Feature Selection for Unbalanced Distribution Hybrid Data Based on k-Nearest Neighborhood Rough Set

Weihua Xu, Ziting Yuan, and Zheng Liu

***Abstract*—Neighborhood rough sets are now widely used to process numerical data. Nevertheless, most of the existing neighborhood rough sets are not able to distinguish class mixture samples well when dealing with classification problems. That is, it cannot effectively classify categories when dealing with data with an unbalanced distribution. Because of this, in this article, we propose a new feature selection method that takes into consideration both heterogeneous data and feature interaction. The proposed model well integrates the ascendancy of δ-neighborhood and k-nearest neighbor. Such heterogeneous data can be handled better than existing neighborhood models. We utilize information entropy theories such as mutual information and conditional mutual information and employ an iterative strategy to define the importance of each feature in decision making. Furthermore, we design a feature extraction algorithm based on the above idea. Experimental results display that the raised algorithm has superior effect than some existing algorithms, particularly the δ-neighborhood rough set model and the k-nearest neighborhood rough set model.**

***Impact Statement*—Feature selection is one of the important topics in machine learning and even artificial intelligence. Inaddition, feature selection using neighborhood rough sets has been proven to be an effective way. However, the sensitivity of most existing algorithms to imbalanced data is an important flaw in practical applications. This paper discusses how to use neighborhood rough sets to solve the problem of feature extraction when the distribution of heterogeneous data is unbalanced.Because the distribution of real-world data is not always uniform, feature selection algorithms can be applied in a wider range of fields, such as fraud identification, recommendation systems, etc. The algorithm for research on unbalanced data in this paper can enable researchers and even industry professionals to obtainmore effective results when dealing with problems in practical applications.**

***Index Terms*—Feature selection, neighborhood mutual information, neighborhood rough set (NRS), unbalanced distribution.**

I. Introduction

THE growth rate of data scale far exceeds the ability of human analysis and application in the era of information explosion. Rough set theory, as a method in the field of mathematics, can play its advantage when dealing with ambiguous imprecise data. This theory requires no prior information other than the original data. Thus, it is more objective when dealing with uncertain problems. As we all know, attribute reduction is one of the most widely used neighborhood rough set theories (NRSTs) in information systems. Attribute reduction can keep its classification and decision-making ability unchanged and remove irrelevant and redundant features. In this way, key attributes are extracted, and the purpose of simplifying the information system is achieved.

According to the research and discussion of researchers at home and abroad, many data mining methods have emerged in recent years. Realistic problem data, such as text voice or image, usually contain more features, but too much characteristics will lead to poor interpretability, and the slow calculation model in fitting problems, such as feature selection, can maintain data classification under the constant ability to effectively remove redundancy and unrelated features of the data and, thus, become an important machine learning pretreatment process of pattern recognition, data mining, and artificial intelligence.

In 1982, Pawlak [1] proposed the rough set theory first. Because rough sets can play its advantages in the processing of fuzzy inaccurate data, the theory has been widely used over the years, making the rapid development in data mining, pattern recognition and artificial intelligence, decision support [2], and other fields. The most obvious difference between rough set theory and other data processing methods is that it does not require to offer any prior information other than the original data being processed, but only needs to deal with object data, which is more objective when dealing with imprecise and uncertain problems. However, the classical rough set theory can only deal with character data in the process of processing data. Actual samples tend to be mixed and diverse and generally have numerical attributes. Therefore, we have to discretize it first. In the process of discretization, data will inevitably be lost. However, the integrity of the data and the accuracy will be affected. Lin [3] put forward the concept of neighborhood rough set (NRS), which can improve this problem. Hu et al. [4] studied the calculation model of rough set for mixed data knowledge discovery and proposed a more systematic NRS model. The NRS puts forward the notion of neighborhood and granulates the domain of each sample by the distance and neighborhood radius between samples. Thus, it obtains the neighborhood relationship between samples, so as to judge whether its attributes can effectively process numerical and mixed data. Meanwhile, it can avoid the problems caused by discretization. Therefore, extended NRS models were investigated extensively

Manuscript received 18 May 2022; revised 19 September 2022 and 21 December 2022; accepted 8 January 2023. Date of publication 16 January 2023; date of current version 8 January 2024. This work was supported by the National Natural Science Foundation of China under Grant 61976245. This paper was recommended for publication by Associate Editor Hani Hagras upon evaluation of the reviewers' comments. *(Corresponding author: Weihua Xu.)*

The authors are with the College of Artificial Intelligence, Southwest University, Chongqing 400715, China (e-mail: chxuwh@gmail.com; yzt1133@163.com; 1017749590@qq.com).

Digital Object Identifier 10.1109/TAI.2023.3237203

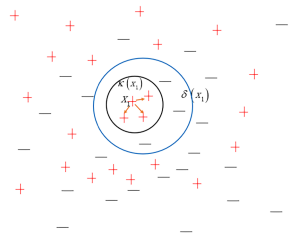

Fig. 1. Neighborhoods of a sample in a high-density region.

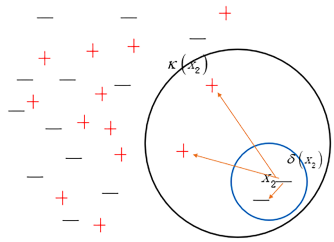

Fig. 2. Neighborhoods of a sample in a sparse-density region.

in recent years, such as local NRS [5], neighborhood-based decision-theoretic rough set [6], fuzzy NRS [7], pseudo-label NRS [7], NRS with nominal metric embedding [8], neighborhood multigranularity rough set [9], [10], [11], etc. Moreover, NRST-based methods have been applied effectively in related fields, such as image annotation [12], data classification [13], [14], [15], gene selection [11], [16], [17], and incremental learning [8], [18].

In practical applications, in fact, a uniformly distributed sample space rarely exists. For example, samples may be very densely or sparsely distributed in certain regions of the sample space. Under these circumstances, neither of the existing NRSs, i.e., δ-neighborhood rough set and k-nearest neighborhood rough set (kNNRS), can classify the samples well into the class to which they belong. For example, in a sample space, there are two types of samples, positive and negative, which are marked as "+" and "−," respectively. The blue circles represent δ-neighborhood, and the black circles represent k-nearest neighbor (kNN) of the sample.

As shown in Fig. 1, the sample x_1 is located in a densely distributed area, so three samples of both positive and negative categories are included in its neighborhood $\delta(\mathrm{x}_1)$. Therefore, according to the NRST, the sample x_1 is classified to the boundary area according to the voting principle. Actually, we can better determine the class label of x_1 if we use three nearest neighbors (3NNs). This is because the nearest-neighbor strategy limits the number of samples in high-density regions, thus limiting the size of the neighborhood. Furthermore, it can better guarantee that the samples in $\kappa(\mathrm{x}_1)$ always belong to the same class.

In another case, the kNN model has similar shortcomings when determining the classification labels in areas with sparse sample distribution. For example, as shown in Fig. 2, the granule $\kappa(\mathrm{x}_2)$ of the 3NNs of x_2 involves three samples, including two positive class samples that far away from x_2 and one negative class sample that is closer. According to the principle of majority voting, this may cause x_2 to be incorrectly assigned to the positive class label. However, at this time, the $\delta(\mathrm{x}_2)$ neighborhood can classify x_2 into the negative class label well. And the weakness of the kNN model can be overcome.

In summary, the above two sample distributions show that the δ-neighborhood model and the kNN model have advantages and disadvantages in different situations. This leads to poor classification accuracy of these two models when encountering imbalanced data.

In another aspect, the theoretical basis of mutual information standard is information entropy in information theory. Information theory has been widely used in feature selection algorithms. The feature selection algorithm studied in this article is based on the mutual information criterion, and the theoretical basis of the proposed new algorithm is information entropy. At present, many feature selection algorithms based on mutual information have been proposed. The earliest mention of the concept of mutual information is the mutual information maximum algorithm [19], which lacks the measurement of the relationship between features. It only computes the mutual information between candidate features and class labels to measure the correlation. This algorithm does not find the existence of redundant features. Battiti et al. [20] proposed the mutual information feature selection algorithm for this problem [19], which introduced a first-order incremental search algorithm and used a greedy selection method to select the most relevant k features from an initial set of n features. Although the algorithm measures the redundancy that exists between features, the redundancy term may become large as the number of selected features increases. Therefore, some extraneous features need to be removed. In response to this finding, Peng et al. [21] proposed a feature selection method called minimal redundancy maximal relevance (mRMR). The mRMR feature selection evaluation criterion uses the mutual information between candidate features and class labels to measure the correlation. At the same time, the mutual information of the candidate features and the selected features is used to measure the redundancy. The mRMR algorithm prefers to select candidate features that have as much correlation with class labels as possible and, at the same time, have as little redundancy with selected features as possible. The conditional mutual information maximization algorithm selects those candidate features that have the minimum conditional mutual information with the class labels given all the selected features [22]. Wan et al. [23] established a neighborhood-mutual-information-based feature selection algorithm, which considers the characteristic of interaction in the NRS. In summary, many scholars use the information entropy and its extension as the feature measure to select feature subset [24], [25], [26].

In addition, unbalanced distribution and mixed data are ubiquitous in model building for practical applications. However, the existing interactive feature selection methods using mutual information do not consider these cases. Inspired by this problem, a feature selection method that considers the interaction of imbalanced distributed data and mixed data is urgently needed, which is also the focus of this article.

The main contributions of this article are as follows.

1) To address the problem of imbalanced data, we proposed a k-nearest model based on neighborhood entropy. At the same time, the neighborhood information entropy

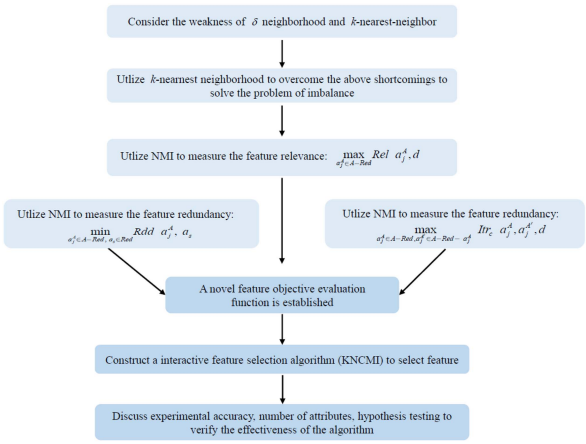

Fig. 3. Process of this article.

and neighborhood conditional mutual information are discussed. Then, the neighborhood conditional mutual information is used as the measure of attribute importance.

2) Afterward, an attribute reduction algorithm for the kNNRS based on neighborhood conditional mutual information is proposed.
3) Using 20 datasets from UCI, four comparative algorithms are carried out. In addition, the effectiveness and classification accuracy of the proposed algorithm is confirmed in the classification accuracy. The effect of neighborhood granularity parameters ε and k on classification accuracy is also discussed in the subsequent experiments.

