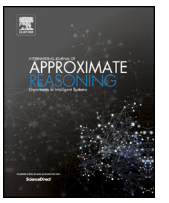
International Journal of Approximate Reasoning 172 (2024) 109218

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

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



Incremental reduction of imbalanced distributed mixed data based on *k*-nearest neighbor rough set

Weihua Xu *, Changchun Liu

College of Artificial Intelligence, Southwest University, Chongqing 400715, PR China

ARTICLE INFO

Keywords:
Incremental feature selection
Neighborhood mutual information
Neighborhood rough set
Unbalanced mixed-type data

ABSTRACT

Incremental feature selection methods have garnered significant research attention in improving the efficiency of feature selection for dynamic datasets. However, there is currently a dearth of research on incremental feature selection methods specifically targeted for unbalanced mixed-type data. Furthermore, the widely used neighborhood rough set algorithm exhibits low classification efficiency for imbalanced data distribution and performs poorly in classifying mixed samples. Motivated by these two challenges, we investigate the use of an incremental feature reduction algorithm based on *k*-nearest neighbors and mutual information in this study. Firstly, we enhance the capabilities of the neighborhood rough set model by incorporating the concept of *k*-nearest neighbors, thereby improving its ability to handle samples with varying densities. Subsequently, we apply information entropy theory and combine neighborhood mutual information with the maximum relevance minimum redundancy criterion to construct a novel feature importance evaluation function. This function is utilized as the evaluation metric for feature selection. Finally, an incremental feature selection algorithm is designed based on the above static algorithm. Experiments were conducted on twelve public datasets to evaluate the robustness of the proposed feature metrics and the performance of the incremental feature selection algorithm. The experimental results validated the robustness of the proposed metrics and demonstrated that our incremental algorithm is effective and efficient in feature reduction for updating unbalanced mixed data.

1. Introduction

In today's age of expansive information growth, the swift progress of big data has led to escalated requirements for handling intricate and high-dimensional data. In the context of data feature extraction, the presence of redundant or irrelevant high-dimensional data presents a significant challenge. Rough set theory is a specialized theory that deals primarily with the analysis of incomplete data, the representation of imprecise knowledge, and the processes of learning and generalization. The primary aim is to extract decision or classification rules for a problem by using knowledge parsimony, without compromising the classification ability. Therefore, it has some advantages in dealing with vague and imprecise data. Firstly, it is crucial to note that the theory solely depends on the raw information. As a result, it provides a higher level of objectivity when confronted with uncertain situations. Secondly, attribute simplification keeps its classification and decision-making power intact while removing irrelevant and redundant features.

* Corresponding author.
E-mail addresses: chxuwh@gmail.com (W. Xu), 1559901036@qq.com (C. Liu).

https://doi.org/10.1016/j.ijar.2024.109218
Received 21 November 2023; Received in revised form 10 March 2024; Accepted 26 May 2024

Available online 4 June 2024
0888-613X/© 2024 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Extensive research has been carried out both domestically and internationally on the topic of data feature selection. High-dimensional datasets usually have more noisy and irrelevant features, and the determination of the ideal characteristic subset from a provided dataset is considered a crucial area of exploration for numerous training applications. By reviewing the previous theories and literature, we can clearly understand the development and improvement of rough set theory.

Rough sets [1] is a theory proposed by Pawlak that is employed to analyze the representation, learning, induction, and other aspects of incomplete data and imprecise knowledge. However, with the advent of the data era, data tends to exist in mixed types, no longer only in character types in the past. The conventional rough set theory results in incomplete and unreliable data when dealing with mixed data types. Lin's [2] proposal of the concept of neighborhood rough set (NRS) in the literature can improve this problem. Hu [3] proposed a neighborhood rough set model that is more suitable for mixed data. Its main idea is to establish the characteristics of the samples by dividing them by the distance between them and the neighborhood δ. It can handle numerical and mixed data effectively, thus avoiding the problems caused by discretization. On the basis of this, numerous scholars have achieved considerable improvement. Qian et al. [4] proposed the local rough set model, one of the newer models in the field of rough set theory. It does not need to approximate the target concept by all the objects in the argument domain, but only the objects in the target concept need to be considered. This effectively reduces the computation time of the approximation operator. Wang et al. [5] utilized a semi-supervised approach to attribute approximation with this model, greatly reducing the complexity of the algorithm. Zhang et al. [6] proposed a generalized MG-DTRS model called adaptive multi-granularity decision theory rough set (AMG-DTRS), which overcomes the inherent weaknesses of the MG-DTRS model by adaptively obtaining a pair of probability thresholds through compensation coefficients. Yin et al. [7] proposed parameterized multi-label fuzzy coverage relationships and fuzzy coverage entropy measures, and constructed a robust multi-label feature selection (RMSMC) model that considers feature multi-correlation. This model can effectively capture the intrinsic information of multi-label data. In addition, many other types of neighborhood rough sets have been extensively studied. For instance pseudoscalar neighborhood rough sets [8,9], neighborhood rough sets with nominal metric embeddings [10,11], neighborhood multigranularity rough sets [12–14], etc. The NRS based theory has been widely applied in feature selection [15–17], multi-label feature selection [18–20], hyperspectral classification [21,22], image annotation [23] and credit rating [24]. Thus, NRS can efficiently handle mixed-type data and it is the theoretical basis of this paper.

Information theory plays a crucial role in handling uncertain information by effectively quantifying and processing various forms of uncertainty through the concepts of entropy, probabilistic models, coding strategies, and error correction techniques [25]. It finds wide applications in domains such as data compression, communication systems, and machine learning. Pawlak [26] proposed 3 uncertainty measures, namely accuracy, roughness, and approximate accuracy. Over the past few years, numerous scholars have achieved significant advancements in this field. Liang et al. [27] propose precision, coarseness, and approximate precision based on knowledge granularity by introducing knowledge granularity. Wang [28] unveiled the limitations of the fine-set model in the probabilistic phase and proposed an associated monotonic uncertainty measure, thereby laying a robust foundation for attribute reduction. Shu et al. [29] introduced an incremental feature selection algorithm that considers the increase in the number of samples. Each of the above methods represents a clear advance in neighborhood-based algorithmic technology entropy. The neighborhood entropy theory will also be an important theoretical basis of this article.

Another significant theoretical foundation of this paper is mutual information, which is a vital method for quantifying the differentiation of attribute features and has garnered considerable research attention from scholars in recent years. Hu et al. [30] introduced the concepts of neighborhood entropy, neighborhood conditional entropy, and neighborhood mutual information. They used neighborhood mutual information to assess the correlation between features and decision attributes. The application of mutual information has been enhanced by Lin et al. [20], who introduced an improved algorithm specifically designed for feature metrics in the context of multi-label data. However, the aforementioned approaches overlook the interactions and dependencies among attributes. To address this limitation, Wan et al. [31] proposed the incorporation of feature interaction by considering it within the context of a neighborhood rough set. The algorithm ensures the stability and reliability of the neighborhood rough set by constructing a multi-neighborhood radius set that incorporates mixed data. Moreover, this study introduces a novel feature objective evaluation function, known as MRmRMI, which is subsequently utilized in the feature selection algorithm. Experimental results demonstrate that the proposed approach exhibits superior performance in hybrid data feature classification. Notably, the model effectively reduces information loss while simultaneously improving the classification accuracy of the data. Building upon these findings, Xu [16] proposes a feature selection method for unbalanced distributed mixed data based on the *k*-nearest neighbor rough set. This model integrates neighborhood delta and *k*-nearest neighbors to enable effective feature selection on datasets that are unevenly distributed. However, none of the aforementioned approaches address the challenge of efficiently performing attribute approximation as the data sample size increases. Given the algorithmic cost challenges arising from unbalanced distributions and the increasing amount of mixed data, there is an immediate need for an incremental feature selection approach that effectively tackles the interaction between these data types. With this research's primary focus on addressing this issue, the paper aims to make the following key contributions:

1) We have designed a neighborhood rough set model that incorporates k-nearest neighbors, the proposed model combines the advantages of both *k*-nearest neighbors and δ-neighborhoods to effectively handle unbalanced mixed data distributions. The strategies of this model are in line with the human way of thinking and correspond to the requirements of the practical application.
2) By combining mutual information with the principles of maximum relevance and minimum redundancy, we design a robust feature importance assessment function. This function is the basis of the feature selection method and heuristic feature selection strategy in this paper.

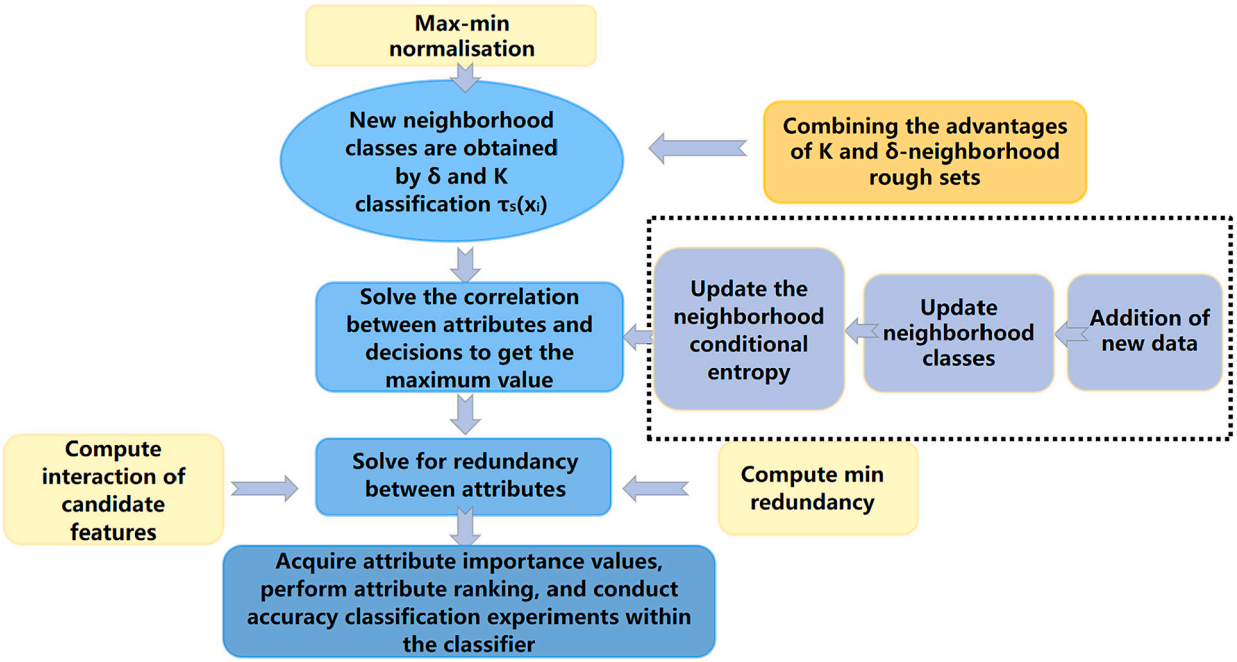


Fig. 1. Flowchart of the algorithm model in this paper.

3) Based on the above static algorithm, we design an incremental feature selection algorithm (IFS-KNCMNI) for improving the efficiency of feature selection for unbalanced mixed data.
4) Experiments were conducted on 12 UCI datasets to evaluate the performance of the proposed algorithms in terms of classification accuracy, computation time, and number of feature reductions. By comparing our algorithm with five existing algorithms, the superiority of our approach was demonstrated.

This paper follows a structured organization. Section 2 provides the theoretical foundations of rough sets, *k*-nearest neighbor rough set, and information theory. Section 3 introduces various characteristic measures of *k*-nearest neighbor rough set. Section 4 presents both the non-incremental and incremental algorithms. Section 5 includes the experimental setup, analysis of data results, and hypothesis testing. Section 6 concludes the paper and outlines future perspectives. Finally, the detailed flow of the paper is illustrated in Fig. 1.

