Information Sciences 734 (2026) 122997

Contents lists available at ScienceDirect

Information Sciences

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

A novel method of feature selection and information fusion for multi-source ordered information systems based on k-nearest neighbor rough sets

Hao Yuan, Weihua Xu*

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

ARTICLE INFO

Keywords:
Multi-source ordered information system
Feature differences
k-nearest neighbor rough set
Feature selection

ABSTRACT

Feature selection in multi-source ordered decision information systems faces significant challenges, including heterogeneous data sources, high redundancy, and the absence of a unified framework for evaluating feature importance. Existing methods often fail to account for the differential contributions of various information sources, resulting in information loss and degraded classification performance. To address these issues, this paper proposes a novel feature selection method based on *k*-nearest neighbor rough sets. First, the multi-source data is decomposed into individual single-source datasets, and a feature importance ranking mechanism based on intra-class and inter-class differences is applied to each subset. Second, the consensus degree between each information source and the system center is computed to identify the optimal source and determine the weighting coefficients. These weights are subsequently used to aggregate the single-source rankings into a unified global feature importance sequence. Third, within the selected optimal source, the dependency relationship of *k*-nearest neighbor rough sets is employed to evaluate the dependency gain from feature inclusion, thereby effectively eliminating redundant and interfering features. This process achieves substantial reduction in both feature dimensionality and computational overhead while preserving critical discriminative information. To validate the proposed approach, we present the Algorithm 1 and conduct comprehensive experiments on 12 public datasets. The results demonstrate that Algorithm 1 outperforms eight comparative methods. The algorithm effectively performs feature selection while maintaining or even enhancing classification performance, demonstrating particular suitability for data analysis and knowledge discovery in complex multi-source environments.

1. Introduction

The proliferation of digital technology has triggered an exponential surge in information volume, placing unprecedented strain on information processing capabilities [1]. Feature selection, alternatively referred to as attribute reduction, aims to reduce data dimensionality [2] and enhance model interpretability [3]. Moreover, recent approaches utilizing fuzzy multigranularity measures [4] have contributed significantly to big data processing and application [5]. Granular Computing (GrC) [6] has demonstrated significant advantages in extracting value from massive data. The design philosophy of GrC draws upon human problem-solving strategies, namely, decomposing complex problems into multiple simpler sub-problems and addressing them sequentially. Specifically, GrC partitions raw data within the data space into several “granules” based on specific relationships, with each granule containing some data points that share common characteristics. Through this approach, originally complex problems are transformed into discrete,

* Corresponding author.
Email addresses: sanjiu123456789@email.swu.edu.cn (H. Yuan), chxuwh@gmail.com (W. Xu).

https://doi.org/10.1016/j.ins.2025.122997
Received 2 September 2025; Received in revised form 5 December 2025; Accepted 12 December 2025
Available online 17 December 2025

tractable representations. By comprehensively processing these granules, valuable information and knowledge can be extracted from the data. This methodology has been widely applied across various domains, including knowledge discovery [7], decision support [8], and feature selection [9].

Rough set theory, introduced by Pawlak [10] based on information granularity, is grounded in relation and set theories. It has been widely applied in uncertainty computation and incomplete data processing. The theory classifies elements in a set via equivalence relations [11], forming partitions that delineate the data space. The approximate space [12] is constructed from indiscernible equivalence classes. Within this space, upper and lower approximations [13] define the boundaries of vague sets. With the increasing complexity of data forms and the diversification of data types, the demand for more advanced data processing techniques has surged. In this context, scholars have now proposed a wide range of innovative feature selection algorithms based on rough sets. These algorithms aim to address the challenges posed by the intricate nature of modern data and to enhance efficiency and effectiveness in data analysis and application. Concerning the application and extension of rough sets, Wang et al. proposed feature selection and classification based on directed fuzzy rough sets [14], Chen et al. developed a spectral feature selection approach with Kernelized fuzzy rough sets [15], and Yuan et al. constructed a feature selection method using zentropy-based uncertainty measures for heterogeneous data [16]. To meet the specific processing and selection requirements for different data types, a matrix-driven feature selection method for ordered data was established [17] by Xu et al., and Zhang et al. proposed feature selection for generalized multigranulation fuzzy neighborhood rough sets based on composite entropy [18].

Information systems [19] have long constituted a core research subject within rough set theory, providing both the foundational structure for knowledge representation and the principal context for methodological development in the field of information science. Feature selection, as a critical component of information system analysis and processing, plays an indispensable and key role. Nevertheless, classical rough set theory based on equivalence relations exhibits certain limitations since it is restricted to discrete data. When dealing with numerical data, discretization is necessary [20], but this process can lead to information loss and compromise the accuracy of subsequent analysis. To address this issue effectively, many scholars have devoted themselves to research and have successfully proposed the neighborhood rough set model (NRS) based on similarity relations. By incorporating parameters into distance measurement, this model can effectively control the neighborhood radius or specify the number of neighbors. This approach significantly enhances tolerance to inter-object variations. Crucially, for numerical data, prior discretization is unnecessary, thereby preventing data loss and offering a more efficient and accurate method for feature selection. For example, Xia et al. developed an accurate and efficient neighborhood rough set for feature selection [21], Yang et al. proposed a three-way classifier based on granular-ball neighborhood rough sets and uncertainty [22], Zhang et al. proposed graph-driven feature selection for interval-valued data based on granular-rectangular neighborhood rough sets [23].

During the procedure of feature selection, ranking features by importance is paramount. Given a large number of raw features, identifying those with the most significant impact on the target variable is essential. This ranking serves as an effective basis for feature evaluation and selection, as it quantifies each feature's relevance to the target and its contribution to model predictions. Currently, numerous methods exist to evaluate the importance of features in information systems, including Laplacian Score [24], PCA-Based Feature Selection [25], Fisher Score [26] etc. Inspired by linear discriminant analysis [27,28], whose fundamental principle is to reduce intra-class variation and increase inter-class separation using linear projection, we propose a feature importance ranking method based on the difference of features. Upon retaining features with higher importance, we subsequently eliminate redundant features iteratively using the dependency relationship [29] of k-nearest neighbor rough sets [30] integrated with a new distance metric.

In real-world scenarios, data often exhibit complex characteristics such as multiple sources, heterogeneity, high dimensionality, and ordinal properties, which far exceed the effective representation capabilities of traditional single-decision and two-dimensional information systems. Although multi-source information systems have become increasingly prevalent in fields such as healthcare and finance, research on feature selection for such systems remains scarce. Most existing feature selection methods are confined to single-source information systems, and their theoretical assumptions and algorithmic processes are difficult to adapt directly to multi-source ordered decision information systems. To address the challenge of multi-source information fusion, mainstream methods generally adopt the strategy of "extraction-then-integration" that is, extracting so-called "key" information from each information source, and then simplifying the multi-source system into a single-source system for subsequent processing [31,32]. Such methods ignore the differences in the degree of influence of different information sources on the overall system during the dimensionality reduction process, which is prone to information loss and reduces the generalization ability and robustness of the model.

To surmount these limitations, this paper proposes a joint feature selection and information fusion method for multi-source ordered information systems. Unlike conventional approaches, this method holistically integrates all information sources during the feature selection phase. First, for each single source, a feature importance evaluation mechanism based on inter-class and intra-class differences is established, thereby generating an internal feature ranking. Second, by assigning weights to different sources, a weighted fusion strategy is employed to consolidate the ranking results from multiple sources into a unified global feature importance sequence. This approach effectively preserves the discriminative information from each source. Finally, based on the results of fused ranking, the most representative information sources are selected to construct a k-nearest neighbor rough set model. By leveraging the change in dependency, redundant and interfering features are removed to accomplish feature selection in a multi-source environment.

The framework of the proposed method is illustrated in Fig. 1. The key contributions of this paper are summarized as follows:

(1) Diversity quantification model: For multi-source ordered data, this paper introduces a novel "inter-class separation degree and intra-class compactness" into a cross-source comparable difference index DI. This approach not only mitigates the limitations

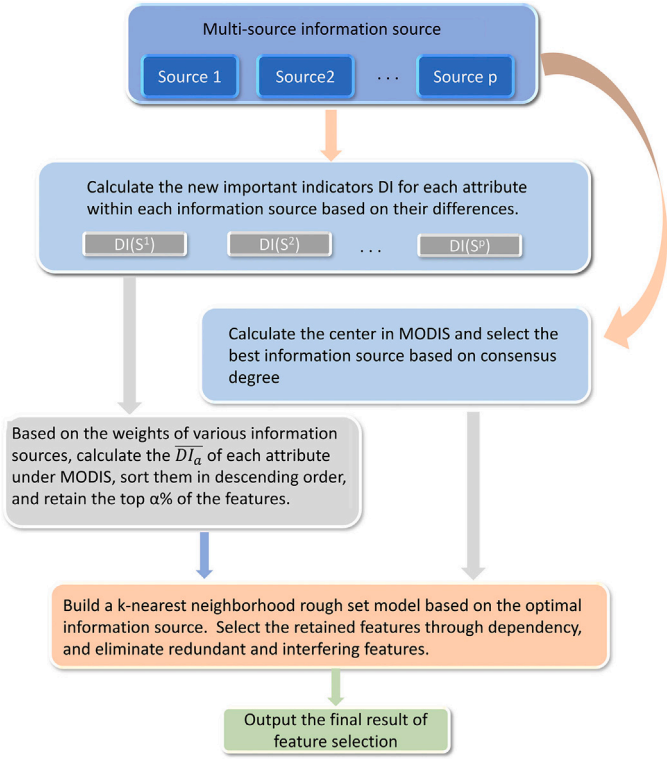


Fig. 1. The structure of this paper.

inherent in single-source measurement perspectives but also provides an interpretable and weighted global metric for the subsequent feature ranking and selection process, laying a solid theoretical foundation for multi-source information fusion.