The rest of this article is organized as follows. Section II retrospects the fundamental concepts of NRSs and presents information measurement methods for neighborhood decision-making systems. The neighborhood radius for different feature types is computed in Section III. Going a step further, we present an uncertainty measure to address the kNNRS of data with imbalanced distributions. Section IV details feature correlation, redundancy, and interactivity under the new neighborhood uncertainty measure. And a feature objective evaluation function, i.e., k-nearest Max-Relevance min-Redundancy Max-Interaction (KNMRmRMI), and a corresponding interactive feature selection algorithm (KNCMI) are constructed using the new NRS. Experimental comparisons and results are presented in Section V. Finally, Section VI concludes this article. The process of this article can be clearly seen from Fig. 3.

II. Related Work and Foundations

In this section, some fundamental concepts are reviewed, such as NRST and uncertainty measurement.

A. Neighborhood Rough Set Theory

In this section, we review some fundamental concepts related to δ-neighborhood rough sets, kNNRS, and the information-measurement-based NRS models. In order to address the issue that the classical rough set is not convenient for processing the dataset with numerical attributes, the basis of the NRS model is that the definition of the neighborhood concept in different ways will constitute different NRS models.

Regarding NRST, the distance formula is used to evaluate the similarity between different samples. In terms of distance metric learning, as introduced in Section II-A, a positive-semidefinite matrix $M(M \geq 0)$ is obtained. Then, there is distance metric on B, which is recorded as d_B. For any $x, y, z \in U$, it satisfies the following:

1) nonnegativity: $d_B(x, y) \geq 0, d_B(x, x) = 0$;
2) symmetry: $d_B(x, y) = d_B(y, x)$;
3) triangular inequality: $d_B(x, z) \leq d_B(x, y) + d_B(y, z)$.

As we all know, classical metrics include Manhattan distance function, Euclidean distance function, and Chebyshev distance function, among which the Euclidean distance function effectively reflects the basic information of unknown data.

The Euclidean distance is often used to calculate the distance between samples. Given two random samples with a real-valued attribute set B, the formula for calculating the Euclidean distance between two samples is as follows:

$$d_B(x_i, x_j) = \sqrt{\sum_{a \in B} (a(x_i) - a(x_j))^2}.$$

Here, $a(x_i)$ represents the value of x_i with respect to the attribute a.

NRSs are proposed to solve the problem that classical rough sets are inconvenient to deal with datasets with numerical features.

Given a neighborhood decision system $NDS = (U, A, D, \Delta, \delta)$, usually written more simply as $NDS = (U, A, D, \delta)$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a sample set named universe, $C = \{a_1, a_2, \ldots, a_m\}$ is a conditional attribute set that describes the samples, $D = \{d\}$ is a decision attribute set that contains one decision attribute, Δ is represented as the distance over the relation BR, usually using the Euclidean distance, and δ is a neighborhood parameter with $0 \leq \delta \leq 1$.

Given a neighborhood decision system $NDS = (U, A, D, \delta)$ with $B \subseteq A$, the similarity relation resulting by B is defined as

$$NR_\delta(B) = \{(x_i, x_j) \in U \times U \mid \Delta_B(x_i, x_j) \leq \delta\}.$$

Given a neighborhood decision system $NDS = (U, A, D, \delta)$ with $B \subseteq A$, for any $x \in U$, the neighborhood class of x with respect to B is described as

$$\delta_B(x_i) = \{x_j \in U \mid \Delta_B(x_i, x_j) \leq \delta\}.$$

If the decision values of all the samples in $\delta(x_i)$ are the same, then x_i is consistent in the δ-neighborhood; otherwise, it is called inconsistent sample.

B. k-Nearest-Neighbor Rough Set

Given a neighborhood decision system $NDS = (U, A, D, \delta)$ with $B \subseteq A$, the kNNs of $x_i \in U$ in terms of B are defined as

$$\kappa_B(x_i) = \left\{x_i^1, x_i^2, \ldots, x_i^k \mid d_B(x_j, x_i) > d_B\left(x_i^h, x_i\right), x_j \neq x_i^l, l, h = 1, 2, \ldots, k\right\}.$$

It can be seen that $\kappa_B(x_i)$ is the set of k samples closest to x_i. This means that the number of samples contained in $\kappa_B(x_i)$ is fixed as κ.

The family of kNN information granules $\{\kappa_B(x_i) \mid x_i \in U\}$ can form a coverage of the universe U.

Analogously, we can get a binary relation K_B, as follows:

$$K_B = \{(x_i, x_j) \in U \times U \mid x_j \in \kappa_B(x_i)\}.$$

Clearly, K_B is reflexive. However, it is obvious that it does not satisfy symmetry and transitivity.

C. Information Measurements in the Neighborhood Decision System

The dataset in the classification issues can be formally defined as a neighborhood decision system $NDS = (U, A, D, \Delta, \delta)$. The decision attribute set is $D = \{d\}$ in the single decision attribute classification learning task. Various information entropies have been widely used in attribute reduction [27], [28]. Some information measurements in the NDS are expressed as follows.

Definition 2.1. (Neighborhood information entropy [29]): Given an $NDS = (U, A, D, \Delta, \delta)$, for $\delta \geq 0$, $\forall B \subseteq A$, the neighborhood relation on B is expressed as NR_B^δ. Then, the neighborhood of $x_i \in U$ obtained from B is $NR_B^\delta(x_i)$, which is abbreviated as $\delta_B(x_i)$. The neighborhood entropy of the sample set with respect to B is defined as

$$NE_\delta(B) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_B(x_i)|}{|U|}$$

where the neighborhood uncertainty of sample x_i makes up the neighborhood entropy (i.e., average uncertainty) of the sample set, which is expressed as

$$NE_\delta^{x_i}(B) = -\log_2\frac{|\delta_B(x_i)|}{|U|}.$$

Definition 2.2. (Neighborhood joint entropy [29]): Given a $NDS = (U, A, D, \Delta, \delta)$, for $\delta \geq 0$, $\forall B, C \in A$, the neighborhood joint entropy of A and B is defined as

$$NE_\delta(B, C) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{BUC}(x_i)|}{|U|}.$$

Definition 2.3. (Neighborhood conditional entropy [29]): Given an $NDS = (U, A, D, \Delta, \delta)$, for $\delta \geq 0$, $\forall B, C \in A$, under the condition that B is known, the information entropy of A is conveyed as the conditional entropy of A with regard to B, which is defined as

$$NE_\delta(B \mid C) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{B\cup C}(x_i)|}{|\delta_C(x_i)|}.$$

Proposition 2.1: Given an $NDS = (U, A, D, \Delta, \delta)$, for $\delta \geq 0$, $\forall B, C \in A$, then $NE_\delta(B \mid C) = NE_\delta(B, C) - NE_\delta(C)$.

Proof: From Definitions 2.1 and 2.2, we obtain

$$\begin{aligned}
&NE_\delta(B, C) - NE_\delta(C)\\
&= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{BUC}(x_i)|}{|U|} + \frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_C(x_i)|}{|U|}\\
&= -\frac{1}{|U|}\sum_{i=1}^{|U|}\left(\log_2\frac{|\delta_{BUC}(x_i)|}{|U|} - \log_2\frac{|\delta_C(x_i)|}{|U|}\right)\\
&= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\left(\frac{|\delta_{BUC}(x_i)|}{|U|}\cdot\frac{|U|}{|\delta_C(x_i)|}\right)\\
&= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{BUC}(x_i)|}{|\delta_C(x_i)|}\\
&= NE_\delta(B \mid C).
\end{aligned}$$

That is, $NE_\delta(B \mid C) = NE_\delta(B, C) - NE_\delta(C)$ holds.

It can be seen from the formula that the neighborhood conditional entropy reflects the amount of extra uncertainty in B after introducing the feature subset C. It can be calculated by the neighborhood joint entropy provided by B and C and the neighborhood uncertainty of C.

Definition 2.4. (Neighborhood mutual information [29]): Given an $NDS = (U, A, D, \Delta, \delta)$, for $\delta \geq 0$, $\forall B, C \in A$, the neighborhood mutual information of B and C is defined as

$$NMI_\delta(B; C) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_B(x_i)| \cdot |\delta_C(x_i)|}{|U \parallel \delta_{B\cup C}(x_i)|}$$

where $\delta_B(x_i)$ and $\delta_C(x_i)$ denote the neighborhoods on B and C, respectively. And $\delta_{B\cup C}(x_i)$ represents the δ-neighborhood of x_i on $B \cup C$.

Given the selected feature subset C, neighborhood mutual information can be used to measure the additional contribution of the feature subset B to determine the magnitude of the reduced classification uncertainty. According to Definitions 2.1−2.4 and Proposition 2.1, the relationships between uncertainty measures in the NRS, such as neighborhood entropy, neighborhood conditional entropy, and neighborhood mutual information, can be obtained, as shown in Proposition 2.2.

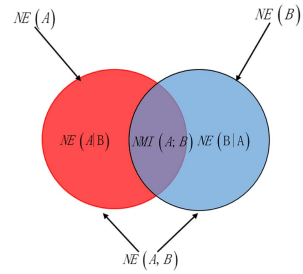

Fig. 4. Relationship between the neighborhood information entropy (NE) and the mutual information (NMI).

Fig. 4 shows the relationship between neighborhood information entropy and neighborhood mutual information in information theory. Neighborhood information entropy NE represents the uncertainty degree of a random variable. NMI represents the interaction between any two random variables. For variable A, its neighborhood information entropy $NE(A)$ deducts its own conditional information entropy $NE(B|A)$ under the conditions of other variables. The value of mutual information $NMI(A; B)$ between two variables can be obtained.

Proposition 2.2: Given an $NDS = (U, A, D, \Delta, \delta)$, for $\delta \geq 0$, $\forall B, C \in A$, we have the following.

1) $NMI_\delta(B;C) = NMI_\delta(C;B)$.
2) $NMI_\delta(B;C) = NE_\delta(B) + NE_\delta(C) - NE_\delta(B,C)$
3) $NMI_\delta(B;C) = NE_\delta(B) - NE_\delta(B \mid C) = NE_\delta(C) - NE_\delta(B \mid C)$.

Proof:

1) According to Definition 2.4, we have

$$\begin{aligned} &NMI_\delta(B;C) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_B(x_i)|\cdot|\delta_C(x_i)|}{|U \parallel \delta_{B\cup C}(x_i)|} \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_C(x_i)|\cdot|\delta_B(x_i)|}{|U \parallel \delta_{C\cup B}(x_i)|} \\ &= NMI_\delta(C;B). \end{aligned}$$

2) From Definitions 2.1 and 2.2, we obtain

$$\begin{aligned} &NE_\delta(B) + NE_\delta(C) - NE_\delta(B,C) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_B(x_i)|}{|U|} - \frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_C(x_i)|}{|U|} \\ &\quad + \frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{BUC}(x_i)|}{|U|} \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\left(\frac{|\delta_B(x_i)|}{|U|}\cdot\frac{|\delta_C(x_i)|}{|U|}\cdot\frac{|U|}{|\delta_{BUC}(x_i)|}\right) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_B(x_i)|\cdot|\delta_C(x_i)|}{|U \parallel |\delta_{B\cup C}(x_i)|} \\ &= NMI_\delta(B;C). \end{aligned}$$

3) According to Definitions 2.1 and 2.3, we have

$$\begin{aligned} &NE_\delta(A) - NE_\delta(A \mid B) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_A(x_i)|}{|U|} + \frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{AUB}(x_i)|}{|\delta_B(x_i)|} \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_A(x_i)|\cdot|\delta_B(x_i)|}{|U|\,|\delta_{A\cup B}(x_i)|}, \end{aligned}$$

$$\begin{aligned} &NE_\delta(B) - NE_\delta(B \mid A) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_B(x_i)|}{|U|} + \frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{B\cup A}(x_i)|}{|\delta_A(x_i)|} \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_B(x_i)|\cdot|\delta_A(x_i)|}{|U|\,|\delta_{B\cup A}(x_i)|}. \end{aligned}$$

Therefore, we have $NMI_\delta(B;C) = NE_\delta(B) - NE_\delta(B \mid C) = NE_\delta(C) - NE_\delta(B \mid C)$.

