Information Sciences 584 (2022) 301–324

Contents lists available at ScienceDirect

Information Sciences

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

Attribute reduction based on overlap degree and *k*-nearest-neighbor rough sets in decision information systems

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, China
[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 8 February 2021
Received in revised form 16 August 2021
Accepted 24 October 2021
Available online 5 November 2021

Keywords:
k-nearest-neighbor rough sets
Attribute reduction
Overlap degree
Neighborhood rough sets

ABSTRACT

The *k*-nearest-neighbor rule is a popular classification technique, and rough set theory is an effective mathematical tool to deal with the uncertainty of data. Rough set models based on *k*-nearest-neighbor relations have a strong ability to approximate decisions, but the calculation is very time-consuming. In this paper, we model the overlap degree of objects from different categories in advance to accelerate the attribute reduction and improve the classification performance of the selected attributes. Firstly, we define the coincidence degree (*CD*) and distance (*DIS*) of objects from different categories to measure the coverage and distance of between-class objects. Secondly, we combine *CD* and *DIS* to define the overlap degree (*OD*) to pre-sort attributes, then use *k*-nearest-neighbor rough sets to filter inconsistent and redundant attributes. The pre-sort operation based on *OD* can greatly reduce the number of searches for attributes and ensure that the attributes with high separability should be selected first. Furthermore, we design a fast reduction algorithm (*OD&KNN*) to obtain a reduct with the ability to approximate decisions as well as the original attributes but with lower *OD*. Comparing experimental results and time complexity of *OD&KNN* with state-of-the-art algorithms, *OD&KNN* is more efficient for high-dimensional data while ensuring classification accuracy.

1. Introduction

In the early 1980s, Pawlak proposed rough set theory (RST) [29], from the perspective of information granules to approximate concepts, which has been widely used to measure the uncertainty and incompleteness of information systems. In order to discover the knowledge of information systems with fuzzy attributes and concepts, Dubois and Prade combined fuzzy sets [43] and rough sets (RS) to propose rough fuzzy sets (RFS) and fuzzy rough sets (FRS) [11,12]. The RFS and FRS models can only deal with fuzzy systems or fuzzy concepts, but they can not characterize the degree of inclusion between information granules and concepts. To describe the degree of inclusion, Ziarko introduced the precision parameter β to RS, then proposed the variable precision rough sets (VPRS)[45]. To solve the problem of ordering attribute values in the evaluation of bankruptcy risk, Greco et al. presented dominance relations to assess the level of risk, and established the dominance-based rough sets for bankruptcy evaluation [14]. With increasing number of data types, rough sets based on various relations have been studied in the literatures, such as neighborhood rough sets (NRS) [19,41], *k*-nearest neighborhood

* Corresponding author.
E-mail addresses: humeng24@sina.com (M. Hu), cctsang@must.edu.mo (E.C.C. Tsang), ytguosx@sina.com (Y. Guo).

https://doi.org/10.1016/j.ins.2021.10.063
0020-0255/

rough sets [18,37] and fuzzy neighborhood rough sets [35]. The classical rough set model and its extended models are applied to attribute reduction [23,33,34], rule extraction [15], gene data expression [31,32] and decision analysis [16,17,26].

Attribute reduction based on RS aims to find the smallest attribute subset that can keep the positive region unchanged in most cases. Different relations produce different information granules, and different information granules induce different positive regions. Therefore, for an information system, we can get different reducts by using different relations. Reducts are not unique for a given information system, so the reducts obtained by different search strategies are different. To get a reduct with high classification accuracy, researchers have studied attribute reduction based on neighborhood rough sets [1,24], fuzzy rough sets [5,25,27], dominance-based rough sets [30] and others [6,22,42].

NRS is a rough set model for dealing with information systems with real-valued attributes. Hu et al. [19] defined the neighborhood relation of heterogeneous features, then applied the relation to define a measure for evaluating the importance of feature subsets. In addition, they used a forward selection strategy to find an optimal feature subset. After that, Hu et al. [18] proposed k-nearest-neighbor relations and δ-neighborhood relations to perform attribute reduction. To find the upper and lower approximations for dynamic data in neighborhood systems, Zhang et al. [44] proposed four methods of updating approximations to efficiently model the knowledge of dynamic neighborhood systems. Chen et al. [7] combined dominance and neighborhood relations to define a novel rough set model, and designed a parallel reduction algorithm by using the neighborhood dominance relation matrix. In addition, Chen et al. [8] divided the boundary region into lower and upper boundary regions to define the importance of attributes in neighborhood systems, and designed a reduction algorithm based on particle swarm optimization. Wang et al. [35] studied attribute reduction based on fuzzy neighborhood rough sets in fuzzy decision systems. Then they defined neighborhood discrimination index to reduce running time of attribute reduction [36] and used neighborhood self-information to improve the classification accuracy of reducts [38]. Other attribute reduction methods based on neighborhood can be found in [20,21,28,39].

The above series of neighborhood rough sets based on neighborhood information granules have the following disadvantage. For information systems with different distribution densities for different attributes (even when rescaled), the neighborhood parameter that controls the size of information granules should be different for the low-density sample distribution region and the high-density sample distribution region. To further improve the effectiveness of attribute reduction methods based on neighborhood information granules, researchers have proposed various improved models. Hu et al. [18] used the k-nearest-neighbor relation to granulate information systems with mixed attributes, and designed a forward attribute reduction based on variable precision k-nearest-neighbor algorithm (FarVPKNN). They found that the classification performance of the reduct obtained by the k-nearest-neighbor relation is better than that of by the δ-neighborhood relation in most cases. Wang et al. [37] combined unit neighborhood information granules and k-nearest-neighbor information granules to define k-nearest neighborhood information granules, and used the defined granules to design the k-nearest neighborhood algorithm (NNRS) for attribute reduction.

Both FarVPKNN and NNRS need to repeatedly sort samples in the reduction process of continuously selecting the relatively important attributes, so the time complexity is very high, and the greedy search strategies (sequentially forward selection and sequentially backward elimination [20]) are also very inefficient. Meanwhile, in the aforementioned attribute reduction methods, the attribute evaluation functions mainly utilize the consistency of conditional attributes and decision attributes in information granules, and do not consider the separability of decision information granules for different conditional attributes. However, the separability of selected attributes is closely related to their classification performance in classification tasks. To solve the above problems, we will improve the k-nearest-neighbor attribute reduction rule from the search strategy and the selection of high-quality attributes in approximate decisions and separability. In this work, we use the coincidence degree (CD) and the distance (DIS) of objects from different categories to define the overlap degree (OD) of objects from different categories for each of the attributes, and employ the overlap degree to pre-sort attributes. Starting with the attribute with the highest OD, we use the dependency of k-nearest-neighbor rough sets to remove redundant attributes one by one. Finally, we design an algorithm ($OD\&KNN$) to perform attribute reduction based on OD and k-nearest-neighbor rough sets. Compared with several existing attribute reduction algorithms, $OD\&KNN$ has higher computational efficiency in terms of the dimensionality of the data. Experimental results show that $OD\&KNN$ is effective and efficient.

The main contributions of this paper are as follows: 1) We define a measure to evaluate the separability of attributes with respect to decisions, then combine the dependency degree to propose an attribute selection approach to capture attributes with both high separability and strong approximation ability. 2) We design fast attribute reduction algorithms for low dimensional data and high-dimensional data based on pre-sorted attribute sets by using OD, respectively. From Table 6, we know that $OD\&KNN$ is computationally more efficient in terms of the dimensionality of the data. 3) We develop an efficient attribute search strategy under the constraints of multi-metric, which provides a new way for a fast attribute reduction of complex data.

The paper is organized as follows. In Section 2, we review NRS and point out its weaknesses. In Section 3, the k-nearest-neighbor rough set model is introduced and its advantages are analyzed. In Section 4, we define some measures to evaluate the overlap degree of objects from different categories for single attribute, and combine the proposed measures and k-nearest-neighbor rough sets to design an attribute reduction algorithm ($OD\&KNN$). In Section 5, we employ public datasets to verify effectiveness and efficiency of $OD\&KNN$. Finally, we summarize this paper in Section 6.

2. Preliminaries

In this section, we briefly review NRS; more detailed descriptions can be found in [19,41]. Meanwhile, we analyze the weakness of NRS in terms of concept description.

2.1. Neighborhood rough sets

Let $S = (U, A)$ be an information system, where $U = \{x_1, x_2, \cdots, x_n\}$ is a nonempty and finite object set which is called the universe. $A = \{a_1, a_2, \cdots, a_m\}$ is an attribute set, to characterize the objects of the universe. In $S = (U, A)$, if $A = C \cup D$, where C is a conditional attribute set and D is a decision attribute set, then S is a decision information system which is usually denoted as $S = (U, C, D)$. Let the decision partition U/D be $\{D_1, D_2, \cdots, D_r\}$.

In a decision information system $S = (U, C, D)$, for any $B \subseteq C$ and a given neighborhood threshold δ, the dependency degree of D w.r.t. B in S is

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

where $POS_B^\delta(D) = \bigcup_{i=1}^{r} R_B^\delta(D_i)$ is the decision positive region of D w.r.t. B. $\overline{R}_B^\delta(D_i) = \{x | \delta_B(x) \cap D_i \neq \varnothing, x \in U\}$ and $R_B^\delta(D_i) = \{x | \delta_B(x) \subseteq D_i, x \in U\}$ are upper and lower approximations of D_i w.r.t. B, respectively, where $\delta_B(x) = \{y | d_B(x, y) \leqslant \delta\}$ is the δ-neighborhood of x w.r.t. B and d_B is a distance function.

$\gamma_B^\delta(D)$ is the proportion of consistent objects in U, which can be used to measure the approximation ability of B w.r.t. D in S. The larger the $\gamma_B^\delta(D)$ is, the stronger the approximation ability of B is. There are two factors that will affect $\gamma_B^\delta(D)$ for a given decision information system. One is B which characterizes objects of universe. The more coordinated B w.r.t. D is, the greater the $\gamma_B^\delta(D)$ is. That is to say, when the consistency of the conditional attribute subset with respect to the decision attribute set is larger, the attribute subset has stronger approximation ability. The other is the neighborhood threshold δ. The size of neighborhood information granules can be controlled by adjusting δ, and then the ability of attribute subset B to approximate decision D can be improved.

2.2. Weakness of neighborhood rough sets

How to determine neighborhood threshold δ is a key problem of neighborhood rough sets. First of all, for data with different distribution densities for different attributes (even when rescaled), neighborhood parameter δ should be set with different values according to different distribution densities. The weakness of NRS is that the same neighborhood threshold is used to granulate data with different distribution densities for different attributes, which may lead to strong approximation ability but weak classification ability. Furthermore it is difficult to get a suitable threshold for different attribute subsets. Moreover, the selection of neighborhood parameters will directly affect the approximation ability of the selected attribute subset to the decision. If δ is too large, the size of neighborhood information granule will be too large, which will reduce the ability of B to approximate D. If δ is too small, the size of neighborhood information granule will be too small, which will improve the ability of B to approximate D, but it will lead to overfitting. Therefore, it is more difficult to choose δ for data with different distribution densities for different attributes. In practice, many data have different dimensions for different attributes, and the density distribution of data is also different. For example, the dimension of neighborhood threshold of information granules formed by cities within 100 km around London is measured in kilometer. The dimension of neighborhood threshold of information granules formed by hotels within 500 m around Oxford University is measured in meter. We can eliminate the dimension of data by normalization, but the difference of distribution density of data for different attributes still exists. It is unreasonable to use a neighborhood threshold to calculate the neighborhood information granules of an object for each attribute. Next, we use an example to illustrate the influence of neighborhood parameters on the approximation ability of attribute subsets.

Example 2.1 A given decision information system $S = (U, C, D)$ with 12 objects, 4 conditional attributes and 1 decision attribute is shown in Table 1a. From Table 1a, we know that $U/D = \{D_1, D_2\}$, where $D_1 = \{x_1, x_2, \cdots, x_6\}$ and $D_2 = \{x_7, x_8, \cdots, x_{12}\}$. The shortest distance between two objects with different decisions is 0.053852 for attribute subset $B_1 = \{a_1, a_2\}$. The shortest distance between two objects with different decisions is 0.20881 for attribute subset $B_2 = \{a_3, a_4\}$. From Figs. 1(a), (b) and (c), we can see that the best threshold value is 0.053852 for $B_1 = \{a_1, a_2\}$. When δ is less than 0.053852, B_1 has the strongest ability to approximate D. However, when the size of information granules is too small, it will lead to overfitting. When δ is more than 0.053852, the ability of B_1 to approximate D will decrease. From Figs. 1 (d), (e) and (f), we can see that the best threshold value is 0.20881 for $B_2 = \{a_3, a_4\}$. When δ is less than 0.20881, B_2 has the strongest ability to approximate D. If the size of information granules is too small, it will lead to overfitting. When δ is more than 0.20881, the ability of B_2 to approximate D will decrease.

We take δ equal to 0.05, 0.08 and 0.22 to calculate the dependency degree for $B_1 = \{a_1, a_2\}$ and $B_2 = \{a_3, a_4\}$, respectively. When $\delta = 0.05, \gamma_{B_1}^{0.05}(D) = 1$ and $\gamma_{B_2}^{0.05}(D) = 1$. When $\delta = 0.08, \gamma_{B_1}^{0.08}(D) = 0.3333$ and $\gamma_{B_2}^{0.08}(D) = 1$. When $\delta = 0.22, \gamma_{B_1}^{0.22}(D) = 0$ and $\gamma_{B_2}^{0.22}(D) = 0.6667$. From the above results, we can see that the ability of B_1 to approximate D is not better than that

Table 1
A decision information system.