(2) Ranking - dependency collaborative feature selection algorithm (MKD-UFS): leveraging the global DI ranking, the algorithm initially retains the top α% of the most distinctive features. Subsequently, the optimal information sources from multi-source information systems are construct a k-nearest neighbor rough set model based on these sources. In each iteration, redundant and noisy attributes are eliminated to establish a "ranking-driven and dependency-pruning" collaborative dimensionality reduction mechanism. This mechanism synergizes the complementarity among sources with the discriminative power within sources, thereby achieving a balance between discriminability and simplicity.

(3) Systematic experimental validation: Benchmarking against 8 representative methods on 12 publicly available UCI datasets, MKD-UFS achieves superior or competitive results in terms of classification accuracy, feature reduction rate, and computational efficiency. This confirms the generalizability and effectiveness of the proposed differentiation model and the ranking-dependency collaborative strategy.

The structure of this paper is organized as follows: In Section 2, we introduce the relevant knowledge of multi-source ordered information systems, and then define the k-nearest rough set model of the new distance metric. In Section 3, we define the difference measure (DI) used to evaluate the importance of features from two perspectives: intra-class differences and inter-class differences. In Section 4, we propose an algorithm for feature importance ranking in multi-source ordered information systems, and combine the ranking retained results with the k-nearest neighbors dependency for feature selection. In Section 5, we conduct comparative experiments on the proposed algorithm with the other eight algorithms on 12 datasets, verifying the effectiveness of our algorithm. Finally, we summarize this article in Section 6.

2. PRELIMINARIES

This section formally reviews the basic notions of multi-source ordered decision information systems and the k-nearest neighbor rough set with new metric.

2.1. Multi-source ordered decision information system

An ordered decision information system (ODIS) is a specialized information system where each attribute has an inherent order or preference relationship. This system is commonly used in decision-making processes where the order of attributes is crucial. An ordered decision information system can be represented as a quadruple $ODIS = (U, A \cup \{d\}, V, f)$, where:

- $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects or instances.
- $A = \{a_1, a_2, \ldots, a_m\}$ is a nonempty finite set of attributes or features.
- $d = \{d_1, d_2, \ldots, d_r\}$ is the decision attribute used to classify or make decisions about objects.
- V is the domain of the attributes, representing the set of possible values that each attribute can take.
- $f : U \times A \longrightarrow V$ is an information function that provides the value of each object for each attribute.

In an ordered decision information system, each attribute a_i has an inherent ordered relationship, typically denoted as $\leq$ *or* $\geq$. This order can be natural (e.g. the magnitude of the numerical attribute) or defined based on expert knowledge (e.g., preference relationships).

A multi-source ordered information system can be represented as a quintuple $MODIS = (U, A_p \cup \{d\}, V_p, f_p, w_p)$, $A_p = \{a_1^p, a_2^p, \ldots, a_m^p\}, p = 1, 2, \ldots, s$, where: $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of objects or instances. $A_p, p = 1, 2, \ldots, s$ are p different sets of attributes, each provided by a different source of information. d, V, f are the same functions as in $ODIS$. $w_i, i = 1, 2, \ldots, p$ are the weights of each information source, indicating their importance or credibility in the decision-making process. Let $S_k = (s_{ij}^k)_{n*m}$ be the kth information source IS_k, and $C = (c_{ij})_{n*m}$ denotes the center of the $MODIS$, the consensus degree([33]) of S_k is defined as:

$$WCD_k = 1 - \frac{1}{mn} d(S_k, C). \tag{1}$$

where $d(S_k, C)$ is the Manhattan distance between S_k and C

$$d(S_k, C) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left| s_{ij}^p - c_{ij} \right| \tag{2}$$

The calculation for information source center is

$$C = \sum_{k=1}^{s} \frac{S_k}{s} \tag{3}$$

2.2. *k-nearest neighbor rough set*

Given a decision information system $DIS = (U, A \cup \{d\}, V, f)$, U is the universe, A is conditional feature set, d is decision feature and f is mapping function of attributes. Let x is an object of U, and B is a subset of A, the Euclidean distance d_B is commonly used to assess the difference between objects and determine neighborhood granules in a DIS, and is defined as follows:

$$d_B(x, y) = \sqrt{\sum_{b \in B} |f(x, b) - f(y, b)|^2}. \tag{4}$$

Then the k-nearest neighbor neighborhood rough set of x over B is defined as follows:

$$\kappa_B(x) = \left\{ y \in U \mid \left| d_B(x, y) - d_B(x, y') \right| \leq \left| d_B(x, y_i) - d_B(x, y') \right|, \; y_i, y' \in U, \left| \kappa_B(x) \right| = k \right\} \tag{5}$$

Here, $\kappa_B(x)$ represents the first k objects that are closest to x based on the set of attributes B.

Definition 1. In a $DIS = (U, A \cup \{d\}, V, f)$,x is an object of U, B is a subset of A, $U/d = \{d_1, d_2, d_3 \ldots, d_r\}$. Given a k-nearest neighbor neighborhood relation R, the lower and upper approximations of d_i with respect to B in relation to R are defined as follows:

$$\begin{aligned} \overline{R_B}^{\kappa}(d_i) &= \left\{ x \mid \kappa_B(x) \cap d_i \neq \emptyset, x \in U \right\} \\ \underline{R_B}^{\kappa}(d_i) &= \left\{ x \mid \kappa_B(x) \subseteq d_i, x \in U \right\} \end{aligned} \tag{6}$$

Then the pair $< \overline{R_B}^{\kappa}(di), \underline{R_B}^{\kappa}(di) >$ is called the rough set of neighborhood k-nearest. For convenience, approximation sets are referred to as the lower and upper approximations in this paper.

Property 1: Given a $DIS = (U, A \cup \{d\}, V, f)$, for $B \subseteq A$,

(1) $\emptyset \subseteq \underline{R_B^{\kappa}}(di) \subseteq R_B(di) \subseteq \overline{R_B^{\kappa}}(di) \subseteq U$;
(2) For $P \subseteq B$, $\underline{R_P^{\kappa}}(di) \subseteq \underline{R_B^{\kappa}}(di)$, $\overline{R_B^{\kappa}}(di) \subseteq \overline{R_P^{\kappa}}(di)$;
(3) For $X \subseteq di$, $\underline{R_P^{\kappa}}(X) \subseteq \underline{R_B^{\kappa}}(di)$, $\overline{R_B^{\kappa}}(X) \subseteq \overline{R_P^{\kappa}}(di)$;
(4) For $|\kappa_B^1(x)| \leq |\kappa_B^2(x)|$, $\underline{R_B^{\kappa^2}}(di) \subseteq \underline{R_B^{\kappa^1}}(di)$, $\overline{R_B^{\kappa^2}}(di) \subseteq \overline{R_B^{\kappa^1}}(di)$;

Table 1
An ordered decision information system.

U	a_1	a_2	a_3	a_4	d
x_1	0.95	0.90	0.94	0.92	0
x_2	0.84	0.82	0.86	0.83	0
x_3	0.79	0.81	0.83	0.80	0
x_4	0.75	0.77	0.76	0.78	1
x_5	0.72	0.73	0.72	0.74	1

In order to characterize the precision of kNRS, the accuracy measure of d_i of B is proposed as follows:

$$\gamma_B(di) = \frac{|\underline{R_B^\kappa}(di)|}{|di|} \tag{7}$$

Therefore, the accuracy of d on B can be defined as follows:

$$\gamma_B(d) = \frac{\sum_{i=1}^{r} |\underline{R_B^\kappa}(di)|}{|U|} \tag{8}$$

Here, $\gamma_B(d)$ indicates the capacity of the conditional attribute set B to approximate the decision attribute set d, with a value ranging between 0 and 1.

Example 1. From Table 1, $U/d = \{\{x_1, x_2, x_3\}, \{x_4, x_5\}\}$, where $d_1 = \{x_1, x_2, x_3\}$ and $d_2 = \{x_4, x_5\}$. For $|\kappa_B(x)| = 2$, let $B = \{a_1, a_2, a_4\}$. Then the lower approximation of d_i with respect to B in relation to R can be obtained as follows: $\underline{R_B^\kappa}(d_1) = \{x_1, x_2, x_3\}$ and $\underline{R_B^\kappa}(d_2) = \{x_4\}$. Also,the accuracy of d on B is $\gamma_B(d) = \frac{|\underline{R_B^\kappa}(d_1)| + |\underline{R_B^\kappa}(d_2)|}{|U|} = 0.80$.

3. Feature ranking based difference

To identify the effective features within the dataset, we implement a differentiation-based feature selection strategy. This method evaluates each feature using a designed scoring function, ranks them according to their scores, and selects the top-ranked features to form a subset. The number of features to retain can be determined manually. We use the difference index (DI) as a scoring function to assess the importance of the features. DI quantifies both the inter-class difference (IR_a) and the intra-class difference (IA_a) associated with a specific feature. To eliminate the influence of varying scales, we rank the features by the ratio of IR_a to IA_a. A smaller IA_a combined with a larger IR_a indicates greater feature importance. Based on the final difference-based feature ranking, we select the top $\alpha\%$ of features as the extracted ones, effectively excluding less important features.

Definition 2. Given a $DIS = (U, A \cup \{d\}, V, f)$, the intra-difference ($IA_a$) of objects d_i relation to a can be defined as follows:

$$IA_a(d_i) = \frac{1}{n} \sum_{j=1}^{n} \left| \frac{x_j}{\overline{x_i}} - 1 \right|, \tag{9}$$

where $n = |d_i|$, $\overline{x_i}$ represents the mean values in the objects d_i for the attribute a. The IA_a is a variation of the mean absolute deviation of the values in class d_i for the attribute a, which can avoid the impact of different data scales. It reflects the fluctuation of values within the class and the concentration of the data.