Proposition 2.2(1) indicates that the neighborhood mutual information is symmetric, that is, the information shared by B and C is the same as that shared by C and B. Proposition 2.2(2) shows that the amount of information shared or duplicated by B and C can be obtained from the difference between the information provided by each of B and C and the information provided by them jointly. In addition to the calculation of the neighborhood mutual information shown in Proposition 2.2(2), Proposition 2.2(3) displays the reduction in the uncertainty of B (or C) under the condition that C (or B) is known.

Definition 2.5 is similar to the introduction of conditional entropy into information measures in entropy theory. The neighborhood conditional mutual information is also introduced, that is, the neighborhood mutual information of B and R under the condition of known C.

Definition 2.5. (Neighborhood conditional mutual information [30]): Given an $NDS = (U, A, D, \Delta, \delta)$, for $\delta \geq 0$, $\forall B, C, R \in A$, under the condition that C is known, the neighborhood conditional mutual information of B and R is expressed as

$$NCMI_\delta(B;R \mid C) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{B\cup C}(x_i)|\cdot|\delta_{R\cup C}(x_i)|}{|\delta_C(x_i)|\cdot|\delta_{B\cup C\cup R}(x_i)|}$$

where $\delta_C(x_i), \delta_{B\cup C}(x_i), \delta_{R\cup C}(x_i)$, and $\delta_{B\cup C\cup R}(x_i)$ represent the neighborhoods of x_i on C, $B \cup C$, $R \cup C$, and $B \cup C \cup R$, respectively.

The meaning of neighborhood conditional mutual information is that when C is known, the uncertainty of B is reduced due to the knowledge of R.

Proposition 2.3: Given an $NDS = (U, A, D, \Delta, \delta)$, for $\delta \geq 0$, $\forall B, C, R \in A$, we have $NCMI_\delta(B;R \mid C) = NE_\delta(B,C) + NE_\delta(R,C) - NE_\delta(B,\mathrm{R},\mathrm{C}) - NE_\delta(C)$.

Proof: From Definition 2.2, we can deduce that

$$NE_\delta(B,\mathrm{R},\mathrm{C}) = \frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{B\cup R\cup C}(x_i)|}{|U|}.$$

According to Definitions 2.1 and 2.2, we have

$$\begin{aligned} &NE_\delta(B,C) + NE_\delta(R,C) - NE_\delta(B,R,C) - NE_\delta(C) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{BUC}(x_i)|}{|U|} - \frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{R\cup C}(x_i)|}{|U|} \\ &\quad + \frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{B\cup R\cup C}(x_i)|}{|U|} + \frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_C(x_i)|}{|U|} \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\left(\frac{|\delta_{B\cup C}(x_i)|}{|U|}\cdot\frac{|\delta_{R\cup C}(x_i)|}{|U|}\right. \\ &\quad \left.\cdot\frac{|U|}{|\delta_{B\cup R\cup C}(x_i)|}\cdot\frac{|U|}{|\delta_C(x_i)|}\right) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_{B\cup C}(x_i)|\cdot|\delta_{R\cup C}(x_i)|}{|\delta_C(x_i)|\cdot|\delta_{B\cup C\cup R}(x_i)|}. \end{aligned}$$

Therefore, we have $NCMI_\delta(B;R \mid C) = NE_\delta(B,C) + NE_\delta(R,C) - NE_\delta(B,\mathrm{R},\mathrm{C}) - NE_\delta(C)$.

Proposition 2.3 expresses that neighborhood conditional mutual information can be obtained by neighborhood joint entropy and neighborhood entropy.

Proposition 2.4: Given an $NDS = (U, A, D, \Delta, \delta)$, for $\delta \geq 0, \forall B, C, R \in A$, under the condition that B is known, we have

$$NCMI_\delta(B; R \mid C) = NCMI_\delta(R; B \mid C).$$

Proof: It can be proved according to Definition 2.5.

By analogy with the symmetry of neighborhood mutual information in Proposition 2.2(1), given a feature subset C, the information quantity of B obtained from R is equivalent to that of R obtained from B, that is, the information provided by them is mutual.

Proposition 2.5: Given an $NDS = (U, A, D, \Delta, \delta)$, for $\delta \geq 0, \forall B, C, R \in A$, under the condition that B is known, if A and R are independent of each other, then $NCMI_\delta(B; R \mid C) = 0$.

Proof: If B and R are independent of each other, it can be calculated from Definition 2.5 that $NCMI_\delta(B; R \mid C) = 0$.

Similar to mutual information in information theory, when B and R are independent under the condition of known attribute set C, the value of neighborhood conditional mutual information is 0.

III. kNNRS for Unbalanced Hybrid Data

In this section, the processing method of hybrid data is introduced in Section III-A. Section III-B explains the disadvantages of rough sets with different neighborhoods in handling unbalanced data. The solution to this problem is shown in Section III-C.

A. Hybrid Data Processing

Data in practical applications usually include numerical, categorical [31], and hybrid data [32]. The heterogeneous Euclidean overlap metric [4], [33] is introduced to deal with the mixed form of data. Similarly, the heterogeneous Chebyshev overlap metric (HCOM) is defined to realize the intersection operation in neighborhood relations.

Definition 3.1. (Hybrid distance function in the NRS): For different types of data, the HCOM distance function is computed by

$$HCOM_A(x, y) = \sum_{j=1}^{|A|} \left(d^\tau_{\{a_j\}}(x, y) \right)^{\frac{1}{\tau}}$$

where

$$d_{\{a_j\}}(x, y) = \begin{cases} \left| \frac{a(x_i) - a(x_j)}{\sigma_a} \right|, & \text{if a is a numerical feature;} \\ 1, & \text{if } f_j \text{ is a categorical feature} \\ & \quad \text{and a } (x_i) \neq a\,(x_j)\,; \\ 0, & \text{if } f_j \text{ is a categorical feature} \\ & \quad \text{and a } (x_i) = a\,(x_j)\,; \\ 1, & \text{if the feature value of } x \text{ or } y \\ & \quad \text{is unknown with respect to } a. \end{cases}$$

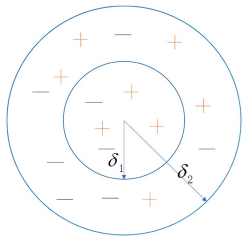

Fig. 5. Classification results of different granules.

Specially, if a_j is a numerical feature, the above formula can be reduced to $HCOM_A(x, y) = \Delta_A(x, y), \tau = +\infty$ according to the distance function shown in Section II-A.

Thus, given a neighborhood decision system $NDS = (U, A, D, \delta)$ with $B \subseteq A$, for any $x \in U$, the neighborhood class of x with respect to B is described as

$$\delta_B\,(x_i) = \{x_j \in U | HCOM_B\,(x_i, x_j) \leq \delta\}\,.$$

In a classification task, we expect the differences between samples belonging to the same class to be as small as possible, while the differences between samples belonging to different classes should be large enough. This makes it easier for us to classify. The features that describe the samples determine how easy it is to distinguish between different classes of samples. We perform feature selection on those features that make samples of different classes easily distinguishable. However, for some complex classification problems, there will always be some undistinguishable samples [34]. We would incorrectly classify these samples into the same class. They consist of samples belonging to different classes but with small differences in eigenvalues.

The size of the neighborhood δ in NRS reflects the tolerance of noise data in its sample space, that is, the degree of allowable quantization error. Fig. 5 visually depicts the inconsistency of classification under different radius granules.

For example, the meanings of the two types of samples marked "+" and "−" are the same as in Fig. 5. In a binary classification problem, given two observation problems with different granularities δ_1 and δ_2, $\delta_1 < l < \delta_2$, it can be seen from Fig. 5 that the classification of sample x under the granularity δ_1 is consistent, but the classification under the granularity δ_2 with larger error tolerance is inconsistent. Neighborhoods with different granularities determine whether the classification results of the samples are consistent. Therefore, the choice of granularity also has a great impact on the performance of the model.

B. k-Nearest Neighborhood Rough Set

Given a k-nearest neighborhood decision system $KNDS = (U, A, D, \Delta, \delta, K)$, usually written more simply as $KNDS = (U, A, D, \delta, K)$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a sample set named universe, $C = \{a_1, a_2, \ldots, a_m\}$ is a conditional attribute set that describes the samples, $D = \{d\}$ is a decision attribute set that contains only one decision attribute, Δ is represented as the distance over the relation R, usually using the Euclidean

distance, and δ is a neighborhood parameter with $0 \leq \delta \leq 1$. The parameter K means taking the K samples closest to x_i. The optimal parameter K is obtained experimentally. The proposed new NRS can better handle the problems caused by imbalanced data.

Definition 3.2. (kNN [35]): Given a k-nearest neighborhood decision system $KNDS = (U, A, D, \delta, K)$ with $B \subseteq A$, the kNN of $x_i \in U$ in terms of B is defined as

$$\tau_B(x_i) = \{x_j \in U \mid x_j \in \delta_B(x_i) \cap \kappa_B(x_i)\}.$$

The kNN $\tau_B(x_i)$ means taking the intersection of $\delta_B(x_i)$ and $\kappa_B(x_i)$. Therefore, the kNN $\tau_B(x_i)$ can overcome the weakness of δ-neighborhood and kNN for describing the classification of samples in a sample space where the sample density is not uniform, as shown in Figs. 1 and 2.

Definition 3.3. (kNN information entropy): Given a $KNDS = (U, A, D, \delta, K)$, $\forall B \subseteq A$, the neighborhood relation on B is expressed as NR_B^τ. Then, the neighborhood of $x_i \in U$ obtained from B is $NR_B^\tau(x_i)$, which is abbreviated as $\tau_B(x_i)$. The neighborhood entropy of the sample set with respect to B is defined as

$$NE_\tau(B) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\tau_B(x_i)|}{|U|}.$$

Definition 3.4. (kNN joint entropy): Given a $KNDS = (U, A, D, \delta, K)$, $\forall B, C \in A$, the neighborhood joint entropy of B and C is defined as

$$NE_\tau(B, C) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\tau_{BUC}(x_i)|}{|U|}.$$

Definition 3.5. (kNN conditional entropy): Given a $KNDS = (U, A, D, \delta, K)$, $\forall B, C \in A$, under the condition that C is known, the information entropy of B is expressed as the conditional entropy of B with respect to C, which is defined as

$$NE_\tau(B \mid C) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\tau_{B\cup C}(x_i)|}{|\tau_C(x_i)|}.$$

Proposition 3.1: Given a $KNDS = (U, A, D, \delta, K)$, $\forall B, C \in A$, $NE_\tau(B \mid C) = NE_\tau(B, C) - NE_\tau(C)$.

