Applied Intelligence (2025) 55:1076
https://doi.org/10.1007/s10489-025-06936-x

Feature selection based on fuzzy rough fitting model with nominal distribution metric

Jin Qian[1] · Shaowei Yan[1] · Ying Yu[1] · Yongting Ni[1] · Duoqian Miao[2]

Received: 3 March 2025 / Accepted: 7 October 2025 / Published online: 4 November 2025
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

Abstract

Fuzzy rough sets constitute a significant granular computing model in knowledge discovery and have been widely applied to feature selection. However, in heterogeneous and nominal data, most existing fuzzy rough set methods rely on Hamming distance to measure dissimilarity between nominal attribute values, which fails to fully capture their underlying relationships and their impact on decision outcomes. To address this limitation, we propose a fuzzy rough fitting model with nominal distribution metric embedding (FR-NDM). First, the concept of nominal distribution and decision probability is defined, and suitable forms of the nominal distribution metric (NDM) are constructed for diverse data distribution scenarios. Second, a heterogeneous fuzzy information granule with dual-parameter adjustment is developed to accommodate complex data structures. Additionally, fitting approximation operators are established by introducing the judgment condition to ensure that samples attain the maximum membership degree within their respective decision categories. Third, a forward search feature selection algorithm is designed based on FR-NDM. Finally, the proposed method is evaluated on 24 public datasets and compared with 8 state-of-the-art feature selection methods. Experimental results demonstrate the superior performance of our approach.

Keywords Fuzzy rough sets · Feature selection · Nominal distribution metric · Fitting approximation operator

1 Introduction

Granular computing (GrC) is a methodological tool for analyzing uncertain and imprecise problems, such as reasoning, decision making and control [1, 2]. Rough set [3], as an important mathematical model in GrC, has been applied in many fields [4–6], but its adaptability to complex data structures is limited because it uses equivalence relation to generate information granules.

In contrast to the classical rough set model [7], fuzzy rough sets (FRS) introduces a fuzzy similarity relation to characterize the similarity degree between samples [8]. Therefore, compared with rough set, FRS does not need to discretize numerical attributes, which reduces the risk of information loss to a certain extent. At present, FRS has been successfully applied to multi-attribute decision making [9], machine learning [10], outlier detection [11] and has good performance in feature selection tasks [12].

Feature selection [13], also known as attribute reduction, is an effective method to reduce the dimension of data. Jensen and Shen [14] first used FRS for attribute reduction, defined the dependency function to calculate the importance

Shaowei Yan contributed equally to this work.

✉ Jin Qian
qjqjlqyf@163.com

Shaowei Yan
yanshaowei16@163.com

Ying Yu
yuyingjx@163.com

Yongting Ni
niyongtingting@Gmail.com

Duoqian Miao
dqmiao@tongji.edu.cn

[1] School of Information and Software Engineering, East China Jiaotong University, 808 East Shuanggang Street, Nanchang 330013, Jiangxi, China

[2] Department of Computer Science and Technology, Tongji University, 1239 Siping Road, Shanghai 201804, China

of each attribute in the dataset, and optimized the algorithm in the following years [15, 16]. In recent years, the research on FRS-based feature selection methods is mainly distributed in two aspects: model improvement and application. For instance, in terms of model improvement, Qian et al. [17] introduced the granule ball model into FRS and proposed a fuzzy rough feature selection method based on granule ball computing. Wang et al. [18] proposed directed distance to analyze the distribution of samples in different classes and designed directed fuzzy approximation operators to construct directed FRS. In [19], a weighted fuzzy approximation operator was proposed to better characterize the membership degree of the sample to the decision class. Yuan et al. [20] redefined the importance of candidate attributes and proposed a FRS model that can perform attribute reduction in an unsupervised environment. Wang et al. [21] reconstructed the fuzzy approximation and designed a FRS model that can fit decision of the sample. Qiu et al. [22] proposed a new FRS for hierarchical feature selection based on Hausdorf distance, which greatly reduces the computational complexity. Based on intuitionistic fuzzy rough sets [23], Tan et al. [24, 25] designed feature subset search algorithms based on intuitionistic fuzzy positive region and intuitionistic fuzzy entropy respectively. Additionally, in order to deal with complex situations such as non-uniform data distribution and noise, a series of robust FRS have been developed [26–28]. In terms of application, Kong et al. [29] proposed a distributed FRS feature selection method and successfully applied it to large-scale data, considering that FRS requires a large amount of computing and memory resources. Bai et al. [30] constructed a kernelized FRS for online streaming feature selection on hierarchical categorical data. Hu et al. [31] integrated multiple kernel functions and constructed a multi-kernel FRS attribute reduction model to solve the high-dimensional problem of multimodal data.

It should be noted that most existing FRS-based feature selection methods currently employ only the Hamming distance to measure differences between nominal values [20, 21, 24, 25, 32]. That is, they utilize a 0/1 equivalence relation to represent fuzzy similarity, even when kernel functions are applied [31, 33]. Employing the Hamming distance not only fails to uncover potential relationships between different nominal values, but also makes the computation of fuzzy similarity relations among samples under attribute subsets more restrictive. As is well known, most FRSs utilize the intersection operator to calculate fuzzy similarity relations, which further amplifies the shortcomings of the Hamming distance. The following example reveals this limitation.

Table 1 Fragment of the credit approval dataset

Index	$A1$	$A2$	$A3$	$A4$	Class
1	a	17.83	11	u	+
2	b	23.17	11.125	u	+
3	a	20.75	9.54	u	−
4	b	18.08	6.75	y	−

Table 1 shows a small part of the Credit Approval dataset[1], which is heterogeneous data containing numerical and nominal attributes. Here we use x_{Index} to denote different samples and $\mathcal{R}_B$ to denote the fuzzy similarity relation induced by $B \subseteq A = \{A1, A2, A3, A4\}$. Let $B = \{A1, A3\}$, since $\mathcal{R}_{A1}(x_1, x_2) = 0$, then $\mathcal{R}_B(x_1, x_2) = \min\{\mathcal{R}_{A1}(x_1, x_2), \mathcal{R}_{A3}(x_1, x_2)\} = 0$ can be easily obtained, however $\mathcal{R}_B(x_1, x_3) \neq 0$. This result is counter-intuitive because x_1 is of the same class as x_2, while x_1 does not belong to the same class as x_3. Such a rigorous fuzzy similarity computation approach may affect the construction of information granules, thereby influencing the FRS's evaluation of feature significance.

To the best of our knowledge, few studies have focused on the fuzzy relationships among nominal data in FRS research. For instance, Wang et al. [34] introduced variable parameters to adjust inter-sample similarity and designed an FRS model for categorical data; however, their model required discretization of numerical attributes when processing heterogeneous data. More recently, Li et al. [35] employed symmetric deviation to mine similarity relationships among unlabeled samples under nominal attributes. In addition, Luo et al. [36] incorporated the relative object dissimilarity measure (RODM) into neighborhood rough sets to evaluate inter-sample distances in nominal data; however, their methodology was specifically designed for nominal attribute analysis.