Then the IA_a of DIS for attribute a is defined as follows:

$$IA_a(DIS) = \sum_{i=1}^{r} IA_a(d_i). \tag{10}$$

The term $IA_a(DIS)$ reflects the degree of difference among the internal data of the decision class itself under the condition attribute a. A lower value signifies smaller discrepancies among data instances belonging to the same decision class under attribute a. This implies a higher degree of data concentration within the class, indicating greater stability and reliability.

Definition 3. In $DIS = (U, A \cup \{d\}, V, f)$, the inter-class difference between the objects d_i and d_j in relation to attribute a can be defined as follows:

$$IR_a(d_i, d_j) = |\overline{x_i} - \overline{x_j}|, \tag{11}$$

where $\overline{x_i}, \overline{x_j}$ are the mean values of the objects d_i and d_j. In this formulation, the mean value serves as a centroid for each equivalence class. Pairwise differences reflect the distances separating these equivalence classes, where a larger value indicates greater dissimilarity between them.

Table 2
$IA_a(DIS)$, $IR_a(DIS)$, $DI_a(DIS)$ for each attribute.

	a_1	a_2	a_3	a_4
$IA_a(DIS)$	0.0901	0.0714	0.0718	0.0812
$R_a(DIS)$	0.125	0.0933	0.1366	0.0899
$DI_a(DIS)$	1.3862	1.306	1.817	1.108

Then IR_a of DIS for attribute a is defined as follows:

$$IR_a(DIS) = \sum_{d_i \neq d_j} IR_a(d_i, d_j) \tag{12}$$

The term $IR_a(DIS)$ reflects the differences among the different data in the decision class under the condition attribute a. A higher value corresponds to greater data divergence across different decision classes under attribute a, implying that the separability between distinct decision categories is enhanced.

Definition 4. Let $DIS = (U, A \cup \{d\}, V, f)$, $a \ni A$, $d = \{d_1, d_2, \ldots, d_r\}$, the difference index of DIS for attribute a is defined as follows:

$$DI_a(DIS) = \frac{IR_a(DIS)}{IA_a(DIS)} \tag{13}$$

The value of DI_a is directly proportional to the numerator and inversely proportional to the denominator. A larger DI_a value indicates greater stability of data within the same decision class associated with that attribute, as well as enhanced separability between different decision categories. Consequently, such an attribute holds higher importance; conversely, a smaller DI_a value signifies lower attribute importance.

Example 2. From Table 1, this ordered decision information system encompasses the universe $U = \{x_1, x_2, \ldots, x_5\}$, and attributes $A = \{a_1, a_2, a_3, a_4\}$. Subsequently, the values $IA_a(DIS)$, $IR_a(DIS)$, $DI_a(DIS)$ for each attribute can be obtained, as illustrated in Table 2. Based on the descending order of $DI_a(DIS)$ values, the resulting order (C) is $\{a_3, a_1, a_2, a_4\}$. Accordingly, we have derived the attribute ranking by importance in descending order. In this scenario, we designate the retention rate α as 75 %. As a result, the retained features are retention(C) $= \{a_3, a_1, a_2\}$, with attribute a_4 being directly filtered out.

4. Feature selection algorithm

In the previous section, we ranked features based on their importance levels, placing the more important ones at the front. We did not take into account the relationships among the features. In order to eliminate redundant and interfering features among the sorted features, we use the dependency relationship of k-nearest neighbor rough sets to establish the significance indicator for attribute a.

Given $DIS = (U, A \cup \{d\}, V, f)$, $a \in A, B \subseteq A$, the significance of attribute a in relation to d is defined as follows:

$$SIG(a, B, d) = \gamma^{\kappa}_{B\cup\{a\}}(d) - \gamma^{\kappa}_{B}(d) \tag{14}$$

In the context of the decision d, $SIG(a, B, d)$ serves as a metric to quantify the incremental importance of the attribute a compared to the attribute subset B. According to Property 1.(2), $SIG(a, B, d)$ has monotonicity and is always greater than or equal to 0. However, this pertains exclusively to the theoretical data set, which is devoid of any noise. In real-world data scenarios, noise is frequently present. If the added attributes contain noise or irrelevant interference information that interferes with the decision, originally consistent equivalence classes may be erroneously subdivided, generating conflicting subclasses (that is, the decision values in the subclasses are inconsistent), thereby leading to the reduction of the positive domain. Therefore, redundant and interfering features must be excluded, and only the features with $SIG(a, B, d) > 0$ are selected.

For an $ODIS$, after obtaining the feature ranking through the feature difference, it only needs to further invoke formula 14 to eliminate redundant and interfering features, thereby obtaining the final result. However, for a $MODIS$, due to the presence of multiple information sources, the result of feature ranking from one information source cannot be considered the result of the whole $MODIS$. The importance degree of features within one information source is only relevant to itself. Therefore, it is necessary to compute the importance degree DI_a of each feature within each information source (DI^i_a), subsequently perform a weighted summation to obtain the degree of feature importance $\overline{DI_a}$ under the entire $MODIS$, and then obtain the ranking result in descending order. Finally, $SIG(a, B, d)$ is applied to prune the feature set and obtain the final selection.

Given a $MODIS = (U, A_p \cup \{d\}, V_p, f_p, w_p)$, for $A_p = \{a^p_1, a^p_2, \ldots, a^p_m\}, p = 1, 2, \ldots, s$. The $\overline{DI_{a_i}}$ of the attribute a_i in $MODIS$ can be defined as follows:

$$\overline{DI_{a_i}} = \sum_{p=1}^{s} w_p \cdot DI_{a_i}(S^p) \tag{15}$$

Leveraging the feature ranking results, we proceed to select an optimal information source to construct the k-nearest neighbor rough set and subsequently employ $SIG(a, B, d)$ to eliminate redundant or interfering features. To facilitate source selection, the WCP of each information source is calculated using formulas 1-3. The information source with the largest WCP is chosen as the optimal

Table 3
An example of $MODIS$ is given.

U	$S_1, w_1 = 0.25$				$S_2, w_2 = 0.5$				$S_3, w_3 = 0.25$				d
	a_1^1	a_2^1	a_3^1	a_4^1	a_1^2	a_2^2	a_3^2	a_4^2	a_1^3	a_2^3	a_3^3	a_4^3	
x_1	0.30	0.26	0.05	0.54	0.11	0.61	0.13	0.24	0.67	0.70	0.61	0.30	2
x_2	0.47	0.63	0.97	0.90	0.32	0.85	0.66	0.54	0.20	0.19	0.79	0.29	1
x_3	0.91	0.52	0.52	0.18	0.29	0.73	0.39	0.80	0.65	0.30	0.14	0.40	2
x_4	0.95	0.42	0.86	0.67	0.25	0.56	0.86	0.22	0.31	0.24	0.58	0.24	1
x_5	0.62	0.27	0.89	0.20	0.40	0.31	0.38	0.95	0.74	0.72	0.69	0.10	2
x_6	0.40	0.99	0.73	0.44	0.84	0.97	0.38	0.95	0.94	0.50	0.89	0.19	1
x_7	0.56	0.41	0.72	0.40	0.44	0.67	0.82	0.89	0.59	0.96	0.99	0.21	2

Table 4
Feature ranking calculation process.

	a_1	a_2	a_3	a_4
$DI_a(S1)$	0.008942	0.5450	0.5377	0.5205
$DI_a(S2)$	0.1818	0.4974	0.2823	0.1928
$DI_a(S3)$	0.2267	0.3323	0.1356	0.0238
$\overline{DI}_a$	0.1498	0.4680	0.3094	0.2324

Table 5
The center of $MODIS$.

U	a_1	a_2	a_3	a_4
x_1	0.36	0.52	0.45	0.36
x_2	0.33	0.37	0.87	0.62
x_3	0.68	0.52	0.35	0.48
x_4	0.50	0.41	0.77	0.18
x_5	0.44	0.32	0.74	0.42
x_6	0.73	0.82	0.67	0.53
x_7	0.72	0.73	0.87	0.41

Table 6
Feature ranking calculation process.

step	C	B	$\gamma_B(d)$	selected features
1	a_2	$\{a_2\}$	0.2857	$\{a_2\}$
2	a_3	$\{a_2, a_3\}$	0.2857	$\{a_2\}$
3	a_4	$\{a_2, a_4\}$	0.1428	$\{a_2\}$

information source. This selected source is utilized not only to construct the rough set but also serves as the basis for evaluating classification performance in the experimental validation.

Example 3. Consider a multi-source medical risk assessment system (MODIS) for evaluating the risk of cardiovascular diseases in patients. The system includes 7 patients($U = \{x_1, x_2, \ldots, x_7\}$). Each patient is measured for four physiological indicators (a_1: blood pressure risk, a_2: blood sugar risk, a_3: blood lipid risk, a_4: heart rate variability risk) by three different medical detection devices (information sources S_1, S_2, S_3). The numerical range is from 0 to 1. A value of 1 indicates the highest risk. The weights of each information source are as follows: $w_1 = 0.25$, $w_2 = 0.5$, $w_3 = 0.25$, reflecting the credibility of the equipment in the diagnostic process. The decision attribute d represents the final risk assessment result (1 = low risk, 2 = high risk). The MODIS is shown in Table 3.