U	C				D
	a_1	a_2	a_3	a_4	d
x_1	0.06	0.08	0.11	0.91	1
x_2	0.03	0.10	0.32	0.92	1
x_3	0.03	0.14	0.31	0.43	1
x_4	0.06	0.11	0.45	0.41	1
x_5	0.05	0.13	0.67	0.68	1
x_6	0.09	0.14	0.81	0.12	1
x_7	0.08	0.03	0.23	0.69	2
x_8	0.12	0.07	0.39	0.21	2
x_9	0.10	0.03	0.61	0.89	2
x_{10}	0.14	0.09	0.78	0.93	2
x_{11}	0.13	0.07	0.74	0.41	2
x_{12}	0.14	0.03	0.91	0.43	2

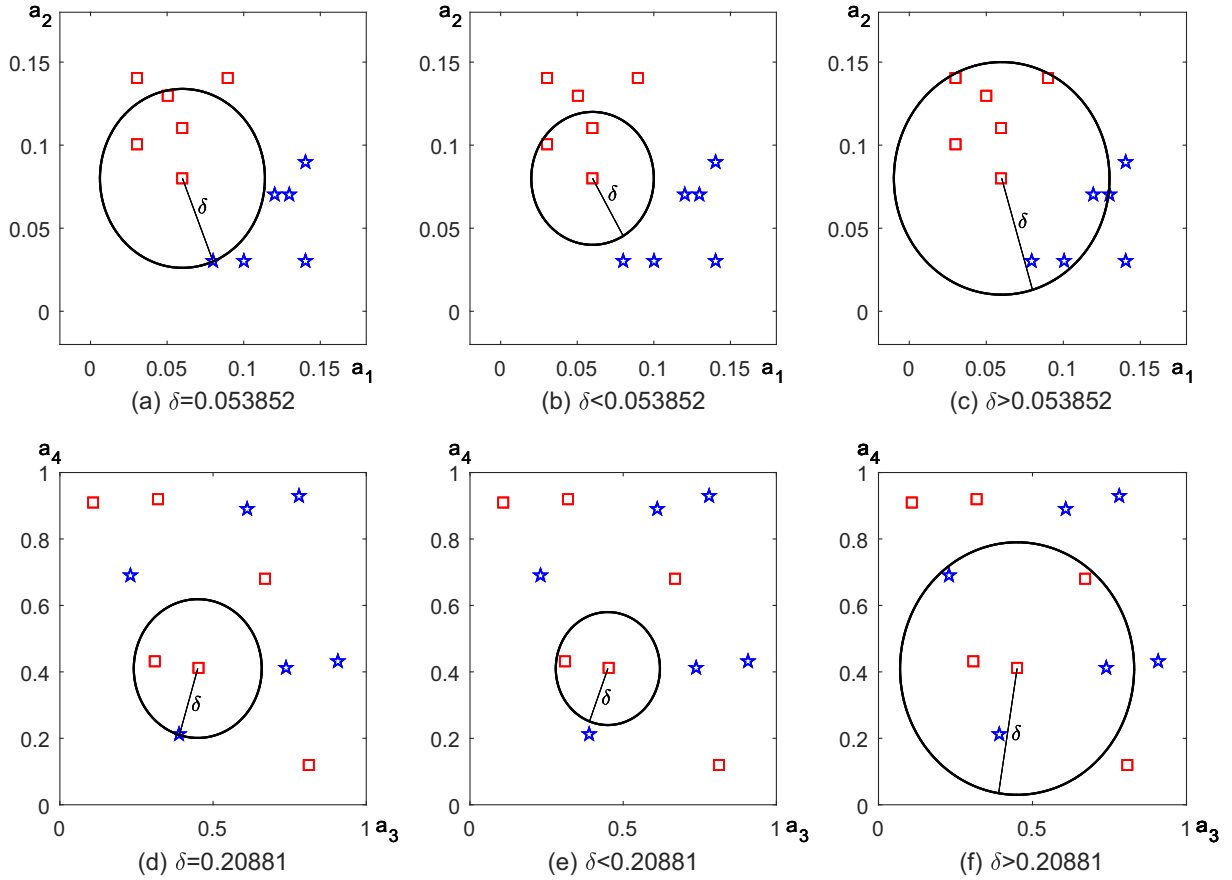


Fig. 1. Size of granules for different δ in $\{a_1, a_2\}$ and $\{a_3, a_4\}$.

of B_2 no matter what value δ takes. Fig. 2 is the classification results using the KNN algorithm and RBF-SVM algorithm, where $k = 3$ in KNN, $C = 1$ and $\delta = 1$ in RBF-SVM. From Fig. 2, the separability of B_1 is significantly better than that of B_2. As can be seen from Figs. 2 (a) and (b), all objects can be classified correctly using KNN and SVM algorithms without overfitting for B_1. From Figs. 2(c) and (d), there are some objects that have been misclassified using KNN and SVM algorithms for B_2. Moreover, there are overfitting for B_2. From the above discussion, we can see that the ability of B_2 to approximate D is better than that of B_1 in neighborhood rough sets, but the separability of objects for B_1 is better than that of B_2.

At present, many research results have shown that the selection of attributes has certain advantages in classification accuracy based on the dependency degree of conditional attribute subsets with respect to decisions. From the results of the above cases, there is no positive correlation between the approximation ability of attribute subsets and the separability

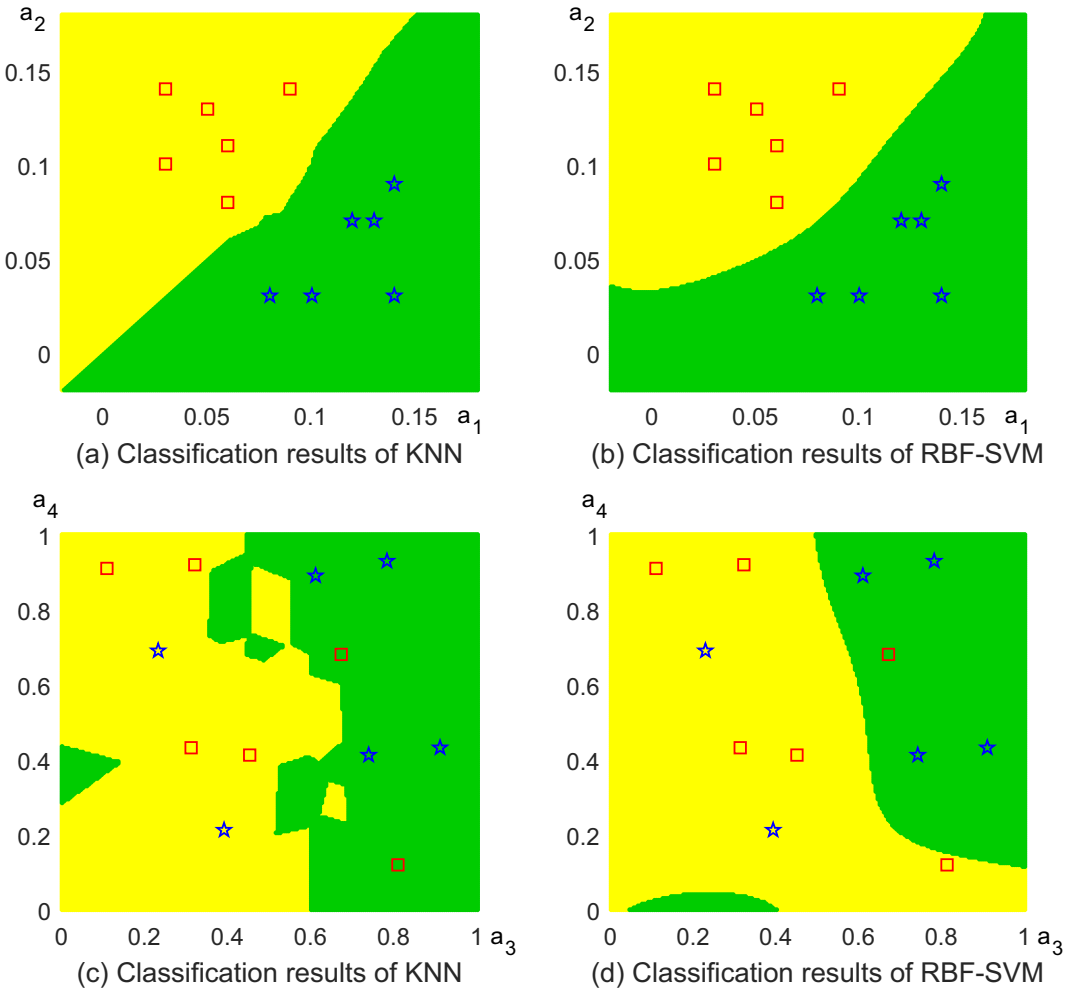


Fig. 2. Classification results for $\{a_1, a_2\}$ and $\{a_3, a_4\}$.

of attribute subsets for decisions. However, the separability of attributes can well describe the classification performance of the attributes for decisions. Therefore, we consider the separability and approximation ability of attributes for decisions simultaneously to ensure the performance of the selected attributes in the classification task. In addition, we will design an efficient search strategy to achieve fast reduction.

3. K-Nearest-neighbor rough sets

In this section, we systematically study k-nearest-neighbor rough set model and analyze its important properties. Meanwhile, we use the corresponding measures to evaluate the importance of attribute subsets.

Definition 1. Let $S = (U, C, D)$ be a decision information system. $\forall x \in U, B \subseteq C$, the k-nearest-neighbor class $top_B^k(x)$ of object x for attribute subset B is

$$top_B^k(x) = \underset{a \in B}{\cap} top_a^k(x), \tag{1}$$

where $top_a^k(x)$ denotes the first k objects (including x itself) that are closest to object x for attribute a in universe U, and k is a given positive integer.

The k-nearest-neighbor class is also called the k-nearest-neighbor information granule. The size of k-nearest-neighbor information granules is controlled by parameter k. All the k-nearest-neighbor information granules in $S = (U, C, D)$ form a cover on U. $x \in top_B^k(x)$. When $k = 1, top_B^k(x) = \{x\}$, the size of information granules is the smallest. When $k = |U|, top_B^k(x) = U$, the size of information granules is the largest. The k-nearest-neighbor relation R_B^k is

$$R_B^k = \{(x,y)|y \in top_B^k(x), x \in U, y \in U\}. \tag{2}$$

Therefore, relation matrix R_B^k is not symmetric.

Property 1. Given a decision information system $S = (U, C, D), B_1 \subseteq C$ and $B_2 \subseteq C, k$ is a positive integer. R_B^k is a k-nearest-neighbor relation, we have

$$R_{B_1 \cup B_2}^k = R_{B_1}^k \cap R_{B_2}^k.$$

Proof 1. According to Definition 1, there are $top_{B_1 \cup B_2}^k(x) = \underset{a \in B_1 \cup B_2}{\cap} top_a^k(x) = top_{B_1}^k(x) \cap top_{B_2}^k(x)$, so $R_{B_1 \cup B_2}^k = R_{B_1}^k \cap R_{B_2}^k$.

According to Property 1, we can independently calculate the k-nearest-neighbor relation of each attribute on U, then the k-nearest-neighbor relation of any attribute subset can be obtained by the intersection. This property is helpful to develop a heuristic search algorithm for attribute reduction, which can reduce repeated relation calculation. The previous δ-neighborhood relations usually do not satisfy the intersection operation between the relation for the attribute subset and the relations of corresponding single attribute. In attribute reduction based on δ-neighborhood relations, it is necessary to repeatedly calculate the distance of samples for many single attributes when calculating the relation of the attribute subset in each loop. This will lead to a lot of repeated calculations and is also time-consuming. However, in the calculation of the k-nearest-neighbor relation, the relations of all single attributes need to be calculated only one time.

Definition 2. Let $S = (U, C, D)$ be a decision information system. $\forall X \subseteq U$, the upper and lower approximations of X for B in S are

$$\begin{aligned} \overline{R}_B^k(X) &= \{x|top_B^k(x) \cap X \neq \varnothing, x \in U\}; \\ R_B^k(X) &= \{x|top_B^k(x) \subseteq X, x \in U\}. \end{aligned} \tag{3}$$

If $\overline{R}_B^k(X) = R_B^k(X)$, then X w.r.t. k-nearest-neighbor relation R_B^k is accurate, otherwise X w.r.t. R_B^k is rough. $R_B^k(X) \subseteq X \subseteq \overline{R}_B^k(X)$. The boundary region of X w.r.t. B in S is

$$BN_B^k(X) = \overline{R}_B^k(X) - R_B^k(X). \tag{4}$$

The size of $BN_B^k(X)$ reflects the roughness of X w.r.t. R_B^k. The larger the size of $BN_B^k(X)$ is, the rougher the X w.r.t. R_B^k is.

Definition 3. Let $S = (U, C, D)$ be a decision information system and $U/D = \{D_1, D_2, \cdots, D_r\}$. $\forall B \subseteq C$ and a given positive integer k, the upper and lower approximations of D w.r.t. B are

$$\begin{aligned} \overline{R}_B^k(D) &= \bigcup_{i=1}^{r} \overline{R}_B^k(D_i); \\ R_B^k(D) &= \bigcup_{i=1}^{r} R_B^k(D_i). \end{aligned} \tag{5}$$

Based on $\overline{R}_B^k(D)$ and $R_B^k(D)$, the decision boundary region and decision positive region of D w.r.t. B are

$$\begin{aligned} BN_B^k(D) &= \overline{R}_B^k(D) - R_B^k(D); \\ POS_B^k(D) &= cup_{i=1}^{r} R_B^k(D_i). \end{aligned} \tag{6}$$

$|BN_B^k(D)|$ reflects the roughness of decision D w.r.t. R_B^k. The larger the size of $BN_B^k(D)$ is, the rougher the D w.r.t. R_B^k is. $|POS_B^k(D)|$ reflects the consistency of decision D w.r.t. R_B^k.

Property 2. Let $B \subseteq C$ and $U/D = \{D_1, D_2, \cdots, D_r\}$, we have

(1) $\overline{R}_B^k(D) = U$;
(2) $POS_B^k(D) \cap BN_B^k(D) = \varnothing$;
(3) $POS_B^k(D) \cup BN_B^k(D) = \overline{R}_B^k(D)$.

Proof 2. (1) There are $D_i \subseteq \overline{R}_B^k(D_i)$ and $\bigcup_{i=1}^{r} D_i = U$. From Definition 3, we have $U \subseteq \overline{R}_B^k(D)$ and $\overline{R}_B^k(D) \subseteq U$. So $\overline{R}_B^k(D) = U$.

(2) By $BN_B^k(D) = \overline{R}_B^k(D) - R_B^k(D)$ and $POS_B^k(D) = \bigcup_{i=1}^{r} R_B^k(D_i) = R_B^k(D)$, there is $POS_B^k(D) \cap BN_B^k(D) = \varnothing$.

(3) According to $POS_B^k(D) \cap BN_B^k(D) = \varnothing$ and $BN_B^k(D) = \overline{R}_B^k(D) - POS_B^k(D)$, so $POS_B^k(D) \cup BN_B^k(D) = \overline{R}_B^k(D)$.

Definition 4. Let $S = (U, C, D)$ be a decision information system. $\forall B \subseteq C$ and a given positive integer k, the dependency degree of D w.r.t. B is defined as

$$\gamma_B^k(D) = \frac{|POS_B^k(D)|}{|U|}. \tag{7}$$

$\gamma_B^k(D)$ reflects the ability of B to approximate D. As $POS_B^k(D) \subseteq U$, we have $0 \leqslant \gamma_B^k(D) \leqslant 1$. The larger the $\gamma_B^k(D)$ is, the stronger the approximation ability of B w.r.t. D is. If $\gamma_B^k(D) = 1$, then B w.r.t. D is consistent for a given k; otherwise, B w.r.t. D is partially consistent.