Distance measurement methods for nominal data have long been a focus of research in machine learning [37–40], and different measures can significantly affect algorithm performance. Previous studies show that the value difference metric (VDM) is an effective distance measure for nominal values, where the distance is calculated based on the conditional probability difference between different nominal values [41]. The inverted specific-class distance measure (ISCDM) estimates the conditional probability of the test sample's attribute value using the decision category of the training sample and then computes the distance between the new sample and the training sample via a standard negative operator [42]. Compared with VDM, ISCDM

[1] Credit Approval: https://archive.ics.uci.edu/dataset/27/credit+approval

avoids misclassification and addresses VDM's sensitivity to noise and missing values. While VDM and ISCDM require a known decision class, several effective nominal measures have been proposed for unsupervised environments, including attribute-value similarity measures based on attribute relationships [43], frequency-probability-based distance measures for categorical data [44], and coupled nominal similarity measures [45]. Additionally, some studies transform nominal values across multiple attribute dimensions into numerical vectors for inter-sample distance calculation [46, 47].

Inspired by these research advances, this paper proposes a novel nominal distance metric tailored for the FRS model framework to effectively extract fuzzy similarity relations among samples. First, the decision probability is defined according to the distribution of nominal values in decision classes. On this basis, a nominal distribution metric (NDM) is constructed. Meanwhile, two variants of NDM are introduced to accommodate complex and variable data. Subsequently, the NDM is integrated into the FRS model, yielding the proposed FR-NDM model. This model introduces an adjustable parameter to regulate fuzzy similarity relations under nominal attributes, with an analogous approach applied to numerical attributes, thereby generating heterogeneous fuzzy information granules with dual adjustable parameters. To address the limitation that classical fuzzy approximation operators cannot guarantee samples to obtain the maximum membership degree in their corresponding decision categories, a decision judgment condition is incorporated into the fuzzy approximation process, thus enabling the FR-NDM model to achieve better data fitting capability. Finally, we develop a forward search algorithm based on FR-NDM that employs fuzzy dependency functions for feature selection.

The rest of this paper is organized as follows. Section 2 reviews the definitions of fuzzy rough sets and introduces the concept of feature selection. In Section 3, we propose a fuzzy rough fitting model with nominal distribution metric embedding and design the corresponding feature selection algorithm. The experimental analysis is carried out in Section 4. Finally, the paper is concluded in Section 5.

2 Preliminaries

This section reviews the basic concepts of fuzzy rough sets and feature selection.

2.1 Fuzzy rough sets

Generally, a decision information system (*DIS*) can be represented by a quadruple $DIS = \langle U, A, V, f \rangle$, where: 1) U is a non-empty finite set of samples; 2) $A = C \cup D$ denotes the attribute set, with C being the set of conditional attributes (i.e., all features) and D the decision attribute; 3) $V = \bigcup_{a \in A} V_a$ and V_a is the value set of attribute a; 4) $f : U \times A \to V$ is an information function mapping each sample-attribute pair to a value, with $f(x, a) \in V_a$ for all $x \in U$ and $a \in A$. Moreover, the decision attribute D can partition the universe U into h equivalence classes, denoted as $U/D = \{D_1, D_2, ..., D_h\}$.

Definition 1 [20]Let $U = \{x_1, x_2, ..., x_n\}$ and $B \subseteq A$, then a fuzzy similarity relation $\mathcal{R}_B$ can be induced on B, which satisfies the following property:

$$\begin{cases} \text{Reflexivity} : \mathcal{R}_B(x, x) = 1 \\ \text{Symmetry} : \mathcal{R}_B(x, y) = \mathcal{R}_B(y, x) \end{cases}$$

Definition 2 [20]Let $U = \{x_1, x_2, ..., x_n\}$ and $B \subseteq A$, then a fuzzy partition generated by $\mathcal{R}_B$ at U is $U/\mathcal{R}_B = \{[x_1]_{\mathcal{R}_B}, [x_2]_{\mathcal{R}_B}, ..., [x_n]_{\mathcal{R}_B}\}$, where $[x_i]_{\mathcal{R}_B}$ is called the fuzzy information granule generated by $\mathcal{R}_B$, and $[x_i]_{\mathcal{R}_B}(x_j) = \mathcal{R}_B(x_i, x_j)$. The degree to which x_j belongs to $[x_i]_{\mathcal{R}_B}$ is positively correlated with the value of $\mathcal{R}_B(x_i, x_j)$.

Definition 3 [20]Let $U = \{x_1, x_2, ..., x_n\}$, $B \subseteq A$. $\mathcal{R}_B$ is a fuzzy similarity relation induced by B on the U, then the fuzzy relation matrix can be expressed as:

$$M(\mathcal{R}_B) = \begin{bmatrix} r_{11}^B & r_{12}^B & \cdots & r_{1n}^B \\ r_{21}^B & r_{22}^B & \cdots & r_{2n}^B \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1}^B & r_{n2}^B & \cdots & r_{nn}^B \end{bmatrix} \tag{1}$$

Based on the above definition of fuzzy information granule and membership degree, here we have $[x_i]_{\mathcal{R}_B}(x_j) = \mathcal{R}_B(x_i, x_j) = r_{ij}^B$. Typically, r_{ij}^B is computed using conjunctive formulas, i.e., $r_{ij}^B = \wedge_{a \in B}\{r_{ij}^a\}$.

Definition 4 [21] Let $U/D = \{D_1, D_2, ..., D_h\}$, and $B \subseteq A$. $\mathcal{R}_B$ is a fuzzy similarity relation induced by B on the U. For any $x \in U$, the fuzzy decision for sample x is defined as follows:

$$\widetilde{D}_i(x) = \frac{|[x]_{\mathcal{R}_B} \cap D_i|}{|[x]_{\mathcal{R}_B}|}, i = 1, 2, .., h \tag{2}$$

where $\widetilde{D}_i(x)$ denotes the degree of membership, and the degree of sample x belonging to D_i is proportional to the value of $\widetilde{D}_i(x)$. $|[x]_{\mathcal{R}_B}|$ denotes the cardinality of fuzzy information

granule, which is calculated by $|[x]_{\mathcal{R}_B}| = \sum_{y \in U} \mathcal{R}_B(x, y)$ and satisfies $1 \leq |[x]_{\mathcal{R}_B}| \leq n$.

Definition 5 [21]Let $U/D = \{D_1, D_2, ..., D_h\}$ and $B \subseteq A$, $\mathcal{R}_B$ is a fuzzy similarity relation induced by B on U. A sample fuzzy decision induced by B and D is $\{\widetilde{D}_1, \widetilde{D}_2, ..., \widetilde{D}_h\}$, then the fuzzy upper and lower approximations of decision D with respect to B are defined as follows, respectively:

$$\overline{\mathcal{R}_B} D_i(x) = \sup_{y \in U} \min\{\mathcal{R}_B(x, y), \widetilde{D}_i(y)\} \tag{3}$$

$$\underline{\mathcal{R}_B} D_i(x) = \inf_{y \in U} \max\{1 - \mathcal{R}_B(x, y), \widetilde{D}_i(y)\} \tag{4}$$

Then the fuzzy rough set of decision D induced by B is defined as $\left(\underline{\mathcal{R}_B} D, \overline{\mathcal{R}_B} D\right)$.

2.2 Feature selection

Feature selection aims to reduce information redundancy in the data and improve decision efficiency. The detailed definitions are given below.