2. Related work

In this section, a comprehensive discussion is provided on the fundamental concepts that underpin this research, encompassing rough set theory, neighborhood rough set theory, and uncertainty measurement.

2.1. *k*-nearest neighbor rough set

When dealing with incomplete and uncertain information, rough set theory serves as a mathematical method that operates solely based on the available data. It does not rely on any prior information beyond the provided data, making it an effective approach for handling such scenarios. The concept was initially introduced by the Polish scholar Pawlak in 1982. Its primary objective is to utilize established information or knowledge to estimate uncertain or imprecise target concepts. Later, scholars extended the traditional rough set theory and introduced the neighborhood rough set(NRST) as a means to overcome its limitations and enhance its capabilities.

Consider a decision table represented by $DT = (U, F, D)$, where the universe U is a non-empty and finite set of objects, denoted as $\{x_1, x_2, ..., x_n\}$. The set of conditional attributes F is also a finite and non-empty set, represented as $\{f_1, f_2, ..., f_m\}$. Similarly, the set of decision attributes is denoted as $D = \{d_1, d_2, ..., d_r\}$ and is also a non-empty and finite set.

The similarity relation that arises from a neighborhood decision table $NDT = (U, F, D, \delta)$ with $S \subseteq F$ can be defined as follows:

$$NSR_\delta(S) = \{(x_i, x_j) \in U \times U \mid d_S(x_i, x_j) \leq \delta\}.$$

The neighborhood class of an element x in the neighborhood decision table $NDT = (U, F, D, \delta)$ with $S \subseteq F$ is defined as follows:

$$\delta_S(x_i) = \{(x_i, x_j) \in U \times U \mid d_S(x_i, x_j) \leq \delta\},$$

where δ is the neighborhood parameter,and $\delta \in [0,1]$. If all samples within the $\delta_S(x_i)$ neighborhood exhibit the same decision values, x_i is deemed consistent within the δ neighborhood. Conversely, if the decision values of the samples within the neighborhood vary, x_i is classified as an inconsistent sample. In this context, the distance between two samples, denoted as $d_S\left(x_i, x_j\right)$, is used as a measure to assess the similarity between samples. The Euclidean distance formula is commonly employed in determining the distance between two data points. The formula can be expressed as follows:

$$d_S\left(x_i, x_j\right) = \sqrt{\sum_{f \in S} \left(f\left(x_i\right) - f\left(x_j\right)\right)^2}.$$

NRST has three properties that apply to the distance metric $d_S\left(x_i, x_j\right)$.

(1) $\forall x, y \in U, d_S(x,y) \geq 0, d_S(x,y) = 0 \; if \; and \; only \; if \; x = y$;
(2) $\forall x, y \in U, d_S(x,y) = d_S(y,x)$;
(3) $\forall x, y, z \in U, d_S(x,z) \leq d_S(x,y) + d_S(y,z)$.

As the distance function of NRST, $d_S(x,y)$ must satisfy non-negativity, symmetry, and triangular inequality.
For any x in the neighborhood decision table $NDT = (U, F, D, \delta)$, and given $S \subseteq F$, $k_S(x_i)$ is defined similarly to $\delta_S(x_i)$.

$$k_S(x_i) = \left\{x_i^1, x_i^2, ..., x_i^n \mid d_S(x_j, x_i) > d_S(x_i^a, x_i), x_j \neq x_i^b, b, a = 1, 2, ..., n\right\}.$$

When using k-nearest neighbors (KNN) and formulas involving $k_S(x_i)$, the formula refers to the set of n samples that are closest to a given point x_i. The set $k_S(x_i)$ always includes a fixed number of samples, which is denoted as k. Furthermore, it is necessary to define the binary relation k_S:

$$k_S = \left\{(x_i, x_j) \in U \times U \mid x_j \in k_S(x_i)\right\}.$$

The information granularity based on k-nearest neighbors is defined by the upper approximation, lower approximation, and boundary domain, as follows:

(1) $\overline{K}_S(D_j) = \left\{x_i \in U \mid k_S(x_i) \cap D_j \neq \phi\right\}$;
(2) $\underline{K}_S(D_j) = \left\{x_i \in U \mid k_S(x_i) \subseteq D_j\right\}$;
(3) $KR_S(D_j) = \overline{K}_S(D_j) - \underline{K}_S(D_j)$.

In this context, D_j represents the partitioning of decision classes. Next, the positive domain of decision D and the dependency on attribute S are defined as follows:

(1) $KPOS_S(D) = \bigcup_{D_j \in U/D} \underline{K}_S(D_j)$;
(2) $\gamma_S^k(D) = \left|KPOS_S(D)\right| / |U|$.

The k-nearest neighbor rough set model is the name given to the rough approximation described by the equation mentioned earlier. After introducing the basic theory, the next section will cover the utilization of information theory in k-nearest neighbor rough set.

2.2. Information theory in neighborhood decision table

Mutual information [16] is widely recognized for its robustness in noisy data environments, which makes it a valuable information measurement in neighborhood decision table. Here are some other information measurements commonly used in NDT. In a neighborhood decision table $NDT = (U, F, D, \delta)$, where $\delta \geq 0$ and S is any subset of attributes from F, the neighborhood relation created by the attribute subset S is denoted as NR_S^δ. The $\delta_S(x_i)$ represents the neighborhood class for every x_i belonging to U. The definition of neighborhood entropy for the set of samples related to attribute subset S is as follows:

$$NE_\delta(S) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{\left|\delta_S(x_i)\right|}{|U|}.$$

The average level of uncertainty for a sample is expressed as:

$$NE_\delta^{x_i}(S) = -\log_2 \frac{\left|\delta_S(x_i)\right|}{|U|}.$$

In addition to the fundamental concept of information entropy, we have extended the notion of neighborhood entropy by incorporating the principles of joint entropy and conditional entropy from information theory. This expansion allows us to capture more comprehensive and intricate relationships within the neighborhood context. In a neighborhood decision table $NDT = (U, F, D, \delta)$, where $\delta \geq 0$ and for any subsets S_1 and S_2 from the set F. The neighborhood relationships determined by the conditional attribute

sets S_1 and S_2 are $N_{S_1}^{\delta}$ and $N_{S_1 \cup S_2}^{\delta}$, respectively. Similarly, the neighborhood classes under the neighborhood relations $N_{S_1}^{\delta}$ and $N_{S_1 \cup S_2}^{\delta}$ are δ_{S_1} and $\delta_{S_1 \cup S_2}$. The joint entropy of the neighborhoods of S_2 and S_1 is defined as:

$$NE_{\delta}(S_1, S_2) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{\left|\delta_{S_1 \cup S_2}(x_i)\right|}{|U|}.$$

When the information entropy of S_2 is known, the neighborhood conditional information entropy of attribute set S_1 with respect to S_2 can be defined as follows:

$$NE_{\delta}(S_1 \mid S_2) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{\left|\delta_{S_1 \cup S_2}(x_i)\right|}{\left|\delta_{S_2}(x_i)\right|}.$$

Based on the entropy relationship in information theory, we have derived the relationship formulas for the three types of entropy.

Proposition 2.1. *For a given $NDT = (U, F, D, \delta)$, for $\delta \geq 0, \forall S_1, S_2 \in F$. Then $NE_{\delta}(S_1 \mid S_2) = NE_{\delta}(S_1, S_2) - NE_{\delta}(S_2)$.*

Proof. According to the previous existing formula, we can get:

$$\begin{aligned}
&NE_{\delta}(S_1, S_2) - NE_{\delta}(S_2) \\
&= -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{\left|\delta_{S_1 \cup S_2}(x_i)\right|}{|U|} + \frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{\left|\delta_{S_2}(x_i)\right|}{|U|} \\
&= -\frac{1}{|U|} \sum_{i=1}^{|U|} \left(\log_2 \frac{\left|\delta_{S_1 \cup S_2]}(x_i)\right|}{|U|} - \log_2 \frac{\left|\delta_{S_2}(x_i)\right|}{|U|} \right) \\
&= -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{\left|\delta_{S_1 \cup S_2}(x_i)\right|}{\left|\delta_{S_2}(x_i)\right|} \\
&= NE_{\delta}(S_1 \mid S_2).
\end{aligned}$$

As demonstrated in the previous proof process, the calculation of the uncertainty measure for attribute subset S_1 requires the utilization of both the joint information entropy of S_1 and S_2 and the neighborhood information entropy of S_2. By incorporating these two entropy measures, we can effectively capture the interplay and information dynamics between the two subsets. This approach allows us to comprehensively assess the level of uncertainty or information content associated with attribute subset S_1, considering its relationship with S_2 and the overall neighborhood context.

Mutual information serves as a metric to evaluate the correlation and dependence between two random variables. It is calculated by comparing the joint probability distribution with the respective marginal probability distributions of the variables [32]. Just like the previous expansion of entropy, similar reasoning can also be done with mutual information. These extensions provide a more comprehensive understanding of the relationships and dependencies between random variables. In a neighborhood decision table $NDT = (U, F, D, \delta)$, where $\delta \geq 0$ and for any subsets S_1 and S_2 from the set F. Similarly, the sets of attributes S_1 and S_2 determine the neighborhood relationships $N_{S_1}^{\delta}$ and $N_{S_2}^{\delta}$, respectively. Under the neighborhood relations $N_{S_1}^{\delta}$ and $N_{S_2}^{\delta}$, the neighborhood classes are denoted as δ_{S_1} and δ_{S_2}. The symbol $\delta_{S_1 \cup S_2}(x_i))$ denotes the neighborhood of x_i within the attribute set $S_1 \cup S_2$ with a radius of δ. The definition of the neighborhood mutual information between S_1 and S_2 is as follows:

$$NMI_{\delta}(S_1; S_2) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{\left|\delta_{S_1}(x_i)\right| \cdot \left|\delta_{S_2}(x_i)\right|}{|U| \left|\delta_{S_1 \cup S_2}(x_i)\right|}.$$

We give three relevant formulas and the proof process.

Proposition 2.2. *For a given $NDT = (U, F, D, \delta)$, for $\delta \geq 0, \forall S_1, S_2 \in F$. Then the following equation holds:*

(1) $NMI_{\delta}(S_1; S_2) = NMI_{\delta}(S_2; S_1)$;
(2) $NMI_{\delta}(S_1; S_2) = NE_{\delta}(S_1) + NE_{\delta}(S_2) - NE_{\delta}(S_1, S_2)$;
(3) $NMI_{\delta}(S_1; S_2) = NE_{\delta}(S_1) - NE_{\delta}(S_2 \mid S_1) = NE_{\delta}(S_2) - NE_{\delta}(S_1 \mid S_2)$.

Proof. (1) Based on the NMI formula provided above, we can obtain that:

$$\begin{aligned}
&NMI_\delta(S_1;S_2)\\
&=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\delta_{S_1}(x_i)\right|\cdot\left|\delta_{S_2}(x_i)\right|}{|U|\left|\delta_{S_1\cup S_2}(x_i)\right|}\\
&=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\delta_{S_2}(x_i)\right|\cdot\left|\delta_{S_1}(x_i)\right|}{|U|\left|\delta_{S_2\cup S_1}(x_i)\right|}\\
&=NMI_\delta(S_2;S_1).
\end{aligned}$$

(2) Based on the NE formula provided above, we can obtain that:

$$\begin{aligned}
&NE_\delta(S_1)+NE_\delta(S_2)-NE_\delta(S_1,S_2)\\
&=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\delta_{S_1}(x_i)\right|}{|U|}-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\delta_{S_2}(x_i)\right|}{|U|}-\left(-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\delta_{S_1\cup S_2}(x_i)\right|}{|U|}\right)\\
&=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\left(\frac{\left|\delta_{S_1}(x_i)\right|}{|U|}\cdot\frac{\left|\delta_{S_2}(x_i)\right|}{|U|}\cdot\frac{|U|}{\left|\delta_{S_1\cup S_2}(x_i)\right|}\right)\\
&=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\delta_{S_1}(x_i)\right|\cdot\left|\delta_{S_2}(x_i)\right|}{|U|\left|\delta_{S_1\cup S_2}(x_i)\right|}\\
&=NMI_\delta(S_1;S_2).
\end{aligned}$$

(3) Combining Proposition 2.1 and Proposition 2.2(2) can be obtained as established.