According to formula Definitions 2–4, we can compute $DI_{a_i}(S^p)$ for the attribute a_i from the source of information pth. Using formula 15, we can obtain the final DI_{a_i} for the entire $MODIS$. Then, sort the attributes in descending order, placing the most important attributes at the top. The ranking result is $\{a_2, a_3, a_4, a_1\}$. Here, we set $\alpha = 0.75$ and obtain the retention(A)$= \{a_2, a_3, a_4\}$. Next, select the one with the largest WCP as the best information source. For the selection of the best information source, first calculate the information source center of $MODIS$ using formula 3, which is shown in Table 5. Next, the Manhattan distances between each information source and the information source center were further calculated, and the results are $d(S_1, C) = 4.94, d(S_2, C) = 4.69, d(S_3, C) = 5.29$. Finally, using formula 1, we compute the $WPC_1 = 0.8233$, $WPC_2 = 0.8325$, $WPC_3 = 0.8110$. Therefore, S_2 is selected as the best source of information to carry out the rough analysis. Finally, we use formula 14 under S_2 to eliminate redundant and interfering attributes to obtain the selected set Red. In this case, we set the value of $k = 2$.

We can calculate $\gamma_{a_2}(d) = 0.2857$. Therefore, we retain a_2. Similarly, when $B = \{a_2, a_3\}$, we calculate $\gamma_{\{a_2,a_3\}}(d) = 0.2857$, and obtain $SIG = 0$. We can consider a_3 to be redundant and thus do not choose a_3. When $B = \{a_2, a_4\}$, we calculate $\gamma_{\{a_2,a_4\}}(d) = 0$, and $SIG < 0$. At this point, we can consider that a_4 is interfering with the classification of the decision making and, therefore, do not choose a_4. Thus, the final selection result is a_2. All of these are shown in Tables 4,6.

Algorithm 1 A feature selection difference-based algorithm for multi-source ordered decision information.

Input: $MODIS = \left(U, A_p : \left\{a_1^p, a_2^p, \ldots, a_m^p\right\} \cup \{d\}, V_p, f_p, w_p\right)$ for $p = 1, 2, \ldots, s$
Output: Selected feature set Red
1: Initialize $Red \longleftarrow \emptyset$
2: $DI(MODIS) \longleftarrow \{\}$
3: **for** $p = 1$ to s **do**
4: **for all** $a \in A_p$ **do**
5: Compute $DI_a(S^p)$ by formula (13)
6: **end for**
7: Get $DI(DIS_p) = [DI_{a_1}(S^p), \ldots, DI_{a_m}(S^p)]$
8: $DI(MODIS) \longleftarrow DI(S^p) \cup DI(MODIS)$
9: **end for**
10: $\overline{DI} = \sum_{p=1}^{s} DI(MODIS)[p] * w_p$ Calculate the final importance of features.
11: Order attributes in descending order by $\overline{DI}$ and denote as result order(A).
12: Put the top $\alpha\%$ attributes for order(A) into retention(A).
13: **for all** $a \in retention(A)$ **do**
14: Compute $\gamma_{Red}(d)$ and $\gamma_{Red\cup\{a\}}(d)$.
15: $SIG(a, Red, d) = \gamma_{Red\cup\{a\}}(d) - \gamma_{Red}(d)$
16: $Red \longleftarrow Red \cup \{a\}$, where $SIG(a, Red, d) > 0$.
17: **end for**
18: **if** $Red = \phi$ **then**
19: $Red = retention(A)[0]$ Red is the top-ranked attribute.
20: **end if**
21: Return Red

In Algorithm 1, Steps 3–10 are dedicated to evaluating the importance of each feature across multiple data sources, including computing the divergence indicator (DI) for each feature and aggregating the DI values from all data sources. The resulting time complexity is O(s × m × n + s × m), where s is the number of data sources, m is the number of features per data source and n is the number of instances. Steps 11–13 rank the features in descending order of their importance and retain a certain proportion of them, incurring a sorting complexity of O(mlogm) for the sorting operation. Subsequently, steps 14–17 select non-redundant features from the retained features by computing the dependency change (SIG) for each feature relative to the currently selected feature set, which contributes a time complexity of O(m × n). Steps 18–21 perform a final validation on the selected feature set with a time complexity of O(1). In general, the algorithm's time complexity is O(s × m × n).

5. Experimental decision and analysis

In this section, we first conduct a set of experiments to validate the effectiveness of the differential measurement index DI. Subsequently, we employ a series of comparison algorithms to demonstrate the superiority of the feature selection method proposed in this study.

5.1. Experimental design

All the algorithms discussed herein, including the proposed algorithm and the comparison algorithms, were implemented in Python 3.8 using PyCharm 2023. The experimental evaluations were conducted on a computing workstation with the following specifications: an AMD Ryzen 7 6800H processor clocked at 4.7 GHz, Radeon series graphics, 16.0 GB of RAM, and a 64-bit Windows 11 operating system. To demonstrate the generalizability of the proposed method across different tasks and domains, we carefully selected 12 publicly available datasets from the UCI Machine Learning Repository, covering a wide range of application areas such as engineering, finance, medicine, and image recognition. These datasets exhibit significant diversity in terms of sample size, feature dimensionality, and number of classes. Specifically, the sample sizes range from 178 (Wine) to 21,263 (Superconductivity), the number of features varies from 8 (Concrete, Rice) to 166 (Musk), and the number of classes ranges from 2 (e.g., Rice, Musk) to 26 (e.g., Letter Recognition), encompassing binary, multiclass, and multi-category classification tasks, which ensures strong representativeness. Detailed characteristics of the datasets are presented in Table 7.

To guarantee that the datasets faithfully represent a multi-source ordered information decision system, a series of specific data processing procedures is implemented: (1) First, all datasets undergo a filtering process to remove non-numeric features, retaining only those attributes that are orderable within their domain. (2) Second, max-min normalization is applied across all datasets to ensure scale consistency. (3) Finally, as the datasets used in this experiment are originally single-source two-dimensional tables, multi-source information systems are constructed by introducing white noise to 50 % of the data for each information source. The formulation below illustrates the construction process, where each information source is derived by introducing white noise to a

Table 7
The summary of datasets.

NO.	Datesets	Object	Attributes	Class
Data1	Concrete	1030	8	4
Data2	Wine	178	13	3
Data3	istanbul stock exchange	536	10	4
Data4	hungarian chickenpox case	521	20	3
Data5	rice	3810	8	2
Data6	statlog	6635	36	6
Data7	superconductivity-data	21,263	81	3
Data8	waveform	5000	21	3
Data9	pen-based recognition	10,992	16	10
Data10	letter recognition	20,000	16	26
Data11	taiwanese bankruptcy prediction	6819	96	2
Data12	musk	6598	166	2

random half of the dataset. The specific operation is defined as follows:

$$V_i(x,a) = \begin{cases} v(x,a) + n_i, \text{if } x \text{in } U_{\text{random}} \\ v(x,a), \text{else} \end{cases} \tag{16}$$

where, $v(x,a)$ denotes the value of object x with respect to attribute a, the set U_{random} consists of 50 % of the dataset, and the values n_i are normally distributed with parameters set to a mean of 0 and a variance of 0.1. Here, the weight for each information source is evenly distributed. The sum of the weights of the other information sources is 1.

First, to validate the effectiveness of DI, we ranked the features in descending order based on their difference scores. Subsequently, we retained different proportions of features and compared the classification accuracy. To investigate the impact of DI on classification accuracy, we set the range of the feature retention rate α from 0.5 to 1, with a step size of 0.05. When $\alpha = 1$, the retained feature set is identical to full original feature set.

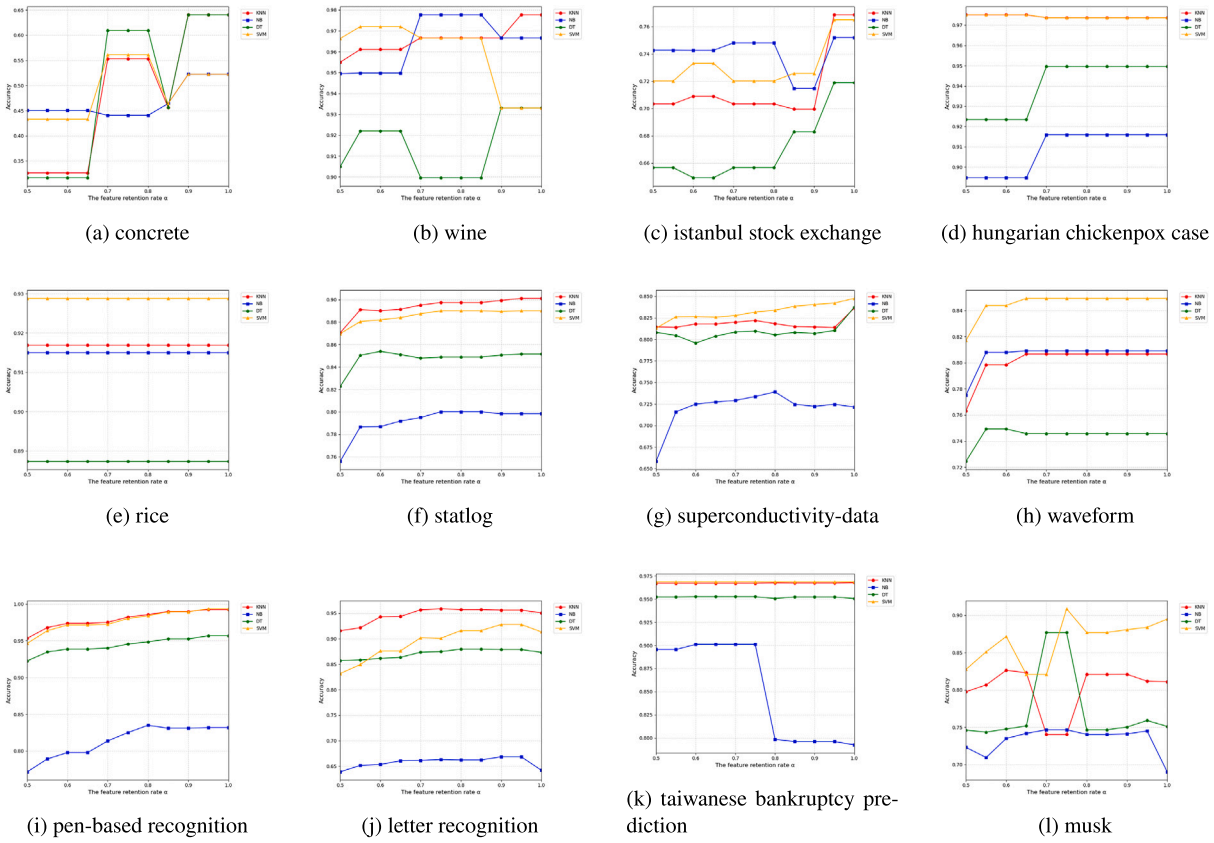
Second, although the proposed MKD-UFS method is specifically designed for feature selection in multi-source ordered information systems, existing algorithms tailored to this specific domain remain scarce. As a result, we selected eight representative single-source feature selection methods as baselines for comparison, including Laplacian score for feature selection (LS)([24]), principal component analysis (PCA)([25]), Attribute Reduction Based on Neighborhood Conditional Mutual Information (KNCMI)([34]), rough set based on the relative stability of local redundancy (RSLRS)([35]), matrix-based feature selection approach using entropy for ordered set([17]), feature selection for dynamic interval-valued ordered data(HFS-IVO)([36]), feature selection using the weighted dominance-based neighborhood rough sets(WDNCE-HAR)([37]), Zentropy-Based uncertainty measure for heterogeneous feature selection(Ze-HFS)([16]). These methods are widely used in the field of feature selection and represent mainstream research directions, encompassing diverse theoretical frameworks such as fuzzy rough sets, information entropy, graph structures, and statistical measures. While they were originally designed for single-source systems, their core ideas—such as feature importance evaluation and redundancy removal—are conceptually comparable to our approach. Therefore, they serve as reasonable benchmarks to demonstrate the effectiveness, classification performance, and efficiency advantages of the proposed MKD-UFS method.