Definition 6 [21]Let $U/D = \{D_1, D_2, ..., D_h\}$ and $B \subseteq A$, the fuzzy positive region of sample $x \in U$ with respect to D is defined as follows:

$$POS_B(D) = \bigcup_{i=1}^{h} \underline{\mathcal{R}_B} D_i \tag{5}$$

Then we can get the fuzzy dependency function of D with respect to B:

$$r_B(D) = \frac{\sum_{x \in U} POS_B(D)(x)}{|U|} = \frac{\sum_{x \in U} \bigcup_{i=1}^{h} \underline{\mathcal{R}_B} D_i(x)}{|U|} \tag{6}$$

Let R denote the selected feature subset, which is an empty set in the initial state. In the first round of calculation, a maximum fuzzy dependency function value $r_{R \cup \{a_i\}}(D)$ will be obtained, at which point a_i will be added to R.

Definition 7 [21] For any $a \in C - R$, the significance of attribute a is defined as follows:

$$Sig(a, R, D) = r_{R \cup \{a\}}(D) - r_R(D) \tag{7}$$

This definition is a measure of the incremental classification capability of an information system. If attribute a is not redundant, then $Sig(a, R, D) > 0$.

3 Fuzzy rough fitting model with nominal distribution metric embedding

In this section, firstly, the metric method between nominal values is defined, then the concept of fuzzy information granule is extended and the fuzzy approximation is modified to obtain the improved fuzzy dependence function, and finally the feature selection algorithm is designed.

3.1 Nominal distribution metric

In a decision information system, samples under nominal attributes exhibit distinct distribution patterns across different decision classes. In other words, their decision tendencies vary, and such tendencies can be quantified through probability calculations.

Figure 1 illustrates a simple example involving a decision information system with eight samples and one nominal attribute. Initially, the classical crisp relation is considered for computing the fuzzy similarity relation among samples. The Hamming distance between nominal values A and S is 1, resulting in a fuzzy similarity of 0 between x_1 and x_3, which is reasonable since their decision categories differ. However, x_5 and x_3 share the same decision categories, yet their fuzzy similarity is also 0, which is counterintuitive. To address this, we introduce the concept of nominal distribution, which comprehensively records the occurrences of nominal values across different decision categories. As shown in Fig. 1, the same nominal value appears with varying frequencies in different decision categories, allowing us to compute the decision probability for each nominal value. Based on these probabilities, a refined fuzzy similarity between nominal values can be derived. Unlike the binary (0 or 1) similarity obtained from crisp relations, this fuzzy similarity is represented as a precise numerical value, effectively reflecting the strength of the fuzzy similarity relationship between different nominal values. This more refined fuzzy similarity measure can better characterize the decision uncertainty of samples, enabling FRS to yield more accurate assessments when evaluating feature significance. The computational formulation of the nominal distribution metric is presented as follows.

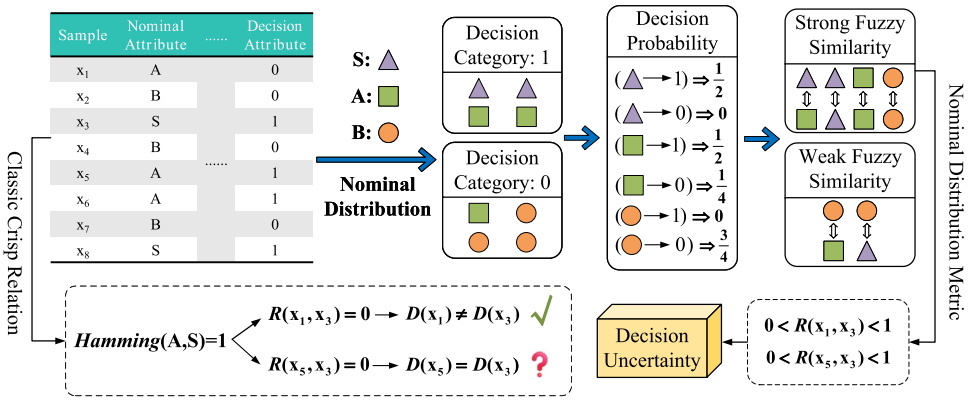


Fig. 1 A figure for illustrating nominal distribution metric

Definition 8 Let $U/D = \{D_1, D_2, ..., D_h\}$, $B \subseteq A$ and a nominal attribute $c \in B$. V_{ci} denotes the nominal value of the ith sample under c, i.e., $f(x_i, c) \to V_{ci}$. The nominal distribution metric is defined as follows:

$$\delta_c(x_i, x_j) = \frac{1}{h}\sum_{k=1}^{h}\left|\mathcal{F}_{D_k}^{V_{ci}}/\mathcal{N}_{D_k} - \mathcal{F}_{D_k}^{V_{cj}}/\mathcal{N}_{D_k}\right|^{\mathcal{P}} \tag{8}$$

where $\mathcal{F}_{D_k}^{V_{ci}}$ denotes the number of occurrences of the nominal value V_{ci} in the equivalence class D_k, $\mathcal{N}_{D_k}$ represents the number of samples in the equivalence class D_k, and $\mathcal{P}$ serves as the adjustment index with $\mathcal{P} \in (0, 1]$.

We refer to $\mathcal{F}_{D_k}^{V_{ci}}/\mathcal{N}_{D_k}$ as the decision probability of the nominal value V_{ci} in the equivalence class D_k, and denote it as $\xi_{D_k}^{V_{ci}}$ for convenience.

Remark 1 $\xi_{D_k}^{V_{ci}} = 1$ does not imply that V_{ci} is entirely distributed within the equivalence class D_k; the decision probabilities of D_k in other equivalence classes partitioned by D on U should also be taken into account. If $\xi_{D_k}^{V_{ci}} = 0$, then it can be considered that V_{ci} is not associated with equivalence class D_k.

Unlike numerical values, nominal values have no dimensions and cannot be normalized. Meanwhile, the nominal values under different attributes are also complex and changeable. Below we discuss two extreme cases and present corresponding solutions.

If two nominal values exhibit extremely low occurrence frequencies, the difference in their decision probabilities will also be negligible, consequently leading to the condition where $\delta_c(x_i, x_j) \approx 0$. It is certain that $\delta_c(x_i, x_j) \approx 0$ indicates minimal dissimilarity (i.e., strong similarity). However, this result becomes counterintuitive when the two nominal values exhibit significantly divergent decision tendencies (i.e., distinct distributions across different decision categories). To solve this problem, $\mathcal{P}$ needs to be set according to the distribution state of the nominal value itself, and the definition is as follows.

Definition 9 Let $U/D = \{D_1, D_2, ..., D_h\}$, $B \subseteq A$ and a nominal attribute $c \in B$. The distribution state of the nominal value in attribute c is used as the adjustment index, and the nominal distribution metric is defined as follows:

$$\delta_c^1(x_i, x_j) = \frac{1}{h}\sum_{k=1}^{h}\left|\xi_{D_k}^{V_{ci}} - \xi_{D_k}^{V_{cj}}\right|^{\frac{\mathcal{N}_{V_{ci}}+\mathcal{N}_{V_{cj}}}{n}} \tag{9}$$

where $\mathcal{N}_{V_{ci}}$ represents the number of occurrences of the nominal value V_{ci} in attribute c.

If $\mathcal{N}_{V_{ci}}$ and $\mathcal{N}_{V_{cj}}$ are too small, $(\mathcal{N}_{V_{ci}} + \mathcal{N}_{V_{cj}})/n$ will magnify the results of decision probability difference, and the ability to distinguish between nominal values with different decision trends will be improved.