Proof: It can be proved using Definitions 3.3 and 3.4.

Neighborhood conditional entropy can be calculated by the neighborhood joint entropy provided by B and C and the neighborhood uncertainty of C.

Definition 3.6. (kNN mutual information): Given a $KNDS = (U, A, D, \delta, K)$, $\forall B, C \in A$, the neighborhood mutual information of B and C is defined as

$$NMI_\tau(B; C) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\tau_B(x_i)| \cdot |\tau_C(x_i)|}{|U \parallel \tau_{B\cup C}(x_i)|}$$

where $\tau_B(x_i)$ and $\tau_C(x_i)$ denote the neighborhoods on B and C, respectively; $\delta_{B\cup C}(x_i)$ represents the τ neighborhood of x_i on $B \cup C$.

According to Definitions 3.3–3.6 and Proposition 3.1, the relationships between uncertainty measures in the NRS, such as neighborhood entropy, neighborhood conditional entropy, and neighborhood mutual information, can be obtained, as shown in Proposition 3.2.

Proposition 3.2: Given a $KNDS = (U, A, D, \delta, K)$, $\forall B, C \in A$, we have the following:

1) $NMI_\tau(B; C) = NMI_\tau(C; B)$;
2) $NMI_\tau(B; C) = NE_\tau(B) + NE_\tau(C) - NE_\tau(B, C)$;
3) $NMI_\tau(B; C) = NE_\tau(B) - NE_\tau(B \mid C) = NE_\tau(C) - NE_\tau(B \mid C)$.

Proof: It can be proved using Definitions 3.3–3.6.

Definition 3.7. (kNN conditional mutual information): Given a $KNDS = (U, A, D, \delta, K)$, $\forall B, C, R \in A$, under the condition of known C, the neighborhood conditional mutual information of B and R is defined as

$$NCMI_\tau(B; R \mid C) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\tau_{B\cup C}(x_i)| \cdot |\tau_{R\cup C}(x_i)|}{|\tau_C(x_i)| \cdot |\tau_{B\cup C\cup R}(x_i)|}$$

where $\tau_C(x_i)$, $\tau_{B\cup C}(x_i)$, $\tau_{R\cup C}(x_i)$, and $\tau_{B\cup C\cup R}(x_i)$ represent the neighborhoods of x_i on C, $B \cup C$, $R \cup C$, and $B \cup C \cup R$, respectively.

Proposition 3.3: Given a $KNDS = (U, A, D, \delta, K)$, $\forall B, C, R \in A$, we have

$NCMI_\tau(B; R \mid C) = NE_\delta(B, C) + NE_\tau(R, C) - NE_\tau(B, \mathrm{R}, \mathrm{C}) - NE_\tau(C)$.

Proof: It is obviously true according to Proposition items (1) and (2) and the definition of the boundary region.

Proposition 3.4: Given a $KNDS = (U, A, D, \delta, K)$, $\forall B, C, R \in A$, under the condition that C is known, we have

$$NCMI_\tau(B; R \mid C) = NCMI_\tau(R; B \mid C).$$

Proof: It can be proved according to Definition 3.7.

By analogy with the symmetry of neighborhood mutual information in Proposition 3.2(1), given a feature subset C, the information quantity of B obtained from R is equivalent to that of R obtained from B, that is, the information provided by them is mutual.

Proposition 3.5: Given a $KNDS = (U, A, D, \delta, K)$, $\forall B, C, R \in A$, under the condition that C is known, if B and R are independent of each other, then $NCMI_\tau(B; R \mid C) = 0$.

Proof: If B and R are independent of each other, it can be calculated from Definition 3.7 that $NCMI_\tau(B; R \mid C) = 0$.

IV. Evaluation of Feature Significance Based on the kNNRS

In this section, a feature selection method that considers imbalanced data and feature interactions is proposed in the framework of kNNRSs. The relevance between features and classes and the redundancy and interaction between features are comprehensively explored and redefined in Section IV-A–IV-C. Based on these defined feature correlations, a new feature objective evaluation function, i.e., KNMRmRMI, is constructed in Section IV-D. Section IV-D also proposes a new neighborhood-conditional-mutual-information-based interaction feature selection algorithm.

The information metric in the NRS in Section III is used to describe the relationship between features, including the

relevance between features and classes, the redundancy between features, and the pairwise interactions between features.

A. Feature Relevance Measure

Many researchers use relevance between features and classes as a criterion for evaluating feature importance [21], [36], [37]. The core connotation is that the stronger the correlation between the feature and the class, the stronger the ability of the feature to distinguish samples.

Features that have greater relevance to classes provide more distinguishable information for class division in information theory. The aforementioned mutual information has been widely used to measure the correlation between features and classes.

Definition 4.1. (Relevance, Rel): Given a $KNDS = (U, A, D, \delta, K)$, $U = \{x_i \mid i \in \{1, \ldots, n\}\}, A = \{a_j \mid j \in \{1, \ldots, m\}\}$, $Red \subseteq A$ is the selected feature subset, $a_j^A \in A - Red$ is the current candidate feature, and the relevance between a_j^A and the decision class d is defined as

$$\begin{aligned} Rel\left(a_j^A, d\right) &= NMI_\tau\left(a_j^A; d\right) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\tau_{a_j^A}\left(x_i\right)\right| \cdot \left|\tau_d\left(x_i\right)\right|}{|U|\left|\tau_{\left\{a_j^A\right\}\cup\{d\}}\left(x_i\right)\right|} \end{aligned}$$

where $\tau_{a_j^A}(x_i)$ and $\tau_d(x_i)$ denote the neighborhood of x_i on a_j^A and d respectively, and $\tau_{\{a_j^A\}\cup\{d\}}(x_i)$ represents the neighborhood of x_i on $\{a_j^A\} \cup \{d\}$.

We choose features that have the largest neighborhood mutual information between features and classes. The selected feature is called the feature with the greatest relevance. This feature selection method is called the maximum-relevance (Max-Relevance, MR) criterion [38], [39]. And the MR criterion is formalized as

$$\max_{a_j^A \in A - Red} Rel\left(a_j^A, d\right).$$

By using the MR criterion to pick features, we can acquire a descending order sorted by the magnitude of the correlation between each feature and the class. In the first step of feature selection, we select the most relevant feature in the ranking as the first selected feature.

B. Feature Redundancy Measure

The above ranking-based methods only consider selecting the features most relevant to the class. But the redundancy between the features and the selected features is ignored, which reduces the classification performance to a certain extent. Thus, the neighborhood mutual information can be a good measure of the class-independent redundancy between the feature and the selected feature. Some researchers consider eliminating the redundancy between features to improve the classification performance of the algorithm [36], [40], [41].

Definition 4.2. (Redundancy, Rdd): Given a $KNDS = (U, A, D, \delta, K)$, $a_j^A \in A - Red$ is the current candidate feature, $a_s \in Red$ is the selected feature, the pairwise redundancy of class independence between a_j^F and a_s is defined as

$$\begin{aligned} sRdd\left(a_j^A, a_s\right) &= NMI_\tau\left(a_j^A; a_s\right) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\tau_{a_j^A}\left(x_i\right)\right| \cdot \left|\tau_{a_s}\left(x_i\right)\right|}{|U|\left|\tau_{\left\{a_j^A\right\}\cup a_s\}}\left(x_i\right)\right|} \end{aligned}$$

where $\tau_{a_j^A}(x_i)$ and $\tau_{a_s}(x_i)$ denote the neighborhood of x_i on a_j^A and a_s, respectively, and $\tau_{\{a_j^A\}\cup\{a_s\}}(x_i)$ represents the neighborhood of x_i on $\{a_j^A\} \cup \{a_s\}$.

In the second step of feature extraction, in order to reduce the redundancy in the selected feature subset, the minimum-redundancy criterion [38] is defined as

$$\min_{a_j^A \in A - Red, a_s \in Red} Rdd\left(a_j^A, a_s\right).$$

C. Feature Interaction Measure

The goal of most feature selection methods is to select the subset of features with the greatest correlation to the category and least redundancy with the selected features while preserving the categorical information. So far, little has been done to study the interactions between features. Then, the ubiquitous interactions between features are ignored, which would change the final reduction set. Going a step further, the final classification accuracy will be affected. Some scholars proposed to use a single neighborhood to solve the interaction between features [23], [42]. Therefore, in this subsection, we explore the interactions between features and seek suitable measures in $KNDS$.

When solving classification tasks, there are multiple interactions due to the differences between individual features or the different discriminative abilities of different features for the final classification. Depending on the different interacting object, the interaction between features can be divided into two situations: 1) the interaction between the current candidate feature and the feature selected according to the aforementioned criteria (measured as Definition 4.3) and 2) the current candidate feature and the remaining candidates' interactions between features (measured as Definition 4.4).

From an information theory perspective, the interaction between the current candidate feature and the selected features can be expressed as the amount of information contributed by adding a new feature to classification when a feature is known. According to Definition 3.7, the conditional mutual information provides a good way to measure these situations. The two interaction situations are defined as follows.

Definition 4.3. (Interaction of selected features, Itr_S): Given a $KNDS = (U, A, D, \delta, K)$, under the condition that the selected feature a_s is known, the neighborhood conditional mutual information of the current candidate feature a_j^A and the decision class d is defined as

$$\begin{aligned} Itr_S\left(a_j^A, a_s, d\right) &= NCMI_\tau\left(a_j^A; d \mid a_s\right) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\tau_{\left\{a_j^A\right\}\cup\{d\}}\left(x_i\right)\right| \cdot \left|\tau_{\{a_s\}\cup\{d\}}\left(x_i\right)\right|}{\left|\tau_{a_s}\left(x_i\right)\right| \cdot \left|\tau_{\left\{a_j^A\right\}\cup\{a_s\}\cup\{d\}}\left(x_i\right)\right|} \end{aligned}$$

where $\tau_{a_s}(x_i)$, $\tau_{\{a_j^A\}\cup\{d\}}(x_i)$, $\tau_{\{a_s\}\cup\{d\}}(x_i)$, and $\tau_{\{a_j^A\}\cup\{a_s\}\cup\{d\}}(x_i)$ represent the neighborhood of x_i on a_s, $\{a_j^A\}\cup\{d\}$, $\{a_s\}\cup\{d\}$, and $\{a_j^A\}\cup\{a_s\}\cup\{d\}$, respectively.

Definition 4.3 is used to measure how much the uncertainty of classification can be reduced by the current candidate feature a_j^A when the selected feature a_s is known, that is, the current contribution of this feature.

Definition 4.4. (Interaction of candidate features, Itr_C): Given a $KNDS = (U, A, D, \delta, K)$, $a_j^A \in A - Red$ is the current candidate feature, and $a_j^{A'} \in A - Red - \{a_j^A\}$ is feature in the remaining candidate feature subset, under the condition that a_j^A is known, the interaction of class dependence between $a_j^{A'}$ and the decision class d is defined as

$$\begin{aligned} &Itr_C\left(a_j^A, a_j^{A'}, d\right) \\ &= NCMI_\tau\left(a_j^{A'}; d \mid a_j^A\right) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\tau_{\{a_j^{A'}\}\cup\{d\}}(x_i)\right|\cdot\left|\tau_{\{a_j^A\}\cup\{d\}}(x_i)\right|}{\left|\tau_{a_j^A}(x_i)\right|\cdot\left|\tau_{\{a_j^{A'}\}\cup a_j^A\}\cup\{d\}}(x_i)\right|} \end{aligned}$$

where $\tau_{a_j^A}(x_i)$, $\tau_{\{a_j^{A'}\}\cup\{d\}}(x_i)$, $\tau_{\{a_j^A\}\cup\{d\}}(x_i)$, and $\tau_{\{a_j^{A'}\}\cup a_j^A\}\cup\{d\}}(x_i)$ represent the neighborhood of x_i on a_j^A, $\{a_j^{A'}\}\cup\{d\}$, $\{a_j^A\}\cup\{d\}$, and $\{a_j^{A'}\}\cup\{a_j^A\}\cup\{d\}$, respectively.

