.5210% and 2.3320%, respectively. For classification accuracies of KNN and RBF-SVM, WNRS performs best in most cases. Therefore, the performance of WNRS is obviously better than those of others algorithms.

Running time is an important index to evaluate the feasibility of reduction algorithms. Running time of four reduction algorithms is shown in Table 8. From Table 8, we know that the running time of WNRS is slightly worse than that of NDI. The main reason is that compared with NRS, NSI and WNRS algorithms, NDI does not need to calculate neighborhood similarity classes. The running time of WNRS is obviously better than that of NRS and NSI. The main reason is that in the WNRS algorithm, attributes that are highly related to decisions are easier to select, which makes it possible to quickly find an informative

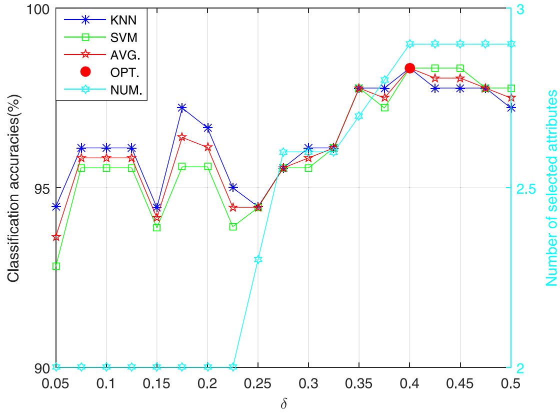


Fig. 10. Number of selected attribute and accuracy varying with neighborhood δ (Lung-Cancer).

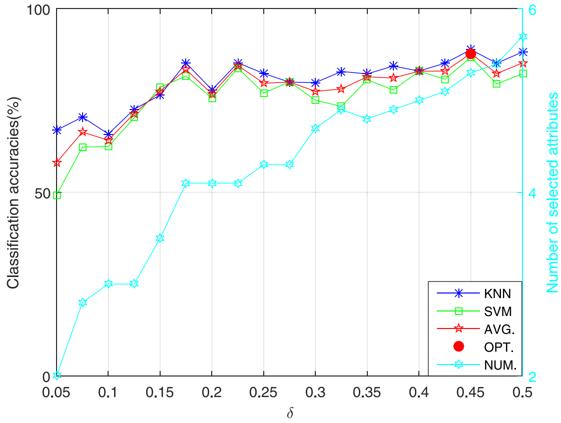


Fig. 11. Number of selected attribute and accuracy varying with neighborhood δ (Prostate-Cancer).

and compact attribute subset. Obviously, both WNRS and NDI are good reduction algorithms under the condition of ensuring both computational efficiency and classification accuracy. When there is a higher demand for classification accuracy, our optimal choice is WNRS.

The size of selected attribute subsets is an important issue for attribute reduction. The purpose of attribute reduction is to find a small and informative attribute subset. The average size of the selected attribute subsets with 10-fold cross validation is shown in Table 9. From Table 9, we can see that the average size of selected attribute subsets by WNRS (14.24) is the smallest when compared with those of the other reduction algorithms. The proposed WNRS algorithm has higher compression ratio, which has obvious advantages in dimension reduction.

WNRS based attribute reduction method can quickly select fewer attributes with discernment information to obtain effective classification performance.

6. Conclusion and future work

Removing low correlation and redundant attributes before classification and regression can improve the performance and computational efficiency. The traditional attribute reduction based on neighborhood rough sets only uses the same weights when computing neighborhood relations and does not fully mine the internal knowledge of attributes and decisions in advance. In this work, we first calculate the correlation coefficients of attributes with respect to decisions to assign different weights to attributes. Then we introduce weights into neighborhood relations to define weighted neighborhood rough set model and employ the dependency of weighted neighborhood relation to measure the significance of attribute subsets. Finally, a greedy attribute subset selection based on weighted neighborhood rough sets (WNRS) algorithm is designed to find an attribute subset which is highly relevant and highly dependent on decisions. Experimental results show that WNRS can achieve high classification performance against other state-of-the-art methods.

This paper mainly studies correlation coefficient between attributes and decisions, and it does not mine correlation coefficient between attributes. In the future, we will assign weights to attributes by using correlation coefficient between attributes. Decision information tables with mixed attribute values such as categorical, real-valued and interval-valued, will be studied.

CRediT authorship contribution statement

Meng Hu: Conceptualization, Methodology, Writing - original draft. **Eric C.C. Tsang:** Validation, Investigation, Supervision. **Yanting Guo:** Designing and performing the experiment, Investigation. **Degang Chen:** Writing - review & editing. **Weihua Xu:** Software, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the Macau Science and Technology Development Fund (No. 0019/2019/A1 and No. 0075/2019/A2), the Natural Science Foundation of China (No. 61976245 and No. 61772002).