In a neighborhood decision table $NDT=(U,F,D,\delta)$, where $\delta\geq 0$ and for any subsets S_1, S_2 and S_3 from the set F. The neighborhood conditional mutual information entropy of S_1 and S_3 under the attribute set S_2 is defined as:

$$NCMI_\delta(S_1;S_3\mid S_2)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\delta_{S_1\cup S_2}(x_i)\mid\cdot\mid\delta_{S_3\cup S_2}(x_i)\right|}{\left|\delta_{S_2}(x_i)\mid\cdot\mid\delta_{S_1\cup S_2\cup S_3}(x_i)\right|}.$$

Proposition 2.3. *For a given $NDT=(U,F,D,\delta)$, for $\delta\geq 0,\forall S_1,S_2,S_3\in F$. Then $NCMI_\delta(S_1;S_3\mid S_2)=NE_\delta(S_1,S_2)+NE_\delta(S_3,S_2)-NE_\delta(S_1,S_3,S_2)-NE_\delta(S_2)$.*

Proof.

$$\begin{aligned}
&NE_\delta(S_1,S_2)+NE_\delta(S_3,S_2)-NE_\delta(S_1,S_3,S_2)-NE_\delta(S_2)\\
&=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\delta_{S_1\cup S_2}(x_i)\right|}{|U|}-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\delta_{S_3\cup S_2}(x_i)\right|}{|U|}\\
&+\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\delta_{S_1\cup S_3\cup S_2}(x_i)\right|}{|U|}+\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\delta_{S_2}(x_i)\right|}{|U|}\\
&=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\left(\frac{\left|\delta_{S_1\cup S_2}(x_i)\right|}{|U|}\cdot\frac{\left|\delta_{S_3\cup S_2}(x_i)\right|}{|U|}\cdot\frac{|U|}{\delta_{S_1\cup S_3\cup S_2}(x_i)}\cdot\frac{|U|}{\left|\delta_{S_2}(x_i)\right|}\right)\\
&=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\delta_{S_1\cup S_2}(x_i)\mid\cdot\mid\delta_{S_3\cup S_2}(x_i)\right|}{\left|\delta_{S_2}(x_i)\mid\cdot\mid\delta_{S_1\cup S_2\cup S_3}(x_i)\right|}\\
&=NCMI_\delta(S_1;S_3\mid S_2).
\end{aligned}$$

The formula introduced in this section reflects that neighborhood mutual information entropy (NMI) represents the degree of interdependence between two attribute sets, and neighborhood conditional mutual information entropy ($NCMI$) can reflect the degree of independence between two attribute sets.

3. Information measure of *k*-nearest neighbor rough set

This section presents the information entropy calculation method in the context of combining *k*-nearest neighbor rough sets and δ neighborhood rough sets. Firstly, we introduce the distance formula calculation method for mixed data. Secondly, we introduce the formula for information entropy in this context. Finally, we discuss the degree of interaction and dependence between features by integrating a formula for evaluating feature importance.

3.1. Distance measurement under hybrid data

Real-world applications often involve diverse types of data, such as numerical, categorical, and mixed data. To handle this heterogeneity and perform intersection operations in neighborhood relations, the Heterogeneous Chebyshev-Overlap Metric (HCOM) is defined. HCOM enables the calculation of distances between instances with different types of attributes, facilitating the analysis and processing of mixed data in neighborhood-based operations. The HCOM distance function expression for mixed-type data is as follows:

$$HCOM_F(x,y) = \sum_{i=1}^{|F|} \left(d_{f_i}^{\tau}(x,y) \right)^{\frac{1}{\tau}},$$

where $d_i(x,y) = \begin{cases} \left| f(x,f_i) - f(y,f_i) \right|, & \text{if } f_i \text{ represents a numerical feature;} \\ 1, & \text{if } f_i \text{ represents a categorical feature and } f(x,f_i) \neq f(y,f_i); \\ 0, & \text{if } f_i \text{ represents a categorical feature and } f(x,f_i) = f(y,f_i); \\ 1, & \text{if the feature value of either } x \text{ or } y \text{ is unknown with respect to } f_i. \end{cases}$

In particular, when f_i is a numerical feature, the above formula can be simplified to $HCOM_F(x,y) = d_f(x,y)$, where $\tau = +\infty$, based on the distance function described in section 2.1. This simplification allows for a direct calculation of the distance between instances x and y using the specific numerical distance function. The neighborhood class of instance x in relation to set S can be characterized as:

$$\delta_S(x_i) = \left\{ x_i \in U \mid HCOM_S(x_i, x_j) \leq \delta \right\}.$$

We know that the size of the neighborhood δ in NRS (Neighborhood Rough Set) has an impact on classification accuracy. The objective of classification is to minimize the dissimilarities among samples belonging to the same class while increasing the dissimilarities between samples from different classes, thereby achieving accurate classification. Nevertheless, when dealing with complex samples, the basic definition of neighborhood δ in NRS may produce insufficient classification results. Consequently, selecting the appropriate granularity becomes crucial for the classification model.

3.2. Calculation of k-nearest neighbor rough set entropy

Let's consider a *k*-nearest neighborhood decision table, denoted as $KNDT = (U, F, D, \triangle, \delta, K)$, where $\forall S \subseteq F$, the universe U is defined as the set of all instances or objects, denoted as $U = \left\{ x_1, x_2, \ldots x_n \right\}$. F denotes the feature set, consisting of attributes that describe the instances, denoted as $F = \left\{ f_1, f_2, \ldots f_m \right\}$. D represents the decision set, comprising possible decision classes, denoted as $D = \left\{ d_1, d_2, \ldots d_r \right\}$. $\triangle$ signifies the distance or similarity measure used to calculate the distance between instances. δ represents the neighborhood parameter, which determines the size of the neighborhood radius, and $\delta \in [0,1]$. K indicates the number of nearest neighbors used in the classification process. The neighborhood of an instance x_i, which belongs to the universe U, with respect to set S, can be described as:

$$\tau_S(x_i) = \left\{ x_j \in U \mid x_j \in \delta_S(x_i) \cap k_S(x_i) \right\},$$

where $\tau_S(x_i)$ is the intersection of $\delta_S(x_i)$ and $k_S(x_i)$, and combining the advantages of both. Similar to the expansion principles of entropy discussed in section 2, entropy is also expanded in the context of the *k*-nearest neighbor rough set. For a given $KNDT = (U, F, D, \triangle, \delta, K)$, where $\forall S \subseteq F$, the neighborhood information entropy of the sample set with respect to set S is defined as follows:

$$NE_{\tau}(S) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{\left| \tau_S(x_i) \right|}{|U|}.$$

In the case of $KNDT$, where S_1 and S_2 are subsets of the attribute set F, the neighborhood joint entropy of S_1 and S_2 can be defined as follows:

$$NE_{\tau}(S_1, S_2) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{\left| \tau_{S_1 \cup S_2}(x_i) \right|}{|U|}.$$

In particular, when the information entropy of S_2 is known, the conditional entropy of S_1 relative to S_2 can be mathematically expressed as:

$$NE_\tau(S_1 \mid S_2) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\tau_{S_1\cup S_2}(x_i)\right|}{\left|\tau_{S_1}x_i\right|}.$$

The neighborhood mutual information between S_1 and S_2 is defined as:

$$NMI_\tau(S_1;S_2) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\tau_{S_1}(x_i)\right|\cdot\left|\tau_{S_2}(x_i)\right|}{|U|\cdot\left|\tau_{S_1\cup S_2}(x_i)\right|}.$$

Similar to the approach used in Proposition 2.2 to prove the entropy formula, we can apply the same principle to derive three formulas for mutual information. These formulas provide a quantitative measure of the information shared and the degree of dependence between random variables.

Proposition 3.1. *For a given $KNDT = (U, F, D, \bigtriangleup, \delta, K)$, f or $\delta \geq 0, \forall S_1, S_2 \in F$. Then the following equation holds:*

(1) $NMI_\tau(S_1;S_2) = NMI_\tau(S_2;S_1)$;
(2) $NMI_\tau(S_1;S_2) = NE_\tau(S_1) + NE_\tau(S_2) - NE_\tau(S_1,S_2)$;
(3) $NMI_\tau(S_1;S_2) = NE_\tau(S_1) - NE_\tau(S_2 \mid S_1) = NE_\tau(S_2) - NE_\tau(S_1 \mid S_2)$.

Proof. According to the principle of Proposition 2.2, the appeal certification process is similar.

For a given $KNDT = (U, F, D, \bigtriangleup, \delta, K)$, $\forall S_1, S_2, S_3 \subseteq F$. The neighborhood conditional mutual information of S_1 and S_3, given the knowledge of S_2, is defined as:

$$NCMI_\tau(S_1;S_3 \mid S_2) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\tau_{S_1\cup S_2}(x_i)\mid\cdot\mid\tau_{S_3\cup S_2}(x_i)\right|}{\left|\tau_{S_2}(x_i)\mid\cdot\mid\tau_{S_1\cup S_2\cup S_3}(x_i)\right|}.$$

3.3. Feature relevance measure in k-nearest neighborhood decision table

Indeed, it is well-established that a stronger correlation between features and classes indicates a higher ability of those features to distinguish samples. [33,34] In information theory, features that exhibit stronger relevance to classes are considered more informative in categorizing them into distinct categories. Mutual information is a widely used metric for quantifying the correlation between features and classes and finding applications in various domains and tasks. For a given $KNDT = (U, F, D, \bigtriangleup, \delta, K)$, the relationship between the selected feature subset and the decision set can be described as the correlation between them. The formula is as follows:

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

where $\tau_{f_j}^F(x_i)$ denotes the neighborhood of x_i on f_j^F, $\tau_d(x_i)$ denotes the neighborhood of x_i on d and $\tau_{\{f_j{}^F\}\cup\{d\}}(x_i)$ denotes the neighborhood of x_i on two samples. In order to select features for analysis, we employ a prioritization approach based on the concept of neighborhood mutual information between the features and the classes. Specifically, we focus on identifying the feature that exhibits the highest neighborhood mutual information, which is referred to as the maximum-relevance criterion (MR). This selection criterion, also known as MR, is formally defined by [35]. The redundancy between a selected feature subset and the decision set refers to the degree of overlap or duplication in the information provided by the features regarding the decision or outcome being predicted. The formula for calculating redundancy between two features is as follows:

$$\begin{aligned}&Rdd(f_j^F, f_s) = NMI_\tau(f_j^F; f_s)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|\tau_{f_j}^F(x_i)\mid\cdot\mid\tau_{f_s}(x_i)\right|}{\left|U\mid\cdot\mid\tau_{\{f_j{}^F\}\cup\{f_s\}}(x_i)\right|}.\end{aligned}$$

Where f_s is the selected feature, to address the issue of redundancy between selected and unselected features, we can utilize the concept of neighborhood mutual information. This metric quantifies redundancy while disregarding class information. Our objective is to enhance the chosen feature set and eliminate any redundancies, thus we have decided to employ the minimum redundancy criterion (mR). The aim of the mR criterion is to select characteristics that offer non-redundant and distinctive information. With the mR criterion, we can balance the relevance of characteristics and their redundancy concerning the chosen features. By applying this method, we can ensure that the selected feature subset consists of informative and non-redundant features, leading to an

improvement in the system's classification performance [36]. This allows us to choose the most relevant features for decision-making while reducing redundancy in the selected features.