Definition 4.4 can be used to measure the contribution of the feature $a_j^{A'}$ to the final classification in the remaining subset of candidate features given that a_j^A is known.

The third step in measuring the importance of features is that we pick the feature with the maximum interaction with the remaining candidate features. The maximum-interaction criterion is formalized as follows:

$$\max_{a_j^A\in A-Red, a_j^{A'}\in A-Red-\{a_j^A\}} Itr_c\left(a_j^A, a_j^{A'}, d\right).$$

D. Original Feature Selection Algorithm Based on kNN Conditional Mutual Information

The detailed descriptions of the heuristic feature selection algorithm for $KNDS$ are given as algorithm KNCMI. Consequently, based on the discussions of feature correlations in Sections IV-A–IV-C, an original feature objective evaluation function L_{KNCMI} for KNMRmRMI is set up as follows:

$$\mathcal{L}_{\text{KNCMI}} = \arg\max_{a_j^A\in A-Red}\mathcal{J}_{\text{sig}}\left(a_j^A\right)$$

where

$$\begin{aligned} &\mathcal{J}_{\text{sig}}\left(a_j^A\right) \\ &= NMI_\tau\left(a_j^A; d\right) - \frac{1}{|Red|}\sum_{a_s\in Red} NMI_\tau\left(a_j^A; a_s\right) \\ &\quad + \frac{1}{|A-Red|-1}\sum_{a_j^{A'}\in A-Red-\{a_j^A\}} NCMI_\tau\left(a_j^{A'}; d \mid a_j^A\right). \end{aligned}$$

In this function for evaluating feature importance, the neighborhood mutual information is used to measure the relevance between features and classes ($Rel(a_j^A, d)$; see Definition 4.1) and the redundancy between the feature and the selected features ($Rdd(a_j^A, a_s)$; see Definition 4.2). Similarly, the neighborhood conditional mutual information is adopted to characterize the interaction between the feature and the candidate features ($Itr_C(a_j^A, a_j^{A'}, d)$; see Definition 4.4).

Algorithm 1: Feature Selection Algorithm Based on kNN Conditional Mutual Information.

Input: A $KNDS = (U, A, D, \delta, K)$ with $U = \{x_1, x_2, \ldots, x_n\}$ and $A = \{a_1, a_2, \ldots, a_m\}$; The neighborhood adjustment parameter ε (its value range is [0.5, 1.0] in steps of 0.1); neighborhood adjustment parameter k (its value range is $[0.01N, 0.1N]$ in steps of $0.01N$), N is the number of samples.

Output: A reduct feature subset Redbest.

```
for j ← 1 : m do
    Compute ℵ_A^τ; // Referring to Definition 3.1 and 3.2, the radius of the δ-neighborhood and k-nearest neighborhood is obtained.
end for
for each a_j^A ∈ A do
    Compute Rel(a_j^A,d); // Compute the relevance between the current candidate feature a_j^A and the class d.
end for
The feature a_s with the maximum relevance is selected according to the MR criterion;
Red ← {a_s};
A ← A\{a_s};
for each a_j^A ∈ A do
    for each a_s ∈ Red do
        Compute Rdd(a_j^A, a_s); // Compute the redundancy between the current candidate feature a_j^A and the selected feature a_s.//
    end for
    for each a_j^{A'} ∈ A − Red − {A_j^A} do
        calculate Itr_A(a_j^A, a_j^{A'}, d); // Compute the interaction between the current candidate feature a_j^A and the feature a_j^{A'} in the remaining candidate feature subset with regard to the class d.//
    end for
    Compute J_sig(a_j^A); // Compute the significance of the current candidate feature a_j^A according to feature objective evaluation function.//
    Select the feature a with the greatest importance(Referring to J_sig(a_j^A));
    Update Red ← Red ∪ {a};
    A ← A\{a};
end for
The best feature subset Redbest is selected by using the different classifiers; // The Redbest contains fewer features and has higher classification accuracy. //
returnRedbest
```

TABLE I
DESCRIPTION OF THE 20 DATASETS

No.	Datasets	Abbreviation	Samples	Attributes	Classes	Data type
1	Tic-Tac-Toe Endgame	Tic-Tac-Toe	958	9	2	Categorical
2	Zoo	Zoo	101	16	7	Categorical
3	Soybean(Large)	Soy	683	35	19	Categorical
4	Sonar, mines versus rocks	Sonar	208	60	2	Numerical
5	Divorce Predictors	Divorce	170	54	2	Numerical
6	Blood Transfusion Service Center	Blood	748	5	2	Numerical
7	HCV data	Hcv	615	13	5	Numerical
8	Wine	Wine	178	13	3	Numerical
9	Wisconsin diagnostic breast cancer	Wdbc	569	31	2	Numerical
10	Iris	Iris	150	4	3	Numerical
11	Glass Identification	Glass	214	7	7	Numerical
12	Ionosphere	Iono	351	33	2	Numerical
13	Letter	Letter	3349	17	5	Numerical
14	Image Segmentation	segment	210	19	7	Numerical
15	Credit approval	Credit	653	15	2	Hybrid
16	Statlog (German Credit Data)	German	1000	20	2	Hybrid
17	Statlog(heart)	heart	270	13	2	Hybrid
18	Dermatology	derm	366	34	6	Hybrid
19	South German Credit	SGC	1000	21	2	Hybrid
20	Teaching Assistant Evaluation	Tae	151	5	3	Hybrid

The KNMRmRMI feature evaluation function can be applied to measure the effectiveness of a feature or a subset of features for classification, that is to say, the amount of information contribution. It can also be understood as the discriminative power of a feature or a subset of features to distinguish different classes. The greater the information contribution of a feature, the more important this feature is. The purpose of this function is to make the final selected subset of features the most representative and informative and to achieve a tradeoff of relevance, redundancy, and interactivity.

The algorithm mainly includes the following three steps. In the first stage, according to the distributions of attribute values, the kNN of the samples under different features is calculated (steps 1–3). In the second phase, the neighborhood mutual information is employed to measure the relevance between features and classes (steps 4–6) and pairwise redundancy of class independence between the feature and the selected features (steps 11–13). Neighborhood conditional mutual information is adopted to measure the interaction of class dependence between the feature and the candidate features (steps 14–16). Moreover, the significance of features is calculated by the feature objective evaluation function (KNMRmRMI), and the feature with greatest classification performance is selected in order (steps 17–20). In the final step, the best reduct feature subset with the highest amount of information and distinguishing ability is selected via using the wrapper feature selection algorithm (step 22).

Next, we will further analyze the time and space complexity of the KNCMI algorithm. In the first "for" loop (steps 1–3), the computation of object's δ-neighborhood and kNN radius sets has the linear complexity $O(2m)$. And the computational complexity of the neighborhood relation matrix is $O(2mn^2)$. For each feature, in the second "for" loop (steps 4–6), the computational complexity of the relevance between the current candidate feature and the class is $O(m)$. In the next double "for" loop (steps 10–21), the computational complexity is $O(m^3)$. In general, the number of samples is larger than the number of features. To sum up, the time complexity of the algorithm is $O(2mn^2)$ and the space complexity is $O(n^2)$.

V. EXPERIMENTAL ANALYSIS

In this section, a series of comparative experiments are performed. This part uses the KNCMI algorithm to select the appropriate neighborhood radius for different datasets and designs different comparative experiments to prove the efficiency of the KNCMI algorithm in feature selection.

A. Experimental Introduction

To verify the efficiency of the KNCMI algorithm in feature selection, this experiment selects 20 datasets with different dimensions as the experimental objects. The 20 datasets are selected from UCI Machine Learning Repository, including three categorical datasets, 11 numerical datasets, and six hybrid datasets. The descriptions of 20 datasets are shown in Table I.

In order to evaluate the effectiveness and robustness of NMD, the proposed feature selection algorithm KNCMI is compared with existing feature selection algorithms, which will be clear shortly. The comparing algorithm includes some information-theory-based feature selection algorithms and some naive forward feature selection based on NRS, kNNRS. Here we give a brief description of each comparing method.

1) *δ-neighborhood rough set [4]:* This attribute reduction algorithm proposed on NRST works well to reduce numerical and categorical trees in a large number of conditional attributes.
2) *kNNRS [35]:* Focusing on the intersection of the δ-neighborhood and kNN [43], the algorithm has better extraction and classification ability when the feature is heterogeneous data.
3) *An interaction feature selection algorithm based on neighborhood conditional mutual information (NCMI _ IFS) [23]:* The NCMI _ IFS algorithm combines the advantages of the NRS to deal with hybrid and uncertain data and information theory to measure feature correlations, which can achieve higher and more stable classification performance.
4) *Hybrid kernel-based fuzzy complementary mutual information (HKCMI) [39]:* Focusing on fuzzy complementary