Consider another case, assuming that (8) is used to calculate the difference between two nominal values V_{ci} and V_{cj}. If the occurrence count of V_{ci} is very small (i.e., $\mathcal{N}_{V_{ci}}$ is minimal), the situation $\delta_c(x_i, x_j) \approx \frac{1}{h}\sum_{k=1}^{h}\left|\xi_{D_k}^{V_{cj}}\right|^{\mathcal{P}}$ may arise. This occurs because when $\mathcal{N}_{V_{ci}}$ is sufficiently small, the decision probability $\xi_{D_k}^{V_{ci}}$ for any $D_k \in U/D$ becomes extremely small and approaches zero. However, when the decision trends of V_{ci} and V_{cj} are consistent—that is, when their distributions across decision categories are similar—then for any $D_k \in U/D$, $\xi_{D_k}^{V_{ci}} \approx \xi_{D_k}^{V_{cj}}$, and consequently, $\delta_c(x_i, x_j)$ should approach 0 rather than

approximately equal $\frac{1}{h}\sum_{k=1}^{h}\left|\xi_{D_k}^{V_{cj}}\right|^{\mathcal{P}}$. Obviously, the above results will cause measurement errors. Therefore, the relative distribution state of two different nominal values should be used as the adjustment index to reduce the measurement error.

Definition 10 Let $U/D=\{D_1,D_2,...,D_h\}$, $B\subseteq A$ and a nominal attribute $c\in B$. Taking the relative distribution state of two nominal values as the adjustment index, another nominal distribution metric is defined as follows:

$$\delta_c^2(x_i,x_j)=\frac{1}{h}\sum_{k=1}^{h}\left|\left[\xi_{D_k}^{V_{ci}}\right]^{\mathcal{P}_i}-\left[\xi_{D_k}^{V_{cj}}\right]^{\mathcal{P}_j}\right| \tag{10}$$

where $\mathcal{P}_i=\frac{\mathcal{N}_{V_{ci}}}{\mathcal{N}_{V_{ci}}+\mathcal{N}_{V_{cj}}}$ and $\mathcal{P}_j=\frac{\mathcal{N}_{V_{cj}}}{\mathcal{N}_{V_{ci}}+\mathcal{N}_{V_{cj}}}$.
Equations 9 and 10 can be thought of as exponential functions with constraints. For (9), $\mathcal{P}$ will amplify the metric results to some extent. For (10), $\mathcal{P}_i$ and $\mathcal{P}_j$ will preferentially amplify the distribution probability for nominal values with lower occurrences.

Property 1 For any nominal attributes $c\in A$ in *DIS*, there are always $0\le\delta_c^1(x_i,x_j)\le 1$ and $0\le\delta_c^2(x_i,x_j)\le 1$.

Proof Given a *DIS*, there is always $\mathcal{N}_{D_k}$ larger than or equal to $\mathcal{F}_{D_k}^{V_{ci}}$, thus $0\le\xi_{D_k}^{V_{ci}}\le 1$. Furthermore, we have $0\le\left|\xi_{D_k}^{V_{ci}}-\xi_{D_k}^{V_{cj}}\right|\le 1$. Since the range of the adjustment index is $(0,1]$, $0\le\left|\xi_{D_k}^{V_{ci}}-\xi_{D_k}^{V_{cj}}\right|^{\mathcal{P}}\le 1$. Obviously, $0\le\sum_{k=1}^{h}\left|\xi_{D_k}^{V_{ci}}-\xi_{D_k}^{V_{cj}}\right|^{\mathcal{P}}\le h$. Dividing both sides by h yields $0\le\delta_c(x_i,x_j)\le 1$. Since $(\mathcal{N}_{V_{ci}}+\mathcal{N}_{V_{cj}})/n\in(0,1]$, it follows that $0\le\delta_c^1(x_i,x_j)\le 1$, which completes the proof. For $\delta_c^2(V_{ci},V_{cj})$, note that $0\le\left[\xi_{D_k}^{V_{ci}}\right]^{\mathcal{P}_i}\le 1$, then the proof proceeds as above. □

Property 2 For any nominal attributes $c\in A$ in *DIS*, the nominal distribution metric satisfies $\delta_c^1(x_i,x_j)+\delta_c^1(x_j,x_p)\ge\delta_c^1(x_i,x_p)$.

Proof Let $\mathcal{X}=\xi_{D_k}^{V_{ci}}-\xi_{D_k}^{V_{cj}}$, $\mathcal{Y}=\xi_{D_k}^{V_{cj}}-\xi_{D_k}^{V_{cp}}$ and $\mathcal{Z}=\xi_{D_k}^{V_{ci}}-\xi_{D_k}^{V_{cp}}$. In general, we have $|\mathcal{X}|+|\mathcal{Y}|\ge|\mathcal{X}+\mathcal{Y}|=|\mathcal{Z}|$, and the inequality still holds in the case of summation. Therefore, the original $\delta_c^1(x_i,x_j)+\delta_c^1(x_j,x_p)\ge\delta_c^1(x_i,x_p)$ is proved. The same goes for δ_c^2. □

In general, (8) is capable of capturing differences between nominal values, but due to the complexity and diversity of the data, we introduce two variants of nominal distribution metrics: δ_c^1 (9) and δ_c^2 (10). These three metrics can be flexibly combined with FRS model, as discussed in the next section.

3.2 The proposed model

Since nominal attributes often coexist with numerical attributes in data, we extend *DIS* to *HDIS*, which is defined as follows.

Definition 11 Given a heterogeneous decision information systems $HDIS=\langle U,A,V,f\rangle$, we denote by $B^{\psi}\subseteq A$ the set of nominal attributes and $B^{\phi}\subseteq A$ the set of numerical attributes such that $B^{\psi}\cap B^{\phi}=\varnothing$ is satisfied.

Definition 12 Let $U=\{x_1,x_2,...,x_n\}$, for a nominal attribute set $B^{\psi}\subseteq A$, the fuzzy similarity relation induced by B^{ψ} on U is defined as follows:

$$\mathcal{R}_{B^{\psi}}(x_i,x_j)=\bigwedge_{a\in B^{\psi}}\{\mathcal{R}_a^{\sigma}(x_i,x_j)\} \tag{11}$$

and

$$\mathcal{R}_a^{\sigma}(x_i,x_j)=\begin{cases}1-\delta_a(x_i,x_j),\delta_a(x_i,x_j)\le\sigma\\0,\delta_a(x_i,x_j)>\sigma\end{cases} \tag{12}$$

where σ is an adjustable parameter and δ_a can be replaced by δ_a^1 or δ_a^2 to adapt to different *HDIS*.
In the above definition, the parameter σ is used to regulate the fuzzy relationship between the nominal values, and then control the size of the fuzzy information granule. Besides, three nominal distribution metrics can be used to accommodate different data.

Remark 2 For a nominal attribute $c\in B^{\psi}$, when $\sigma=0$, $\mathcal{R}_c^{\sigma}(x_i,x_j)\in\{0,1\}$.

Remark 2 shows that when σ is set to 0, the nominal distribution metric degrades to the Hamming distance, that is, the Hamming distance is a special case of the nominal distribution metric.