3.4. Feature interaction measure in k-nearest neighborhood decision table

To measure the extent of interaction between attributes, researchers use various methods that consider relationships and dependencies between features. These methods aim to capture how attributes interact with each other and how these interactions impact classification accuracy. The mutual information-based interaction measure is a commonplace tool for assessing attribute interaction. This metric quantifies the amount of information that is shared among several attributes, indicating their degree of interdependence or interaction. Subsequently, we will introduce techniques for calculating the degree of interaction between these properties. For a given $KNDT = (U, F, D, \triangle, \delta, K)$, when the selected feature f_s is already known, we define the neighborhood conditional mutual information between the current candidate feature f_j^F and the decision class d as follows:

$$Itrs(f_j^F, f_s, d) = NMI_\tau(f_j^F; d \mid f_s)$$
$$= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2 \cdot \frac{\left|\tau_{\{f_j{}^F\}\cup\{d\}}(x_i)\right| \cdot \left|\tau_{\{d\}\cup\{f_s\}}(x_i)\right|}{\left|\tau_{f_s}(x_i)\right| \cdot \left|\tau_{\{f_j{}^F\}\cup\{d\}\cup\{f_s\}}(x_i)\right|}.$$

$ItrS$ is a metric used to quantify the contribution or reduction in uncertainty of classification provided by the current candidate feature f_j^F when the selected feature f_s is known. It represents the degree of interaction between the candidate feature and the selected feature by utilizing the principles of information theory [37].

In the study referenced as [38], the researchers have leveraged the concepts from information theory to measure the amount of information contributed by adding new features to the classification task, given the knowledge of the existing features. This information-based approach allows for the estimation of the degree of interaction between the current candidate feature and the selected feature, providing insights into their combined impact on the classification process. We define the interaction of class dependence between $f_j^{F'}$ and d, given that f_j^F is known, as follows:

$$Itrc(f_j^F, f_j^{F'}, d) = NMI_\tau(f_j^{F'}; d \mid f_j^F)$$
$$= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2 \cdot \frac{\left|\tau_{\left\{f_j^{F'}\right\}\cup\{d\}}(x_i)\right| \cdot \left|\tau_{\{d\}\cup\left\{f_j^F\right\}}(x_i)\right|}{\left|\tau_{f_j^F}(x_i)\right| \cdot \left|\tau_{\{f_j{}^F\}\cup\{d\}\cup\left\{f_j^{F'}\right\}}(x_i)\right|}.$$

This formula uses information theory principles to quantify the impact of features $f_j^{F'}$ on the further selection of remaining candidate features. The first formula calculates the interaction degree between the candidate feature f_j^F and the selected features f_s, taking into account their mutual information and conditional probabilities. It quantifies how the candidate feature f_j^F interacts with the already selected features, providing information about their joint contribution to the classification task. The second formula calculates the interaction degree between the candidate feature f_j^F and the remaining set of candidate features $f_j^{F'}$. It measures the interaction between the candidate feature f_j^F and the other candidate features that have not been selected yet. This interaction degree helps assess how the candidate feature f_j^F interacts with the remaining features in terms of their information sharing and joint impact on the classification task.

Algorithm $KNCMI$ provides a heuristic feature selection approach for $KNDT$, and its detailed descriptions are presented below. To establish an objective evaluation function for the KNMRmRMI (k-nearest neighborhood max-relevance min-redundancy max-interaction) approach, which considers the feature correlations discussed in sections 3.1-3.3, we introduce the original feature evaluation function LKNCMI. This function is defined as follows:

$$\iota_{sig}(f_j^F) = NMI_\tau(f_j^F; d) - \frac{1}{|RED|}\sum_{f_s \in Red} NMI_\tau(f_j^F; f_s) + \frac{1}{|F - RED| - 1}\sum_{f_j^{F'} \in F - Red - \left\{f_j^F\right\}} NCMI_\tau(f_j^{F'}; d \mid f_j^F).$$

In the LKNCMI feature evaluation function, $NMI_\tau(f_j^F; f_s)$ and $NCMI_\tau(f_j^{F'}; d \mid f_j^F)$ are the previously introduced redundancy and interaction degrees, respectively. The $\iota_{sig}(f_j^F)$ denotes the attribute importance of the candidate feature f_j^F. This measure quantifies the relevance or significance of each attribute in the classification task. al We can obtain a list of attributes sorted by their importance score by calculating their importance using $\iota_{sig}(f_j^F)$ and ranking them accordingly. By using this ranking, the optimal feature combination that creates the chosen feature subset, designated as denotes, can be chosen. Therefore, the LKNCMI function can identify groups of features that collaboratively produce the most pertinent information for the classification task that decreases redundancy and accounts for attribute interactions.

Table 1
Dataset case.

	f_1	f_2	f_3	f_4	f_5	f_6	d
x_1	0.34	0.33	0.28	0.49	0.65	0.78	1
x_2	0.37	0.56	0.61	0.63	0.70	0.55	1
x_3	0.45	0.57	0.43	0.55	0.83	0.21	1
x_4	0.35	0.57	0.33	0.63	0.71	0.65	1
x_5	0.42	0.68	0.57	0.47	0.23	0.1	2
x_6	0.20	0.38	0.71	0.73	0.50	0.30	2
x_7	0.36	0.76	0.34	0.52	0.72	0.43	1
x_8	0.06	0.16	0.07	0.17	0.26	0.17	2

4. Incremental neighborhood conditional entropy of the hybrid data

When the number of data samples increases, it is possible to update the distance between each attribute sample. This can lead to the preservation of previously calculated distances. In addition to this, model computation time can be saved by updating the neighborhood entropy. Next, we describe the mechanism for updating the domain class and neighborhood conditional entropy after data addition.

In a neighborhood decision table $NDT = (U, F, D, \delta)$, where $\delta \geq 0$ and S are any subset of attributes from F. The $U = \{x_1, x_2, ..., x_m\}$ represents a complete set of samples in the data set. Each instance is denoted x_i, where i ranges from 1 to m. That is, there are m original samples. We introduced the distance formula for mixed data earlier, so we can get the domain classes under the feature subset S_1. The original neighborhood class is $\delta_{S_1}(x_i) = \left\{ x_i \in U \mid HCOM_{S_1}(x_i, x_j) \leq \delta \right\}$. When n samples are added, at which point new samples $U' = \{x_1, x_2, ..., x_m, x_{m+1}, ..., x_n\}$. At this stage, the updated formula for neighborhood classification is outlined as follows:

$$\delta'_{S_1}(x_i) = \delta_{S_1}(x_i) \bigcup \delta_{S_1}(x_i)', i \leq m,$$

where $\delta'_{S_1}(x_i)$ denotes the updated neighborhood class, $\delta_{S_1}(x_i)$ denotes the originally computed neighborhood class. To avoid double counting, we only need to calculate the distance to the newly added sample. Thus the specific equation for $\delta_{S_1}(x_i)'$ is expressed as follows:

$$\delta_{S_1}(x_i)' = \left\{ x_i \in U \mid HCOM_{S_1}(x_i, x_j) \leq \delta \right\}, i \leq m, m \leq j \leq n.$$

Next, we'll give a simple case study of updating neighborhood classes. A simple data table is shown in Table 1. Given a neighborhood decision table $NDT = (U, F, D, \delta), for\ \delta \geq 0, \forall S_1, S_2 \subseteq F$. The original sample $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, the set of features $F = \{f_1, f_2, f_3, f_4, f_5, f_6\}$. Two new samples, x_7 and x_8, have been added. We set δ to 0.4. The subset of features $S = \{f_1, f_2, f_3, f_4\}$. Calculating from the distance formula we know that $HCOM_S(x_1, x_1) = 0$, $HCOM_S(x_1, x_2) = 0.42$, $HCOM_S(x_1, x_3) = 0.30$, $HCOM_S(x_1, x_4) = 0.28$, $HCOM_S(x_1, x_5) = 0.46$, $HCOM_S(x_1, x_6) = 0.51$. So $\delta_s = \{x_1, x_3, x_4\}$. When new samples are added, the $HCOM_S(x_1, x_7) = 0.43$ and $HCOM_S(x_1, x_8) = 0.50$. So $\delta_S(x_1)'$ is an empty set and Neighborhood classes do not change. Incremental computation demonstrates the time savings from this step as we get larger amounts of data.

Then, we will introduce the formula update of neighborhood entropy. In a neighborhood decision table $NDT = (U, F, D, \delta)$, where $\delta \geq 0$ and for any subsets S_1 and S_2 from the set F. The new neighborhood entropy of the set of samples concerning S_1 is defined as follows:

$$NE'_\delta(S_1) = -\frac{1}{|U'|} \sum_{i=1}^{|U'|} \log_2 \frac{\left|\delta_{S_1}(x_i)\right| + \left|\delta_{S_1}(x_i)'\right|}{|U'|},$$

where U' denotes the dataset after the addition of new samples, $\left|\delta_S(x_i)'\right|$ denotes the number of new samples belonging to the neighborhood class. Similarly to the principle, we also improve the formulae in neighborhood conditional entropy and joint entropy. Simply add the new number of neighborhood classes without recalculating the previous number of neighborhood classes. Next, we give new formulas for neighborhood conditional entropy and joint entropy, respectively. The formula for the conditional entropy of the neighborhood after a new sample is added is:

$$NE'_\delta(S_1 \mid S_2) = -\frac{1}{|U'|} \sum_{i=1}^{|U'|} \log_2 \frac{\left|\delta_{S_1 \cup S_2}(x_i)\right| + \left|\delta_{S_1 \cup S_2}(x_i)'\right|}{\left|\delta_{S_2}(x_i)\right| + \left|\delta_{S_2}(x_i)'\right|}.$$

The neighborhood mutual information entropy after adding new samples is as follows:

$$NMI'_\delta(S_1; S_2) = -\frac{1}{|U'|} \sum_{i=1}^{|U'|} \log_2 \frac{(\left|\delta_{S_1}(x_i)\right| + \left|\delta_{S_1}(x_i)'\right|) \cdot (\left|\delta_{S_2}(x_i)\right| + \left|\delta_{S_2}(x_i)'\right|)}{|U'| \cdot (\left|\delta_{S_1 \cup S_2}(x_i)\right| + \left|\delta_{S_1 \cup S_2}(x_i)'\right|)}.$$

In the context of the k-nearest neighbor rough set, assuming that the value of k is already set, we focus on updating the original neighborhood entropy by considering the k-nearest neighbors. The neighborhood entropy of the updated sample set, denoted as S, can be defined as follows:

$$NE'_{\tau}(S) = -\frac{1}{|U'|}\sum_{i=1}^{|U'|}\log_2\frac{\left|\tau'_S(x_i)\right|}{|U'|}$$
$$= -\frac{1}{|U'|}\sum_{i=1}^{|U'|}\log_2\frac{(\left|\delta_{S_1}(x_i)\right| + \left|\delta_{S_1}(x_i)'\right|) \cap \left|k_{U'}(x_i)\right|}{|U'|}.$$

In this formula, the value of k is determined based on the sample size's magnitude. Specifically, k is established within the range of [0.01n, 0.1n], where n denotes the number of samples in the dataset. This range offers flexibility in determining the appropriate k value based on the dataset size. The conditional information entropy of S_1 relative to S_2 can be expressed as:

$$NE'_{\tau}(S_1 \mid S_2) = -\frac{1}{|U'|}\sum_{i=1}^{|U'|}\log_2\frac{\left|\tau'_{S_1\cup S_2}(x_i)\right|}{\left|\tau'_{S_1}x_i\right|}$$
$$= -\frac{1}{|U'|}\sum_{i=1}^{|U'|}\log_2\frac{\left|\left(\delta_{s_1\cup s_2}(x_i)\cup\delta_{s_1\cup s_2}(x_i)'\right)\cap k_{U'}(x_i)\right|}{\left|\left(\delta_{s_1}(x_i)\cup\delta_{s_1}(x_i)'\right)\cap k_{U'}(x_i)\right|}.$$