To understand the calculation process of k-nearest-neighbor rough sets and its difference from NRS, we calculate the dependency degrees of $B_1 = \{a_1, a_2\}$ and $B_2 = \{a_3, a_4\}$, using the information in Example 2.1. Given $k = 5$, we get $top_{a_1}^5(x_1) = \{x_1, x_2, x_4, x_5, x_7\}, top_{a_2}^5(x_1) = \{x_1, x_2, x_8, x_{10}, x_{11}\}, top_{a_3}^5(x_1) = \{x_1, x_2, x_3, x_7, x_8\}$ and $top_{a_4}^5(x_1) = \{x_1, x_2, x_7, x_9, x_{10}\}$, so $top_{B_1}^5(x_1) = top_{a_1}^5(x_1) \cap top_{a_2}^5(x_1) = \{x_1, x_2\}$ and $top_{B_2}^5(x_1) = top_{a_3}^5(x_1) \cap top_{a_4}^5(x_1) = \{x_1, x_2, x_7\}$. The k-nearest-neighbor information granules of all objects in universe for B_1 and B_2 are shown in Table 2. The k-nearest-neighbor relations for B_1 and B_2 are

$$R_{B_1}^5 = \begin{pmatrix} 1&1&0&0&0&0&0&0&0&0&0&0\\ 1&1&0&1&1&0&0&0&0&0&0&0\\ 0&1&1&1&1&0&0&0&0&0&0&0\\ 1&1&0&1&1&0&0&0&0&0&0&0\\ 0&1&1&1&1&0&0&0&0&0&0&0\\ 0&0&0&1&0&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&1&0&1&0&0&0\\ 0&0&0&0&0&0&0&1&0&1&1&0\\ 0&0&0&0&0&0&1&1&1&0&1&0\\ 0&0&0&0&0&0&0&1&0&1&1&0\\ 0&0&0&0&0&0&0&1&0&1&1&0\\ 0&0&0&0&0&0&0&1&1&0&1&1 \end{pmatrix},$$

and

$$R_{B_2}^5 = \begin{pmatrix} 1&1&0&0&0&0&1&0&0&0&0&0\\ 0&1&0&0&0&0&1&0&0&0&0&0\\ 0&0&1&1&0&0&0&1&0&0&0&0\\ 0&0&1&1&0&0&0&1&0&0&0&0\\ 0&0&0&0&1&0&0&0&1&0&0&0\\ 0&0&0&0&0&1&0&0&0&0&1&0\\ 1&1&0&0&0&0&1&0&0&0&0&0\\ 0&0&1&1&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&0&1&1&0&0\\ 0&0&0&0&0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&0&0&0&1&1 \end{pmatrix}.$$

Table 2
K-Nearest-Neighbor information granules for B_1 and B_2.

U	B_1	B_2
x_1	$\{x_1, x_2\}$	$\{x_1, x_2, x_7\}$
x_2	$\{x_1, x_2, x_4, x_5\}$	$\{x_2, x_7\}$
x_3	$\{x_2, x_3, x_4, x_5\}$	$\{x_3, x_4, x_8\}$
x_4	$\{x_1, x_2, x_4, x_5\}$	$\{x_3, x_4, x_8\}$
x_5	$\{x_2, x_3, x_4, x_5\}$	$\{x_5, x_9\}$
x_6	$\{x_4, x_6\}$	$\{x_6, x_{11}\}$
x_7	$\{x_7, x_9\}$	$\{x_1, x_2, x_7\}$
x_8	$\{x_8, x_{10}, x_{11}\}$	$\{x_3, x_4, x_8\}$
x_9	$\{x_7, x_8, x_9, x_{11}\}$	$\{x_9, x_{10}\}$
x_{10}	$\{x_8, x_{10}, x_{11}\}$	$\{x_{10}\}$
x_{11}	$\{x_8, x_{10}, x_{11}\}$	$\{x_{11}\}$
x_{12}	$\{x_8, x_9, x_{11}, x_{12}\}$	$\{x_{11}, x_{12}\}$

The relation matrices $R^5_{B_1}$ and $R^5_{B_2}$ are not symmetric. $U/D = \{D_1, D_2\}$, where $D_1 = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ and $D_2 = \{x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\}$. From Definition 2, we can get

$$\overline{R}^5_{B_1}(D_1) = \{x_1, x_2, x_3, x_4, x_5, x_6\},$$
$$\overline{R}^5_{B_1}(D_2) = \{x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\},$$
$$R^5_{B_1}(D_1) = \{x_1, x_2, x_3, x_4, x_5, x_6\},$$
$$R^5_{B_1}(D_2) = \{x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\},$$

and

$$\overline{R}^5_{B_2}(D_1) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\},$$
$$\overline{R}^5_{B_2}(D_2) = U,$$
$$R^5_{B_2}(D_1) = \varnothing,$$
$$R^5_{B_2}(D_2) = \{x_9, x_{10}, x_{11}, x_{12}\}.$$

Therefore, we have $POS^5_{B_1}(D) = U$ and $POS^5_{B_2}(D) = \{x_9, x_{10}, x_{11}, x_{12}\}$. According to the Definition 4, we get $\gamma^5_{B_1}(D) = 1$ and $\gamma^5_{B_2}(D) = 0.3333$. From the view of the dependency degree of k-nearest-neighbor rough sets, the approximation ability of B_1 is better than that of B_2. From Fig. 2, we find that the separability of B_1 is also better than that of B_2. From the above discussion, we know that the dependency degree of k-nearest-neighbor rough sets is better than that of neighborhood rough sets to evaluate the approximation ability and separability of attribute subsets.

Property 3. (Type-I monotonicity). Let $S = (U, C, D)$ be a decision information system. $\forall B_1 \subseteq B_2 \subseteq C$ and a given positive integer k, we have

(1) $R^k_{B_1} \supseteq R^k_{B_2}$;
(2) $\forall X \subseteq U, \overline{R}^k_{B_1}(X) \supseteq \overline{R}^k_{B_2}(X),\ R^k_{B_1}(X) \subseteq R^k_{B_2}(X)$;
(3) $POS^k_{B_1}(D) \subseteq POS^k_{B_2}(D), \gamma^k_{B_1}(D) \leqslant \gamma^k_{B_2}(D)$.

Proof 3. (1) According to $B_2 = B_1 \cup (B_2 - B_1)$ and Property 1, we have $R^k_{B_2} = R^k_{B_1} \cap R^k_{B_2 - B_1}$, so $R^k_{B_1} \supseteq R^k_{B_2}$.

(2) From $B_1 \subseteq B_2$ and Definition 1, there are $top^k_{B_2}(x) \subseteq top^k_{B_1}(x), \forall x \in U$. If $x \in \overline{R}^k_{B_2}(X)$, we have $top^k_{B_2}(x) \cap X \neq \varnothing$, then $top^k_{B_1}(x) \cap X \neq \varnothing$, we obtain $x \in \overline{R}^k_{B_1}(X)$ by Definition 2, so $\overline{R}^k_{B_1}(X) \supseteq \overline{R}^k_{B_2}(X)$; if $x \in R^k_{B_1}(X)$, we have $top^k_{B_1}(x) \subseteq X$, then $top^k_{B_2}(x) \subseteq X$, we know $x \in R^k_{B_2}(X)$, so $R^k_{B_1}(X) \subseteq R^k_{B_2}(X)$.

(3) According to (2), $\forall D_i \in U/D$, we have $R^k_{B_1}(D_i) \subseteq R^k_{B_2}(D_i)$, so $POS^k_{B_1}(D) = \bigcup\limits_{D_i} R^k_{B_1}(D_i) \subseteq \bigcup\limits_{D_i} R^k_{B_2}(D_i) = POS^k_{B_2}(D)$, namely $POS^k_{B_1}(D) \subseteq POS^k_{B_2}(D)$; then from Definition 4 we have $\gamma^k_{B_1}(D) \leqslant \gamma^k_{B_2}(D)$.

From Property 3, we find that with the gradual increase of the number of attributes, the dependency function is monotonic and nondecreasing. The purpose of attribute reduction is to find a minimum attribute subset with the same characterizing ability as the original attribute set. The monotonicity of dependency function can be used to construct a greedy search algorithm. With the gradual increase of the attributes that cause the greatest change in dependency, we are committed to quickly finding a minimum attribute subset that has the same or almost the same approximation ability as the original conditional attribute set.

Property 4. (Type-II monotonicity). Let $S = (U, C, D)$ be a decision information system. $\forall B \subseteq C$ and two given positive integers k_1 and $k_2, k_1 \leqslant k_2$, we have

(1) $R^{k_1}_B \subseteq R^{k_2}_B$;
(2) $\forall X \subseteq U, \overline{R}^{k_1}_B(X) \subseteq \overline{R}^{k_2}_B(X)\ R^{k_1}_B(X) \supseteq R^{k_2}_B(X)$;
(3) $POS^{k_1}_B(D) \supseteq POS^{k_2}_B(D), \gamma^{k_1}_B(D) \geqslant \gamma^{k_2}_B(D)$.

Proof 4. (1) As $k_1 \leqslant k_2, \forall a \in B$, there is $top^{k_1}_a \subseteq top^{k_2}_a$, then $top^{k_1}_B \subseteq top^{k_2}_B$, so $R^{k_1}_B \subseteq R^{k_2}_B$.

(2) $\forall x \in U$, by $k_1 \leqslant k_2$, there is $top^{k_1}_B(x) \subseteq top^{k_2}_B(x)$. If $x \in \overline{R}^{k_1}_B(X)$, we can get $top^{k_1}_B(x) \cap X \neq \varnothing$, which implies conclusion $top^{k_2}_B(x) \cap X \neq \varnothing$, namely $x \in \overline{R}^{k_2}_B(X)$, so $\overline{R}^{k_1}_B(X) \subseteq \overline{R}^{k_2}_B(X)$. If $x \in R^{k_2}_B(X)$, we get $top^{k_2}_B(x) \subseteq X$, further deduce the conclusion $top^{k_1}_B(x) \subseteq X$, namely $x \in R^{k_1}_B(X)$, so $R^{k_1}_B(X) \supseteq R^{k_2}_B(X)$.

(3) According to (2), $\forall D_i \in U/D$, we have $R_B^{k_1}(D_i) \supseteq R_B^{k_2}(D_i)$. From Definitions 3 and 4, there are $POS_B^{k_1}(D) \supseteq POS_B^{k_2}(D)$ and $\gamma_{B_1}^k(D) \geqslant \gamma_{B_2}^k(D)$.

Property 4 shows that the dependency degree is closely related to the size of information granules. However, for a given attribute set, the size of information granules is controlled by parameter k. By conclusions (1) and (3) of Property 4, we know that with the increase of k, the size of information granules does not decrease, while the dependency degree does not increase. That is to say, the smaller the k is, the stronger the approximation ability of attribute subsets is; otherwise, the weaker the approximation ability is. But if k is too small, it will lead to the phenomenon of early convergence. The parameter k of k-nearest-neighbor rough sets is very important for attribute reduction. Later, we will discuss how to set it.

From Properties 3 and 4, it can be seen that the ability of an attribute set to approximate decisions depends not only on the attribute set of characterizing objects of universe, but also on the size of parameter k of k-nearest-neighbor rough sets. Both attribute sets and parameter k can control the granularity of information granules.

4. Attribute reduction based on overlap degree and K-nearest-neighbor rough sets

The computation of k-nearest-neighbor information granules is very time-consuming, and the efficiency of heuristic search strategy is also very low. To solve the problem of computational efficiency, we will improve the efficiency of reduction from the search strategy point of view. We will define some measures to pre-evaluate degree of importance of the single attribute, and pre-sort attributes based on the degree of importance. Starting from the attribute with the lowest importance, we use dependency degree to judge whether the attribute can be removed one by one.

The purposes of attribute reduction are to retain the attributes with high separability and strong approximation ability, and to remove the trivial attributes. If the coincidence degree (CD) of the raw data from different categories is high, and CD of the reduced data from different categories is low, then the degree of importance of the retained attributes is high. If the distance (DIS) of the raw data from different categories is long and the DIS of reduced data from different categories is short, then the degree of importance of the retained attributes is high. From Fig. 3, we can see that the yellow area is the CD and the red line is the DIS of objects from different categories. In the process of search reducts, we need to select attributes which can decrease the CD and increase the DIS.

Definition 5. Let $S = (U, C, D)$ be a decision information system, $a \in C, x \in U, D_i, D_j \in U/D$, the coincidence degree (CD) of objects between D_i and D_j for a is defined as

$$CD_a(D_i, D_j) = \frac{|\left[m_{D_i}^a, M_{D_i}^a\right] \cap \left[m_{D_j}^a, M_{D_j}^a\right]|}{|\left[m_{D_i}^a, M_{D_i}^a\right] \cup \left[m_{D_j}^a, M_{D_j}^a\right]|}, \tag{8}$$

where $m_{D_i}^a = \min_{x \in D_i} f(x, a)$ and $M_{D_i}^a = \max_{x \in D_i} f(x, a)$. When $|\left[m_{D_i}^a, M_{D_i}^a\right] \cup \left[m_{D_j}^a, M_{D_j}^a\right]| = 0$, we set $CD_a(D_i, D_j) = \frac{|\left[m_{D_i}^a, M_{D_i}^a\right] \cap \left[m_{D_j}^a, M_{D_j}^a\right]| + eps}{|\left[m_{D_i}^a, M_{D_i}^a\right] \cup \left[m_{D_j}^a, M_{D_j}^a\right]| + eps}$, where eps is a very small positive number. $CD_a(D_i, D_j)$ represents the coincidence degree of objects between D_i and D_j for a. The smaller the $CD_a(D_i, D_j)$ is, the higher the coincidence degree of objects between D_i and D_j is. $0 \leqslant CD_a(D_i, D_j) \leqslant 1$.

Definition 6. Let $S = (U, C, D)$ be a decision information system, $a \in C, U/D = \{D_1, D_2, \cdots, D_r\}$, the coincidence degree of S for a is defined as

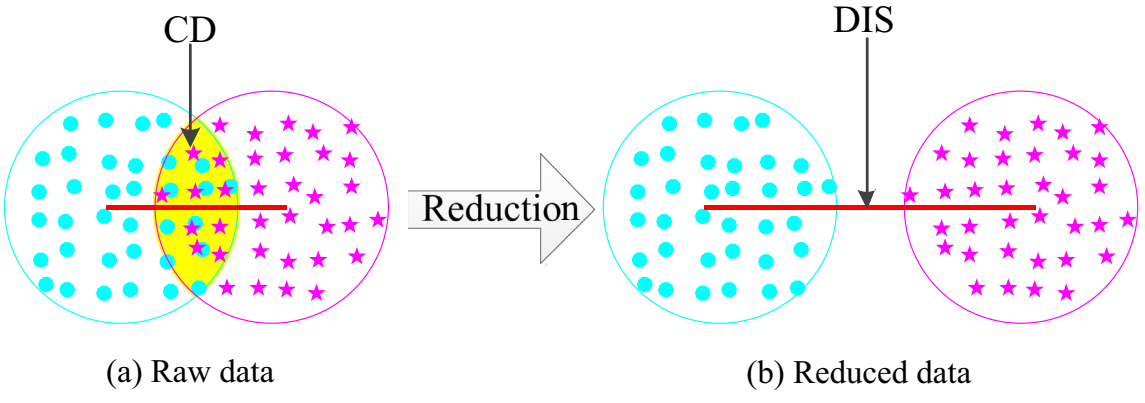


Fig. 3. Overlap degrees of objects from different categories on raw and reduced data.

$$CD_a(S) = \sum_{D_i \neq D_j} CD_a(D_i, D_j). \tag{9}$$

$CD_a(S)$ represents the coincidence degree of objects between different decisions. Besides coincidence degree can reflect the ability to distinguish objects with different decisions, the distance between objects can also measure the ability to distinguish objects with different decisions. Next, we will introduce the distance of objects with different decisions to evaluate the importance of attributes.