All authors read and approved the final manuscript.

References

[1] Z. Pawlak, Rough sets, Int. J. Comput. Inform. Sci. 11 (5) (1982) 341–356.

[2] Y.Y. Yao, Relational interpretations of neighborhood operators and rough set approximation operators, Inform. Sci. 111 (1–4) (1998) 239–259.

[3] Y.Y. Yao, Rough sets, neighborhood systems and granular computing, IEEE Conf. Electr. Computer Engg. Canada, 1999, pp. 1553–1558.

[4] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17 (1990) 191–209.

[5] D. Dubois, H. Prade, Putting rough sets and fuzzy sets together, intelligent decision support, in: R. Slowinski (Ed.), Handbook of Applications and Advances of the Rough Set Theory, Kluwer Academic, 1992, pp. 203–232.

[6] E.C.C. Tsang, D. Chen, D.S. Yeung, X.Z. Wang, J.W.T. Lee, Attributes reduction using fuzzy rough sets, IEEE Trans. Fuzzy Syst. 16 (5) (2008) 1130–1141.

[7] W. Wu, M. Shao, X. Wang, Using single axioms to characterize (S, T)-intuitionistic fuzzy rough approximation operators, Int. J. Mach. Learn. Cybern. 10 (1) (2019) 27–42.

[8] S. Greco, B. Matarazzo, R. Slowinski, Rough approximation by dominance relations, Int. J. Intell. Syst. 17 (2) (2002) 153–171.

[9] W. Xu, Y. Li, X. Liao, Approaches to attribute reductions based on rough set and matrix computation in inconsistent ordered information systems, Knowl.-Based Syst. 27 (2012) 78–91.

[10] A.H. Attia, A.S. Sherif, G.S. El-Tawel, Maximal limited similarity-based rough set model, Soft Comput. 20 (8) (2016) 3153–3161.

[11] R. Slowinski, D. Vanderpooten, A generalized definition of rough approximations based on similarity, IEEE Trans. Knowl. Data Eng. 12 (2) (2000) 331–336.

[12] D. Chen, L. Zhang, S. Zhao, Q. Hu, P. Zhu, A novel algorithm for finding reducts with fuzzy rough sets, IEEE Trans. Fuzzy Syst. 20 (2) (2012) 385–389.

[13] L. Chen, D. Chen, H. Wang, Fuzzy kernel alignment with application to attribute reduction of heterogeneous data, IEEE Trans. Fuzzy Syst. 27 (7) (2019) 1469–1478.

[14] C. Wang, Y. Huang, M. Shao, et al., Fuzzy rough set-based attribute reduction using distance measures, Knowl.-Based Syst. 164 (2019) 205–212.

[15] X. Zhang, C. Mei, D. Chen, et al., Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy, Pattern Recognit. 56 (2016) 1–15.

[16] X. Zhang, C. Mei, D. Chen, Y. Yang, J. Li, Active incremental feature selection using a fuzzy-rough-set-based information entropy, IEEE Trans. Fuzzy Syst. 28 (5) (2019) 901–915.

[17] D. Chen, W. Zhang, D. Yeung, E.C.C. Tsang, Rough approximations on a complete completely distributive lattice with applications to generalized rough sets, Inform. Sci. 176 (13) (2006) 1829–1848.

[18] C. Luo, T. Li, H. Chen, D. Liu, Incremental approaches for updating approximations in set-valued ordered information systems, Knowl.-Based Syst. 50 (50) (2013) 218–233.

[19] Y. Guo, E.C.C. Tsang, W. Xu, et al., Local logical disjunction double-quantitative rough sets, Inform. Sci. 500 (2019) 87–112.

[20] Y. Guo, E.C.C. Tsang, M. Hu, et al., Incremental updating approximations for double-quantitative decision-theoretic rough sets with the variation of objects, Knowl.-Based Syst. 189 (2020) 105082.

[21] C. Luo, T. Li, Y. Huang, H. Fujitad, Updating three-way decisions in incomplete multi-scale information systems, Inform. Sci. 476 (2019) 274–289.

[22] C. Luo, T. Li, H. Chen, H. Fujita, Z. Yi, Incremental rough set approach for hierarchical multicriteria classification, Inform. Sci. 429 (2018) 72–87.

[23] W. Wu, W. Zhang, Neighborhood operator systems and approximations, Inform. Sci. 144 (2002) 201–217.

[24] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inform. Sci. 178 (18) (2008) 3577–3594.

[25] Q. Hu, W. Pedrycz, D. Yu, et al., Selecting discrete and continuous features based on neighborhood decision error minimization, IEEE Trans. Syst. Man Cybern. B 40 (1) (2010) 137–150.