Similarly, when a new sample is added, the k-nearest neighbor mutual information of the sample under the feature sets S_1 and S_2 can be defined as:

$$NMI'_{\tau}(S_1;S_2) = -\frac{1}{|U'|}\sum_{i=1}^{|U'|}\log_2\frac{\left|\tau'_{S_1}(x_i)\right|\cdot\left|\tau'_{S_2}(x_i)\right|}{|U'|\cdot\left|\tau'_{S_1\cup S_2}(x_i)\right|}$$
$$= -\frac{1}{|U'|}\sum_{i=1}^{|U'|}\log_2\frac{\left|\left(\delta_{s_1}(x_i)\cup\delta_{s_1}(x_i)'\right)\cap k_{U'}(x_i)\right|\cdot\left|\left(\delta_{s_2}(x_i)\cup\delta_{s_2}(x_i)'\right)\cap k_{U'}(x_i)\right|}{|U'|\cdot\left|\left(\delta_{s_1\cup s_2}(x_i)\cup\delta_{s_1\cup s_2}(x_i)'\right)\cap k_{U'}(x_i)\right|},$$

the function $NMI'_{\tau}(S_1;S_2)$ represents the degree of mutual dependence between two feature subsets. A higher value indicates a stronger correlation between the two subsets. $NMI'_{\tau}(S_1;S_2)$ is the core component of the feature evaluation function and a key element of the algorithm model. When adding a new sample, we need to calculate the number of new neighborhood classes after combining the k nearest neighbors, which corresponds to $\tau'_{S_1}(x_i)$ and $\tau'_{S_2}(x_i)$. The formula $(\delta_{s_1}(x_i)\cup\delta_{s_1}(x_i)'\cap k_{U'}(x_i))$ represents the change in the neighborhood class, effectively utilizing the initially calculated neighborhood class $\delta_{s_1}(x_i)$, and only requiring the calculation of the neighborhood class for the added sample $\delta_{s_1}(x_i)'$. The $k_{U'}(x_i)$ is determined solely by the sample size. As the number of samples increases, we only need to adjust the k value in the model.

To summarize, in order to leverage the benefits of k-nearest neighbors, we perform intersection processing between the traditional neighborhood set and the k-nearest neighbors. However, as the sample size increases, the appropriate value of k may change. Therefore, during each update, the focus is on updating the traditional neighborhood classes for the calculation of new samples and then performing union processing with the neighborhood classes from the non-updated stage. This approach of selectively updating saves computational resources and reduces computational time compared to recalculating the entire neighborhood classes.

To verify the computational effectiveness of our update mechanism, we have developed two feature selection algorithms: Algorithm 1, a non-incremental feature selection algorithm, and Algorithm 2, an improved incremental algorithm. In the following algorithm flow chart, we will explain these two algorithms in detail, leveraging our previous theoretical knowledge.

Next, we will detail the main flow of Algorithm 1 and the algorithmic time complexity it consumes. The algorithm is divided into three main core steps. First input data set U, which contains n samples and m attributes. We performed a min-max normalization of the data, i.e., data pre-processing. Then the mixed data distances were calculated for each sample value in each column and the neighborhood classes were divided under K and δ to form a new neighborhood relation matrix (steps 3-8). The time complexity here is $O(mn)$. In the second step, the correlation between the feature to be selected f_j^F and the decision d is calculated. Then the feature with the maximum value f_j^F is taken as the first selected feature f_s(steps 9-14). The time complexity here is O(m). In the third step, loop the number of features f_j^F. Calculate the redundancy between the alternative features and the other selected features. Next, calculate the interaction between the current candidate feature f_j^F and the feature $f_j^{F'}$ in the remaining candidate feature subset, taking into account the class d. This interaction analysis aims to assess the mutual influence and relationship between the two features when considering their impact on the assigned class. Furthermore, feature importance is determined using the feature objective evaluation function KNMRmRMI, which helps quantify the relevance of each attribute. The features are sorted according to their importance values and the number of selected features is determined by evaluating the classification accuracy of the classifier (steps 15-29). The time complexity of this process is $O(mn)$, and the computational complexity is $O(m^3)$. Considering that the selected

Algorithm 1: The feature selection algorithm based on *k*-nearest neighborhood conditional mutual information (KNCMI).

Input : A $KNDT = (U, F, D, \delta, K)$ with $U = \left(x_1, x_2, ..., x_n\right)$, $F = \left(f_1, f_2, ...f_m\right)$,$\delta \in [0.5, 1.0]$. It is in steps of 0.1, $K \in [0.01n, 0.1n]$, n is the number of samples.
Output : The best subset of features Redbest

```
for δ ← 0.5 to 1.0 do
    for K ← 0.01N to 0.1N do
        Pre-processing the data;
        for each f_j^F ∈ F do
            for each i ∈ n do
                Compute the domain classes under K and δ.
            end
        end
        for each f_j^F ∈ F do
            Compute Rel (f_j^F, d);//Calculate the maximum correlation of features.
        end
        f_s ← max Rel (f_j^F, d);
        Red ← f_s;
        F ← F \ f_s;
        for each f_j^F ∈ F do
            for each f_s ∈ Red do
                Compute Rdd(f_j^F, f_s);//Calculating feature redundancy
            end
            for each f_j^F' ∈ F − Red − {f_j^F} do
                Compute ItrS(f_j^F, f_j^F', d);
            end
            Compute l_sig (f_j^F); //The maximum value of l_sig (f_j^F) is the most important attribute.
            Update Red ← Red ∪ {f};
            F ← F \ {f};
        end
        return Redbest;//The best subset after reductiony
    end
end
Finally, selecting the best feature subset Redbest and the best combination of δ and K by using different classifiers;
```

data set usually contains more samples than the number of feature attributes and the increased computational load, the overall time complexity of the algorithm can be approximated as $O(2mn^2)$.

Correspondingly, we will analyze Algorithm 2 in detail next and give the time complexity of each step. The basic core idea of the algorithm remains unchanged, and again, after going to data pre-processing, the relationship between each attribute and decision attribute d is computed to calculate the desired attribute importance. The main change is that when new additions to the data come in, we first update the good U, and then after normalization, we update the neighborhood relationship matrix, if the number of rows or columns of the loop is under the original Ma, it is directly assigned to the new matrix and goes to the next loop, this eliminates the need to recalculate the time between the old samples and only the relationship between the new samples and the old samples (steps 2-11). The time complexity of this step of Algorithm 1 is $O(mn^2)$. After updating the conditional entropy. Consistent with the steps of Algorithm 1, since the mutual information calculation still requires sample values under each column attribute.

5. Experimental analysis

This section describes a set of experiments performed to evaluate the effectiveness of our proposed iterative algorithm. This collection contains 12 datasets from the UCI repository. Table 2 gives a summary of each dataset, where the largest sample dataset is Electrical grid data, which has a size of 1000*14. The smallest sample data set is Sonar, its size is 208*60. The overall 12 data sets include three types of numerical, categorical, and mixed data. During data preprocessing, we normalize numerical features to the range [0, 1]. This experiment was conducted on a Windows 10 PC equipped with Intel(R) Core(TM) i5-8300H CPU @ 2.30 GHz and 8 GB RAM. Pycharm2020 is used as the integrated development environment, and Python is used to implement the algorithm of this article and other comparison algorithms.

This paper examines the performance of Algorithm 2 in this paper by comparing the following three parts, namely calculation time, feature subset size reduction, and classification accuracy. In addition, in order to prove the effectiveness of our proposed incremental algorithm, we selected four contrasting algorithms to compare time and accuracy.

Algorithm 2: The incremental feature selection algorithm based on *k*-nearest neighborhood conditional mutual information (IFS-KNCMNI).

Input : A $KNDT = (U_1, F, D, \delta, K)$ with $U_1 = (x_1, x_2, ..., x_n)$, $F = (f_1, f_2, ...f_m)$, new data U_2, select the best combination of δ and K in Algorithm 1, Original Neighborhood Matrix M_f.

Output : A new reduce feature subset $Redbest'$

Firstly, Update datasets and neighborhood classes;
Update $U \leftarrow U_1 \cup \{U_2\}$;
for $i \in U.column$ **do**
    The x is the value of the sample under each attribute.
    for $j \in U.row$ **do**
        for $q \in length(x)$ **do**
            Compute new martix M_f' ;
            //If the number of rows or columns of the loop is under the original M_f, it is directly assigned to the new matrix and goes to the next loop. Otherwise, the value is calculated according to Algorithm 1.
        end
    end
end
Secondly, update the neighborhood conditional entropy;
for *each* $i \in F$ **do**
    Compute new $NE'_\tau(f \mid d)$;
end
Lastly, Update the Redbest. //Reference Algorithm 1 Step 9 to Step 21.
return $Redbest'$;

Table 2
Dataset description.

No.	Datasets	Cases	Features	Classes	Data type
1	Sonar	208	60	2	Numerical
2	WDBC	569	31	2	Numerical
3	Australian	690	14	2	Hybrid
4	Blood	748	5	2	Numerical
5	German-Credit	1000	20	2	Hybrid
6	Flare	1066	11	6	Categorical
7	Car	1728	6	4	Categorical
8	Segment	2310	20	7	Numerical
9	wine	4998	11	2	Numerical
10	Page-blocks	5472	10	5	Numerical
11	Twonorm	7400	20	2	Numerical
12	Electrical Grid data	10000	14	2	Numerical

5.1. Feature subset size

In the following, Table 3 illustrates the number of optimal features found by the two algorithms. Column "$Features$" in the table indicates the number of features in the original data set. The remaining two columns represent the best number of features for attribute simplification under Algorithm 1 and Algorithm 2, respectively.

From Table 3, we know that the number of attribute simplifications is approximately the same for both algorithms, which reflects the effectiveness of incremental Algorithm 2.

5.2. Classification accuracy

In this subsection, we will focus on evaluating the classification accuracy of the algorithm, which is considered one of the most effective and direct measures to assess the quality of feature selection algorithms. To ensure reliable results and mitigate the impact of data sparsity and computational randomness, we average the classification accuracy of the same feature selection algorithm across different datasets. The "Average" rows in the results display this averaged classification accuracy. To establish a baseline for experimental comparison, we exploit the average classification accuracy on raw data for *k*-nearest neighbor (KNN) classifiers, (Support Vector Machine) SVM, and (Random Forest) RF. The KNN classifier is evaluated using a 5-fold cross-validation approach, where the original dataset is randomly divided into five subsets. Four subsets are used as the training set, while the remaining subset serves as the test set. This process is repeated five times, with each subset used as the test set once. The classifier is trained using the features selected by the feature selection algorithm on the training set, and its performance is evaluated on the test set. The average performance across the five test sets is considered the final classification performance. By comparing the classification accuracies obtained from different feature selection algorithms, we can assess the effectiveness of our proposed incremental algorithm.

Table 3
Average feature size of the two feature selection algorithms.