Definition 7. Let $S = (U, C, D)$ be a decision information system, $a \in C, D_i, D_j \in U/D$, the distance of objects between D_i and D_j for a is defined as

$$DIS_a(D_i, D_j) = \frac{|\overline{C}_a(D_i) - \overline{C}_a(D_j)|}{f^a_{max}(D_i, D_j) - f^a_{min}(D_i, D_j)}, \tag{10}$$

where $\overline{C}_a(D_i) = \frac{1}{|D_i|}\sum_{x \in D_i} f(x, a), f^a_{max}(D_i, D_j) = \max_{x \in D_i \cup D_j} f(x, a)$ and $f^a_{min}(D_i, D_j) = \min_{x \in D_i \cup D_j} f(x, a)$. If $f^a_{max}(D_i, D_j) = f^a_{min}(D_i, D_j)$, we set $DIS_a(D_i, D_j) = \frac{|\overline{C}_a(D_i) - \overline{C}_a(D_j)|}{f^a_{max}(D_i, D_j) - f^a_{min}(D_i, D_j) + eps}$. $0 \leqslant DIS_a(D_i, D_j) \leqslant 1$. The larger the $DIS_a(D_i, D_j)$ is, the stronger the ability of a to distinguish objects with different decisions is.

Definition 8. Let $S = (U, C, D)$ be a decision information system and $U/D = \{D_1, D_2, \cdots, D_r\}$. $\forall a \in C, D_i, D_j \in U/D$, the distance ($DIS$) of objects in S for a is defined as

$$DIS_a(S) = \sum_{D_i \neq D_j} DIS_a(D_i, D_j), \tag{11}$$

where $DIS_a(D_i, D_j)$ is the distance of objects between D_i and D_j for a. Obviously, $0 \leqslant DIS_a(D_i, D_j) \leqslant \frac{r(r-1)}{2}$. The larger the $DIS_a(S)$ is, the higher the ability of a to distinguish objects is.

Definition 9. Let $S = (U, C, D)$ be a decision information system, $a \in C$, the overlap degree (OD) of objects in S for a is defined as

$$OD_a(S) = \frac{CD_a(S)}{DIS_a(S)}. \tag{12}$$

$OD_a(S)$ is used to evaluate the quality of a in terms of separability. The smaller the overlap degree $OD_a(S)$ is, the higher the separability of a is, and the more important a is.

Further considering Example 2.1, we obtain the CD, DIS and OD of each attribute as shown in Table 3. In Example 2.1, from the view of the dependency degree of k-nearest-neighbor rough sets, we know that the approximation ability of $\{a_1, a_2\}$ is better than that of $\{a_3, a_4\}$. From Table 3, we find that the CD of S for a_1 and a_2 is much less than that of a_3 and a_4 and the DIS of S for a_1 and a_2 is greater than that of a_3 and a_4. Furthermore, the OD of S for a_1 and a_2 is far lower than that of a_3 and a_4. We use OD to pre-sort the attributes, then use the sorted attributes to accelerate the speed of attribute reduction based on k-nearest-neighbor rough sets. We sort the four attributes in descending order by OD to get $SORT(C) = (a_4, a_3, a_2, a_1)$. We use the dependency degree to remove redundant attributes in the sorted attributes one by one. Because $\gamma^5_{\{a_3,a_2,a_1\}}(D) = \gamma^5_{\{a_4,a_3,a_2,a_1\}}(D), a_4$ is removed. $\gamma^5_{\{a_2,a_1\}}(D) = \gamma^5_{\{a_3,a_2,a_1\}}(D)$, so a_3 is also removed. Because $\gamma^5_{\{a_1\}}(D) < \gamma^5_{\{a_2,a_1\}}(D), a_2$ is retained. $\gamma^5_{\{a_2\}}(D) < \gamma^5_{\{a_2,a_1\}}(D)$, so a_1 is also retained.

Based on the above discussion, we can quickly find a reduct by OD and k-nearest-neighbor rough sets, and the reduct has a low OD while keeping the dependency degree unchanged. In other words, the reduct has high separability and strong ability to approximate decisions. We develop a reduction algorithm by using OD and k-nearest-neighbor rough set theory.

Definition 10. Let $S = (U, C, D)$ be a decision information system, $B \subseteq C, a \in B$, the significance of attribute a relative to B and D is defined as

Table 3
CD, DIS and OD of each attribute in Example 2.1.

C	CD_a	DIS_a	OD_a
a_1	0.0909	0.5909	0.1538
a_2	0.0909	0.5758	0.1579
a_3	0.7250	0.2062	3.5152
a_4	0.8765	0.0185	47.3333

$$SIG(a,B,D) = \gamma_B^k(D) - \gamma_{B-\{a\}}^k(D). \tag{13}$$

By the Type-I monotonicity, there is $SIG(a,B,D) \geqslant 0$. When $SIG(a,B,D) = 0$, we call a an unnecessary attribute with respect to B and D. If $SIG(a,B,D) > 0$, then a is a necessary attribute. We develop a reduction algorithm based on OD and k-nearest-neighbor rough sets (OD&KNN), which is shown in Algorithm 1.

Algorithm 1. Attribute reduction based on OD and k-nearest-neighbor rough sets (OD&KNN)

Input: A decision information system $S = (U, C, D)$ and a parameter k.

Output: A reduct red.

1: **for** each $a \in C$
2: Compute $OD_a(S)$ by formula (12);
3: **end for**
4: All attributes are sorted in descending order by $OD_a(S)$, and the result is marked as $SORT(C)$;
5: Initialize: $red \longleftarrow SORT(C)$;/* where attributes in $SORT(C)$ are ordered */
6: Compute $\gamma_{red}^k(D)$ by formula (7);
7: **for** each a in $SORT(C)$ **do**
8: $SIG(a, red, D) = \gamma_{red}^k(D) - \gamma_{red-a}^k(D)$;
9: **if** $SIG(a, red, D) = 0$ **then**
10: $red \longleftarrow red - \{a\}$;/* remove the unnecessary attribute a */
11: **end if**
12: **end for**
13: **return** red.

In Algorithm 1, steps 1–3 compute the overlap degree (OD) of objects in S for each attribute of C, and the time complexity is $O(|U| \times |C|)$. In steps 4–5, the descending sorted attribute set $SORT(C)$ is obtained based on the overlap degree with the time complexity $O(|C| \times log|C|)$ and red is initialized to $SORT(C)$ with the time complexity $O(1)$. Step 6 calculates the dependency degree of D with respect to red namely $\gamma_{red}^k(D)$ with the time complexity $O\left(|U|^2 \times |U/D|\right)$. Steps 7–12 remove the unnecessary attributes sequentially, and the time complexity is $O\left(|U|^2 \times |U/D| \times |C|\right)$. Step 13 is to return a reduct, and the time complexity is $O(1)$. The time complexity of Algorithm 1 is $O\left(|U|^2 \times |U/D| \times |C|\right)$.

For high-dimensional data, we can quickly reduce the dimension of data by using half-division searching strategy. The strategy is presented in Algorithm 2. Algorithm 2 is a subalgorithm of Algorithm 1 for dealing with high-dimensional data. When the dimension of the data is very high, we use Algorithm 2 after the step 4 of Algorithm 1 to quickly reduce the dimension of data.

Algorithm 2. Half-division searching to reduce the dimension of high-dimensional data

Input: $S = (U, C, D)$, $SORT(C)$, k and dim.

Output: New sorted attributes $SORT(C)$.

1: Compute $\gamma_{SORT(C)}^k(D)$ by formula (7);
2: **While** the size of $SORT(C)$ is greater that dim **do**
3: Take the last half of the attributes in $SORT(C)$ as $Half$;
4: Compute $\gamma_{Half}^k(D)$ by formula (7);
5: **if** $\gamma_{Half}^k(D) == \gamma_{SORT(C)}^k(D)$ **then**
6: $SORT(C) \longleftarrow Half$;
7: **else**
8: **break**;
9: **end if**
10: **end while**
11: **return** $SORT(C)$.

5. Experimental analysis

We will verify the effectiveness of the proposed *OD&KNN* on a number of classification datasets. We evaluate the performance of *OD&KNN* by comparing it with four other advanced neighborhood reduction algorithms: k-nearest neighborhood rough sets (NNRS) [37], neighborhood rough sets (NRS) [19], variable precision k-nearest neighbor rough sets (FarVPKNN) [18] and attribute group (AG) [4]. The evaluation metrics of comparative analysis are the running time, the number of selected attributes and classification accuracies of the reduced data.

5.1. Experimental setup

We download eleven datasets from UCI Machine Learning Repository [10] (Nos. 1–8) and ELVIRA Biomedical Dataset Repository [3] (Nos. 9–11). The information of these datasets is outlined in Table 4, and all conditional attributes are normalized to $[0, 1]$ by using the Max–Min normalization. Two classifiers KNN ($k = 3$) and RBF-SVM ($\sigma = 10$ and $C = 10$) are used to evaluate classification performance of reduced data. All reduction experiments adopt 5-fold cross validation. The average result of the 5 results is regarded as final result. All programs are executed in MATLAB 2015B and run in the hardware environment with Inter(R) Core(TM) i7-4790 CPU @ 3.60 GHz 3.60 GHz, with 16 GB RAM.

NNRS is an attribute reduction algorithm based on the dependency of attributes with respect to decisions by using the k-nearest neighborhood information granules, which are induced by unit neighborhoods and k-nearest neighbors simultaneously. NNRS has a neighborhood parameter k and a termination threshold θ that need to be set. NRS is an attribute reduction algorithm by using δ-neighborhood granules to approximate decisions. It has a neighborhood parameter δ that needs to be set. FarVPKNN uses inclusion degrees of k-nearest neighbor information granules and decisions to evaluate the importance of attribute subsets. It has a neighborhood parameter k and a precision parameter β that need to be set. To reduce the number of attribute evaluations and the search range of calculating neighborhood information granules, AG combines the bucket and attribute group to perform attribute selection. AG has a neighborhood parameter δ and the number of groups that need to be set.

Researchers set neighborhood parameters k and δ mainly through experiments to search and observe the performance of reduction algorithms for different parameters. According to the research results of predecessors [18], the value of k is generally set as $0.25N$ and the value of δ is set as 0.25. k is set to the value that makes reduction algorithms achieve approximate optimal performance on most datasets. For example, in the literature [37], the authors search for the optimal value of parameter k from $0.01N$ to $0.1N$ with step $0.01N$, such that the reduction algorithm can select an attribute subset that has the highest classification ability for the reduced data, where N is the number of objects. By observation, they set $k = 0.05N$. In the literature [18], the authors observe the performance of the reduction algorithm for different neighborhood parameters of k (from $0.1N$ to $0.5N$ with step $0.05N$), and find that the performance of the algorithm is better on most datasets when $k = 0.25N$. In the literature [37], the k-nearest neighborhood of an object with respect to an attribute subset is computed for all the attributes of the attribute subset. In the literature [18], the k-nearest neighbor of an object about an attribute subset is to first calculate the k-nearest neighbor of each attribute in the attribute subset, then take the intersection of the k-nearest neighbors of each attribute in the attribute subset. The k-nearest neighbor of this paper is similar to the k-nearest neighbor of reference [18].

In our experiments, we set the parameters of each algorithm as follows: In our *OD&KNN* algorithm, we set $k = 0.25N$ based on previous research results [18]. At the same time, in order to verify the reliability of the results [18], we search k from $0.05N$ to $0.5N$ with step $0.05N$, and the detailed search results of k on all datasets are shown in Figs. 4–14. It should be pointed out that for high-dimensional data (more that 1000), in the above experiments we set $dim = 100$ to reduce the dimension of data in advance. As can be seen from Figs. 4–14, KNN and SVM reach the relatively high accuracy in most cases when $k = 0.25N$. For NNRS, $k = 0.05N$ and termination threshold $\theta = 0.01$ [37]. If $k = 0.25N$, most of the neighborhood information granules cannot be included in any decision class. In such case, the approximate ability of the attributes with respect to decisions cannot be properly described. Therefore, we still set $k = 0.05N$ according to the research results of ref-

Table 4
Description of datasets.

No.	Name	Objects	Attributes	Classes
1	Seeds	210	8	3
2	Wine	178	14	3
3	Australian	690	15	2
4	Pop_failures	540	19	2
5	Segment	2310	20	7
6	Wdbc	569	31	2
7	Wpbc	198	34	2
8	Sonar	208	61	2
9	Leukemia-ALLAML	72	7130	2
10	DLBCL-Harvard	77	7130	2
11	Lung-Cancer-Harvard2	181	12534	2

erence [37]. For FarVPKNN, $k = 0.25N$ and precision $\beta = 0.8$ [18]. For NRS, neighborhood parameter $\delta = 0.25$ [19]. For AG, neighborhood parameter $\delta = 0.25$, for low dimensional data and high-dimensional data (more than 1000), the number of groups is set to $\lceil |AT|/3 \rceil$ and 50, respectively [4].

5.2. Experimental comparison

The running time of five reduction algorithms is presented in Table 5. Out of the 11 cases, NNRS, NRS, FarVPKNN, AG and *OD&KNN* achieve the shortest running time in 0, 0, 1, 5 and 5, respectively. When running time of AG is the least, it is mainly concentrated on low dimensional data, and AG performs the worst on three high-dimensional data. However, on the three high-dimensional data, the running time of *OD&KNN* is always the least. The running time of *OD&KNN* is always less than that of NNRS on all datasets. The running time of *OD&KNN* is $20-70\%$ lower than that of NNRS. The running time of *OD&KNN* is less than that of NRS on ten datasets (except Seeds). The running time of *OD&KNN* is less than that of FarVPKNN on nine datasets (except Australian and Pop_failures). To sum up, our reduction algorithm *OD&KNN* is feasible and efficient in computation time from the experimental results.

The above time comparison intuitively reflects the feasibility of the proposed method *OD&KNN* in the calculation of reduction. Next, we analyze the complexity of the above five algorithms from the theoretical level. Detailed comparison results are shown in Table 6. From Table 6, we can see that the time complexity of NRS and FarVPKNN is the same. It should be pointed out that in the real data, generally speaking, the number of attributes (features) is greater than or far greater than the number of decision classes (categories). In this case, we rank the five algorithms according to the complexity as *OD&KNN* $<$ NNRS $<$ NRS = FarVPKNN <= AG. Moreover, the complexity of *OD&KNN* is about $|U/D|/|C|$ of that of NNRS and is about $1/|C|$ of those of NRS, FarVPKNN and AG.

At the same time, we compare the absolute and relative number of selected attributes which are obtained by reduction algorithms and the separability of these attributes. The absolute and relative numbers of these attributes are shown in Table 7. From Table 7, we know that the absolute average number of selected attributes by *OD&KNN* (6.09) is far less than that of raw data (2453.18). Therefore, the reduction performance of *OD&KNN* is effective and feasible. Comparing the absolute and relative average numbers of attributes retained by the five reduction algorithms on all datasets, it can be seen that *OD&KNN* (6.09, 28.45%) removes more features than NNRS (7.45, 38.29%), NRS (10.18, 47.03%) and AG (12.31, 51.84%), second only to FarVPKNN (4.04, 19.47%).