Remark **3** For a nominal attribute $c \in B^{\psi}$, if $f(x_i, c) \neq f(x_j, c)$ and $\mathcal{R}_c^{\sigma}(x_i, x_j) = 1$, then for any $k \in \{1, 2, ..., h\}$, we have $\xi_{D_k}^{V_{ci}} = \xi_{D_k}^{V_{cj}}$.

Property 3 For $B_1^{\psi} \subseteq B_2^{\psi} \subseteq A$, we have $\mathcal{R}_{B_1^{\psi}}^{\sigma} \supseteq \mathcal{R}_{B_2^{\psi}}^{\sigma}$.

Proof Let $x, y \in U$, and by Definition 12, we have $\mathcal{R}_{B_1^{\psi}}^{\sigma}(x, y) = \wedge_{c \in B_1^{\psi}} \mathcal{R}_c^{\sigma}(x, y) \geq \mathcal{R}_{B_2^{\psi}}^{\sigma}(x, y) = \wedge_{c \in B_2^{\psi}} \mathcal{R}_c^{\sigma}(x, y)$. Thus $\mathcal{R}_{B_1^{\psi}}^{\sigma} \supseteq \mathcal{R}_{B_2^{\psi}}^{\sigma}$. □

Property 4 For $B^{\psi} \subseteq A$, let $\sigma_1 \leq \sigma_2$, then we have $\mathcal{R}_{B^{\psi}}^{\sigma_1} \subseteq \mathcal{R}_{B^{\psi}}^{\sigma_2}$.

Proof Let $x, y \in U$, and by Definition 12, we have $\mathcal{R}_{B^{\psi}}^{\sigma_1}(x, y) = \wedge_{c \in B^{\psi}} \mathcal{R}_c^{\sigma_1}(x, y) \leq \mathcal{R}_{B^{\psi}}^{\sigma_2}(x, y) = \wedge_{c \in B^{\psi}} \mathcal{R}_c^{\sigma_2}(x, y)$. Thus $\mathcal{R}_{B^{\psi}}^{\sigma_1} \subseteq \mathcal{R}_{B^{\psi}}^{\sigma_2}$. □

Definition 13 Let $U = \{x_1, x_2, ..., x_n\}$, for a numeric attribute set $B^{\phi} \subseteq A$, the fuzzy similarity relation induced by B^{ϕ} on U is defined as follows:

$$\mathcal{R}_{B^{\phi}}(x_i, x_j) = \bigwedge_{a \in B^{\phi}} \{\mathcal{R}_a^{\varepsilon_a}(x_i, x_j)\} \tag{13}$$

and

$$\mathcal{R}_a^{\varepsilon_a}(x_i, x_j) = \begin{cases} 1 - d_a(x_i, x_j), d_a(x_i, x_j) \leq \varepsilon_a \\ 0, d_a(x_i, x_j) > \varepsilon_a \end{cases} \tag{14}$$

where $d_a(x_i, x_j) = |f(x_i, a) - f(x_j, a)|$ and $\varepsilon_a = std(a)/\lambda$, $std(\cdot)$ represents the standard deviation and λ is an adjustable parameter.

Definition 14 Let $U = \{x_1, x_2, ..., x_n\}$, $B = B^{\psi} \cup B^{\phi}$, $B \subseteq A$, and the fuzzy similarity relation induced by B is denoted $\mathcal{R}_B^{\sigma,\lambda}$. The heterogeneous fuzzy information granule induced by $\mathcal{R}_B^{\sigma,\lambda}$ of x_i is defined as follows:

$$[x_i]_{\mathcal{R}_B^{\sigma,\lambda}} = \sum_{j=1}^{n} \frac{\mathcal{R}_B^{\sigma,\lambda}(x_i, x_j)}{x_j} \tag{15}$$

Distinct from conventional information granulation approaches, Definition 14 develops heterogeneous fuzzy information granules with dual-parameter adjustment by integrating fuzzy similarity measures of different attribute types. This construction endows the FRS model with enhanced flexibility to handle complex and variable data, while effectively overcoming the limitations of using Hamming distance as discussed in Section 1.

In the fuzzy rough set, because the fuzzy lower approximation will obtain the maximum membership degree of the sample in each decision, it cannot guarantee that the membership degree of the decision category to which the sample belongs reaches the maximum value [21]. To overcome this problem, decision category judgments should be incorporated into the fuzzy approximation.

Input: $HDIS =< U, A, V, f >, \sigma$ and λ
Output: A reduction set R

$R \leftarrow \varnothing$, $flag \leftarrow 1$, $B \leftarrow C - R$;
for $k \leftarrow 1$ *to* $|C|$ **do**
  Calculate $M\left(\mathcal{R}_{a_k}^{\sigma,\lambda}\right)$ by (12) and (14) ;
  Calculate $dep_{a_k}^{\sigma,\lambda}(D)$ by (18);
end
Select feature a_q with the maximum $dep_{a_q}^{\sigma,\lambda}(D)$;
$R \leftarrow R \cup \{a_q\}$, $B \leftarrow B - \{a_q\}$;
while $flag$ **do**
  for $l \leftarrow 1$ *to* $|B|$ **do**
    Calculate $M\left(\mathcal{R}_{R \cup \{a_l\}}^{\sigma,\lambda}\right)$;
    Calculate $\underline{\mathcal{R}_{R \cup \{a_l\}}^{\sigma,\lambda}}(D_i)(x)$ by (16);
    Calculate $dep_{R \cup \{a_l\}}^{\sigma,\lambda}(D)$ by (18);
  end
  Select feature a_p with the maximum $dep_{R \cup \{a_p\}}^{\sigma,\lambda}(D)$;
  Calculate
  $Sig^{\sigma,\lambda}(a_p, R, D) = dep_{R \cup \{a_p\}}^{\sigma,\lambda}(D) - dep_R^{\sigma,\lambda}(D)$ by (19);
  if $Sig^{\sigma,\lambda}(a_p, R, D) > 0$ **then**
    $R \leftarrow R \cup \{a_p\}$, $B \leftarrow B - \{a_p\}$;
  else
    $flag \leftarrow 0$;
  end
end
return R

Algorithm 1 FR-NDM Algorithm.

Definition 15 Let $U/D = \{D_1, D_2, ..., D_h\}$, $B = B^{\psi} \cup B^{\phi}$ and $B \subseteq A$, $\mathcal{R}_B^{\sigma,\lambda}$ is a fuzzy similarity relation induced by B on U. A sample fuzzy decision induced by B and D is $\{\widetilde{D}_1, \widetilde{D}_2, ..., \widetilde{D}_h\}$, the fitting fuzzy lower and upper approximations of decision D with respect to B are defined as follows:

$$\underline{\mathcal{R}_B^{\sigma,\lambda}}(D_i)(x) = \begin{cases} \inf\limits_{y \in U} \max\{1 - \mathcal{R}_B^{\sigma,\lambda}(x, y), \widetilde{D}_i(y)\}, x \in D_i \\ 0, x \notin D_i \end{cases} \tag{16}$$

$$\overline{\mathcal{R}_B^{\sigma,\lambda}}(D_i)(x) = \begin{cases} \sup\limits_{y \in U} \min\{\mathcal{R}_B^{\sigma,\lambda}(x, y), \widetilde{D}_i(y)\}, x \in D_i \\ 0, x \notin D_i \end{cases} \tag{17}$$