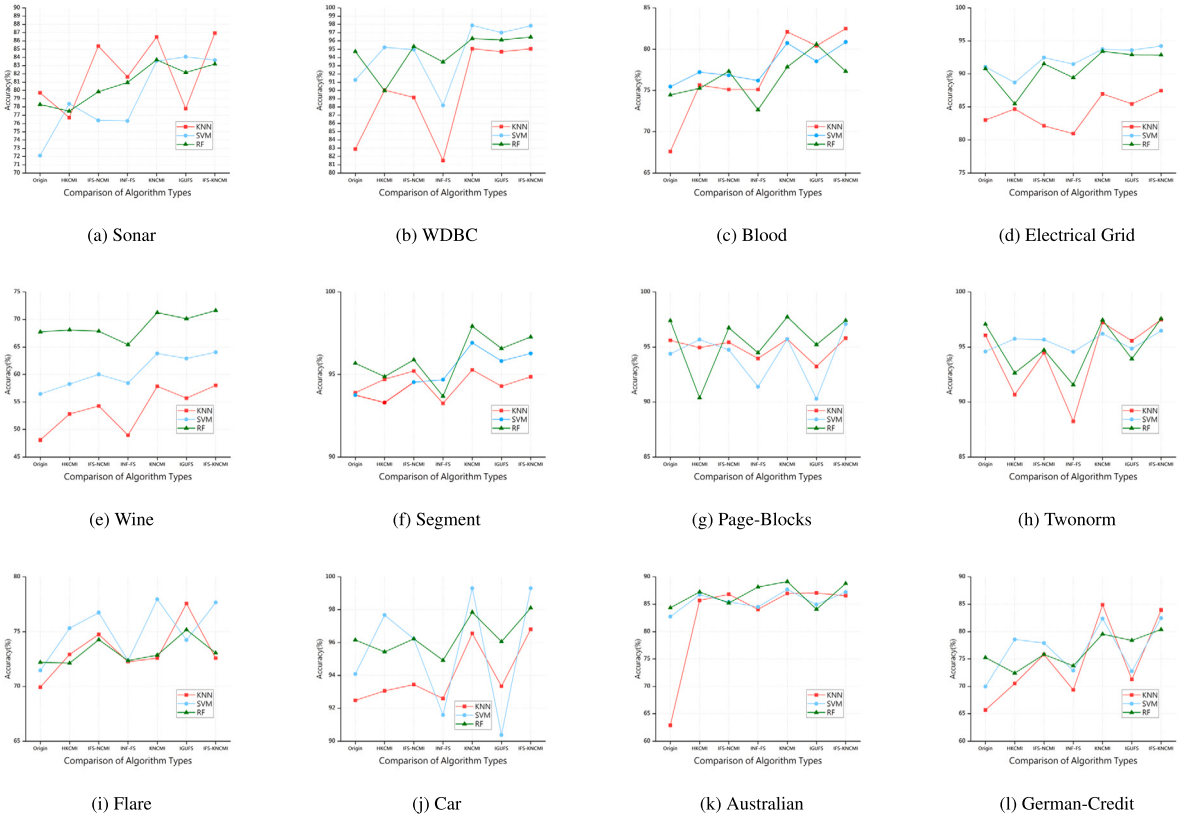
No.	Datasets	Features	KNN		SVM		RF	
			KNCMI	IFS-KNCMI	KNCMI	IFS-KNCMI	KNCMI	IFS-KNCMI
1	Sonar	60	29	30	16	14	33	33
2	WDBC	31	11	10	15	15	13	13
3	Blood	5	4	3	3	3	4	4
4	Segment	20	9	8	11	11	14	13
5	Page-blocks	10	6	6	5	6	5	5
6	Twonorm	20	16	17	14	13	15	14
7	Car	6	5	5	5	5	5	5
8	wine	11	9	8	8	9	8	8
9	Flare	11	8	8	7	7	9	8
10	Australian	14	10	9	8	8	8	9
11	German-Credit	20	11	10	13	13	17	16
12	Electrical Grid data	14	10	9	10	11	10	10

Table 4
Comparison of classification accuracy of seven algorithms on KNN classifier.

Datasets	Original data	HKCMI	IFS-NCMI	INF-FS	KNCMI	IGUFS	IFS-KNCMI
Sonar	79.70 ± 3.94	76.73 ± 2.18	85.37 ± 0.31	81.64 ± 3.19	86.48 ± 3.86	77.78 ± 1.74	$\mathbf{86.95 \pm 3.97}$
WDBC	82.90 ± 2.86	90.05 ± 0.84	89.15 ± 0.98	81.51 ± 2.99	$\mathbf{95.07 \pm 1.19}$	94.71 ± 2.16	95.06 ± 1.23
Blood	67.60 ± 1.34	75.65 ± 0.47	75.12 ± 0.54	75.12 ± 1.78	82.09 ± 3.28	80.42 ± 2.01	$\mathbf{82.49 \pm 2.32}$
Segment	93.89 ± 0.95	94.71 ± 0.73	95.21 ± 0.36	93.25 ± 1.13	$\mathbf{95.28 \pm 0.85}$	94.29 ± 2.38	94.86 ± 0.82
Page-blocks	95.61 ± 0.25	94.95 ± 0.14	95.43 ± 0.56	93.96 ± 0.49	95.72 ± 0.18	93.23 ± 0.87	$\mathbf{95.81 \pm 0.23}$
Twonorm	96.07 ± 0.54	90.67 ± 0.21	94.48 ± 0.92	88.24 ± 0.83	97.21 ± 0.29	95.57 ± 0.78	$\mathbf{97.48 \pm 0.51}$
Flare	69.92 ± 3.40	72.92 ± 2.48	74.75 ± 0.32	72.25 ± 1.84	72.58 ± 2.64	$\mathbf{77.55 \pm 3.13}$	72.58 ± 2.65
Wine	48.08 ± 1.37	52.81 ± 1.12	54.27 ± 1.26	48.94 ± 2.15	57.84 ± 1.36	55.69 ± 2.58	$\mathbf{58.02 \pm 1.17}$
Car	92.48 ± 0.82	93.06 ± 0.94	93.44 ± 1.87	92.59 ± 1.03	96.56 ± 1.04	93.34 ± 1.91	$\mathbf{96.81 \pm 1.19}$
Australian	62.89 ± 1.91	85.68 ± 2.76	86.81 ± 2.28	84.05 ± 3.39	86.95 ± 1.25	$\mathbf{87.05 \pm 2.24}$	86.53 ± 1.41
German-Credit	65.66 ± 2.88	70.52 ± 2.59	75.82 ± 0.34	69.35 ± 3.06	$\mathbf{84.87 \pm 2.35}$	71.27 ± 2.77	83.95 ± 2.33
Electrical Grid	83.04 ± 0.97	84.67 ± 1.09	82.16 ± 1.32	80.97 ± 1.14	86.96 ± 1.22	85.45 ± 1.24	$\mathbf{87.45 \pm 1.49}$
Average	78.15 ± 1.77	81.87 ± 1.30	83.50 ± 0.92	80.16 ± 1.92	86.46 ± 1.62	83.86 ± 1.98	$\mathbf{86.50 \pm 1.60}$

In Table 4, Table 5 and Table 6, we show the score statistics of 12 data sets after applying the algorithm using KNN classifier, SVM classifier and RF classifier respectively. These scores are compared with those obtained from other feature selection algorithms. The aim is to showcase the effectiveness and robustness of our proposed algorithm by comparing it with existing feature selection methods. In total, we include five comparison algorithms for this evaluation.

1) **Hybrid-kernel based fuzzy complementary mutual information**(**HKCMI**)[34]. The HKCMI is a feature selection algorithm that utilizes fuzzy complementary mutual information. It is well-suited for clustering tasks involving datasets with multiple types of attributes. The algorithm selects a subset of features that exhibit high relevance and dependency, taking into account their fuzzy and complementary nature. By incorporating hierarchical k-means clustering, HKCMI enhances the clustering performance of the dataset.
2) **An interaction feature selection algorithm based on neighborhood conditional mutual information** (**IFS-NCMI**)[32]. The IFS-NCMI algorithm synergistically leverages the benefits of NRS in handling mixed and uncertain data with information-theoretic measures of feature relevance. This integration aims to enhance classification performance, resulting in higher accuracy and improved stability.
3) **Infinite feature selection** (**INF-FS**)[39]. The feature selection approach treats subsets of features as paths in a graph, resulting in high classification accuracy and effective removal of redundancy.
4) **Graph-based unsupervised feature selection for interval-valued information system**(**IGUFS**)[40]. An unsupervised feature selection method based on graph theory is employed. The method utilizes the properties of matrix power series to optimize the computation process while efficiently and swiftly performing feature selection by incorporating the principles of maximum relevance and minimum redundancy.
5) **Feature selection for unbalanced distribution hybrid data based on k-nearest neighborhood rough set** (**KNCMI**)[16]. The algorithm takes into account the interaction of heterogeneous data and features, combines well the advantages of δ-neighborhood and k-nearest neighbor, and utilizes mutual information entropy for feature extraction.



(a) Sonar (b) WDBC (c) Blood (d) Electrical Grid

(e) Wine (f) Segment (g) Page-Blocks (h) Twonorm

(i) Flare (j) Car (k) Australian (l) German-Credit

Fig. 2. Comparison of accuracy of different algorithms based on KNN, SVM and RF classifiers. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 5
Comparison of classification accuracy of seven algorithms on SVM classifier.

Datasets	Original data	HKCMI	IFS-NCMI	INF-FS	KNCMI	IGUFS	IFS-KNCMI
Sonar	72.11 ± 4.56	78.37 ± 0.31	76.38 ± 1.34	76.32 ± 3.56	83.56 ± 3.61	$\mathbf{84.07 \pm 5.07}$	83.67 ± 3.67
WDBC	91.28 ± 2.27	95.24 ± 0.65	94.96 ± 1.49	88.20 ± 1.87	$\mathbf{97.88 \pm 0.55}$	97.00 ± 1.04	97.83 ± 0.54
Blood	75.46 ± 2.53	77.23 ± 2.45	76.84 ± 0.96	76.20 ± 2.38	80.75 ± 0.48	78.54 ± 0.67	$\mathbf{80.87 \pm 0.44}$
Segment	93.75 ± 0.73	93.29 ± 0.86	94.53 ± 0.95	94.68 ± 1.45	$\mathbf{96.92 \pm 1.25}$	95.82 ± 1.92	96.28 ± 1.82
Page-blocks	94.38 ± 0.46	95.68 ± 0.24	94.75 ± 0.55	91.39 ± 0.32	95.72 ± 0.18	90.29 ± 0.48	$\mathbf{97.07 \pm 0.23}$
Twonorm	94.59 ± 0.52	95.75 ± 0.37	95.67 ± 0.63	94.56 ± 0.44	96.21 ± 0.23	94.85 ± 0.67	$\mathbf{96.48 \pm 0.28}$
Flare	71.45 ± 2.38	75.32 ± 2.76	76.74 ± 2.87	72.35 ± 3.06	$\mathbf{77.94 \pm 1.94}$	74.23 ± 1.82	77.65 ± 1.93
Wine	56.45 ± 1.19	58.26 ± 0.29	60.02 ± 1.33	58.44 ± 1.23	63.83 ± 0.27	62.92 ± 0.86	$\mathbf{64.06 \pm 0.19}$
Car	94.09 ± 0.71	97.68 ± 0.39	96.23 ± 0.64	91.59 ± 1.11	$\mathbf{99.30 \pm 0.68}$	90.38 ± 1.68	99.30 ± 0.59
Australian	82.75 ± 2.39	86.71 ± 2.62	85.42 ± 2.48	84.49 ± 2.55	$\mathbf{87.68 \pm 2.25}$	84.92 ± 2.56	87.21 ± 2.12
German-Credit	69.97 ± 3.70	78.58 ± 2.80	77.92 ± 0.85	72.87 ± 3.04	82.38 ± 2.19	72.77 ± 4.24	$\mathbf{82.48 \pm 2.79}$
Electrical Grid	91.06 ± 0.94	88.68 ± 0.72	92.47 ± 0.21	91.48 ± 0.41	93.74 ± 0.72	93.60 ± 0.97	$\mathbf{94.21 \pm 0.26}$
Average	82.28 ± 1.86	85.07 ± 1.21	85.16 ± 1.19	82.71 ± 1.78	87.99 ± 1.21	84.95 ± 1.83	$\mathbf{88.09 \pm 1.24}$

We show the scores of various algorithms in Fig. 2. The y-axis displays the algorithm's classification accuracy, and the x-axis shows the algorithm type. The red line illustrates the score trend of the algorithm on the KNN classifier. The blue line represents the score trend of the algorithm using the SVM classifier, while the green line depicts the score trend of the algorithm on the RF classifier. Overall, in the case of SVM, KNN and RF, the INF-FS algorithm has the lowest scores in most cases. On average, the algorithm produces scores that are only about 2 points higher than the raw score. The score of our dynamic algorithm IFS-KNCMI is basically the same as the score of the static algorithm KNCMI. This demonstrates the reliability of our proposed dynamic algorithm as it aligns with the results of the static algorithm. In most cases, our proposed IFS-KNCMI algorithm outperforms other algorithms, demonstrating its advantages. Only in the Flare dataset, the IFS-KNCMI score does not perform as well as the other algorithms, but it still outperforms the original algorithm and the INF-FS algorithm. Therefore, based on the classification scores displayed in the 12 dataset line graphs, we can conclude that our IFS-KNCMI algorithm is superior to previous algorithms.