Finally, we analyze the influence of two parameters, the feature retention rate α and the number of neighborhoods k, on the final classification accuracy under different parameter combinations. Prior to the formal experiments, preliminary studies were conducted to optimize the key parameter configurations. The setting of the parameter k underwent a refinement process. Initially defined as a proportion of the total number of objects, this setting was found to be too sparse in pre-experiments, often resulting in a dependency degree $\gamma_{red}(d)$ of zero and failing to construct a valid rough set. We reasoned that the dimensionality of the feature space is more critical to the formation of neighborhood structures. Consequently, k was redefined as a proportion of the number of features in the currently selected subset. Based on this, its value range was set from 0.05 to 0.5 with a step size of 0.05, designed to encompass a spectrum from local to global neighborhood relations. As for the feature retention rate, its initial range was set empirically from 0.5 to 1.0 (step size 0.05), based on the consideration that retaining a majority portion of the pre-ranked features would preserve most of the essential information while aiming to strike a balance between keeping important features and eliminating redundancy.

To ensure the robustness of the experimental results and mitigate the influence of data randomness, we employed four standard classifiers: KNN, NB, DT and SVM. To a counteract potential bias arising from the division of the training set and the test set, we used fivefold cross-validation to evaluate the final classification accuracy. In the fivefold differential validation, the dataset is divided into five equal parts. Each part is taken as a test set once, and the remaining four parts are used for model training. The classification accuracy in the result report is the average value (μ) plus or minus the standard deviation, expressed in the form of $\mu \pm \sigma$.

5.2. Analysis of comparative experiments:

(1) Effectiveness analysis of DI: Fig. 2 illustrates the impact of attribute ranking based on DI and the feature retention rate on classification accuracy. The results demonstrate that retaining a specific proportion of features can enhance or maintain the classification accuracy within a certain range, which strongly confirms the effectiveness of the DI difference measure proposed in this paper in measuring the importance of attributes. In addition, eliminating attributes of low importance can not only significantly improve the classification accuracy rate, but also effectively shorten the time required for the feature selection



(a) concrete (b) wine (c) istanbul stock exchange (d) hungarian chickenpox case

(e) rice (f) statlog (g) superconductivity-data (h) waveform

(i) pen-based recognition (j) letter recognition (k) taiwanese bankruptcy prediction (l) musk

Fig. 2. Classification accuracy of different feature retention rates α under KNN, NB, DT, and SVM classifiers.

process. When the dataset contains numerous attributes, adjusting different α values helps ensure the stability of classification accuracy. This suggests that, in appropriate scenarios, lowering the α value is a justifiable strategy to balance efficiency and performance.

For most datasets, as α increases, the classification accuracy exhibits an overall upward trend. This phenomenon occurs because the DI index preferentially retains features with large inter-class distance and small intra-class distance; as α increases, more high-DI attributes are included, leading to a monotonic increase in the model's discriminative information. Accuracy naturally rises until the saturation zone is reached. However, once α enters the saturation region ($\approx$0.75 in our tests), the additional features retained are mostly redundancies of medium or low DI. These are later identified as "zero-gain" by the k-neighborhood dependency mechanism and are automatically discarded. As a result, the effective feature set scarcely expands, and the accuracy stabilizes within a high but narrow band, sometimes even declining slightly due to noise accumulation.

However, for a small portion of the data, the classification accuracy rate may undergo sudden changes within a specific range, necessitating careful parameter selection. For example, in Data 12, for DT, the classification accuracy is 75.17 %. When α is set to 0.7, the classification accuracy rate rises to 87.69 %. Then, when $\alpha = 0.75$, the classification accuracy rate remains, but when $\alpha = 0.8$ the classification accuracy rate drops back to 74.66 %. The root cause is that for $\alpha < 0.7$ the DI filter removes pivotal features and $\gamma_{Red}(d)$ remains low; within $[0.7, 0.75]$ all high-DI attributes are kept and the subsequent k-NN dependency purge effectively eliminates redundancy, pushing $\gamma_{Red}(d)$ to its peak; once $\alpha > 0.8$ medium- and low-DI features flood in, the dependency degree judges them as zero-gain, the positive region no longer expands, and increased noise drags accuracy down.

Furthermore, the experimental findings demonstrate the remarkable robustness of MKD-UFS across a wide spectrum of datasets. Its exceptional adaptability to diverse data characteristics and the capacity to consistently maintain high classification accuracy position it as a highly reliable method for feature selection tasks. The method's remarkable efficiency in reducing feature dimensions while effectively preserving crucial information significantly contributes to the overall enhancement of the performance of machine learning models. By skillfully leveraging the DI measure and the flexible feature selection mechanism, MKD-UFS offers a comprehensive and integrated solution for optimizing feature subsets in a multitude of diverse applications.

(2) Analysis of Comparative Experiments: For the single information source selection, the information source used earlier to construct the k-nearest neighbor rough set was employed for experimental classification performance evaluation. The average running time of each algorithm is detailed in Table 8, and the classification accuracy of the selected features using DT and SVM classifiers is presented in Tables 9 and 10. The evaluation results indicate that MKD-UFS possesses significant advantages in

Table 8
Comparison of runtime of different algorithms (S).

Datasets	PCA	LS	KNCMI	RALRS	HAR	HFS-IVO	WDNCE-HAR	Ze-HFS	Ours
Data1	**0.18**	1.3	550.46	0.25	0.51	6.72	1.45	8.3	1.12
Data2	0.47	0.15	932.19	**0.13**	0.16	0.82	0.42	5.74	0.24
Data3	0.47	0.24	431.65	**0.11**	0.19	2.56	0.62	22.97	**0.11**
Data4	0.46	0.18	662.35	**0.14**	0.45	17.35	1.05	34.16	0.51
Data5	**0.45**	0.65	1573.25	0.85	7.42	62.16	21.46	280.4	2.31
Data6	1.98	1.89	3540.9	**1.25**	404.67	10,853.2	12,032.2	562.4	118.42
Data7	19.25	28.33	4761.1	**2.38**	3237.5	35,122.6	21,364.4	2634.2	1608.2
Data8	**1.43**	2.25	7451.2	42.97	63.52	1857.4	350.17	1503.2	18.54
Data9	2.92	4.32	1176.4	2.98	93.99	2126.3	7640.1	2436.7	**2.84**
Data10	18.36	29.69	1263.5	**3.49**	1707.0	25,319.1	11,534.2	1895.3	1123.43
Data11	**1.01**	1.51	11,656.4	27.27	426.22	8651.1	3456.4	846.5	92.59
Data12	3.12	4.31	3152.3	**2.32**	1640.5	10,154.4	7631.2	1546.3	209.22
Average	**4.17**	7.06	3040.77	7.01	631.84	7847.80	5336.13	981.34	264.79

Table 9
Classification accuracy of different algorithms under DT classifier (%).