Further, we can derive an improved fuzzy dependency function based on dual-parameter adjustable heterogeneous fuzzy information granules and fitting fuzzy approximations:

$$dep_B^{\sigma,\lambda}(D) = \frac{\sum\limits_{x \in U} POS_B^{\sigma,\lambda}(D)(x)}{|U|} = \frac{\sum\limits_{x \in U} \bigcup\limits_{i=1}^{r} \underline{\mathcal{R}_B^{\sigma,\lambda}} D_i(x)}{|U|} \tag{18}$$

Thus, we have constructed a fuzzy rough fitting model with nominal distribution metric embedding, which effectively utilizes decision information to measure fuzzy similarity between nominal values while ensuring that samples attain maximum membership degrees within their respective decision categories. For evaluating feature significance, the following formula can be derived by combining with (18):

$$Sig^{\sigma,\lambda}(a,R,D) = dep^{\sigma,\lambda}_{R\cup\{a\}}(D) - dep^{\sigma,\lambda}_{R}(D) \tag{19}$$

where $a \in C - R$, and R represents the already selected feature subset.

3.3 Feature selection algorithm

According to the proposed model, we design a forward search feature selection algorithm called fuzzy rough feature selection with nominal distribution metric embedding (FR-NDM for short). The pseudo-code for FR-NDM is shown in Algorithm 1.

Given an $HDIS =< U, A, V, f >$, assume that $|U| = n$, $|A| = m$, and $|U/D| = h$. In Steps 2 to 5 of Algorithm 1, it is necessary to compute the fuzzy relation matrix and corresponding fuzzy dependency degree for each attribute, which exhibits a time complexity of $O(n^2 \times m \times h)$. Steps 6 and 7 involve selecting the attribute with the maximum dependency degree to add into the reduction set R. From Steps 8 to 21, the algorithm calculates the dependency degree of each remaining attribute in subset B, then selects the attribute with the highest dependency degree to evaluate its significance. If the significance is not greater than 0, the attribute is considered redundant and the algorithm terminates. This procedure demonstrates a worst-case time complexity of $O(m^2 \times h)$. Consequently, the overall time complexity of the FR-NDM algorithm is $O(n^2 \times m \times h + m^2 \times h)$.

3.4 Example analysis

In order to better illustrate the FR-NDM method, an example is given below.

Example 1 Given a heterogeneous decision information system $HDIS =< U, A, V, f >$ in Table 2, where $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, $A = B^{\psi} \cup B^{\phi} = \{a_1, a_2\} \cup \{a_3, a_4\}$ and $U/D = \{D_1, D_2\} = \{\{x_1, x_4\}, \{x_2, x_3, x_5, x_6\}\}$.

Table 2 A *HDIS* sample for feature selection

U	a_1	a_2	a_3	a_4	D
x_1	M	g	0.7	6	1
x_2	L	p	0.2	4	0
x_3	M	p	0.5	4	0
x_4	H	g	0.6	7	1
x_5	L	p	0.7	3	0
x_6	M	g	0.2	5	0

Firstly, we employ (9) as an example to calculate the decision probability. Based on Table 2, we obtain $\mathcal{N}_M = 3$, $\mathcal{N}_L = 2$, $\mathcal{N}_H = 1$, $\mathcal{N}_g = 3$, and $\mathcal{N}_p = 3$. Furthermore, the calculations yield $\xi_0^M = 0.5$, $\xi_1^M = 0.5$, $\xi_0^L = 0.5$, $\xi_1^L = 0$, $\xi_0^H = 0$, $\xi_1^H = 0.5$, $\xi_0^g = 0.25$, $\xi_1^g = 1$, $\xi_0^p = 0.75$ and $\xi_1^g = 0$.

Next, all numerical attributes are normalized using the following equation:

$$f'(x_i, a_k) = \frac{f(x_i, a_k) - \min_{a_k}}{\max_{a_k} - \min_{a_k}} \tag{20}$$

where $\max_{a_k}$ and $\min_{a_k}$ represent the maximum and minimum values in attribute a_k.

Subsequently, the fuzzy similarity relation is computed using (12) and (14), with parameter σ set to 0.7 and parameter λ set to 1. Finally, the fuzzy relation matrix for all attributes is obtained as follows:

$$M\left(\mathcal{R}^{0.7}_{a_1}\right) = \begin{bmatrix} 1 & 0.72 & 1 & 0.69 & 0.72 & 1 \\ 0.72 & 1 & 0.72 & 0 & 1 & 0.72 \\ 1 & 0.72 & 1 & 0.69 & 0.72 & 1 \\ 0.69 & 0 & 0.69 & 1 & 0 & 0.69 \\ 0.72 & 1 & 0.72 & 0 & 1 & 0.72 \\ 1 & 0.72 & 1 & 0.69 & 0.72 & 1 \end{bmatrix}$$

$$M\left(\mathcal{R}^{0.7}_{a_2}\right) = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$M\left(\mathcal{R}^{1}_{a_3}\right) = \begin{bmatrix} 1 & 0 & 0.6 & 0.8 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0.6 & 0 & 1 & 0.8 & 0.6 & 0 \\ 0.8 & 0 & 0.8 & 1 & 0.8 & 0 \\ 1 & 0 & 0.6 & 0.8 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$M\left(\mathcal{R}^{1}_{a_4}\right) = \begin{bmatrix} 1 & 0 & 0 & 0.75 & 0 & 0.75 \\ 0 & 1 & 1 & 0 & 0.75 & 0.75 \\ 0 & 1 & 1 & 0 & 0.75 & 0.75 \\ 0.75 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0.75 & 0.75 & 0 & 1 & 0 \\ 0.75 & 0.75 & 0.75 & 0 & 0 & 1 \end{bmatrix}$$

By employing (16) and (18), the dependency degree of D on the four attributes in Table 2 can be calculatedas follows: $dep^{0.7}_{\{a_1\}}(D) \approx 0.4748$, $dep^{0.7}_{\{a_2\}}(D) \approx 0.7778$, $dep^{1}_{\{a_3\}}(D) \approx 0.6458$ and $dep^{1}_{\{a_4\}}(D) \approx 0.6314$. Note that at this point R is the empty set. The attribute with the highest dependency degree is selected to be added to the reduction set R, resulting in $R = \{a_2\}$ and $B = C - R = \{a_1, a_3, a_4\}$. In the next round of calculation, we obtain: $Sig^{0.7,1}(a_1, R, D)$

$= dep^{0.7,1}_{R\cup\{a_1\}}(D) - dep^{0.7,1}_{R}(D) \approx 0.7617 - 0.7778 = -0.0161$, $Sig^{0.7,1}(a_3, R, D) \approx 1 - 0.7778 = 0.2222$ and $Sig^{0.7,1}(a_4, R, D) \approx 0.7381 - 0.7778 = -0.0397$. Updating the reduction set R at this point, $R = \{a_2, a_3\}$ and $B = \{a_1, a_4\}$ can be obtained.

Similarly, performing the third round of calculation, we can obtain $Sig^{0.7,1}(a_1, R, D) = 1 - 1 = 0$ and $Sig^{0.7,1}(a_4, R, D) = 1 - 1 = 0$. Therefore, the reduction result is $R = \{a_2, a_3\}$.