TABLE II
NUMBER OF ATTRIBUTES SELECTED UNDER EACH CLASSIFIER AND THE ORDER IN WHICH THE ATTRIBUTES ARE SELECTED

Datasets	knn	svm	nb	The order of feature selection
Tic-Tac-Toe Endgame	6	1	9	5, 4, 7, 1, 9, 3, 8, 6
Zoo	12	12	13	13, 4, 8, 10, 9, 2, 1, 3, 14, 12, 5, 6, 16, 11, 7, 15
Soy	21	20	23	15, 22, 1, 29, 3, 13, 21, 14, 7, 28, 26, 8, 18, 19, 4, 30, 11, 24, 31, 23, 16, 2, 12, 32, 6, 27, 5, 9, 33, 20, 17, 35, 34, 10, 25
Sonar	21	51	51	28, 29, 58, 4, 14, 50, 22, 11, 33, 16, 36, 1, 60, 48, 40, 38, 6, 23, 31, 43, 15, 32, 20, 17, 13, 10, 5, 59, 18, 56, 34, 39, 42, 37, 19, 41, 7, 44, 46, 8, 45, 9, 49, 52, 57, 12, 47, 35, 3, 30, 55, 2, 24, 21, 26, 25, 27, 53, 54, 51
Divorce	17	25	3	9, 45, 32, 3, 50, 42, 34, 49, 51, 43, 1, 2, 4, 5, 6, 7, 41, 37, 44, 26, 31, 52, 13, 54, 53, 48, 47, 21, 46, 8, 10, 11, 12, 14, 15, 16, 33, 17, 18, 19, 20, 22, 23, 39, 40, 38, 24, 25, 27, 28, 29, 30, 35, 36
Blood	4	2	2	2, 4, 1, 3
HCV	11	11	11	10, 1, 7, 6, 12, 11, 3, 5, 9, 4, 8, 2
Wine	10	13	13	5, 6, 4, 11, 1, 8, 9, 12, 7, 10, 13, 2, 3
Wdbc	10	17	18	29, 21, 1, 7, 8, 3, 17, 27, 25, 13, 18, 14, 23, 2, 26, 11, 4, 24, 16, 9, 15, 6, 29, 22, 20, 10, 30, 12, 19, 5
Iris	4	4	1	4, 3, 1, 2
Glass	7	7	6	3, 4, 6, 1, 7, 2, 5
Iono	7	33	10	26, 33, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30, 31, 32, 1
Letter	10	16	16	14, 15, 12, 9, 8, 11, 13, 16, 7, 6, 4, 5, 2, 10, 3, 1
segment	7	7	15	18, 4, 5, 3, 11, 2, 13, 8, 1, 16, 17, 12, 14, 15, 19, 9, 7, 10, 6
Credit	4	13	3	2, 9, 13, 10, 4, 1, 12, 7, 5, 11, 6, 14, 8, 15, 3
German	2	10	8	5, 10, 14, 19, 15, 16, 1, 9, 17, 8, 3, 11, 12, 7, 4
heart	8	9	8	5, 6, 9, 2, 12, 13, 3, 11, 7, 10, 4, 1, 8
derm	34	32	29	21, 34, 28, 14, 16, 3, 4, 19, 2, 23, 18, 1, 17, 32, 5, 6, 7, 8, 9, 10, 11, 33, 12, 13, 15, 20, 22, 24, 25, 26, 30, 27, 29, 31
SGC	20	13	13	5, 20, 10, 19, 18, 14, 17, 16, 15, 7, 6, 12, 4, 8, 9, 1, 2, 11, 3, 13
Tae	4	4	5	5, 1, 4, 3, 2
Avg.	10.95	15	12.85	

TABLE III
COMPARISONS OF AVERAGE CLASSIFICATION ACCURACIES (MEAN $\pm$ STD. DEV. %) OF THE KNN CLASSIFIER ON 20 DATASETS

Datasets	Original data	NRS	kNNRS	NCMI_IFS	HKCMI	KNCMI
Tic-Tac-Toe	74.65±0.27	76.44±1.07	76.87±0.83	73.91±1.52	78.13±0.28	**79.67±0.40**
Zoo	90.19±0.52	95.84±0.72	94.79±1.31	95.96±0.87	89.36±1.74	**97.02±0.76**
Soy	61.33±1.73	71.29±0.43	69.51±0.87	70.75±0.32	63.96±0.48	**72.66±0.45**
Sonar	72.61±1.19	81.57±0.33	80.59±0.56	85.37±0.31	81.95±0.21	**86.56±0.69**
Divorce	94.35±0.17	95.27±0.72	96.34±0.41	97.38±0.42	**97.75±0.23**	96.58±0.63
Blood	67.60±1.32	76.25±0.29	72.18±0.96	75.12±0.54	75.65±0.47	**77.39±0.48**
Hcv	79.89±4.33	88.89±1.27	89.37±0.13	88.23±1.73	87.85±1.06	**90.94±1.48**
Wine	86.63±3.71	93.93±2.70	92.77±3.42	96.12±1.13	89.12±1.71	**96.97±1.74**
Wdbc	82.90±2.86	87.79±2.32	85.37±1.63	89.15±0.98	90.05±0.84	**91.54±1.54**
Iris	90.25±0.12	93.42±0.61	93.01±0.93	94.73±0.54	95.12±0.51	**95.90±0.26**
Glass	63.74±1.49	71.85±0.31	71.14±0.48	**73.81±0.36**	60.93±0.75	72.53±0.37
Iono	77.51±2.26	89.63±1.28	87.25±0.49	88.44±0.62	89.29±0.94	**91.15±0.73**
Letter	81.94±3.36	90.13±1.42	90.88±0.83	93.95±0.76	94.01±1.32	**96.28±0.61**
segment	74.29±1.42	80.28±1.29	**84.57±0.23**	83.13±1.76	80.73±0.84	83.75±0.43
Credit	83.32±3.67	84.73±2.97	85.24±3.37	86.38±2.51	85.67±3.93	**88.09±2.38**
German	67.33±3.58	71.56±0.93	77.12±1.46	75.82±0.34	72.61±0.89	**84.32±0.74**
heart	61.03±0.74	73.61±0.89	71.96±1.81	79.62±1.32	69.46±0.77	**81.14±1.92**
derm	91.61±0.53	94.83±0.91	94.36±0.57	**95.53±0.46**	90.27±0.78	95.30±1.2
SGC	67.26±3.75	81.03±0.25	80.45±0.49	82.48±0.92	**87.16±0.54**	83.17±1.28
Tae	50.97±7.39	45.83±10.32	48.12±9.71	55.62±8.92	53.33±6.64	**61.88±4.63**
Avg.	75.97±2.22	82.21±1.55	82.09±1.52	84.08±1.31	81.62±1.25	**86.14±1.14**

mutual information, HKCMI is suitable for the attribute reduction of multiple attribute types on clustering tasks.

Assess the quality of feature selection results by using the average classification accuracy of three widely used machine learning classifiers, namely, kNN, support vector machine (SVM), and naive Bayes (NB). We used fivefold cross validation in evaluation. First, the fivefold crossover is to randomly divide the original dataset into five subsets, four of which are used as the training set, and the remaining one is used as the test set. The experiment can be performed five times each time with a different subset as the test set. Next, a classifier is trained on the training set using the features selected by the feature selection algorithm. Finally, the performance of the classifier trained on the selected features is evaluated on the test set. We use the average of five test sets as the final classification performance.

B. Experimental Results on Real-World Datasets

Classification performance is considered to be one of the most efficient and straightforward methods to examine the quality of the feature selection algorithm, in which the classification accuracy is usually utilized to measure the classification performance. To avoid the experimental results from being affected by the rareness of data and the randomness of computation, the classification accuracies of the same feature selection algorithm on different datasets are averaged, which is displayed in the row labeled as Avg. After fivefold cross validation in evaluation, the average classification accuracies of the original data on three different classifiers are used as the benchmarks for experimental comparisons. The best classification performance result is highlighted in boldface.

In this section, we conducted a total of two experiments. In first experiment, the five feature selection comparison algorithms and the average classification accuracy of this algorithm on the three classifiers are shown in Tables III–V. Specifically, for these comparison feature selection algorithms, in the light of the range of the parameters offered in the original articles, we revise the homologous parameters to select the highest average classification accuracy as the final evaluation result in the table. The parameters for KNCMI algorithms in the three average classification accuracy tables are set as follows: the parameter ε

TABLE IV
COMPARISONS OF AVERAGE CLASSIFICATION ACCURACIES (MEAN ± STD. DEV. %) OF THE SVM CLASSIFIER ON 20 DATASETS

Datasets	Original data	NRS	kNNRS	NCMI_IFS	HKCMI	KNCMI
Tic-Tac-Toe	83.85±0.41	84.13±0.79	83.22±0.94	84.81±0.56	84.95±0.43	**85.34±0.81**
Zoo	91.81±0.56	95.37±0.66	95.62±0.19	94.85±0.87	95.71±0.42	**96.03±1.42**
Soy	80.49±0.33	82.94±0.31	81.77±0.43	81.30±0.52	**84.86±0.64**	83.37±0.72
Sonar	71.42±1.43	76.81±0.95	74.79±1.64	76.38±1.34	**78.37±0.31**	77.44±1.23
Divorce	95.17±0.36	96.33±0.14	96.19±0.52	97.82±0.36	**98.23±0.61**	97.06±0.24
Blood	75.33±0.56	75.87±0.52	77.02±0.51	76.84±0.96	76.29±0.87	**77.39±0.48**
Hcv	87.51±0.68	90.32±0.68	91.67±0.85	91.89±0.62	90.51±0.94	**92.34±0.93**
Wine	86.43±3.95	96.47±1.52	**97.84±0.95**	97.39±0.86	94.28±1.42	97.69±1.77
Wdbc	91.72±2.46	91.63±0.32	93.58±0.83	94.96±1.49	**95.24±0.65**	94.66±1.82
Iris	96.38±0.47	95.84±0.62	94.92±0.77	96.14±0.27	97.11±0.53	**97.72±0.47**
Glass	70.31±1.37	75.23±2.47	77.29±1.64	81.84±1.22	**88.59±1.60**	80.21±2.61
Iono	80.67±0.20	84.53±0.61	83.24±0.67	**87.83±0.64**	86.42±0.57	82.18±1.51
Letter	87.41±0.49	94.29±0.56	94.62±0.67	95.03±0.83	94.72±0.89	**96.28±0.61**
segment	84.29±1.42	80.28±1.29	84.57±0.23	83.13±1.76	90.23±0.62	**91.50±0.44**
Credit	82.75±3.82	90.04±0.71	90.39±0.62	91.74±0.29	90.17±0.43	**92.85±1.42**
German	66.14±1.43	75.34±0.68	74.09±0.79	77.92±0.85	73.21±0.86	**85.47±0.83**
heart	78.94±1.32	80.15±0.48	79.14±0.63	80.83±0.97	79.75±0.96	**82.94±2.34**
derm	93.32±2.55	94.28±0.74	93.82±0.63	95.36±1.44	93.97±0.75	**95.84±0.82**
SGC	67.26±3.74	81.03±0.25	80.45±0.49	82.48±0.92	**89.14±0.32**	86.38±1.74
Tae	50.12±9.13	49.20±11.38	48.92±9.96	53.74±10.33	55.23±8.09	**56.74±9.51**
Avg.	81.07±1.83	84.50±1.28	84.66±1.17	86.11±1.38	86.85±1.09	**87.47±1.59**