Next, we analyze the performance of two classifiers on the data subsets corresponding to the selected attributes to further verify the advantages of the reduction algorithms in attribute selection. The detailed classification accuracy results of the two classifiers are shown in Tables 8,9. The numbers before and after $\pm$ are the average classification accuracy and standard

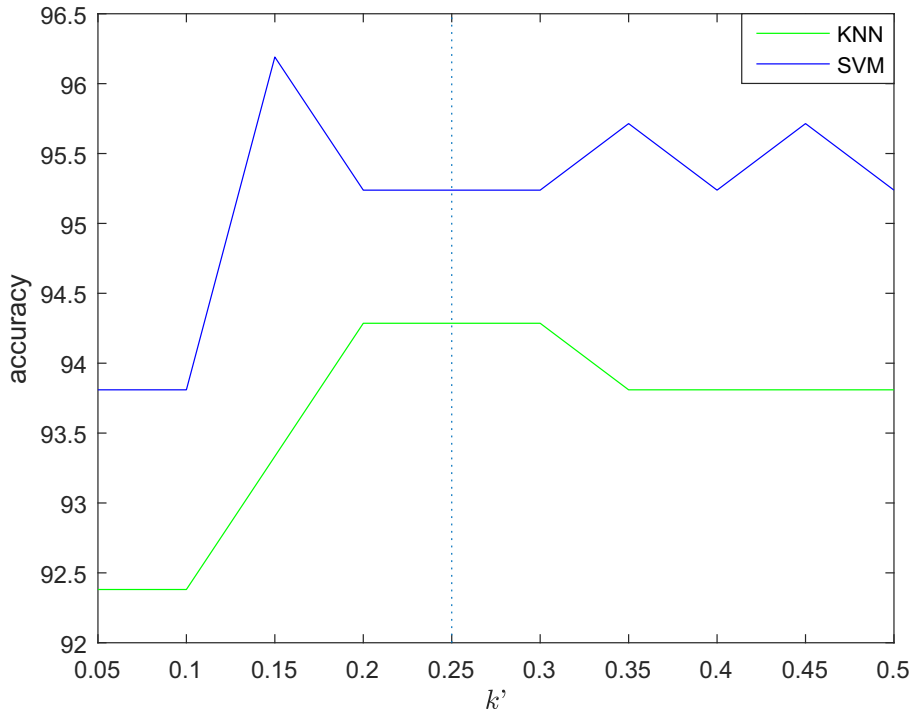


Fig. 4. Classification accuracies of Seeds for different k ($k = k'N$).

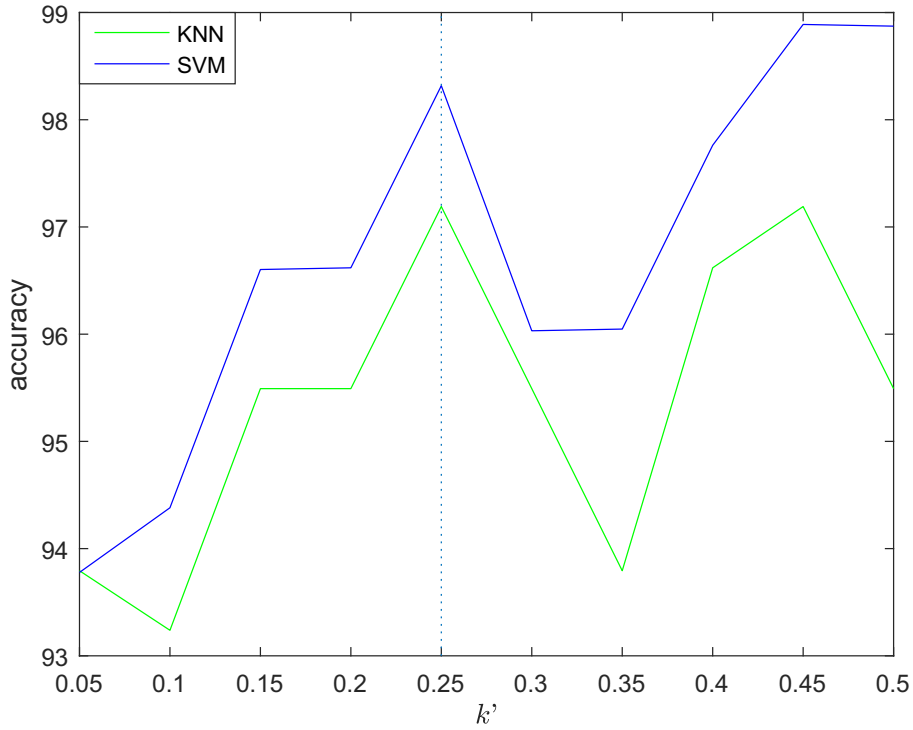


Fig. 5. Classification accuracies of Wine for different k ($k = k\prime N$).

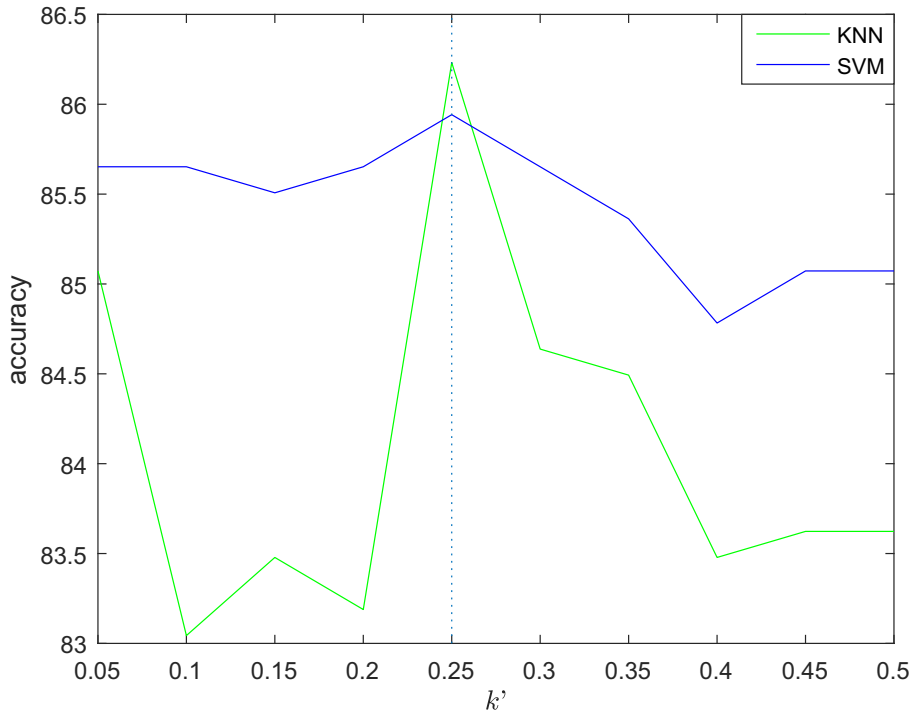


Fig. 6. Classification accuracies of Australian for different k ($k = k\prime N$).

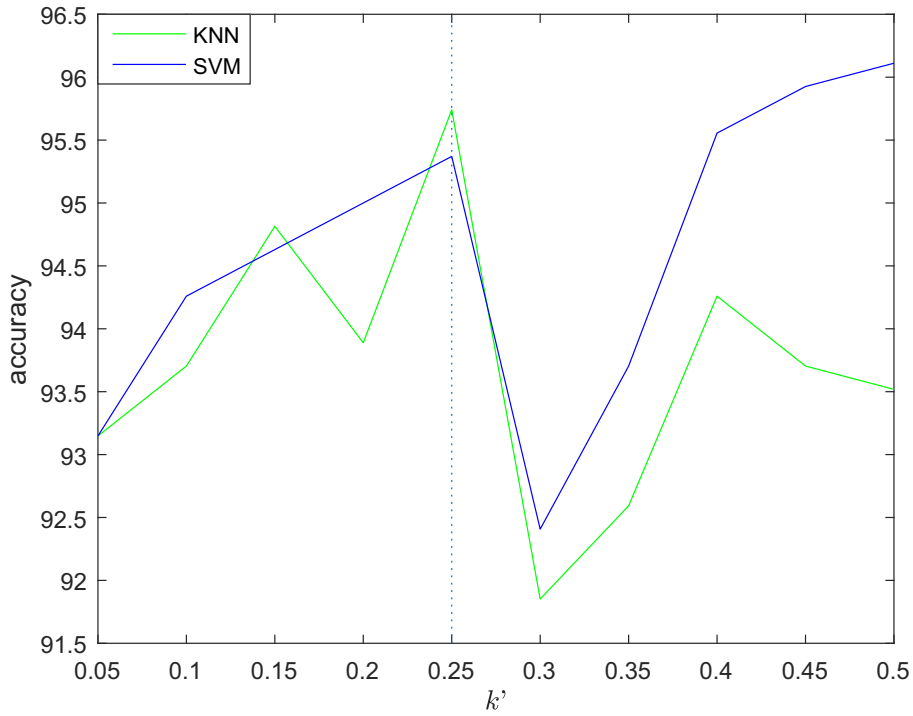


Fig. 7. Classification accuracies of Pop_failures for different k ($k = k\prime N$).

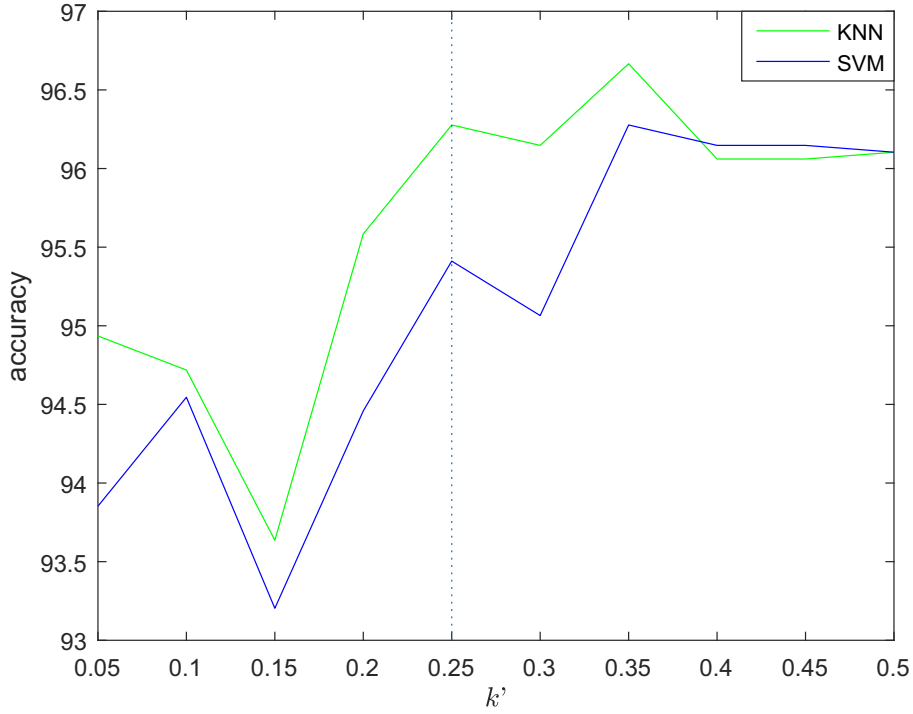


Fig. 8. Classification accuracies of Segment for different k ($k = k\prime N$).

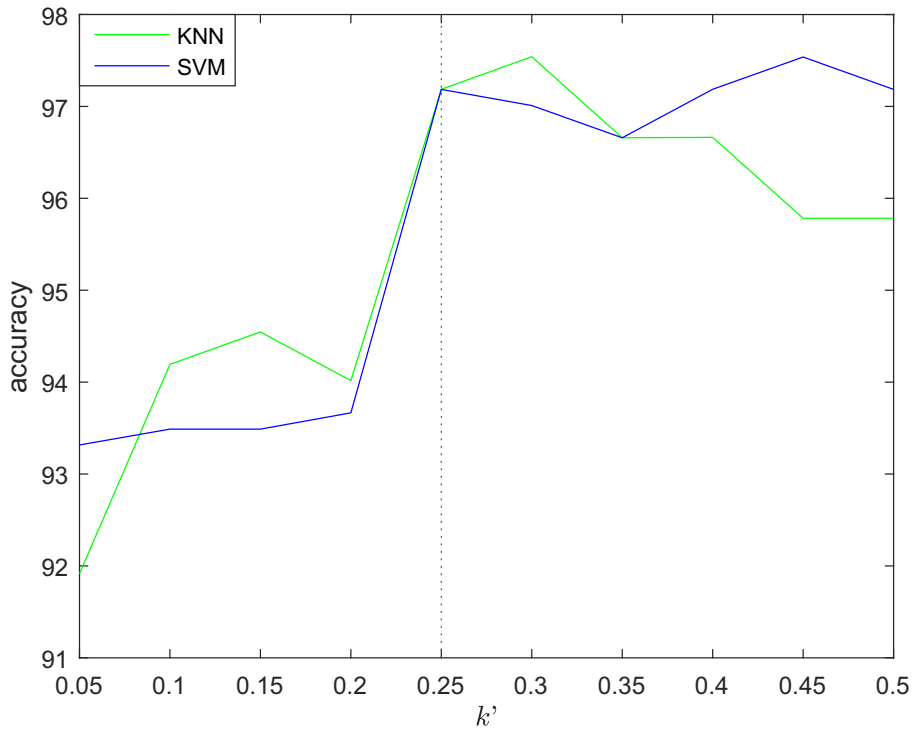


Fig. 9. Classification accuracies of Wdbc for different k ($k = k\prime N$).

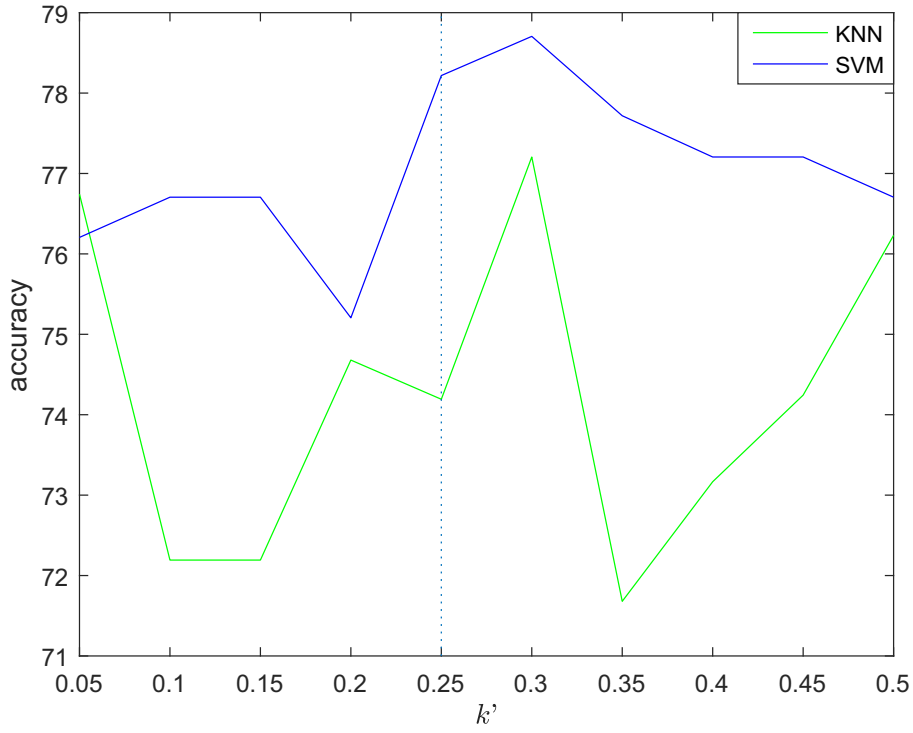


Fig. 10. Classification accuracies of Wpbc for different k ($k = k\prime N$).