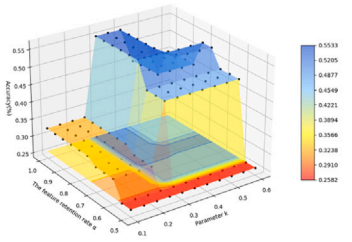
Datasets	RAW	PCA	LS	KNCMI	RALRS	HAR	HFS-IVO	WDNCE-HAR	Ze-HFS	Ours
Data1	62.13 ± 9.21	44.17 ± 3.67	63.78 ± 1.43	62.81 ± 8.58	63.89 ± 8.87	63.89 ± 8.87	28.73 ± 10.5	62.13 ± 9.24	62.13 ± 9.24	**65.43±5.77**
Data2	86.53 ± 4.40	90.91 ± 0.64	89.48 ± 5.86	92.15 ± 3.20	91.53 ± 2.82	88.84 ± 7.2	89.93 ± 4.08	**93.30±3.29**	86.53 ± 4.40	**93.30±3.29**
Data3	69.21 ± 1.77	64.94 ± 4.96	72.66 ± 3.78	74.07 ± 2.76	61.56 ± 1.79	69.21 ± 1.60	58.02 ± 2.33	69.21 ± 1.60	63.25 ± 3.33	**75.83±2.59**
Data4	95.40 ± 2.57	88.61 ± 1.65	93.76 ± 2.19	95.21 ± 1.05	**99.83±0.3**	**99.83±0.3**	96.93 ± 0.9	**99.83±0.3**	92.92 ± 3.08	95.48 ± 2.34
Data5	**89.83±2.0**	88.61 ± 1.65	88.87 ± 1.35	89.01 ± 2.44	88.84 ± 7.90	89.13 ± 2.25	88.81 ± 1.48	88.50 ± 2.04	88.41 ± 1.43	88.21 ± 2.27
Data6	83.66 ± 0.80	69.19 ± 1.13	80.31 ± 1.54	82.70 ± 2.4	81.38 ± 1.90	83.14 ± 1.85	76.34 ± 2.07	82.08 ± 1.48	81.21 ± 3.10	**84.21±0.3**
Data7	83.06 ± 8.66	90.46 ± 0.57	84.81 ± 0.57	65.64 ± 0.3	70.06 ± 3.90	**93.89±7.42**	81.32 ± 1.42	79.21 ± 3.41	64.64 ± 0.30	83.72 ± 3.21
Data8	74.7 ± 1.35	74.12 ± 1.06	63.42 ± 0.62	76.2 ± 1.0	74.74 ± 0.12	74.28 ± 1.58	62.75 ± 1.60	68.64 ± 1.06	64.12 ± 6.3	**76.41±0.86**
Data9	95.47 ± 0.36	90.17 ± 0.61	92.70 ± 0.51	95.56± 7.2	95.43 ± 0.41	95.46 ± 0.43	95.00 ± 0.54	93.58 ± 0.54	94.21 ± 3.27	**95.68±0.47**
Data10	85.51 ± 0.54	84.20 ± 0.72	76.04 ± 0.80	43.51 ± 0.52	87.82 ± 0.40	68.94 ± 14.30	64.23 ± 6.1	62.41 ± 5.21	82.41 ± 0.81	**87.88±1.53**
Data11	94.72 ± 0.27	95.21 ± 0.41	95.19 ± 0.27	94.14 ± 5.1	92.41 ± 4.1	94.99 ± 0.98	92.31 ± 6.31	91.46 ± 4.35	93.21 ± 4.12	**95.23±1.21**
Data12	65.96 ± 14.79	73.31 ± 0.53	77.02 ± 11.46	53.12 ± 0.32	71.99 ± 14.43	68.51 ± 4.31	68.72 ± 3.61	68.52 ± 8.21	67.26 ± 4.85	**77.99±12.58**
Average	82.18 ± 3.89	79.90 ± 1.47	81.46 ± 1.47	76.76 ± 2.66	81.95 ± 3.91	82.50 ± 4.25	75.26 ± 3.41	79.90 ± 3.39	78.35 ± 3.42	**84.08±3.02**

Table 10
Classification accuracy of different algorithms under SVM classifier (%).

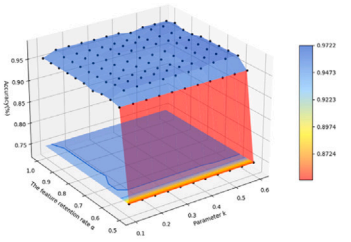
Datasets	RAW	PCA	LS	KNCMI	RALRS	HAR	HFS-IVO	WDNCE-HAR	Ze-HFS	Ours
Data1	51.45 ± 7.34	54.73 ± 1.18	**64.80±3.57**	56.21 ± 11.05	51.45 ± 7.24	51.45 ± 7.24	42.61 ± 7.40	51.45 ± 7.24	51.45 ± 7.24	58.05 ± 7.12
Data2	97.76 ± 2.0	91.64 ± 0.97	97.93 ± 2.75	97.79 ± 2.08	94.98 ± 5.08	97.20 ± 1.75	94.39 ± 2.46	96.65 ± 3.23	97.76 ± 2.08	**98.33±2.22**
Data3	77.43 ± 2.78	70.79 ± 3.04	76.04 ± 2.69	73.70 ± 3.12	69.21 ± 2.27	78.17 ± 3.0	68.10 ± 2.22	78.17 ± 3.0	73.14 ± 3.72	**78.92±1.59**
Data4	97.51 ± 0.46	92.48 ± 0.98	97.60 ± 0.01	97.51 ± 0.40	**99.04±0.6**	98.04 ± 0.60	97.51 ± 0.4	98.04 ± 0.60	94.51 ± 4.60	97.51 ± 0.46
Data5	92.91 ± 0.65	92.48 ± 0.98	92.68 ± 1.12	92.67 ± 2.03	92.65 ± 1.70	92.80 ± 1.8	92.80 ± 1.84	92.91 ± 1.69	91.14 ± 1.47	**93.16±1.94**
Data6	89.16 ± 1.15	76.39 ± 1.35	85.70 ± 1.55	86.74 ± 0.40	87.02 ± 0.90	88.26 ± 1.05	82.57 ± 1.71	87.55 ± 1.06	89.11 ± 2.10	**89.27±2.41**
Data7	87.21 ± 1.21	83.68 ± 0.47	81.07 ± 0.36	62.41 ± 0.14	68.94 ± 0.40	85.19 ± 12.9	64.32 ± 0.45	69.21 ± 1.52	60.91 ± 0.14	**88.22±1.01**
Data8	86.56 ± 0.97	84.02 ± 1.81	73.06 ± 0.72	**86.76±8.30**	83.82 ± 0.91	84.82 ± 0.85	73.56 ± 0.89	78.18 ± 1.24	76.15 ± 1.36	85.33 ± 1.29
Data9	**99.30±0.12**	93.39 ± 0.63	95.38 ± 0.43	98.62 ± 0.00	94.25 ± 0.25	99.30 ± 0.19	97.93 ± 0.34	93.58 ± 0.54	94.21 ± 3.47	**99.30±0.12**
Data10	91.50 ± 0.59	82.18 ± 0.97	70.15 ± 1.01	53.61 ± 5.13	92.59 ± 0.50	77.49 ± 8.92	76.51 ± 2.14	74.16 ± 6.31	74.32 ± 1.46	**92.79±4.53**
Data11	96.68 ± 1.2	**96.81±0.46**	**96.81±0.04**	**96.81±0.04**	92.11 ± 3.1	96.77 ± 0.46	94.85 ± 2.3	92.31 ± 4.45	93.21 ± 4.12	96.68 ± 1.2
Data12	84.99 ± 10.36	81.64 ± 0.97	80.89 ± 10.94	51.42 ± 5.14	**87.70±7.63**	77.49 ± 8.92	77.16 ± 4.9	76.59 ± 3.21	67.26 ± 4.85	84.36 ± 2.65
Average	87.72 ± 2.17	83.51 ± 1.15	84.50 ± 2.09	79.52 ± 3.15	84.23 ± 2.54	85.58 ± 3.97	80.19 ± 2.25	82.40 ± 2.75	80.13 ± 3.05	**88.49±2.21**

feature selection. Across 12 datasets, it achieved the highest accuracy for the DT classifier in 9 cases and performed best for SVM in 6 cases, securing the highest average accuracy among all baseline methods. This performance is attributed to its capability to effectively identify and remove redundant and noisy features. In terms of computational efficiency, PCA and RALLS exhibit the fastest average running time; MKD-UFS ranks fourth in terms of average time consumption due to the computation overhead introduced by k-nearest neighborhood rough sets for redundancy elimination. However, on low-dimensional datasets (such as Data2 and Data3), its time overhead is negligible (<0.3 s). For high-dimensional or multi-class datasets, this algorithm consistently trades an additional 30–120 s of computation for a $3\,\% - 8\,\%$ increase in accuracy, demonstrating that it achieves significant performance gains at a moderate computational cost. In conclusion, MKD-UFS not only comprehensively leads in classification accuracy, but also offers flexibility in running time via parameter adjustment(k and α), successfully achieving a favorable balance between precision and efficiency.

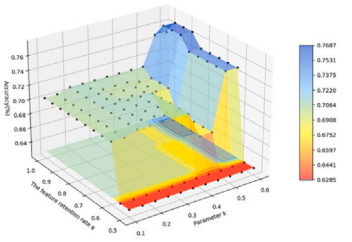
(3) Parameter analysis: To explore how feature retention rate α and neighborhood size k affect the MKD-UFS algorithm's classification accuracy, we plotted $3D$ surface plots for 12 datasets. The results for the first six datasets on KNN are shown in the



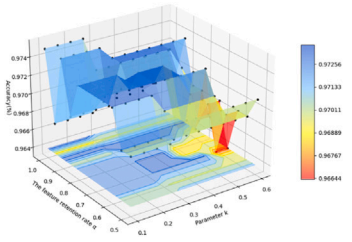
(a) concrete



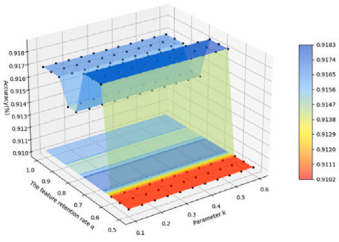
(b) wine



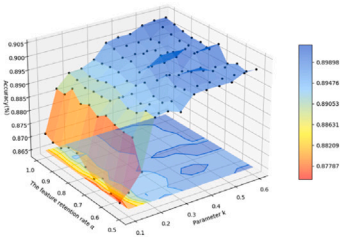
(c) istanbul stock exchange



(d) hungarian chickenpox case

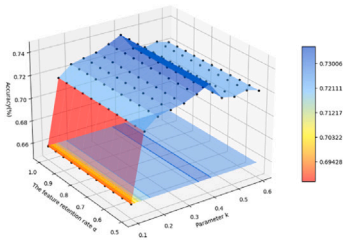


(e) rice

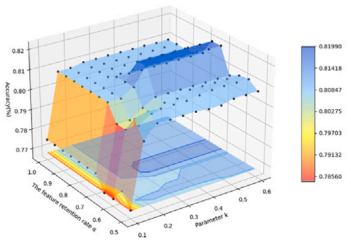


(f) statlog

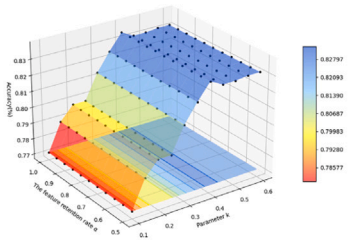
Fig. 3. Classification accuracy of different combinations of α and k in KNN classifier.