TABLE V
COMPARISONS OF AVERAGE CLASSIFICATION ACCURACIES (MEAN ± STD. DEV. %) OF THE NB CLASSIFIER ON 20 DATASETS

Datasets	Original data	NRS	kNNRS	NCMI_IFS	HKCMI	KNCMI
Tic-Tac-Toe	73.26±1.38	79.28±1.57	78.68±1.63	81.42±0.95	80.19±1.85	**82.62±0.91**
Zoo	90.33±0.67	91.24±0.73	90.51±0.36	93.46±0.51	94.09±0.47	**95.12±1.33**
Soy	80.78±0.75	83.29±0.43	79.37±0.69	91.24±0.62	88.36±0.48	**93.12±0.61**
Sonar	63.38±2.71	75.19±2.94	75.63±1.62	74.26±3.31	76.73±2.18	**77.16±3.26**
Divorce	98.03±0.92	98.41±0.36	98.21±0.52	97.86±0.47	98.42±0.71	**98.57±0.64**
Blood	77.31±0.84	78.74±0.73	78.82±0.63	77.98±0.54	79.65±0.81	**80.18±0.83**
Hcv	92.65±0.53	92.54±0.83	94.58±0.72	94.71±0.46	93.89±0.54	**94.81±0.91**
Wine	96.29±1.69	96.88±1.15	96.14±1.43	97.15±1.13	95.46±1.67	**97.82±1.41**
Wdbc	94.32±1.63	**96.43±0.52**	95.42±0.74	94.84±1.06	95.35±0.67	95.86±1.41
Iris	97.76±0.41	96.19±0.87	94.56±0.48	97.23±0.54	97.36±0.38	**98.72±0.53**
Glass	52.56±6.25	76.44±2.37	77.59±3.01	78.81±2.37	78.69±1.59	**79.19±3.78**
Iono	88.57±1.42	89.53±0.74	90.42±0.62	**92.12±0.71**	90.31±0.49	89.27±1.13
Letter	88.53±1.44	91.56±0.85	92.81±0.99	92.87±0.69	91.39±0.74	**93.78±0.94**
segment	82.53±1.96	84.63±1.31	81.72±1.54	87.29±0.82	90.37±0.78	**92.51±0.36**
Credit	84.69±3.87	86.71±0.68	**91.20±1.04**	85.66±0.74	88.23±0.56	90.74±1.21
German	53.26±0.89	76.57±0.74	77.29±0.56	77.35±0.77	75.42±0.58	**79.53±0.61**
heart	85.43±1.73	86.53±0.90	83.97±1.24	84.21±1.86	87.56±0.94	**88.76±1.52**
derm	87.96±2.63	90.48±0.61	89.87±0.42	88.65±1.41	91.17±0.65	**91.33±0.71**
SGC	70.42±2.15	82.15±3.46	80.78±1.88	**86.97±1.73**	86.15±0.96	85.50±2.67
Tae	51.13±10.39	50.13±10.21	48.99±10.19	52.68±11.17	54.19±8.93	**57.44±9.68**
Avg.	80.46±1.91	85.15±1.60	84.83±1.49	81.91±1.59	86.65±1.29	**88.10±1.72**

is set on 0.6 and the parameter K is set on $0.05N$, where N is the number of samples. In the second experiment, we study the effect of the size of the two parameters on the classification accuracy. The different values of the parameter ε and the different values of the parameter K will lead to the different classification performance.

As can be seen from Table II, several attributes are finally selected by each classifier in each dataset. Moreover, the order of feature selection is also given together. We can find that the least features are selected on average under the kNN algorithm, probably due to the unique categorical nature of the kNN algorithm.

From Tables III–V, we make the following observations. Compared with the original data, the classification performance of the KNCMI algorithm is improved well. The average classification accuracies of the proposed algorithm are improved by 13.39%, 7.89%, and 9.50% than the original data, respectively, on the kNN, SVM, and NB classifiers.

Overall, the proposed KNCMI algorithm outperforms other feature selection algorithms on most datasets. For example, when NB is used as a classifier for testing (as shown in Table V), the KNCMI algorithm achieves the best classification performance on 15 datasets. Moreover, when kNN is used as a classifier for testing (as shown in Table III), the KNCMI algorithm achieves the best classification performance on 17 datasets. Although the classification performance of the proposed algorithm on some datasets is not as good as that of the NCMI _ IFS (such as on Glass, derm, Iono, and SGC datasets) and HKCMI (such as on Divorce, Soy, Sonar, Glass, Wdbc, and SGC datasets) algorithms, it achieves the maximum value of the average classification accuracy.

We can discover that for the kNN, SVM, and NB classifiers, the proposed feature selection algorithm performs better than the other four common feature selection algorithms in most datasets. Furthermore, of the 60 comparisons on the three classifiers, 48 outperformed the other algorithms. When using different classifiers to test the performance of the feature selection algorithm, the results of classification accuracies are slightly different. And they also present different test results on different datasets.

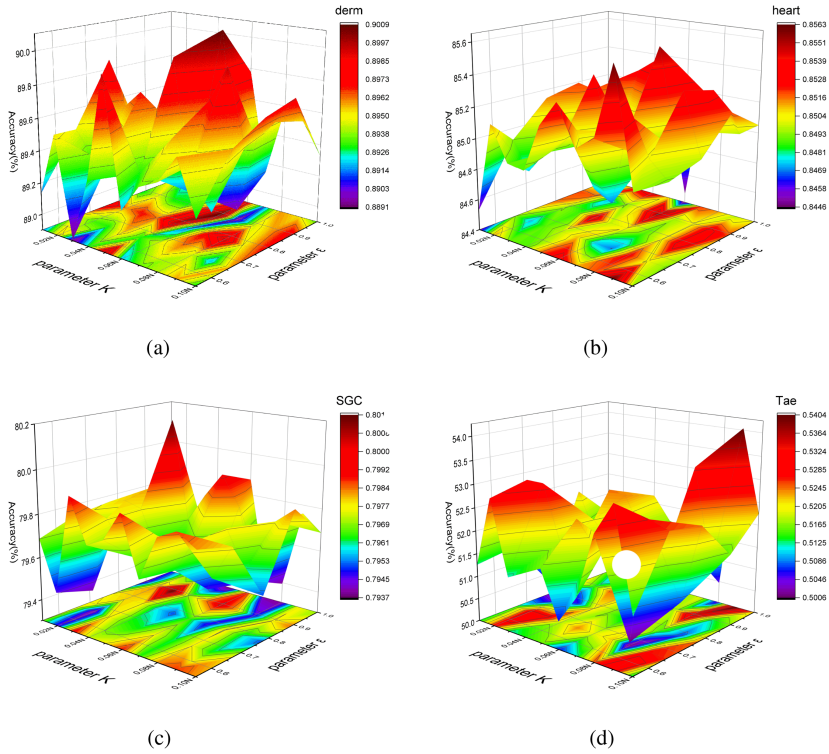

Fig. 6. K-means classification accuracy variations with parameter ε on four hybrid datasets versus the variations with parameter K. (a) derm. (b) heart. (c) SGC. (d) Tae.

For stability analysis, it can be measured by the standard deviation of the accuracy under each classifier. It can be observed that the average standard deviation of the KNCMI algorithm under the kNN classifier is the smallest, which is 1.14%. This shows that the algorithm has strong robustness on the kNN classifier. However, the standard deviation of the SVM and NB classifiers is large, and the stability effect is not good.

To further search the efficiency of our algorithm, Figs. 6 and 7 depict heatmaps of classification accuracy as a function of parameters ε and K on mixed datasets and some representative numerical datasets, respectively. The heatmap results are based on the K-means classifier. The x-axis is the parameter K for the radius of kNNRS. The y-axis is the parameter ε for the δ-neighborhood rough sets, and the z-axis is the classification accuracy of K-means different classifiers.

The values of parameter ε are set from 0.5 to 1.0 in steps of 0.1. The different values of the parameter ε and the different values of the parameter K will lead to the different classification performance. It shows the effectiveness of the proposed algorithm in terms of different values of the parameter.

The changes of values of the parameter ε would relatively impact the classification performances on most datasets, which are shown in Figs. 6 and 7. However, for some datasets, such as Glass, the influences of the parameter K on the classification performances for the K-means classifiers is relatively small, which can be seen from Fig. 7(b).

When the parameters ε and K take different values, the highest classification accuracy corresponding to the K-means classifier is different. For instance, in the derm dataset [see Fig. 6(a)], when the parameter ε is set to 0.9 and the parameter K is set to $0.05N$, the classification accuracy reaches the maximum. In addition, in the Wdbc dataset [see Fig. 7(f)], when the parameter ε is set to 0.6 and the parameter K is set to $0.05N$, the classification accuracy reaches the maximum. Especially, for the Glass datasets [see Fig. 7(b)], the value of parameter K basically does not affect the classification accuracy.

C. Statistical Testing and Analysis of Algorithms

In order to further compare the experimental results of different algorithms, two statistical test methods, i.e., Friedman test and Nemenyi test, were selected to verify the validity of the algorithm comparison.

The Friedman test is a nonparametric statistical test method; its null hypothesis is that all the experimental algorithms have comparable classification performance. The formula is defined as

$$F_F = \frac{(T-1)\chi_F^2}{T(s-1)-\chi_F^2}$$

$$\chi_F^2 = \frac{12T}{s(s+1)}\left(\sum_{i=1}^{s} R_i^2 - \frac{s(s+1)^2}{4}\right)$$

where T and s are the number of experimental datasets and experimental algorithms, respectively, and R_{i} represents the average ranking value of the classification accuracy results of algorithm i on different classifiers.

Table VI represents the five comparison algorithms of the algorithm and the average ranking of the classification accuracy results of the algorithm KNCMI on the kNN, SVM, and NB

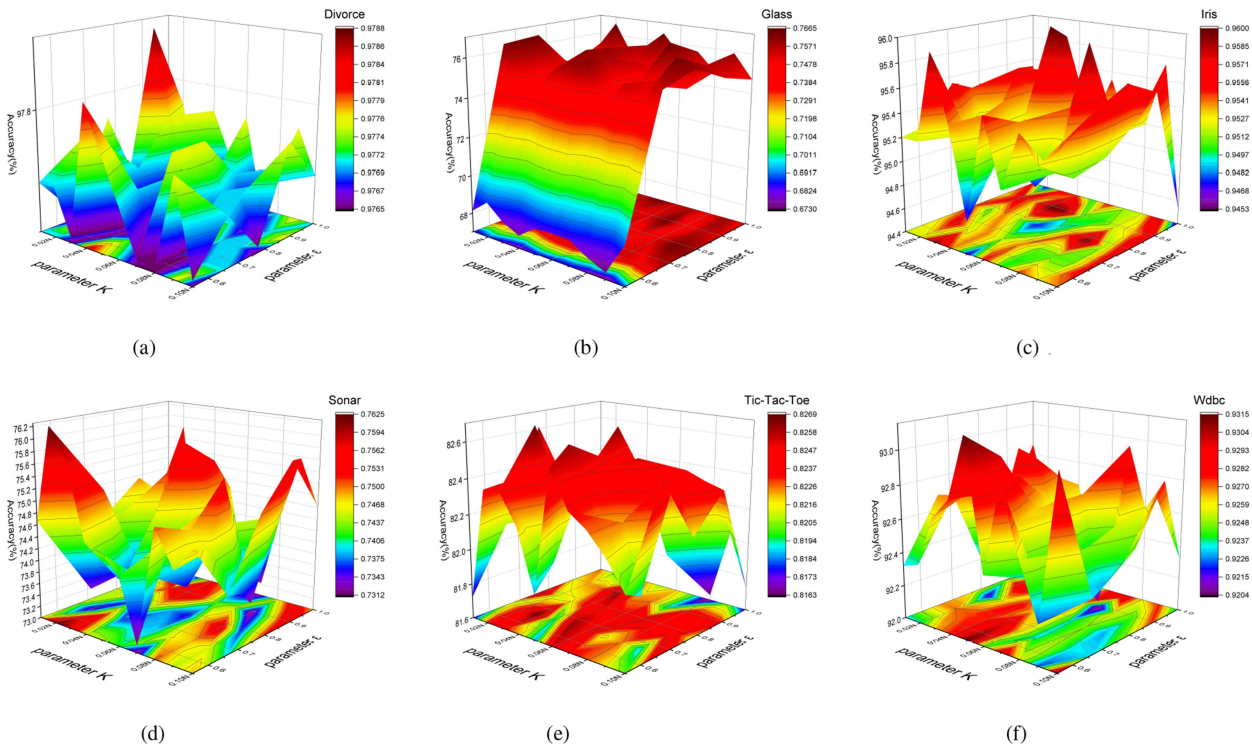

Fig. 7. K-means classification accuracy variations with parameter ε on five numerical and a categorical datasets versus the variations with parameter K. (a) Divorce. (b) Glass. (c) Iris. (d) Sonar. (e) Tic-Tac-Toe. (f) Wdbc.