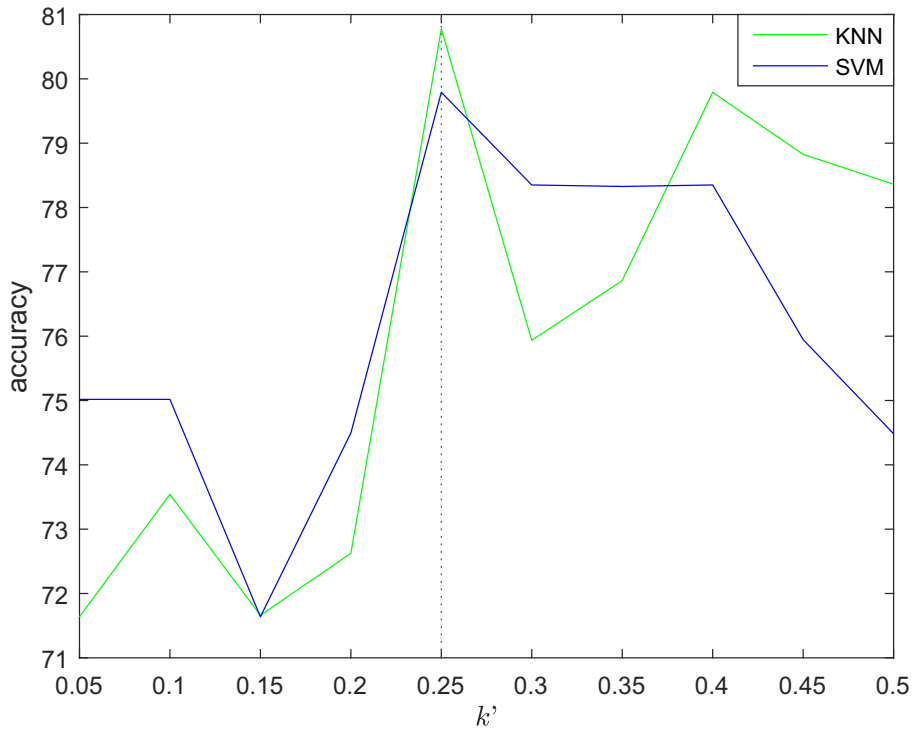


Fig. 11. Classification accuracies of Sonar for different k ($k = k\prime N$).

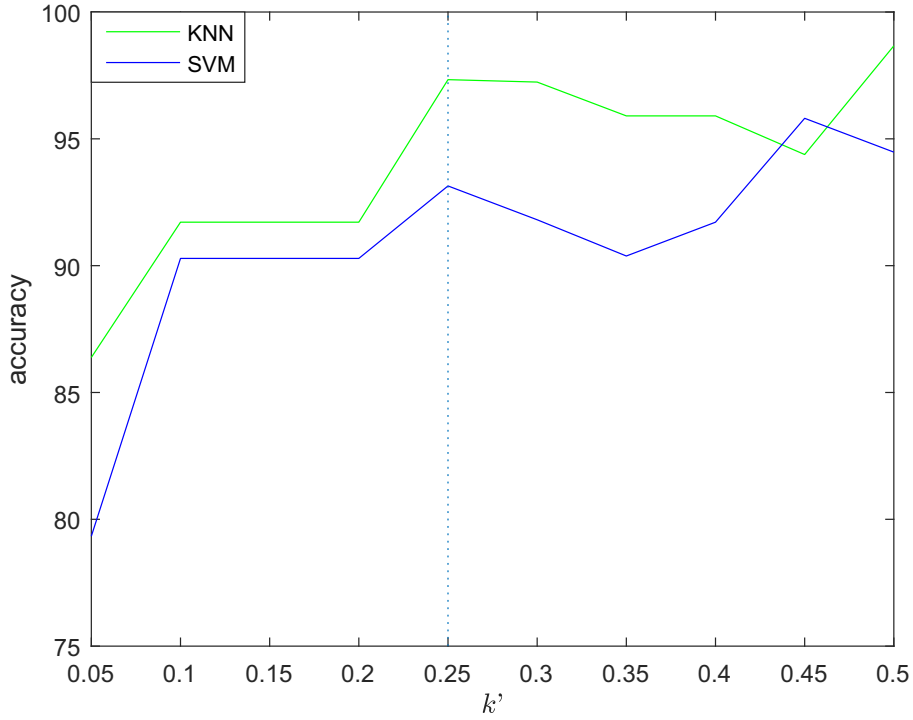


Fig. 12. Classification accuracies of Leukemia-ALLAML for different k ($k = k\prime N$).

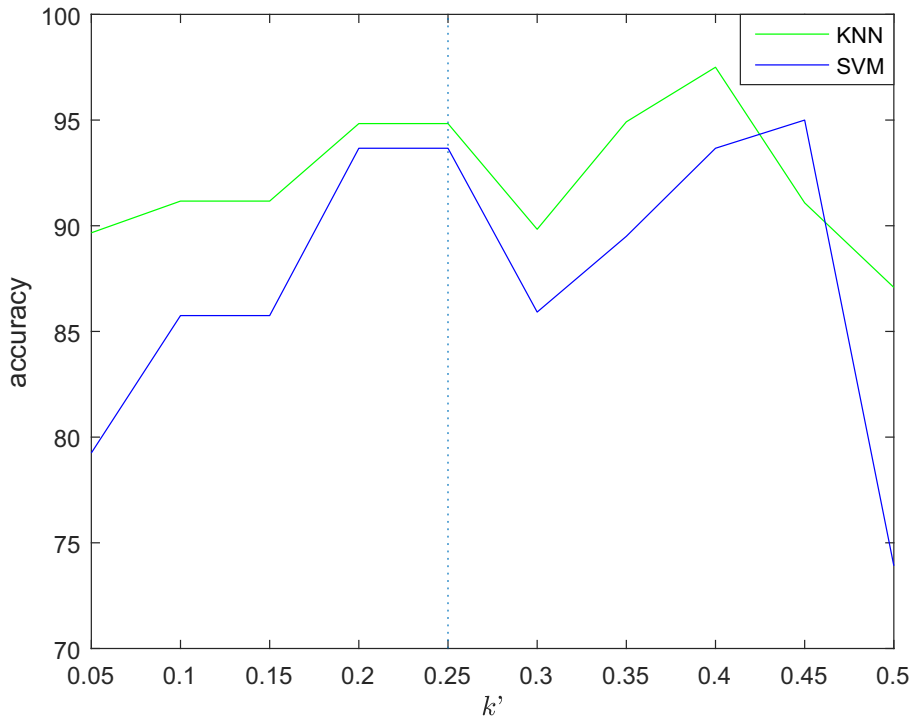


Fig. 13. Classification accuracies of DLBCL-Harvard for different k ($k = k\prime N$).

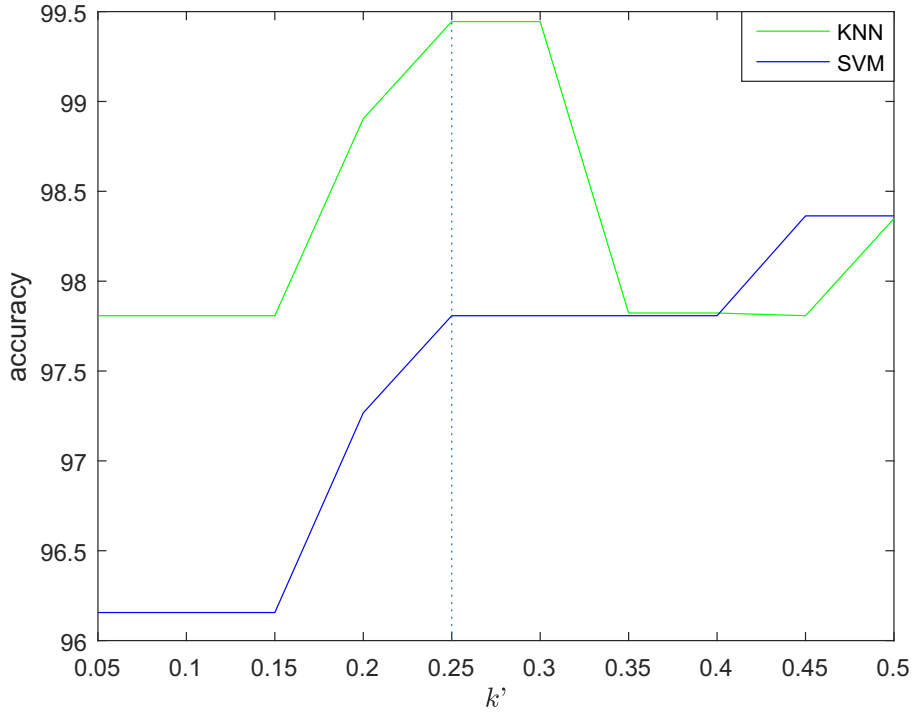


Fig. 14. Classification accuracies of Lung-Cancer-Harvard2 for different k ($k = k\prime N$).

Table 5
Running time of different reduction algorithms (s).

Datasets	NNRS	NRS	FarVPKNN	AG	*OD&KNN*
Seeds	0.04460	0.01478	0.03162	0.01302	0.02012
Wine	0.06778	0.03340	0.04592	0.01324	0.02360
Australian	1.33760	0.97776	0.38186	0.45372	0.45076
Pop_failures	0.47104	0.57770	0.29124	0.13676	0.33520
Segment	15.68770	39.98910	40.03250	15.97070	12.37550
Wdbc	1.03920	4.25570	1.68310	0.91508	0.77436
Wpbc	0.33938	0.25098	0.20656	0.04850	0.09636
Sonar	0.63118	0.37780	0.35120	0.04214	0.27950
Leukemia-ALLAML	5.25990	3.10260	3.53090	6.78900	1.80160
DLBCL-Harvard	6.68570	2.73880	4.26850	6.35440	1.99550
Lung-Cancer-Harvard2	32.79460	22.46360	19.53760	58.37850	15.94250

Table 6
Time complexity of different reduction algorithms.

Algorithms	Stage complexity	Total complexity
NNRS [37]	step 1: $O\left(\lvert U\rvert^2 \times \lvert C\rvert\right)$ step 2: $O(1)$ steps 3–22: $O(\lvert U\rvert \times \lvert C\rvert \times (\lvert U\rvert + log\lvert U\rvert))$ step 23: $O(\lvert C\rvert)$ step 24: $O(1)$ step 25: $O\left(\lvert U\rvert \times \lvert C\rvert^2 \times (\lvert U\rvert + log\lvert U\rvert)\right)$	$O\left(\lvert U\rvert^2 \times \lvert C\rvert^2\right)$
NRS [19]	step 1: $O(1)$ steps 2–5: $O\left(\lvert U\rvert^2 \times \lvert C\rvert \times \lvert U/D\rvert\right)$ step 6: $O(\lvert C\rvert)$ steps 7–12: $O\left(\lvert U\rvert^2 \times \lvert C\rvert^2 \times \lvert U/D\rvert\right)$	$O\left(\lvert U\rvert^2 \times \lvert C\rvert^2 \times \lvert U/D\rvert\right)$
FarVPKNN [18]	steps 1–2: $O\left(\lvert U\rvert^2 \times \lvert C\rvert\right)$ step 3: $O(1)$ step 4: $O\left(\lvert U\rvert^2 \times \lvert C\rvert \times \lvert U/D\rvert\right)$ step 5: $O(\lvert C\rvert)$ step 6: $O\left(\lvert U\rvert^2 \times \lvert C\rvert^2 \times \lvert U/D\rvert\right)$	$O\left(\lvert U\rvert^2 \times \lvert C\rvert^2 \times \lvert U/D\rvert\right)$
AG [4]	step 1: $O(1)$ step 2: $O\left(\lvert U\rvert \times \lvert C\rvert^2 \times L\right)$ step 3: $O\left(\lvert U\rvert^2 \times \lvert C\rvert^2 \times \lvert U/D\rvert\right)$ step 4: $O\left(\lvert U\rvert^2 \times \lvert C\rvert \times \lvert U/D\rvert\right)$ step 5: $O(1)$	$O\left(\lvert U\rvert \times \lvert C\rvert^2 \times max\{\lvert U\rvert \times \lvert U/D\rvert, L\}\right)$
OD&KNN	steps 1–3: $O(\lvert U\rvert \times \lvert C\rvert)$ steps 4–5: $O(\lvert C\rvert \times log\lvert C\rvert)$ step 6: $O\left(\lvert U\rvert^2 \times \lvert U/D\rvert\right)$ steps 7–12: $O\left(\lvert U\rvert^2 \times \lvert C\rvert \times \lvert U/D\rvert\right)$ step 13: $O(1)$	$O\left(\lvert U\rvert^2 \times \lvert C\rvert \times \lvert U/D\rvert\right)$

* $|U|$ is the number of objects. $|C|$ is the number of conditional attributes. $|U/D|$ is the number of decision classes. L is the maximum number of iterations of k-means clustering.

deviation of the 5-fold cross validation, respectively. In the average rows, the number after $\pm$ is standard deviation of average classification accuracy of each dataset. From Tables 8,9, we know that out of the 22 cases, NNRS, NRS, FarVPKNN, AG and *OD&KNN* achieve the highest classification accuracy in 2, 1, 1, 1 and 17 cases, respectively. Under two classifiers and when compared with the performance of raw data, *OD&KNN* performs better than raw data 18 times; NNRS and NRS perform better than raw data 12 times; FarVPKNN performs better than raw data 9 times; and AG performs better than raw data 8 times. Among the five reduction algorithms, we find that *OD&KNN* ranks the first, NRS ranks the second, NNRS ranks the third, AG ranks the fourth, FarVPKNN ranks the last according to the average classification results.

Meanwhile, we use statistical test methods to show the differences among reduction algorithms. First, the commonly used Friedman test [13] is chosen to verify the existence of significant differences. Let N be the number of data sets, k be

Table 7
Number of selected attributes for different reduction algorithms.

Datasets	RAW	The absolute number of attributes retained					The relative number of attributes retained (%)				
		NNRS	NRS	FarVPKNN	AG	*OD&KNN*	NNRS	NRS	FarVPKNN	AG	*OD&KNN*
Seeds	7	6	7	5	7	5	85.71	100.00	71.43	100.00	71.43
Wine	13	7	7	4	8.2	5	53.85	53.85	30.77	63.08	38.46
Australian	14	14	14	1	14	8	100.00	100.00	7.14	100.00	57.14
Pop_failures	18	8	7	1	8.4	6	44.44	38.89	5.56	46.67	33.33
Segment	19	12	18	10	19	11	63.16	94.74	52.63	100.00	57.89
Wdbc	30	10	23	6	28	7	33.33	76.67	20.00	93.33	23.33
Wpbc	33	9	12	6	15.8	7	27.27	36.36	18.18	47.88	21.21
Sonar	60	8	10	5	11.4	6	13.33	16.67	8.33	19.00	10.00
Leukemia-ALLAML	7129	2	5	2.4	7.2	3.4	0.03	0.07	0.03	0.10	0.05
DLBCL-Harvard	7129	3	4	3	8.4	4.6	0.04	0.06	0.04	0.12	0.06
Lung-Cancer-Harvard2	12533	3	5	1	8	4	0.02	0.04	0.01	0.06	0.03
Average	2453.18	7.45	10.18	4.04	12.31	6.09	38.29	47.03	19.47	51.84	28.45

Table 8
Classification accuracies of reduced data by reduction algorithms under KNN.