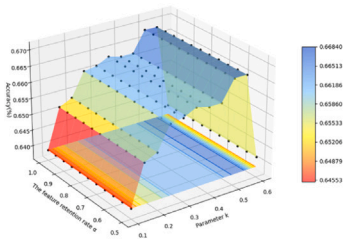
(a) superconductivity-data



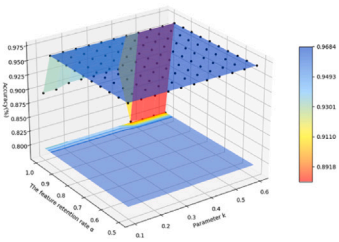
(b) waveform



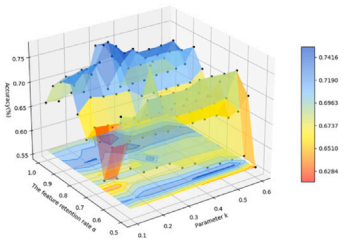
(c) pen-based recognition



(d) letter recognition



(e) taiwanese bankruptcy prediction



(f) musk

Fig. 4. Classification accuracy of different combinations of α and k in NB classifier.

Fig. 3, and the results for the last six datasets on NB are shown in the Fig. 4. These plots reveal that for different datasets, varying combinations of α and k can significantly impact classification accuracy. In Fig. 3, for instance, in Data 3, setting k to 0.5 and α to 1 yields the highest accuracy of 76.87 %. In Data 6, an accuracy of 90.23 % is achieved with $k = 0.25$ and $\alpha = 0.95$, highlighting the algorithm's relative optimality. For most datasets, different α and k combinations lead to varying accuracies. In Fig. 4, the accuracy of Data 8 is 76.88 % when $k = 0.1$ and $\alpha = 0.5$, 77.52 %, when $k = 0.45$ and $\alpha = 0.5$, and 80.92 %, when $k = 0.45$ and $\alpha = 0.7$. However, in Data 11, the $3D$ plot shows a flat plane, indicating that multiple α and k combinations can result in the same accuracy.

Table 11
ANOVA results of data 1 on KNN.

	sum-sq	df	F	PR(>F)	eta-sq
C(k-level)	0.5958	10	4.3047	0.0005	0.4182
C(α-level)	0.6463	10	4.6697	0.0001	0.4537
C(k-level):C(α-level)	1.3840	100	1.00	0.5000	0.9715

Table 12
ANOVA results of data 12 on NB.

	sum-sq	df	F	PR(>F)	eta-sq
C(k-level)	0.2113	10	7.9105	2.944e-9	0.7910
C(α-level)	0.0141	10	0.5303	0.08	0.0530
C(k-level):C(α-level)	0.2671	100	1.00	0.5000	1.000

Furthermore, we conducted a two way ANOVA on the Data 1 of Fig. 3 and the Data 12 of Fig. 4 respectively. The results are presented in Tables 11 and 12. In these tables, sum-sq is sum of squares, indicating the amount of variation caused by the factor. Larger values indicate that the factor contributes more to the total variation. df is the degrees of freedom related to the sum of squares. F indicates the significance of the factors. The larger the F value, the greater the significance of the factor's influence on the dependent variable. PR(>F) indicates the statistical significance of the results. Generally speaking, a P-value less than 0.05 (a significance level of 5 %) is considered statistically significant. eta-sq is the effect size, indicating how much variance is explained. As shown in Table 11, in the KNN results of Data 1, the eta-sq of α is 0.4537 and $p< 0.05$, while the eta-sq of k is 0.4182 and $p = 0.418$ and $p = 0.0005$. Both are below the 0.05 significance threshold. Although the interaction term accounts for 97.15 % of the total sum of squares, the statistical test is not significant ($p = 0.5000$), indicating that the effects of the two factors on the classification accuracy are independent of each other, and the contribution of α is slightly higher. In contrast, for the NB results in Table 2, the eta-sq of k is as high as 0.791 and $p < 0.001$, while the eta-sq of alpha is only 0.053 and $p = 0.08$. The interaction term is also not significant ($p = 0.5000$). Therefore, k is the only significant driving factor, and α can be ignored. It is evident that parameter dominance varies depending on the dataset. In the process of implementation, one ought to initially pinpoint the significant factors via two-factor variance analysis. Subsequently, differentiated optimization should be performed on the significant parameters. This approach serves to avert blind global searches and enhance the optimization efficiency.

(4) Friedman test and Nemenyi test: In our study, we conducted the Friedman test and Nemenyi test to evaluate the classification performance of all algorithms on DT and SVM. The null hypothesis of the Friedman test posits that the performance distributions of all algorithms are identical. Given the conservative nature of the conventional Friedman test([38]), we employed an adjusted formula 17.

$$\tau_F = \frac{(N-1)\tau_{\chi}^2}{N(k-1)-\tau_{\chi}^2} \tag{17}$$

where

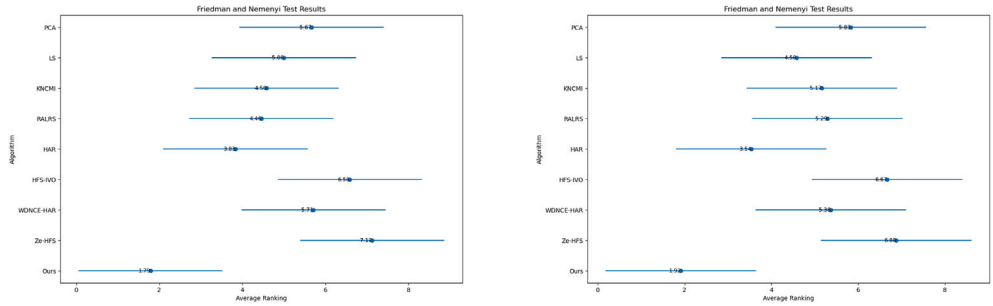
$$\tau_{\chi^2} = \frac{12\,N}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4}\right) \tag{18}$$

Here,k is the number of the algorithms, N is the number of the datasets.τ_F adheres to an F-distribution with (k-1, N-1) degrees of freedom,when N is bid enough. We set α at 0.05 and referred to the critical value of the F-distribution with (8, 10) degrees of freedom, which is 2.159. The computed Friedman Statistic f for DT and SVM were 4.5624 and 6.1654, respectively, both surpassing 2.156. Consequently, we rejected the null hypothesis, indicating that the algorithms exhibit differing performance distributions. The Critical Difference diagram illustrating the average rankings of all algorithms is presented in Fig. 5. It is evident that our algorithm attained the highest ranking on both DT and SVM, outperforming the competing algorithms. This suggests that our algorithm has a significant advantage in terms of classification performance on these datasets. Subsequently, we carried out the Nemenyi test to conduct a post-hoc analysis. The critical value domain was determined using the formula 19.

$$CD = q_{\alpha}\sqrt{\frac{k(k+1)}{6\,N}} \tag{19}$$

With $q_{0.05}$ established at 3.102, the resulting CD was 3.4681. When the performance difference between our algorithm and others exceeded CD, we rejected the hypothesis that the two algorithms possess equivalent performance. Our findings revealed that our algorithm significantly outperformed four other algorithms. Specifically, the performance differences between our algorithm and these four algorithms were all greater than the CD value, providing strong evidence of our algorithm's superior performance.

In conclusion, the combination of the Friedman test and the Nemenyi test strongly demonstrates the superior performance of our algorithm on DT and SVM. These statistical analyses not only confirm the overall superiority of our algorithm but also provide a strict



(a) Average Ranking result On DT (b) Average Ranking result On SVM

Fig. 5. The result of Friedman test and Nemenyi test.

basis for comparing its performance with other algorithms. This comprehensive evaluation ensures that the performance advantage of our algorithm is statistically significant and reliable, making it a robust choice for classification tasks that involve these datasets.

6. Conclusion

This study proposes an innovative feature selection method tailored for multi-source ordered information systems. By integrating the k-nearest neighbor rough set theory with the measurement of feature differences, a feature evaluation framework with a ranking mechanism is constructed. Systematic experiments on 12 public datasets show that compared with 8 existing feature selection algorithms, the proposed method exhibits significant advantages in both classification accuracy and the quality of feature subsets. These results effectively validate the method's effectiveness and robustness in handling complex data structures.