4 Experimental analysis

This section conducts a series of numerical experiments to validate the effectiveness of the proposed method, which can be evaluated from three aspects: 1) whether the proposed method can eliminate redundant features while improving classification accuracy; 2) whether the proposed method is sensitive to parameter settings; 3) whether the proposed method outperforms existing feature selection algorithms.

4.1 Experimental setup and environment

To verify the effectiveness of the FR-NDM algorithm, 24 UCI public datasets[2] are used for feature selection in this section, and the details of each dataset are shown in Table 3. The numerical attributes in all data are normalized using (20) before the experiment to eliminate the effect of different units. In addition, for missing values in the data, we take the original proportion probability for random filling.

We select eight advanced feature selection algorithms to compare with the proposed method. The characteristics and principles of these algorithms are described as follows:

1. Raw data: No feature selection.
2. FScNCE [48]: A three-stage feature selection algorithm based on neighborhood combination entropy, with neighborhood radius δ configured within the range [0, 0.7] using a step size of 0.05.
3. WAFS [19]: A heuristic feature selection algorithm based on weighted fuzzy rough sets, with parameter p configured within the range [0.1, 3] using a step size of 0.1.
4. DRFFS [18]: A heuristic feature selection algorithm based on directed fuzzy rough sets, with parameter δ

Table 3 The datasets information

No.	Dataset		Samples	Features		Classes
	Name	Abbr.		Numerical	Nominal	
1	automobile	Auto	205	15	10	6
2	breast cancer	Cancer	286	1	8	2
3	cardiotocography	Cardio	2126	19	2	3
4	heart failure clinical records	Clinic	299	7	5	2
5	credit approval	Credit	690	6	9	2
6	german	Ger	1000	3	17	2
7	hcv data	HCV	589	11	1	5
8	heart disease	Heart	270	6	7	2
9	horse colic	Horse	368	9	18	2
10	obesity	Obe	2111	8	8	7
11	predict students' dropout and academic success	Predict	4424	19	17	3
12	differentiated thyroid cancer recurrence	Thyroid	383	1	15	2
13	audiology	Audio	226	0	70	24
14	chess	Chess	3196	0	36	2
15	higher education students performance evaluation	Education	145	0	31	8
16	lymphography	Lym	148	0	18	4
17	molecular biology (splice-junction gene sequences)	Molecular	3190	0	60	3
18	mushroom	Mush	8124	0	22	2
19	national poll on healthy aging	NPHA	714	0	14	3
20	risk factor prediction of chronic kidney disease	Risk	200	0	27	2
21	SPECT	SPECT	267	0	22	2
22	soybean	Soybean	266	0	35	15
23	tic-tac-toe endgame	Tic	958	0	9	2
24	zoo	Zoo	101	0	16	7

[2] UCI Datasets: https://archive.ics.uci.edu/

configured within the range [0.01, 0.1] using a step size of 0.01.

5. MRNG [49]: A monotonic attribute reduction algorithm based on relative neighborhood granularity, with neighborhood radius δ configured within the range [0.05, 0.3] using a step size of 0.05.
6. MFIGI [50]: A monotonic attribute reduction algorithm based on fuzzy implication conditional entropy, with parameter ω fixed at 0.001.
7. IFIE [25]: A parameter-free relative attribute reduction algorithm based on intuitionistic fuzzy entropy.
8. FRC [34]: A fuzzy rough attribute reduction algorithm specifically designed for categorical data, with parameter δ configured within the range [0.05, 0.35] using a step size of 0.05.
9. LDP [5]: A parameter-free attribute reduction algorithm based on discernibility pairs.

Among them, FScNCE, MRNG, MFIGI, and IFIE are applicable to both heterogeneous data and nominal data, while WAFS and DRFFS are designed for heterogeneous data, and FRC and LDP are specifically developed for nominal data. Due to the varying applicability of these algorithms to different data types, we conduct grouped comparative experiments to ensure fairness. Specifically, for the heterogeneous data in Table 3, we embed δ_c (i.e., (8)) into the FRS model proposed in Section 3.2 and compared it with FScNCE, WAFS, DRFFS, MRNG, MFIGI, and IFIE. In our method, the parameter $\mathcal{P}$ in δ_c is set to 1, while parameter λ is configured within the range of [0.1, 3] with a step size of 0.1, and parameter σ is set within [0.1, 0.9] with a step size of 0.1. For nominal data, we embed δ_c, δ_c^1 (i.e., (9)) and δ_c^2 (i.e., (10)) into the proposed FRS model, respectively, and conducted comparisons with FRC, LDP, FScNCE, MRNG, MFIGI, and IFIE. For parameter σ, it is set within the range of [0.1, 0.9] with a step size of 0.05.

We use classification experiments to evaluate the performance of the above comparison algorithms as well as the proposed method. Classification experiments are carried out by calling classification and regression trees (CART) and gaussian naive bayes (GNB) in *Scikit-learn* machine learning library, using 10-fold cross validation. The parameters of the classifier are all default values. For nominal attributes in the data, we uniformly apply one-hot encoding before feeding them into the classifier for training. We randomize the data partitioning without fixing random seeds to ensure generalizability, repeat the experiment 10 times and calculate the mean and standard deviation of the classification accuracy as the final result.

In our previous study [51], we identified a primary-key-like feature in the Audi dataset where each feature value is unique. This characteristic causes both MRNG and MFIGI to fail, specifically resulting in their output being restricted to this single feature. Consequently, this outcome completely compromises the classification capability of the GNB classifier as it cannot perform effective probability estimation. Therefore, in our experiments, we will remove such primary-key-like features from the Audi dataset, after which the number of features should be reduced to 69.

All experiments were done using python 3.9.7 with AMD Ryzen 9 CPU @ 2.50 GHz & 16 GB RAM hardware configuration. The version of *Scikit-learn* is 0.24.2.

4.2 Classification results

Tables 4 and 5 show the performance of each feature selection algorithm with different classifiers on heterogeneous data. Among them, the highest classification accuracy for each dataset is highlighted in bold. Results marked with underlines indicate that the maximum classification accuracy among all outputs of the algorithm matches that of the raw data (i.e., no features are eliminated). Additionally, a horizontal line denotes cases where the algorithm fails to