[26] H. Chen, T. Li, X. Fan, C. Luo, Feature selection for imbalanced data based on neighborhood rough sets, Inform. Sci. 483 (2019) 1–20.

[27] S. Liang, X. Yang, X. Chen, et al., Stable attribute reduction for neighborhood rough set, Filomat 32 (5) (2018) 1809–1815.

[28] H. Chen, T. Li, Y. Cai, C. Luo, H. Fujita, Parallel attribute reduction in dominance-based neighborhood rough set, Inform. Sci. 373 (2016) 351–368.

[29] Y. Guo, E.C.C. Tsang, W. Xu, D. Chen, Adaptive weighted generalized multi-granulation interval-valued decision-theoretic rough sets, Knowl.-Based Syst. 187 (2020) 104804.

[30] E.C.C. Tsang, Q. Hu, D. Chen, Feature and instance reduction for PNN classifiers based on fuzzy rough sets, Int. J. Mach. Learn. Cybern. 7 (2016) 1–11.

[31] Z.S. Xu, X.L. Zhang, Hesitant fuzzy multi-attribute decision making based on TOPSIS with incomplete weight information, Knowl.-Based Syst. 52 (2013) 53–64.

[32] S. Vluymans, N.M. Parthaláin, C. Cornelis, Y. Saeys, Weight selection strategies for ordered weighted average based fuzzy rough sets, Inform. Sci. 501 (2019) 155–171.

[33] M. Hashemzadeh, A.G. Oskouei, N. Farajzadeh, New fuzzy C-means clustering method based on feature-weight and cluster-weight learning, Appl. Soft Comput. 78 (2019) 324–345.

[34] M. Hu, E.C.C. Tsang, Y.T. Guo, W.H. Xu, Fast and robust attribute reduction based on the separability in fuzzy decision systems, IEEE Trans. Cybern. (2021) http://dx.doi.org/10.1109/TCYB.2020.3040803, in press.

[35] J. Huang, Y. Wei, J. Yi, M. Liu, An improved kNN based on class contribution and feature weighting, in: 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), IEEE, 2018, pp. 313–316.

[36] X. Xie, X. Qin, C. Yu, X. Xu, Test-cost-sensitive rough set based approach for minimum weight vertex cover problem, Appl. Soft Comput. 64 (2018) 423–435.

[37] Q. Hu, L. Zhang, D. Zhang, W. Pan, S. An, W. Pedrycz, Measuring relevance between discrete and continuous features based on neighborhood mutual information, Expert Syst. Appl. 38 (2011) 10737–10750.

[38] C. Wang, M. Shao, Q. He, Y. Qian, Y. Qi, Feature subset selection based on fuzzy neighborhood rough sets, Knowl.-Based Syst. 111 (1) (2016) 173–179.

[39] C. Wang, Q. Hu, X. Wang, et al., Feature selection based on neighborhood discrimination index, IEEE Trans. Neural Netw. Learn. Syst. 29 (7) (2018) 2986–2999.

[40] L.H. Patil, M. Atique, A novel approach feature selection based on neighborhood positive region (NPR), Int. J. Comput. Appl. 124 (3) (2015) 16–22.

[41] Q. Wang, Y. Qian, X. Liang, et al., Local neighborhood rough set, Knowl.-Based Syst. 153 (2018) 53–64.

[42] A. Mariello, R. Battiti, Feature selection based on the neighborhood entropy, IEEE Trans. Neural Netw. Learn. Syst. 29 (12) (2018).

[43] C. Wang, Y. Huang, M. Shao, et al., Feature selection based on neighborhood self-information, IEEE Trans. Cybern. 50 (9) (2020) 4031–4042.

[44] X. Yang, S. Liang, H. Yu, S. Gao, Y. Qian, Pseudo-label neighborhood rough set: Measures and attribute reductions, Internat. J. Approx. Reason. 105 (2019) 112–129.

[45] Z. Jiang, K. Liu, X. Yang, H. Yu, H. Fujita, Y. Qian, Accelerator for supervised neighborhood based attribute reduction, Internat. J. Approx. Reason. 119 (2020) 122–150.

[46] B. Sang, H. Chen, L. Yang, D. Zhou, T. Li, W. Xu, Incremental attribute reduction approaches for ordered data with time-evolving objects, Knowl.-Based Syst. 212 (2021) 106583.

[47] B. Sang, H. Chen, T. Li, W. Xu, H. Yu, Incremental approaches for heterogeneous feature selection in dynamic ordered data, Inform. Sci. 541 (2020) 475–501.

[48] D. Dua, C. Graff, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2019, http://archive.ics.uci.edu/ml.

[49] A. Cano, A. Masegosa, S. Moral, ELVIRA Biomedical Data Set Repository, 2005, http://leo.ugr.es/elvira/DBCRepository/.