Table 6
Comparison of classification accuracy of seven algorithms on RF classifier.

Datasets	Original data	HKCMI	IFS-NCMI	INF-FS	KNCMI	IGUFS	IFS-KNCMI
Sonar	78.29 ± 5.35	77.48 ± 3.93	79.83 ± 2.43	80.95 ± 1.82	**83.70 ± 3.42**	82.17 ± 5.47	83.45 ± 3.67
WDBC	94.73 ± 1.90	89.96 ± 1.32	95.35 ± 1.29	93.47 ± 1.92	96.29 ± 1.81	96.12 ± 2.27	**96.47 ± 1.54**
Blood	74.46 ± 2.52	75.26 ± 2.05	77.33 ± 1.72	72.66 ± 1.61	77.85 ± 3.47	**80.61 ± 1.53**	77.33 ± 1.52
Segment	95.68 ± 1.11	96.87 ± 1.25	95.89 ± 1.02	93.68 ± 1.29	**97.92 ± 1.04**	96.58 ± 2.21	97.28 ± 1.73
Page-blocks	97.38 ± 0.60	90.37 ± 0.24	96.75 ± 0.48	94.48 ± 0.64	**97.72 ± 0.44**	95.22 ± 0.95	97.40 ± 0.28
Twonorm	97.06 ± 0.55	92.64 ± 0.64	94.73 ± 1.82	91.56 ± 1.91	97.44 ± 0.23	93.93 ± 0.88	**97.56 ± 0.24**
Flare	72.20 ± 2.26	72.12 ± 2.84	74.27 ± 1.96	72.35 ± 3.06	72.86 ± 2.84	**75.18 ± 1.62**	73.05 ± 2.20
Wine	67.76 ± 1.47	68.12 ± 1.61	67.89 ± 1.23	65.43 ± 1.52	71.22 ± 2.12	70.13 ± 2.97	**71.62 ± 1.79**
Car	96.16 ± 0.56	95.43 ± 0.67	96.23 ± 0.64	94.92 ± 1.24	97.85 ± 0.81	96.06 ± 1.28	**98.10 ± 0.42**
Australian	84.33 ± 2.83	87.23 ± 2.52	85.21 ± 2.28	88.12 ± 2.38	**89.13 ± 3.03**	84.05 ± 3.07	88.78 ± 3.46
German-Credit	75.28 ± 3.46	72.43 ± 2.28	75.84 ± 2.89	73.74 ± 2.86	79.56 ± 3.47	78.40 ± 3.42	**80.40 ± 3.04**
Electrical Grid	90.78 ± 0.15	85.45 ± 2.52	91.52 ± 0.49	89.43 ± 1.44	**93.42 ± 0.92**	92.89 ± 0.32	92.87 ± 0.86
Average	85.34 ± 1.89	83.61 ± 1.82	85.90 ± 1.52	84.23 ± 1.81	**87.91 ± 1.96**	86.77 ± 2.16	87.86 ± 1.73

Table 7
Average ranking of algorithm classification accuracy.

Classifiers	HKCMI	IFS-NCMI	INF-FS	KNCMI	IGUFS	IFS-KNCMI	F_F	χ_F^2	P value
KNN	2.50	3.42	1.25	5.0	3.58	**5.08**	13.46	33.02	8.30×10^{-9}
SVM	2.91	2.75	1.58	5.25	2.91	**5.5**	20.08	38.76	1.02×10^{-11}
RF	1.83	3.08	1.83	**5.25**	3.83	5.08	17.40	36.76	1.30×10^{-10}

5.3. Statistical analysis

In this subsection, in order to enhance the comparison of experimental results between different algorithms, the Friedman test was employed as a statistical method to assess the validity of the algorithm comparison. The Friedman test is a non-parametric statistical test, and its null hypothesis states that all the experimental algorithms exhibit the same classification performance. The formula for the Friedman test is defined as follows:

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

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

The first formula here is the calculation formula for the Friedman statistic, and the second formula is the calculation formula for the parameter χ_F^2. T represents the experimental dataset, comprising a total of 12 datasets, while s represents the number of experimental algorithms being compared, totaling 6 algorithms. Here, R_i represents the average rank value of the classification accuracy results for the 6 algorithms across 3 classifiers.

From Table 7, it can be observed that the two algorithms proposed in this study exhibit higher average ranking of accuracy across the three classifiers. Furthermore, by conducting the Friedman hypothesis test, we obtained p-values below 0.05 for all three classifiers, indicating significant superiority or effectiveness of the proposed algorithm compared to the other algorithms being compared.

5.4. Computational time

In order to assess the computational efficiency, we measured the algorithm times for IFS-NCMI, INF-FS, KNCMI, and IFS-KNCMI. However, due to the extended execution time of the HKCMI algorithm, it was not included in the algorithm comparison graph.

To differentiate the computation time among these algorithms, we employed a specific methodology. For each dataset, we extracted 50% of the features as the original feature set, and then gradually added 20%, 40%, 60%, 80%, and 100% of the remaining features to the original set. The figure presents more detailed variation trendlines for each algorithm and the corresponding datasets with dynamic feature sets. In each subgraph, the X coordinate represents the size of the dynamic feature set, while the Y coordinate indicates the time consumption in seconds.

As seen in Fig. 3, the computation time for each algorithm increases with the dataset. However, based on the trend lines of both algorithms, It is found that in most cases, INF-FS (blue line) has the least computation time, while the proposed IFS-KNCMI (red line) has the second least computation time. The IFS-KNCMI has the least time only in the dataset “Australian”. IFS-NCMI is the most time-consuming among the four algorithms.

The primary reason for this is the following. In data increment, IFS-KNCMI(Algorithm 2) performs feature selection by updating the neighborhood information entropy and updating the neighborhood class, while KNCMI (Algorithm 1) is required to start calculating each step from scratch again. Although IFS-NCMI also has a part that utilizes the previous computation, it takes more time

because it divides the dominant neighborhood itself more than the *k*-nearest neighbor-based algorithms do. So the static algorithm KNCMI is also less time-consuming than the IFS-NCMI algorithm. INF-FS is a feature selection algorithm based on graph theory, which can calculate the importance of each feature by the knowledge of the graph, and then sort the attribute importance to select the best combination of features. So there is no need to cycle through the entropy between the features and the feature neighborhood, which greatly saves time.

Repeatedly performing incremental feature selection calculations is very time-consuming. Instead, many recalculations are avoided by using previous results on dynamic datasets instead of recalculating new datasets, which saves the time of repeated calculations. We can also verify this conclusion with the following detailed example.

For example, for the blood dataset, the first time 20% of the data is added, the time for static KNCMI is 6.26 s, while the time required for incremental IFS-KNCMI is 4.09 s, and when the data is added for the fifth time, the time for KNCMI is 18.09 s and the time for IFS-KNCMI is 11.88 s. The efficiency of IFS-KNCMI is about 34% improvement. The INF-FS algorithm exhibited a consistently low growth rate, taking only 3.09 s on its first run and 4.69 s on its fifth run. This provides empirical evidence that the algorithm is highly efficient in terms of time-saving measures. Although the INF-FS algorithm requires less time, its accuracy in KNN classification is only 75.12%, while IFS-KNCMI has an accuracy of 82.49%. Furthermore, IFS-KNCMI is more accurate than INF-FS on SVM. Another example is the "Twononrm" dataset, which is inherently large and presents challenges for static algorithms due to the time-consuming nature of processing it. However, after five updates, it becomes clear that the dynamic algorithm is not only more efficient than KNCMI each time but also becomes increasingly time-efficient over time, as evidenced by the emerging trend. Overall, the IFS-KNCMI algorithm takes less time than the static KNCMI algorithm and much less time than the IFS-NCMI algorithm for the 12 datasets. Except for the "Australian" dataset, IFS-KNCMI takes longer than the INF-FS algorithm, but it is more accurate. In conclusion, the IFS-KNCMI algorithm is an efficient and accurate method for selecting dynamic environmental data. It is an algorithm for feature selection of dynamic environmental data.

Hence, it can be deduced that IFS-KNCMI handles incremental data with greater efficiency and computational superiority over non-incremental KNCMI. This is confirmed by examining the 12 subgraphs presented in Fig. 3.

6. Conclusion and future work

In the era of data explosion, the increasing presence of mixed-type data presents significant challenges to traditional feature selection methods. This study investigates an incremental feature selection method specifically designed for mixed-type data in the context of *k*-nearest neighbors. Experimental evaluations are conducted on a set of 12 public datasets. The findings from the experiments are summarized as follows: (1) The IFS-KNCMI algorithm exhibits superior classification capability for hybrid data. (2) The IFS-KNCMI algorithm demonstrates higher efficiency in incremental hybrid data classification, leading to cost savings in computational power. (3) The proposed incremental feature selection algorithm has been experimentally validated and subjected to hypothesis testing, confirming its reliability.

In our future research, we intend to concentrate on enhancing the IFS-KNCMI algorithm to effectively address the dynamic challenge of data attribute changes. Furthermore, we aim to investigate the integration of the IFS-KNCMI algorithm with the INF-FS algorithm to improve both efficiency and accuracy. These prospective research directions will steer the development of more advanced and efficient feature selection methods, thereby facilitating decision-making and knowledge discovery in constantly evolving data environments.

CRediT authorship contribution statement

Weihua Xu: Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Changchun Liu:** Writing – original draft, Visualization, Software, Methodology, Data curation.