This method demonstrates unique advantages in processing real-world data characterized by high dimensionality, multiple sources, and class imbalance. Instead of relying on prior distribution assumptions, it evaluates feature importance through local neighborhood relationships, thereby effectively addressing the challenge of imbalanced data distribution. This characteristic gives it direct application potential in real-world scenarios such as medical diagnosis (e.g. rare disease identification) and financial risk control (e.g. fraud transaction detection). Taking early breast cancer screening in medical diagnosis as an example, this application scenario typically involves heterogeneous data sources from multi-modal imaging equipment (such as X-ray, ultrasound, and MRI), with healthy samples significantly outnumbering cancerous samples. The proposed method can: (1) effectively integrate multi-source imaging features to overcome the limitations of single data sources; (2) evaluate feature importance based on local neighborhood relationships to reduce sensitivity to imbalanced data distributions; (3) select the most discriminative combinations of imaging biomarkers to provide precise diagnostic support for physicians.

Although the proposed method in this study demonstrates promising application prospects, it still exhibits several noteworthy limitations. First, the current algorithm is primarily designed for static datasets and lacks adaptability to dynamic data streams. Second, the uniform weighting strategy adopted in multi-source data processing fails to fully account for the quality and reliability differences among various data sources. This approach may not fully leverage the value of high-quality data sources and may struggle to effectively mitigate the noise interference introduced by low-quality data sources. Based on the above analysis, future research efforts will focus on the following three directions: (1) Developing feature selection mechanisms for dynamic incremental learning environments, and investigating efficient feature evaluation and update strategies suitable for data streams; (2) Establishing adaptive source weighting allocation mechanisms based on data quality and reliability, achieving intelligent weight allocation by quantitatively evaluating the confidence levels of various data sources; (3) Constructing interpretable feature selection frameworks that deeply integrate domain knowledge, combining expert experience with data-driven methods to enhance the interpretability and practicality of feature selection results. These improvements will significantly enhance the practical value and application potential of the method in real-world dynamic environments, providing more robust technical support for the intelligent processing of multi-source dynamic data.

CRediT authorship contribution statement

Hao Yuan: Writing – review & editing, Writing – original draft, Visualization, Software, Investigation, Formal analysis, Data curation. **Weihua Xu:** Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of Competing Interest

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

Acknowledgments

This paper is supported in part by the National Natural Science Foundation of China(No. 62376229) and Natural Science Foundation of Chongqing, China (No. CSTB2023NSCQ-LZX0027).

Data availability

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

References

[1] Y. Fan, C. Liu, J. Wang, Integrating multi-granularity model and similarity measurement for transforming process data into different granularity knowledge, Adv. Eng. Inform. 37 (2018) 88–102.
[2] D. Cai, C. Zhang, X. He, Unsupervised feature selection for multi-cluster data, Association for Computing Machinery, New York, NY, USA, 2010.
[3] H. Chen, T. Li, X. Fan, C. Luo, Explainable feature selection based on shapley value and mutual information, Appl. Soft Comput. 136 (2023) 110059.
[4] J. Wan, H. Chen, T. Li, Z. Yuan, J. Liu, W. Huang, Interactive and complementary feature selection via fuzzy multigranularity uncertainty measures, IEEE Trans. Cybern. 53 (2) (2023) 1208–1221.
[5] K. Park, M.C. Nguyen, H. Won, Web-based collaborative big data analytics on big data as a service platform, in: 2015 17th International Conference on Advanced Communication Technology (ICACT), 2015, pp. 564–567.
[6] J.T. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: Perspectives and challenges, IEEE Trans. Cybern. 43 (6) (2013) 1977–1989.
[7] X. Zhang, J. Wang, J. Hou, Matrix-based approximation dynamic update approach to multi-granulation neighborhood rough sets for intuitionistic fuzzy ordered datasets, Appl. Soft Comput. 163 (2024) 111915.
[8] J. Yang, G. Wang, Q. Zhang, Y. Chen, T. Xu, Optimal granularity selection based on cost-sensitive sequential three-way decisions with rough fuzzy sets, Knowl. Based Syst. 163 (2019) 131–144.
[9] D. Guo, C. Jiang, R. Sheng, S. Liu, A novel outcome evaluation model of three-way decision: A change viewpoint, Inf. Sci. 607 (2022) 1089–1110.
[10] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982) 341–356.
[11] M. De Cock, C. Cornelis, E.E. Kerre, Fuzzy rough sets: The forgotten step, IEEE Trans. Fuzzy Syst. 15 (1) (2007) 121–130.
[12] D.I.D.I.E.R. DUBOIS, H.E.N.R.I. PRADE, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17 (2–3) (1990) 191–209.
[13] Z.D.Z.I.S.L.A.W. PAWLAK, Rough set theory and its applications to data analysis, Cybern. Syst. 29 (7) (1998) 661–688.
[14] C. Wang, C. Wang, S. An, W. Ding, Y. Qian, Feature selection and classification based on directed fuzzy rough sets, IEEE Trans. Syst. Man Cybern. Syst. 55 (1) (2025) 699–711.
[15] J. Chen, Y. Lin, J. Mi, S. Li, W. Ding, A spectral feature selection approach with kernelized fuzzy rough sets, IEEE Trans. Fuzzy Syst. 30 (8) (2022) 2886–2901.
[16] K. Yuan, D. Miao, W. Pedrycz, W. Ding, H. Zhang, Ze-hfs: Zentropy-based uncertainty measure for heterogeneous feature selection and knowledge discovery, IEEE Trans. Knowl. Data Eng. 36 (11) (2024) 7326–7339.
[17] W. Xu, Y. Yang, Matrix-based feature selection approach using conditional entropy for ordered data set with time-evolving features, Knowl. Based Syst. 279 (2023) 110947.
[18] X. Zhang, W. Zhao, Uncertainty measures and feature selection based on composite entropy for generalized multigranulation fuzzy neighborhood rough set, Fuzzy Sets Syst. 486 (2024) 108971.
[19] J.W. Guan, D.A. Bell, Rough computational methods for information systems, Artif. Intell. 105 (1) (1998) 77–103.
[20] Z. Feng, X. Zhang, Supervised incremental feature selection using regularization vector for dynamic multi-scale interval valued datasets, Pattern Recognit. 170 (2026) 111985.
[21] S. Xia, S. Wu, X. Chen, G. Wang, X. Gao, Q. Zhang, E. Giem, Z. Chen, Grrs: Accurate and efficient neighborhood rough set for feature selection, IEEE Trans. Knowl. Data Eng. 35 (9) (2023) 9281–9294.
[22] J. Yang, Z. Liu, S. Xia, G. Wang, Q. Zhang, S. Li, T. Xu, 3wc-gbnrs++: A novel three-way classifier with granular-ball neighborhood rough sets based on uncertainty, IEEE Trans. Fuzzy Syst. 32 (8) (2024) 4376–4387.
[23] X. Zhang, X. Shen, Graph-driven feature selection via granular-rectangular neighborhood rough sets for interval-valued data sets, Appl. Soft Comput. 170 (2025) 112716.
[24] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in: Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'05, Cambridge, MA, USA, MIT Press, 2005, pp. 507–514.
[25] A. Malhi, R.X. Gao, Pca-based feature selection scheme for machine defect classification, IEEE Trans. Instrum. Meas. 53 (6) (2004) 1517–1525.
[26] Q. Gu, Z. Li, J. Han, Generalized fisher score for feature selection, in: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI'11, Arlington, Virginia, USA, AUAI Press, 2011, pp. 266–273.
[27] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[28] G.J. McLachlan, Discriminant analysis and statistical pattern recognition (1992).
[29] C. Wang, Q. Hu, X. Wang, D. Chen, Y. Qian, Z. Dong, Feature selection based on neighborhood discrimination index, IEEE Trans. Neural Netw. Learn. Syst. 29 (7) (2018) 2986–2999.
[30] P. Zhang, Q. Zhang, J. Liu, D. Wang, H. Ye, X. Zhang, T. Li, Information fusion and feature selection for multi-source data utilizing dempster-shafer evidence theory and k-nearest neighbors, Information Sci. 718 (2025) 122408.
[31] P. Zhang, T. Li, Z. Yuan, C. Luo, K. Liu, X. Yang, Heterogeneous feature selection based on neighborhood combination entropy, IEEE Trans. Neural Netw. Learn. Syst. 35 (3) (2024) 3514–3527.
[32] N.N. Thuy, S. Wongthanavasu, A novel feature selection method for high-dimensional mixed decision tables, IEEE Trans. Neural Netw. Learn. Syst. 33 (7) (2022) 3024–3037.
[33] Y. Shen, X. Ma, J. Zhan, A two-stage adaptive consensus reaching model by virtue of three-way clustering for large-scale group decision making, Inf. Sci. 649 (2023) 119658.
[34] W. Xu, Z. Yuan, Z. Liu, Feature selection for unbalanced distribution hybrid data based on k-nearest neighborhood rough set, IEEE Trans. Artif. Intell. 5 (1) (2024) 229–243.
[35] S. Xia, X. Bai, G. Wang, Y. Cheng, D. Meng, X. Gao, Y. Zhai, E. Giem, An efficient and accurate rough set for feature selection, classification, and knowledge representation, IEEE Trans. Knowl. Data Eng. 35 (8) (2023) 7724–7735.
[36] B. Sang, H. Chen, L. Yang, T. Li, W. Xu, C. Luo, Feature selection for dynamic interval-valued ordered data based on fuzzy dominance neighborhood rough set, Knowl. Based Syst. 227 (2021) 107223.
[37] Y. Pan, W. Xu, Q. Ran, An incremental approach to feature selection using the weighted dominance-based neighborhood rough sets, Int. J. Mach. Learn. & Cyber. 14 (4) (2023) 1217–1233.
[38] M.F. Mridha, M.A.H. Wadud, M.A. Hamid, M.M. Monowar, M. Abdullah-Al-Wadud, A. Alamri, L-boost: Identifying offensive texts from social media post in bengali, IEEE Access 9 (2021) 164681–164699.