Datasets	RAW	NNRS	NRS	FarVPKNN	*AG*	*OD&KNN*
Seeds	91.90 ± 5.22	91.43 ± 4.94	91.90 ± 5.22	91.90 ± 5.22	91.90 ± 5.22	$\mathbf{94.29 \pm 2.71}$
Wine	93.81 ± 3.15	94.95 ± 2.32	96.08 ± 1.49	91.62 ± 5.20	94.37 ± 6.03	$\mathbf{97.19 \pm 2.82}$
Australian	83.04 ± 3.72	83.04 ± 3.72	83.04 ± 3.72	61.45 ± 12.53	83.04 ± 3.72	$\mathbf{86.23 \pm 0.89}$
Pop_failures	91.30 ± 2.97	92.04 ± 2.75	90.93 ± 2.21	88.52 ± 1.68	91.48 ± 2.31	$\mathbf{95.74 \pm 1.40}$
Segment	95.84 ± 0.83	95.89 ± 0.82	95.84 ± 0.83	95.93 ± 1.08	95.84 ± 0.83	$\mathbf{96.28 \pm 0.47}$
Wdbc	95.78 ± 1.81	94.20 ± 2.70	96.13 ± 1.49	92.09 ± 2.73	95.78 ± 1.81	$\mathbf{97.19 \pm 0.74}$
Wpbc	73.74 ± 12.49	72.23 ± 3.87	74.22 ± 10.42	$\mathbf{74.79 \pm 4.40}$	70.19 ± 7.79	74.19 ± 5.76
Sonar	79.35 ± 5.64	$\mathbf{82.23 \pm 6.83}$	77.89 ± 1.95	76.93 ± 1.97	75.45 ± 7.67	80.78 ± 6.04
Leukemia-ALLAML	80.57 ± 11.53	95.81 ± 6.34	91.43 ± 12.78	86.19 ± 10.79	80.67 ± 8.86	$\mathbf{97.33 \pm 3.65}$
DLBCL-Harvard	80.42 ± 15.73	88.42 ± 6.94	92.25 ± 2.66	87 ± 10.03	80.58 ± 7.51	$\mathbf{94.83 \pm 5.28}$
Lung-Cancer-Harvard2	92.27 ± 3.03	98.35 ± 1.51	98.89 ± 1.52	74.01 ± 4.36	91.68 ± 5.22	$\mathbf{99.44 \pm 1.24}$
Average	87.09 ± 7.79	89.87 ± 7.83	89.87 ± 8.02	83.68 ± 10.51	86.45 ± 8.86	$\mathbf{92.14 \pm 8.12}$

Table 9
Classification accuracies of reduced data by reduction algorithms under SVM.

Datasets	RAW	NNRS	NRS	FarVPKNN	*AG*	*OD&KNN*
Seeds	92.38 ± 1.99	92.86 ± 2.92	92.38 ± 1.99	92.86 ± 2.92	92.38 ± 1.99	$\mathbf{95.24 \pm 2.38}$
Wine	96.62 ± 3.70	94.37 ± 2.84	97.19 ± 2.82	90.44 ± 4.28	94.92 ± 3.68	$\mathbf{98.32 \pm 2.49}$
Australian	84.35 ± 2.64	84.35 ± 2.64	84.35 ± 2.64	85.51 ± 2.71	84.35 ± 2.64	$\mathbf{85.94 \pm 2.69}$
Pop_failures	95.37 ± 1.46	94.44 ± 1.73	92.41 ± 0.77	91.48 ± 2.57	91.48 ± 2.73	$\mathbf{95.37 \pm 2.27}$
Segment	95.80 ± 0.79	$\mathbf{95.93 \pm 0.60}$	95.80 ± 0.79	94.07 ± 0.50	95.80 ± 0.79	95.41 ± 0.77
Wdbc	97.54 ± 1.58	96.66 ± 0.75	96.66 ± 2.09	95.96 ± 1.33	$\mathbf{97.54 \pm 1.81}$	97.19 ± 1.91
Wpbc	74.23 ± 10.41	74.22 ± 5.62	76.72 ± 7.43	77.83 ± 6.32	77.22 ± 10.53	$\mathbf{78.22 \pm 8.33}$
Sonar	81.79 ± 6.56	76.41 ± 4.80	73.09 ± 3.01	72.16 ± 6.73	71.65 ± 9.93	$\mathbf{79.79 \pm 3.74}$
Leukemia-ALLAML	65.52 ± 13.15	88.86 ± 10.83	92.95 ± 7.15	88.95 ± 6.19	73.71 ± 13.12	$\mathbf{93.14 \pm 4.72}$
DLBCL-Harvard	75.25 ± 12.94	76.67 ± 8.58	91.00 ± 5.69	83.33 ± 16.23	76.67 ± 5.41	$\mathbf{93.67 \pm 10.83}$
Lung-Cancer-Harvard2	82.90 ± 5.90	96.68 ± 3.05	$\mathbf{98.89 \pm 1.52}$	82.87 ± 5.36	92.79 ± 6.41	97.81 ± 2.27
Average	85.61 ± 10.83	88.31 ± 8.86	90.13 ± 8.51	86.86 ± 7.33	86.23 ± 9.74	$\mathbf{91.83 \pm 7.17}$

the number of reduction algorithms and R_i be the average rank of the i^{th} algorithm on all data sets. F follows a Fisher distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom. The Friedman statistic is defined as $\chi_F^2 = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} R_i^2 - \frac{k(k+1)^2}{4}\right)$ and $F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$. We rank the reduction algorithms under the two classifiers, and the detailed ranking results are shown in Table 10. By the definition of the Friedman statistic, we know $F = 9.8207$ under KNN and $F = 5.6436$ under SVM. These two values of F are all greater than the critical value $F(k-1,(k-1)(N-1)) = F(4,40) = 2.0909$ at $\alpha = 0.1$, so we know that the five algorithms are significantly different under two classifiers. Since the average classification accuracy of the 5-fold cross validation is used for ranking, we use the Friedman statistic to further test whether there are significant differences among the 5-fold cross validation. The classification accuracies and ranks of *OD&KNN* under KNN are shown in Table 11. From Table 11, we obtain that $F = 0.5080 < F(4,40)$ of *OD&KNN* under KNN, i.e., there are no significant differences between the 5-fold cross validation. In the same way, we obtain that there are no significant differences between the 5-fold cross validation of each reduction algorithm under KNN and SVM. The average performance of each reduction algorithm can be

Table 10
Ranks of reduction algorithms under KNN and SVM.

Datasets	KNN					SVM				
	NNRS	NRS	FarVPKNN	*AG*	*OD&KNN*	NNRS	NRS	FarVPKNN	*AG*	*OD&KNN*
Seeds	5	3	3	3	1	2.5	4.5	2.5	4.5	1
Wine	3	2	5	4	1	4	2	5	3	1
Australian	3	3	5	3	1	4	4	2	4	1
Pop_failures	2	4	5	3	1	2	3	4	5	1
Segment	3	4.5	2	4.5	1	1	2.5	5	2.5	4
Wdbc	4	2	5	3	1	4	3	5	1	2
Wpbc	4	2	1	5	3	5	4	2	3	1
Sonar	1	3	4	5	2	2	3	4	5	1
Leukemia-ALLAML	2	3	4	5	1	4	2	3	5	1
DLBCL-Harvard	3	2	4	5	1	4	2	3	5	1
Lung-Cancer-Harvard2	3	2	5	4	1	3	1	5	4	2
Average	3	2.7727	3.9091	4.0455	1.2727	3.2273	2.8182	3.6818	3.8182	1.4545

Table 11
Classification accuracies and ranks of *OD&KNN* under KNN.

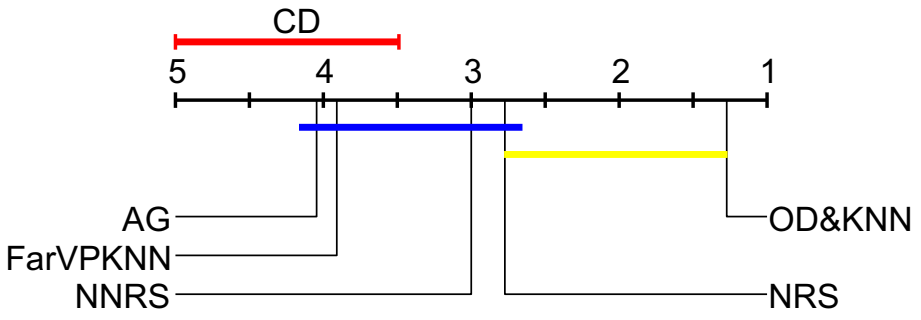
Datasets	1^{st} time	2^{nd} time	3^{rd} time	4^{th} time	5^{th} time
Seeds	92.86(4)	95.24(2.5)	95.24(2.5)	90.48(5)	97.62(1)
Wine	94.29(5)	97.22(3)	94.44(4)	100(1.5)	100(1.5)
Australian	86.96(1.5)	84.78(5)	86.96(1.5)	86.23(3.5)	86.23(3.5)
Pop_failures	93.52(5)	95.37(4)	96.3(2.5)	97.22(1)	96.3(2.5)
Segment	96.75(1)	96.54(2.5)	95.67(5)	95.89(4)	96.54(2.5)
Wdbc	98.25(1)	97.37(2.5)	96.49(4)	96.46(5)	97.37(2.5)
Wpbc	75(2.5)	75(2.5)	66.67(5)	82.5(1)	71.79(4)
Sonar	82.93(2)	88.1(1)	71.43(5)	80.49(4)	80.95(3)
Leukemia-ALLAML	100(2)	93.33(4.5)	93.33(4.5)	100(2)	100(2)
DLBCL-Harvard	93.33(3.5)	87.5(5)	100(1.5)	100(1.5)	93.33(3.5)
Lung-Cancer-Harvard2	100(2.5)	100(2.5)	100(2.5)	97.22(5)	100(2.5)
Average	92.17(2.7273)	91.86(3.1818)	90.59(3.4545)	93.31(3.0455)	92.74(2.5909)

regarded as the final performance of the algorithm. So it is reasonable to rank the performance of the five reduction algorithms by using the average classification accuracy of the 5-fold cross validation.

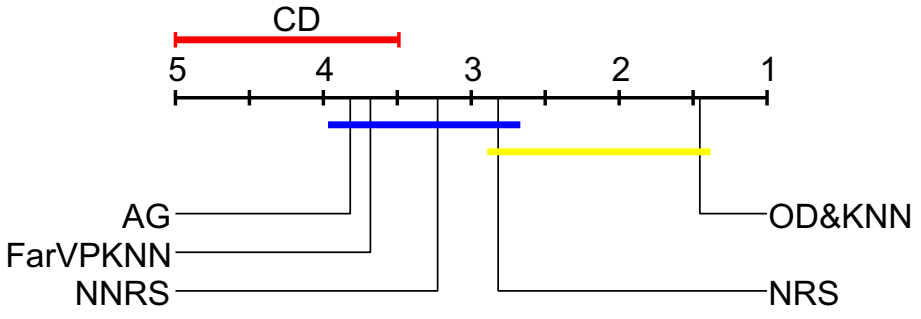
Next, we test the statistical differences between the five reduction algorithms by using Bonferroni-Dunn test [9]. When $k = 5, N = 11, \alpha = 0.1$ and $q_{0.1} = 2.241$, we obtain the critical distance $CD_\alpha = q_\alpha\sqrt{\frac{k(k+1)}{6N}} = 1.5109$. For any two algorithms, if their distance of average ranks exceeds CD_α, then the performance of the two algorithms is significantly different. We use the Bonferroni-Dunn test graph [9] to display intuitively the statistical differences among four algorithms. In Fig. 15, we draw the critical difference ($CD = 1.5109$) by using red line segment, then connect the reduction algorithms whose distances of average ranks are less than 1.5109 at $\alpha = 0.1$. From Fig. 15, for NNRS, NRS, FarVPKNN and AG under KNN and SVM, we did not test significant differences by using Bonferroni-Dunn test. The performance of *OD&KNN* is significantly different from that of NNRS, FarVPKNN and AG under KNN and SVM. We do not know whether there are significant differences between NRS and *OD&KNN* under KNN and SVM.

From [2], we know that the outcome of the comparison between any two reduction algorithms depends also on the performance of the other algorithms by using the *post hoc* test based on mean-ranks. Therefore, we use the Wilcoxon signed-rank test [40] to pairwise compare the differences of *OD&KNN* with respect to NNRS, NRS, FarVPKNN and AG. The information of the Wilcoxon signed-rank test between *OD&KNN* and four other algorithms under KNN and SVM are shown in Tables 12,13, respectively. In the following, we describe in details the comparison between NNRS and *OD&KNN* under KNN. From Table 12 (NNRS vs. *OD&KNN*), we obtain that the sum of ranks with positive sign $W_+ = 3$ and the sum of ranks with negative sign $W_- = 7 + 6 + 9 + 10 + 1 + 8 + 5 + 4 + 11 + 2 = 63$. From [40], we obtain that the critical value of the Wilcoxon signed-rank test $W_{n,\alpha} = 11$ at $n = 11$ and $\alpha = 0.05$. The test statistic $W = min\{W_+, W_-\} = 3 < W_{n,\alpha}$, and the average rank of NNRS is greater than that of *OD&KNN* under KNN. That is, the performance of *OD&KNN* is significantly better than that of NNRS under KNN. In the same way, we get that the performance of *OD&KNN* is significantly better than that of NNRS, NRS, FarVPKNN and AG under both KNN and SVM.

By analyzing the classification accuracy and the number of selected attributes simultaneously, we find that FarVPKNN selects the least number of attributes, but its classification accuracy is lower than that of raw data in most cases. Moreover, among the five reduction algorithms, its classification performance is the worst. This shows that FarVPKNN may lose some necessary attributes for classification tasks in the reduction process. NNRS and NRS select a relatively large number of attributes when compared with *OD&KNN*. The classification performances of NNRS and NRS are comparable, and their classifi-



(a) Comparison of all reduction algorithms under KNN



(b) Comparison of all reduction algorithms under SVM

Fig. 15. Average ranks of four reduction algorithms, there are no significant differences between the algorithms that are connected at $\alpha = 0.1$.

Table 12
Wilcoxon signed-rank test between *OD&KNN* and four other algorithms under KNN.