Declaration of competing interest

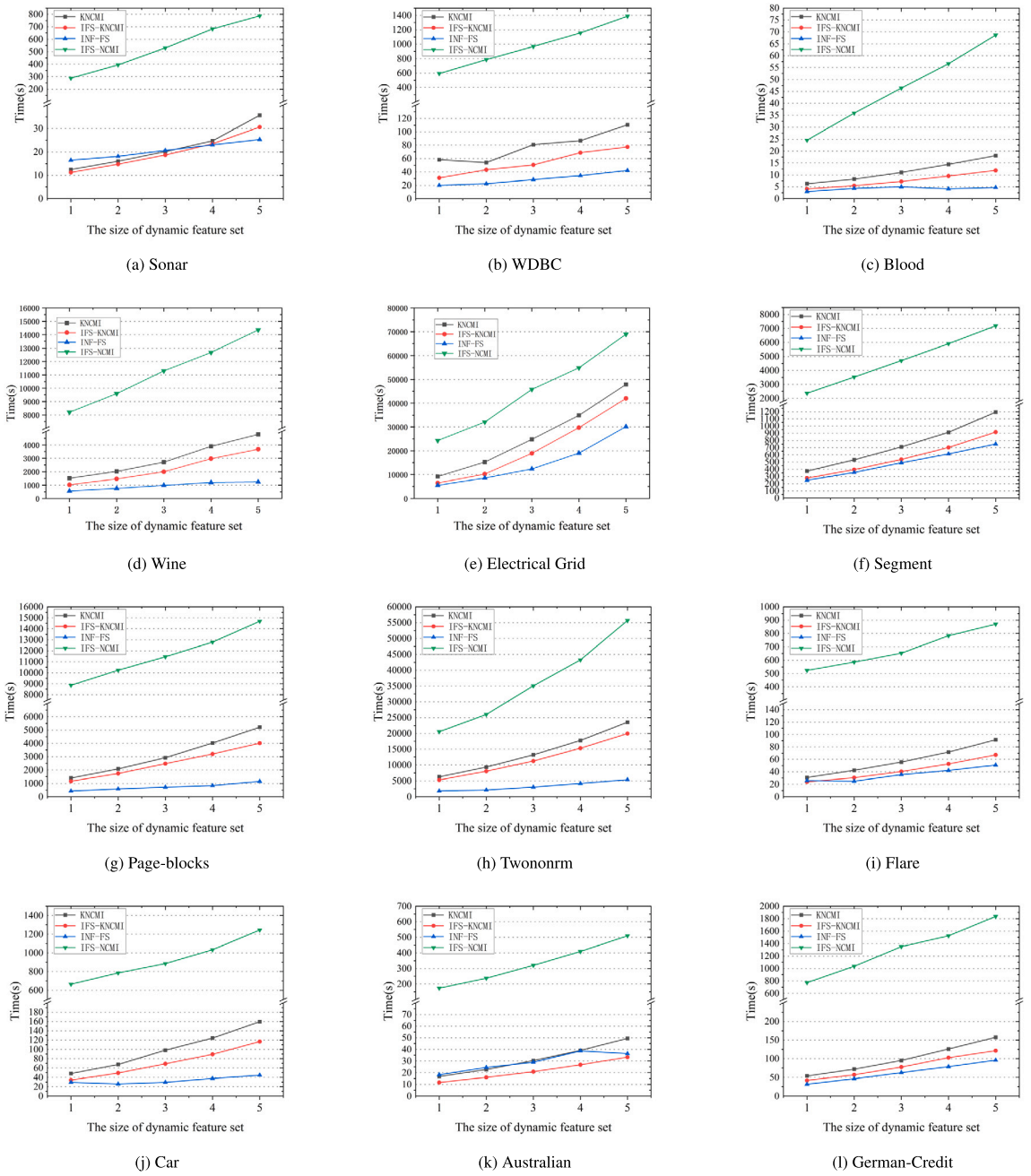
We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

Data availability

No data was used for the research described in the article.



(a) Sonar (b) WDBC (c) Blood

(d) Wine (e) Electrical Grid (f) Segment

(g) Page-blocks (h) Twononrm (i) Flare

(j) Car (k) Australian (l) German-Credit

Fig. 3. Comparison of time consumption between two algorithms.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 62376229, and Natural Science Foundation of Chongqing, China under Grant CSTB2023NSCQ-LZX0027.

References

[1] Zdzisław Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982) 341–356.
[2] T.Y. Lin, Neighborhood systems and approximation in relational databases and knowledge bases, in: Proceedings of the 4th International Symposium on Methodologies of Intelligent Systems, Citeseer, 1988, pp. 75–86.
[3] Qinghua Hu, Daren Yu, Jinfu Liu, Congxin Wu, Neighborhood rough set based heterogeneous feature subset selection, Inf. Sci. 178 (18) (2008) 3577–3594.

[4] Yuhua Qian, Xinyan Liang, Qi Wang, Jiye Liang, Bing Liu, Andrzej Skowron, Yiyu Yao, Jianmin Ma, Chuangyin Dang, Local rough set: a solution to rough data analysis in big data, Int. J. Approx. Reason. 97 (2018) 38–63.
[5] Qi Wang, Yuhua Qian, Xinyan Liang, Qian Guo, Jiye Liang, Local neighborhood rough set, Knowl.-Based Syst. 153 (2018) 53–64.
[6] Pengfei Zhang, Tianrui Li, Chuan Luo, Guoqiang Wang, Amg-dtrs: adaptive multi-granulation decision-theoretic rough sets, Int. J. Approx. Reason. 140 (2022) 7–30.
[7] Tengyu Yin, Hongmei Chen, Jihong Wan, Pengfei Zhang, Shi-Jinn Horng, Tianrui Li, Exploiting feature multi-correlations for multilabel feature selection in robust multi-neighborhood fuzzy β covering space, Inf. Fusion 104 (2024) 102150.
[8] Xibei Yang, Shaochen Liang, Hualong Yu, Shang Gao, Yuhua Qian, Pseudo-label neighborhood rough set: measures and attribute reductions, Int. J. Approx. Reason. 105 (2019) 112–129.
[9] Xiaoyan Zhang, Jianglong Hou, Jirong Li, Multigranulation rough set methods and applications based on neighborhood dominance relation in intuitionistic fuzzy datasets, Int. J. Fuzzy Syst. 24 (8) (2022) 3602–3625.
[10] Keyu Liu, Tianrui Li, Xibei Yang, Hengrong Ju, Xin Yang, Dun Liu, Hierarchical neighborhood entropy based multi-granularity attribute reduction with application to gene prioritization, Int. J. Approx. Reason. 148 (2022) 57–67.
[11] Meng Hu, Eric C.C. Tsang, Yanting Guo, Degang Chen, Weihua Xu, Attribute reduction based on overlap degree and k-nearest-neighbor rough sets in decision information systems, Inf. Sci. 584 (2022) 301–324.
[12] Yuhua Qian, Jiye Liang, Yiyu Yao, Chuangyin Dang, Mgrs: a multi-granulation rough set, Inf. Sci. 180 (6) (2010) 949–970.
[13] Caihui Liu, Duoqian Miao, Jin Qian, On multi-granulation covering rough sets, Int. J. Approx. Reason. 55 (6) (2014) 1404–1418.
[14] Weihua Xu, Yanzhou Pan, Xiuwei Chen, Weiping Ding, Yuhua Qian, A novel dynamic fusion approach using information entropy for interval-valued ordered datasets, IEEE Trans. Big Data (2022).
[15] Binbin Sang, Weihua Xu, Hongmei Chen, Tianrui Li, Active anti-noise fuzzy dominance rough feature selection using adaptive k-nearest neighbors, IEEE Trans. Fuzzy Syst. (2023).
[16] Weihua Xu, Ziting Yuan, Zheng Liu, Feature selection for unbalanced distribution hybrid data based on k-nearest neighborhood rough set, IEEE Trans. Artif. Intell. (2023).
[17] Xiaoyan Zhang, Jirong Li, Incremental feature selection approach to interval-valued fuzzy decision information systems based on λ-fuzzy similarity self-information, Inf. Sci. 625 (2023) 593–619.
[18] Amin Hashemi, Mohammad Bagher Dowlatshahi, Hossein Nezamabadi-Pour, Mfs-mcdm: multi-label feature selection using multi-criteria decision making, Knowl.-Based Syst. 206 (2020) 106365.
[19] Jinghua Liu, Yaojin Lin, Yuwen Li, Wei Weng, Shunxiang Wu, Online multi-label streaming feature selection based on neighborhood rough set, Pattern Recognit. 84 (2018) 273–287.
[20] Yaojin Lin, Qinghua Hu, Jinghua Liu, Jinkun Chen, Jie Duan, Multi-label feature selection based on neighborhood mutual information, Appl. Soft Comput. 38 (2016) 244–256.
[21] Wenbo Yu, Miao Zhang, Yi Shen, Learning a local manifold representation based on improved neighborhood rough set and lle for hyperspectral dimensionality reduction, Signal Process. 164 (2019) 20–29.
[22] Yao Liu, Hong Xie, Yuehua Chen, Kezhu Tan, Liguo Wang, Wu Xie, Neighborhood mutual information and its application on hyperspectral band selection for classification, Chemom. Intell. Lab. Syst. 157 (2016) 140–151.
[23] Ying Yu, Witold Pedrycz, Duoqian Miao, Neighborhood rough sets based multi-label classification for automatic image annotation, Int. J. Approx. Reason. 54 (9) (2013) 1373–1387.
[24] Yao Ping, Lu Yongheng, Neighborhood rough set and svm based hybrid credit scoring classifier, Expert Syst. Appl. 38 (9) (2011) 11300–11304.
[25] Wentao Li, Haoxiang Zhou, Weihua Xu, Xi-Zhao Wang, Witold Pedrycz, Interval dominance-based feature selection for interval-valued ordered data, IEEE Trans. Neural Netw. Learn. Syst. (2022).
[26] Zdzisław Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, vol. 9, Springer Science & Business Media, 1991.
[27] Jiye Liang, Junhong Wang, Yuhua Qian, A new measure of uncertainty based on knowledge granulation for rough sets, Inf. Sci. 179 (4) (2009) 458–470.
[28] Guoyin Wang, Hong Yu, et al., Monotonic uncertainty measures for attribute reduction in probabilistic rough set model, Int. J. Approx. Reason. 59 (2015) 41–67.
[29] Wenhao Shu, Wenbin Qian, Yonghong Xie, Incremental feature selection for dynamic hybrid data using neighborhood rough set, Knowl.-Based Syst. 194 (2020) 105516.
[30] Qinghua Hu, Lei Zhang, David Zhang, Wei Pan, Shuang An, Witold Pedrycz, Measuring relevance between discrete and continuous features based on neighborhood mutual information, Expert Syst. Appl. 38 (9) (2011) 10737–10750.
[31] Jihong Wan, Hongmei Chen, Zhong Yuan, Tianrui Li, Xiaoling Yang, BinBin Sang, A novel hybrid feature selection method considering feature interaction in neighborhood rough set, Knowl.-Based Syst. 227 (2021) 107167.
[32] Lin Sun, Jiucheng Xu, Feature selection using mutual information based uncertainty measures for tumor classification, Bio-Med. Mater. Eng. 24 (1) (2014) 763–770.
[33] Hanchuan Peng, Fuhui Long, Chris Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[34] Huijuan Lu, Junying Chen, Ke Yan, Qun Jin, Yu Xue, Zhigang Gao, A hybrid feature selection algorithm for gene expression data classification, Neurocomputing 256 (2017) 56–62.
[35] Pradipta Maji, Sankar K. Pal, Feature selection using f-information measures in fuzzy approximation spaces, IEEE Trans. Knowl. Data Eng. 22 (6) (2009) 854–867.
[36] Swarnajyoti Patra, Prahlad Modi, Lorenzo Bruzzone, Hyperspectral band selection based on rough set, IEEE Trans. Geosci. Remote Sens. 53 (10) (2015) 5495–5503.
[37] Weihua Xu, Kehua Yuan, Wentao Li, Weiping Ding, An emerging fuzzy feature selection method using composite entropy-based uncertainty measure and data distribution, IEEE Trans. Emerg. Top. Comput. Intell. 7 (1) (2022) 76–88.
[38] Jihong Wan, Hongmei Chen, Tianrui Li, Zhong Yuan, Jia Liu, Wei Huang, Interactive and complementary feature selection via fuzzy multigranularity uncertainty measures, IEEE Trans. Cybern. (2021).
[39] Giorgio Roffo, Simone Melzi, Umberto Castellani, Alessandro Vinciarelli, Marco Cristani, Infinite feature selection: a graph-based feature filtering approach, IEEE Trans. Pattern Anal. Mach. Intell. 43 (12) (2020) 4396–4410.
[40] Weihua Xu, Man Huang, Zongying Jiang, Yuhua Qian, Graph-based unsupervised feature selection for interval-valued information system, IEEE Trans. Neural Netw. Learn. Syst. (2023).