TABLE VI
AVERAGE RANK OF THE CLASSIFICATION ACCURACY OF THE SIX ALGORITHMS

classifier	Original data	NRS	kNNRS	NCMI_IFS	HKCMI	KNCMI
kNN	5.7	3.75	3.95	2.75	3.55	1.3
SVM	5.5	4.35	4.1	2,85	2.5	1.7
NB	5.25	3.85	4.15	3.25	3.15	1.35

TABLE VII
FRIEDMAN TEST STATISTICS

classifier	kNN	SVM	NB
Chi-square value	64.129	59.939	52.32
P value	0.000000000001699	0.00000000001251	0.0000000004639

classifiers. It can be seen that the ranking of the algorithm in this article is obviously better than that of the comparison algorithm, which shows the superiority of the algorithm in this article.

Table VII shows the chi-square value of the Friedman test and the corresponding P value. It can be seen from Table VII that the test P values on the three classifiers are all less than 0.05, so the null hypothesis is rejected. That is, we believe that there are significant differences between these six algorithms.

VI. CONCLUSION AND FUTURE WORK

NRSs are one of the most important feature selection methods today. Aiming at the two major problems of mixed data and feature interaction, this article proposes a novel feature selection method of NRSs. The proposed KNCMI algorithm can not only solve the problem of imbalanced data distribution, but also consider the role of feature interaction when calculating feature importance. This method can greatly improve the accuracy of feature selection and help us filter out really important features. In the experiment, the accuracy of each classifier and the number of feature selections are compared. The experimental results show that the performance of the algorithm is significantly better than that of the comparison algorithm. Finally, a hypothesis test is carried out, and the experimental results also show that there are obvious differences between the algorithms.

In the following work, we can consider how to improve the existing algorithm to solve the dynamic mixed data problem. This will also be the focus of future development.

REFERENCES

[1] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, pp. 341–356, 1982.
[2] T.-L. Tseng, C.-C. Huang, K. Fraser, and H.-W. Ting, "Rough set based rule induction in decision making using credible classification and preference from medical application perspective," *Comput. Methods Programs Biomed.*, vol. 127, pp. 273–289, 2016.
[3] T. Y. Lin, "Neighborhood systems and approximation inrelational databases and knowledgebases," in *Proc. 4th Int. Symp. Methodol. Intell. Syst.*, 1989, pp. 75–86.
[4] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection," *Inf. Sci.*, vol. 178, pp. 3577–3594, 2008.
[5] Q. Wang, Y. Qian, X. Liang, Q. Guo, and J. Liang, "Local neighborhood rough set," *Knowl.-Based Syst.*, vol. 153, pp. 53–64, 2018.
[6] W. Li, Z. Huang, X. Jia, and X. Cai, "Neighborhood based decision-theoretic rough set models," *Int. J. Approx. Reason.*, vol. 69, pp. 1–17, 2016.
[7] W. Changzhong, S. Mingwen, H. Qiang, Q. Yuhua, and Q. Yali, "Feature subset selection based on fuzzy neighborhood rough sets," *Knowl.-Based Syst.*, vol. 111, pp. 173–179, 2016.

[8] C. Hu, L. Zhang, B. Wang, Z. Zhang, and F. Li, "Incremental updating knowledge in neighborhood multigranulation rough sets under dynamic granular structures," *Knowl.-Based Syst.*, vol. 163, pp. 811–829, 2019.
[9] X. Yang, S. Liang, H. Yu, S. Gao, and Y. Qian, "Pseudo-label neighborhood rough set: Measures and attribute reductions," *Int. J. Approx. Reason.*, vol. 105, pp. 112–129, 2019.
[10] K. Liu, T. Li, X. Yang, H. Ju, X. Yang, and D. Liu, "Hierarchical neighborhood entropy based multi-granularity attribute reduction with application to gene prioritization," *Int. J. Approx. Reason.*, vol. 148, pp. 57–67, 2022.
[11] L. Sun, L. Wang, W. Ding, Y. Qian, and J. Xu, "Neighborhood multigranulation rough sets-based attribute reduction using Lebesgue and entropy measures in incomplete neighborhood decision systems," *Knowl.-Based Syst.*, vol. 192, pp. 105–373, 2020.
[12] S. Luo, D. Miao, Z. Zhang, Y. Zhang, and S. Hu, "A neighborhood rough set model with nominal metric embedding," *Inf. Sci.*, vol. 520, pp. 373–388, 2020.
[13] S. Lin, T. Wang, W. Ding, J. Xu, and Y. Lin, "Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification," *Inf. Sci.*, vol. 578, pp. 887–912, 2021.
[14] D. Liu and J. Li, "Safety monitoring data classification method based on wireless rough network of neighborhood rough sets," *Saf. Sci.*, vol. 118, pp. 103–108, 2019.
[15] X. Chu et al., "Neighborhood rough setbased three-way clustering considering attribute correlations: An approach to classification of potential gout groups," *Inf. Sci.*, vol. 535, pp. 28–41, 2020.
[16] Y. Chen, Z. Zhang, J. Zheng, Y. Ma, and Y. Xue, "Gene selection for tumor classification using neighborhood rough sets and entropy measures," *J. Biomed. Informat.*, vol. 67, pp. 59–68, 2017.
[17] L. Sun, X. Zhang, Y. Qian, J. Xu, and S. Zhang, "Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification," *Inf. Sci.*, vol. 502, pp. 18–41, 2019.
[18] Y. W. Liyanage, D.-S. Zois, and C. Chelmis, "Dynamic instance-wise joint feature selection and classification," *IEEE Trans. Artif. Intell.*, vol. 2, no. 2, pp. 169–184, Apr. 2021.
[19] D. D. Lewis, "Feature selection and feature extraction for text categorization," in *Proc. Workshop Speech Natural Lang.*, 1992, pp. 212–217.
[20] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.
[21] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[22] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, 2004.
[23] J. Wan, H. Chen, Z. Yuan, T. Li, X. Yang, and B. B. Sang, "A novel hybrid feature selection method considering feature interaction in neighborhood rough set," *Knowl.-Based Syst.*, vol. 227, 2021, Art. no. 107167.
[24] Y. Chen, Y. Xue, Y. Ma, and F. Xu, "Measures of uncertainty for neighborhood rough sets," *Knowl.-Based Syst.*, vol. 120, pp. 226–235, 2017.
[25] N. Xie, M. Liu, Z. Li, and G. Zhang, "New measures of uncertainty for an interval-valued information system," *Inf. Sci.*, vol. 470, pp. 156–174, 2019.
[26] C. Gao, Z. Lai, J. Zhou, J. Wen, and W. K. Wong, "Granular maximum decision entropybased monotonic uncertainty measure for attribute reduction," *Int. J. Approx. Reason.*, vol. 104, pp. 9–24, 2019.
[27] X. Zhang, C. L. Mei, and D. G. Chen, Y. Yang, and J. Li, "Active incremental feature selection using a fuzzy-rough-set-based information entropy," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 5, pp. 901–915, May 2020.
[28] J. Grover and M. Hanmandlu, "Development of an optimal entropy classifier and prudent learning model," *IEEE Trans. Artif. Intell.*, vol. 3, no. 2, pp. 164–175, Apr. 2022.
[29] Q. Hu, W. Pan, S. An, P. Ma, and J. Wei, "An efficient gene selection technique for cancer recognition based on neighborhood mutual information," *Int. J. Mach. Learn. Cybern.*, vol. 1, pp. 763–770, 2010.
[30] L. Sun and J. Xu, "Feature selection using mutual information based uncertainty measures for tumor classification," *Bio-Med. Mater. Eng.*, vol. 24, pp. 763–770, 2014.
[31] C. Z. Wang, Y. Wang, M. W. Shao, Y. Qian, and D. Chen, "Fuzzy rough attribute reduction for categorical data," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 5, pp. 818–830, May 2020.
[32] Q. H. Hu, L. J. Zhang, Y. C. Zhou, and W. Pedrycz, "Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 1, pp. 226–238, Feb. 2018.
[33] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *J. Artif. Intell. Res.*, vol. 6, pp. 1–34, 1997.
[34] J. H. Dai, Q. H. Hu, H. Hu, and D. Huang, "Neighbor inconsistent pair selection for attribute reduction by rough set approach," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 937–950, Apr. 2018.
[35] C. Wang, Y. Shi, X. Fan, and M. Shao, "Attribute reduction based on k-nearest neighborhood rough sets," *Int. J. Approx. Reason.*, vol. 106, pp. 18–31, 2019.
[36] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, and Z. Gao, "A hybrid feature selection algorithm for gene expression data classification," *Neurocomputing*, vol. 256, pp. 56–62, 2017.
[37] W. F. Gao, L. Hu, Y. H. Li, and P. Zhang, "Preserving similarity and staring decisis for feature selection," *IEEE Trans. Artif. Intell.*, vol. 2, no. 6, pp. 584–593, Dec. 2021.
[38] P. Maji and S. K. Pal, "Feature selection using f-information measures in fuzzy approximation spaces," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 6, pp. 854–867, Jun. 2010.
[39] S. Patra, P. Modi, and L. Bruzzone, "Hyperspectral band selection based on rough set," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5495–5503, Oct. 2015.
[40] D. Wang, F. Nie, and H. Huang, "Feature selection via global redundancy minimization," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 10, pp. 2743–2755, Oct. 2015.
[41] S. Fernndez, J. Trinidad, and J. Ochoa, "A supervised filter feature selection method for mixed data based on spectral feature selection and information-theory redundancy analysis," *Pattern Recognit. Lett.*, vol. 138, pp. 321–328, 2020.
[42] J. Wan, H. Chen, T. Li, Z. Yuan, J. Liu, and W. Huang, "Interactive and complementary feature selection via fuzzy multigranularity uncertainty measures," *IEEE Trans. Cybern.*, vol. 53, no. 2, pp. 1208–1221, Feb. 2023.
[43] Q. Hu, J. Liu, and D. Yu, "Mixed feature selection based on granulation and approximation," *Knowl.-Based Syst.*, vol. 21, pp. 294–309, 2008.

Weihua Xu received the M.Sc. degree in mathematics from the School of Mathematics and Information Sciences, Guangxi University, Nanning, China, in 2004, and the Ph.D. degree in mathematics from the School of Sciences, Xi'an Jiaotong University, Xi'an, China, in 2007.

He is currently a Professor with the College of Artificial Intelligence, Southwest University, Chongqing, China. He is on the Editorial Boards of several international journals. He has authored or coauthored four monographs and more than 180 articles in international journals. His current research interests include granular computing, approximate reasoning, fuzzy sets, rough sets, concept lattices, cognitive computing, and evolutionary computing.

Ziting Yuan received the B.Sc. degree in statistics from the School of Science and Technology Beijing, Beijing, China, in 2021. She is currently working toward the M.Sc. degree in statistics with the School of Artificial Intelligence, Southwest University, Chongqing, China.

Her research interests include feature selection and granular computing.

Zheng Liu is working toward the B.Sc. degree in intelligent science with Southwest University, Chongqing, China.

His research interests include feature selection and granular computing.