Datasets	NNRS vs. *OD&KNN*					NRS vs. *OD&KNN*				
	NNRS	*OD&KNN*	sign	abs1	rank	NRS	*OD&KNN*	sign	abs1	rank
Seeds	91.43	94.29	–	2.86	7	91.90	94.29	–	2.39	6
Wine	94.95	97.19	–	2.24	6	96.08	97.19	–	1.11	5
Australian	83.04	86.23	–	3.19	9	83.04	86.23	–	3.19	9
Pop_failures	92.04	95.74	–	3.70	10	90.93	95.74	–	4.81	10
Segment	95.89	96.28	–	0.39	1	95.84	96.28	–	0.44	2
Wdbc	94.2	97.19	–	2.99	8	96.13	97.19	–	1.06	4
Wpbc	72.23	74.19	–	1.96	5	74.22	74.19	+	0.03	1
Sonar	82.23	80.78	+	1.45	3	77.89	80.78	–	2.89	8
Leukemia-ALLAML	95.81	97.33	–	1.52	4	91.43	97.33	–	5.90	11
DLBCL-Harvard	88.42	94.83	–	6.41	11	92.25	94.83	–	2.58	7
Lung-Cancer-Harvard2	98.35	99.44	–	1.09	2	98.89	99.44	–	0.55	3
Datasets	FarVPKNN vs. *OD&KNN*					AG vs. *OD&KNN*				
	FarVPKNN	*OD&KNN*	sign	abs1	rank	AG	*OD&KNN*	sign	abs1	rank
Seeds	91.9	94.29	–	2.39	3	91.9	94.29	–	2.39	3
Wine	91.62	97.19	–	5.57	6	94.37	97.19	–	2.82	4
Australian	61.45	86.23	–	24.78	10	83.04	86.23	–	3.19	5
Pop_failures	88.52	95.74	–	7.22	7	91.48	95.74	–	4.26	7
Segment	95.93	96.28	–	0.35	1	95.84	96.28	–	0.44	1
Wdbc	92.09	97.19	–	5.10	5	95.78	97.19	–	1.41	2
Wpbc	74.79	74.19	+	0.60	2	70.19	74.19	–	4.00	6
Sonar	76.93	80.78	–	3.85	4	75.45	80.78	–	5.33	8
Leukemia-ALLAML	86.19	97.33	–	11.14	9	80.67	97.33	–	16.66	11
DLBCL-Harvard	87	94.83	–	7.83	8	80.58	94.83	–	14.25	10
Lung-Cancer-Harvard2	74.01	99.44	–	25.43	11	91.68	99.44	–	7.76	9

[1] abs represents the absolute values of differences.

Table 13
Wilcoxon signed-rank test between *OD&KNN* and four other algorithms under SVM.

Datasets	NNRS vs. *OD&KNN*					NRS vs. *OD&KNN*				
	NNRS	*OD&KNN*	sign	abs1	rank	NRS	*OD&KNN*	sign	abs1	rank
Seeds	92.86	95.24	–	2.38	6	92.38	95.24	–	2.86	9
Wine	94.37	98.32	–	3.95	8	97.19	98.32	–	1.13	5
Australian	84.35	85.94	–	1.59	5	84.35	85.94	–	1.59	7
Pop_failures	94.44	95.37	–	0.93	3	92.41	95.37	–	2.96	10
Segment	95.93	95.41	+	0.52	1	95.8	95.41	+	0.39	2
Wdbc	96.66	97.19	–	0.53	2	96.66	97.19	–	0.53	3
Wpbc	74.22	78.22	–	4.00	9	76.72	78.22	–	1.50	6
Sonar	76.41	79.79	–	3.38	7	73.09	79.79	–	6.70	11
Leukemia-ALLAML	88.86	93.14	–	4.28	10	92.95	93.14	–	0.19	1
DLBCL-Harvard	76.67	93.67	–	17.00	11	91	93.67	–	2.67	8
Lung-Cancer-Harvard2	96.68	97.81	–	1.13	4	98.89	97.81	+	1.08	4
Datasets	FarVPKNN vs. *OD&KNN*					AG vs. *OD&KNN*				
	FarVPKNN	*OD&KNN*	sign	abs1	rank	AG	*OD&KNN*	sign	abs1	rank
Seeds	92.86	95.24	–	2.38	5	92.86	95.24	–	2.38	5
Wine	90.44	98.32	–	7.88	9	90.44	98.32	–	7.88	9
Australian	85.51	85.94	–	0.43	2	85.51	85.94	–	0.43	2
Pop_failures	91.48	95.37	–	3.89	6	91.48	95.37	–	3.89	6
Segment	94.07	95.41	–	1.34	4	94.07	95.41	–	1.34	4
Wdbc	95.96	97.19	–	1.23	3	95.96	97.19	–	1.23	3
Wpbc	77.83	78.22	–	0.39	1	77.83	78.22	–	0.39	1
Sonar	72.16	79.79	–	7.63	8	72.16	79.79	–	7.63	8
Leukemia-ALLAML	88.95	93.14	–	4.19	7	88.95	93.14	–	4.19	7
DLBCL-Harvard	83.33	93.67	–	10.34	10	83.33	93.67	–	10.34	10
Lung-Cancer-Harvard2	82.87	97.81	–	14.94	11	82.87	97.81	–	14.94	11

[1] abs represents the absolute values of differences.

cation performances are slightly better than that of raw data. But the performances of NNRS and NRS are worse than that of *OD&KNN* in most cases. This shows that the selected attributes by NNRS and NRS are still redundant. AG selects the largest number of attributes, and its classification performance is worse than that of *OD&KNN* in most cases. That is to say, the selected attributes by AG have more redundant attributes. From the perspective of the size and quality of the selected attribute subset, we know that the proposed *OD&KNN* is more reasonable and superior to NNRS, NRS, FarVPKNN and AG.

From Tables 5–9, Fig. 15 and the corresponding analysis, we know that the proposed *OD&KNN* can achieve better classification performance by quickly capturing fewer attributes with high separability and strong approximation ability.

6. Conclusions

To improve the learning efficiency and the performance of classification tasks, we need to use attribute reduction to remove redundant and inconsistent attributes from raw data.

In this work, we introduce overlap degree (*OD*) of attributes into *k*-nearest-neighbor rough sets to enhance the computational efficiency and classification performance of reduced data. First, *OD* of attributes is applied to sort the attributes. Then we use *k*-nearest-neighbor rough sets to remove redundant attribute sequentially from the sorted attributes. There are two advantages of the proposed *OD&KNN* when compared with existing reduction algorithms. One is that the efficiency of *OD&KNN* is higher than that of other attribute reduction methods based on heuristic search by eliminating the repeated calculation of selecting relatively important attributes in the process of looping search. Compared with several existing attribute reduction algorithms, *OD&KNN* has higher computational efficiency in terms of the dimensionality of the data. This computational advantage is useful, since attribute reduction is most useful in high-dimensional datasets. The other is that the reduct obtained by *OD&KNN* not only keeps the dependency degree unchanged, but also selects the attributes with low overlap degree. Finally, we use public datasets to verify efficiency and feasibility by comparing the performance of *OD&KNN* with other reduction algorithms. The experimental results show that the performance of *OD&KNN* is better than the others in computational efficiency and classification accuracy.

CRediT authorship contribution statement

Meng Hu: Conceptualization, Methodology, Writing - original draft. **Eric C.C. Tsang:** Validation, Investigation, Supervision. **Yanting Guo:** 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.

Acknowledgment

This work is supported by the Macau Science and Technology Development Fund (No. 0019/2019/A1 and No. 0075/2019/A2), the National Natural Science Foundation of China (No. 62106148 and No. 61772002).

References

[1] A.F. Attia, M. Abd Elaziz, A.E. Hassanien, R.A. El-Sehiemy, Prediction of solar activity using hybrid artificial bee colony with neighborhood rough sets, IEEE Trans. Comput. Soc. Syst. (2020), https://doi.org/10.1109/TCSS.2020.3007769.
[2] A. Benavoli, G. Corani, F. Mangili, Should we really use post-hoc tests based on mean-ranks?, J. Mach. Learn. Res. 17 (1) (2017) 152–161.
[3] A. Cano, A. Masegosa, S. Moral, ELVIRA Biomedical Data Set Repository, 2005.http://leo.ugr.es/elvira/DBCRepository/..
[4] Y. Chen, K. Liu, J. Song, H. Fujita, X. Yang, Y. Qian, Attribute group for attribute reduction, Inf. Sci. 535 (Oct. 2020) 64–80.
[5] D. Chen, L. Zhang, S. Zhao, et al, A novel algorithm for finding reducts with fuzzy rough sets, IEEE Trans. Fuzzy Syst. 20 (2) (Apr. 2012) 385–389.
[6] D. Chen, Y. Yang, Z. Dong, An incremental algorithm for attribute reduction with variable precision rough sets, Appl. Soft Comput. 45 (2016) 129–149.
[7] H. Chen, T. Li, Y. Cai, et al, Parallel attribute reduction in dominance-based neighborhood rough set, Inf. Sci. 373 (2016) 351–368.
[8] H. Chen, T. Li, X. Fan, C. Luo, Feature selection for imbalanced data based on neighborhood rough sets, Inf. Sci. 483 (2019) 1–20.
[9] J. Demšar, Statistical comparison of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (Dec. 2006) 1–30.
[10] D. Dua, C. Graff, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2019, URL:http://archive.ics.uci.edu/ml.
[11] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. General Syst. 17 (2–3) (1990) 191–209.
[12] D. Dubois, H. Prade, Fuzzy sets in approximate reasoning, Part 1: Inference with possibility distributions, Fuzzy Sets Syst. 40 (1991) 143–202.
[13] M. Friedman, A comparison of alternative tests of significance for the problem of m ranking, Ann. Math. Stat. 11 (1) (1940) 86–92.
[14] S. Greco, B. Matarazzo, R. Slowinski, A new rough set approach to evaluation of bankruptcy risk, in: C. Zopounidis (Ed.), Operational Tools in the Management of Financial Risks, Kluwer, Dordrecht, The Netherlands, 1998, pp. 121–136.
[15] Y. Guo, E.C.C. Tsang, W. Xu, et al, Local logical disjunction double-quantitative rough sets, Inf. Sci. 500 (2019) 87–112.
[16] Y. Guo, E.C.C. Tsang, W. Xu, et al, Adaptive weighted generalized multi-granulation interval-valued decision-theoretic rough sets, Knowl.-Based Syst. 187 (Jan. 2020), https://doi.org/10.1016/j.knosys.2019.06.012.
[17] 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 (Feb. 2020), https://doi.org/10.1016/j.knosys.2019.105082.
[18] Q. Hu, J. Liu, D. Yu, Mixed feature selection based on granulation and approximation, Knowl.-Based Syst. 21 (4) (2008) 294–304.
[19] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inf. Sci. 178 (18) (2008) 3577–3594.
[20] Q. Hu, W. Pedrycz, D. Yu, et al, Selecting discrete and continuous features based on neighborhood decision error minimization, IEEE Trans. Syst. Man Cybern. Part B 40 (1) (2010) 137–150.
[21] Q. Hu, L. Zhang, D. Zhang, et al, Measuring relevance between discrete and continuous features based on neighborhood mutual information, Expert Syst. Appl. 38 (Sep. 2011) 10737–10750.
[22] M. Hu, E.C.C. Tsang, Y. Guo, W. Xu, Fast and Robust Attribute Reduction Based on the Separability in Fuzzy Decision Systems, IEEE Trans. Cybern. (2021), https://doi.org/10.1109/TCYB.2020.3040803.
[23] M. Hu, E.C.C. Tsang, Y. Guo, D. Chen, W. Xu, A novel approach to attribute reduction based on weighted neighborhood rough sets, Knowledge-Based Syst., vol. 220, art. 106908, 2021.https://doi.org/10.1016/j.knosys.2021.106908..
[24] R.A. Ibrahim, M. Abd Elaziz, D. Oliva et al., An improved runner-root algorithm for solving feature selection problems based on rough sets and neighborhood rough sets, Appl. Soft Comput., art. 105517, 2019.https://doi.org/10.1016/j.asoc.2019.105517..
[25] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Trans. Fuzzy Syst. 17 (4) (Aug. 2009) 824–838.
[26] H. Jiang, J. Zhan, D. Chen, Covering based variable precision $(\mathscr{I},\mathscr{T})$-fuzzy rough sets with applications to multi-attribute decision-making, IEEE Trans. Fuzzy Syst. 27 (8) (2019) 1558–1572.
[27] P. Maji, P. Garai, Fuzzy-rough simultaneous attribute selection and feature extraction algorithm, IEEE Trans. Cybern. 43 (4) (Aug. 2013) 1166–1177.
[28] A. Mariello, R. Battiti, Feature selection based on the neighborhood entropy, IEEE Trans. Neural Networks Learn. Syst. 29 (12) (Dec. 2018) 6313–6322.
[29] Z. Pawlak, Rough sets, Int. J. Comput. Inform. Sci. 11 (5) (1982) 341–356.
[30] B. Sang, H. Chen, T. Li, et al, Incremental approaches for heterogeneous feature selection in dynamic ordered data, Inf. Sci. 541 (2020) 475–501.
[31] L. Sun, X. Zhang, Y. Qian, et al, Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification, Inf. Sci. 502 (Oct. 2019) 18–41.
[32] L. Sun, X. Zhang, Y. Qian, et al, Joint neighborhood entropy-based gene selection method with Fisher score for tumor classification, Appl. Intell. 49 (4) (2019) 1245–1259.
[33] A. Tan, W.Z. Wu, Y. Qian, J. Liang, et al, Intuitionistic fuzzy rough set-based granular structures and attribute subset selection, IEEE Trans. Fuzzy Syst. 27 (3) (2018) 527–539.
[34] E.C.C. Tsang, D.G. Chen, D.S. Yeung, et al, Attributes reduction using fuzzy rough sets, IEEE Trans. Fuzzy Syst. 16 (5) (Oct. 2008) 1130–1141.
[35] C. Wang, M. Shao, Q. He, et al, Feature subset selection based on fuzzy neighborhood rough sets, Knowl.-Based Syst. 111 (2016) 173–179.
[36] C. Wang, Q. Hu, X. Wang, et al, Feature selection based on neighborhood discrimination index, IEEE Trans. Neural Networks Learn. Syst. 29 (7) (Jul. 2018) 2986–2999.
[37] C. Wang, Y. Shi, X. Fan, M. Shao, Attribute reduction based on k-nearest neighborhood rough sets, Int. J. Approximate Reasoning 106 (Mar. 2019) 18–31.
[38] C. Wang, Y. Huang, M. Shao, et al, Feature selection based on neighborhood self-information, IEEE Trans. Cybern. 50 (9) (Sep. 2020) 4031–4042.
[39] Q. Wang, Y. Qian, X. Liang, et al, Local neighborhood rough set, Knowl.-Based Syst. 153 (2018) 53–64.
[40] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945) 80–83.
[41] W. Wu, W. Zhang, Neighborhood operator systems and approximations, Inf. Sci. 144 (2002) 201–217.
[42] S. Yang, H. Zhang, Quantitative dominance-based neighborhood rough sets via fuzzy preference relations, IEEE Trans. Fuzzy Syst. (2019), https://doi.org/10.1109/TFUZZ.2019.2955883.
[43] L.A. Zadeh, Fuzzy sets, Control 8 (3) (1965) 338–353.
[44] J. Zhang, T. Li, D. Ruan, D. Liu, Neighborhood rough sets for dynamic data mining, Int. J. Intell. Syst. 27 (4) (2012) 317–342.
[45] W. Ziarko, Variable precision rough set model, J. Comput. Syst. Sci. 46 (1993) 39–59.