Table 4 Classification accuracy of heterogeneous data under CART classifier

Datasets	Raw data	FScNCE	WAFS	DRFFS	MRNG	MFIGI	IFIE	FR-NDM
Auto	82.46±1.66	86.99±1.74	81.84±1.93	**87.56±1.41**	87.20±1.95	85.35±1.50	81.84±1.93	**87.56±1.66**
Cancer	66.25±1.88	68.63±1.34	67.00±2.70	66.97±2.11	66.25±1.88	65.31±1.89	66.25±1.88	**75.45±0.50**
Cardio	92.24±0.40	92.17±0.23	81.45±0.08	92.37±0.32	92.10±0.32	90.76±0.39	90.92±0.37	**92.59±0.33**
Clinic	77.69±1.56	77.69±1.56	77.69±1.56	77.53±1.63	77.69±1.56	78.02±1.54	**78.65±1.47**	78.48±1.88
Credit	81.47±1.01	81.31±1.40	83.88±0.61	80.88±0.86	81.29±1.01	78.74±1.35	80.85±1.09	**86.83±0.22**
Ger	67.88±1.27	67.02±1.15	67.07±1.08	65.31±1.02	66.97±0.98	66.37±1.02	66.26±1.10	**70.62±0.42**
HCV	92.89±0.49	93.31±0.32	89.30±0.01	93.31±0.63	93.11±0.49	93.33±0.44	92.40±0.55	**93.58±0.44**
Heart	73.64±1.79	50.65±2.24	54.19±0.90	54.60±0.58	48.78±1.57	51.29±2.11	74.16±2.18	**79.61±1.28**
Horse	80.46±1.39	—	77.34±1.74	80.19±1.18	77.31±1.16	65.27±0.96	77.80±1.10	**82.41±1.16**
Obe	93.39±0.34	94.15±0.25	93.39±0.34	95.50±0.37	94.46±0.32	94.28±0.29	87.59±0.51	**95.65±0.27**
Predict	68.37±0.51	67.68±0.52	67.43±0.42	68.38±0.09	64.58±0.42	50.37±0.34	60.53±0.60	**69.09±0.53**
Thyroid	93.23±0.76	93.38±0.86	93.21±0.53	**95.88±0.20**	94.07±0.57	92.61±0.47	92.61±0.63	94.62±0.69
Average	80.83±1.09	79.45±1.08	77.82±0.99	79.87±0.87	78.65±1.02	75.98±1.02	79.15±1.12	**83.87±0.78**

Table 5 Classification accuracy of heterogeneous data under GNB classifier

Datasets	Raw data	FScNCE	WAFS	DRFFS	MRNG	MFIGI	IFIE	FR-NDM
Auto	55.67±2.03	53.92±2.25	53.15±1.75	52.63±1.86	50.16±1.69	53.75±1.39	52.17±1.63	**58.28±1.93**
Cancer	43.17±0.71	43.21±0.58	43.17±0.71	43.01±1.09	43.17±0.71	40.33±0.89	43.17±0.71	**73.03±0.95**
Cardio	51.26±0.71	84.30±0.12	22.66±0.50	81.10±0.14	82.42±0.10	**84.73±0.11**	80.78±0.18	84.52±0.10
Clinic	76.44±0.51	82.13±0.62	76.44±0.51	78.63±0.82	82.13±0.62	**82.81±0.72**	76.44±0.75	82.61±0.61
Credit	71.36±0.72	75.52±1.23	75.91±0.86	80.47±0.92	75.91±0.86	81.20±1.08	72.10±1.35	**86.37±0.00**
Ger	67.46±0.77	67.90±1.30	68.40±0.86	69.85±0.17	67.24±0.56	64.48±0.64	64.90±0.86	**73.09±0.49**
HCV	88.63±1.47	92.67±0.28	83.43±0.62	92.43±0.19	93.34±0.30	**93.80±0.25**	86.85±1.35	93.31±0.35
Heart	82.16±1.20	35.16±2.23	82.16±1.20	55.15±0.57	33.71±2.44	33.22±1.79	81.81±1.51	**84.53±0.81**
Horse	40.64±1.01	—	78.86±0.32	78.94±0.64	78.70±0.25	71.09±0.79	80.00±0.56	**84.59±0.32**
Obe	53.80±0.61	54.13±0.59	53.80±0.61	60.18±0.33	59.86±0.50	55.11±0.26	54.13±0.59	**62.35±0.27**
Predict	23.54±0.17	33.02±0.12	23.42±0.16	67.86±0.00	24.61±0.64	29.75±1.07	22.84±0.12	**72.39±0.11**
Thyroid	92.31±0.51	92.99±0.98	93.31±0.44	94.93±0.41	92.99±0.98	94.25±0.37	94.72±0.34	**95.74±0.23**
Average	62.2±0.87	62.97±0.94	62.89±0.71	71.26±0.6	65.35±0.8	65.38±0.78	67.49±0.83	**79.23±0.51**

output a feature subset. In fact, only FScNCE produces no output on the Horse dataset, as it fails to identify any features with internal significance greater than 0 during the initialization phase, resulting in an empty set. In subsequent calculations, we substitute the original classification accuracy of the Horse dataset for the missing FScNCE result. According to Tables 4 and 5, the following experimental results are obtained.

1. Under the CART classifier, FR-NDM achieves the highest classification accuracy on 10 datasets. Among them, the Cancer dataset has the highest classification accuracy improvement of 13.89%. In addition, DRFFS and IFIE attain the highest classification accuracy on 2 and 1 datasets, respectively.
2. Under the GNB classifier, FR-NDM achieves the highest classification accuracy on 9 datasets. Among them, the classification accuracy of the Predict dataset is improved the most, which is 207.52%. Besides, MFIGI attains the highest classification accuracy on 3 datasets.
3. Some algorithms fails to effectively eliminate redundant features. For instance, under the CART classifier, MRNG is unable to select the optimal feature subset for both the Cancer and Clinic datasets. Similarly, under the GNB classifier, WAFS dose not improve the classification performance for the Cancer, Clinic, Heart, and Obe datasets.
4. Our method achieves superior average classification accuracy compared to other algorithms under both classifiers, with scores of 83.87% and 79.23% respectively.

Comparative analyses on 12 heterogeneous datasets demonstrate the superior performance of the proposed method over other algorithms. First, the introduced nominal distribution metric enhances the fuzzy lower approximation's capability to characterize decision uncertainty in heterogeneous data samples, thereby improving the evaluation of feature significance. Second, the dual-parameter adjustable fuzzy heterogeneous information granules exhibit strong adaptability across diverse datasets while providing classifiers with more flexible candidate feature subsets.

Tables 6 and 7 present the actual performance of different feature selection algorithms on nominal data. Here, FR-NDM, FR-NDM^1, and FR-NDM^2 represent fuzzy rough fitting models embedded with δ_c, δ_c^1 and δ_c^2 respectively, where the highest classification accuracy for each dataset is indicated in bold font. From Tables 6 and 7, the following analytical conclusions can be drawn.

1. Under the CART classifier, FR-NDM, FR-NDM^1, and FR-NDM^2 achieve the highest classification accuracy on 5, 4, and 3 datasets respectively. Among these, FR-NDM shows the most significant classification performance improvement (33.48%) on the NPHA dataset, while FR-NDM^1 and FR-NDM^2 attain their maximum performance enhancements on the Education dataset, with improvements of 49.49% and 38.45% respectively. Overall, our methods outperform other algorithms on 11 datasets. Additionally, FRC attains the highest classification accuracy on 1 dataset.
2. Under the GNB classifier, FR-NDM, FR-NDM^1, and FR-NDM^2 achieve the highest classification accuracy on 3, 6, and 5 datasets respectively. Notably, all three methods obtain their maximum classification performance improvement on the NPHA dataset, with accuracy gains of 97.18%, 60.92%, and 111.00% respectively. Collectively, our methods demonstrate superior performance compared to other algorithms across 9 datasets. Moreover, FRC and IFIE achieve the highest classification accuracy on 2 and 1 datasets, respectively.
3. FR-NDM, FR-NDM^1, and FR-NDM^2 all achieve higher average classification accuracy than other algorithms under both classifiers. Meanwhile, their overall performance remain comparable. For instance, under

Table 6 Classification accuracy of nominal data under CART classifier

Datasets	Raw data	FRC	LDP	FScNCE	MRNG	MFIGI	IFIE	FR-NDM	FR-NDM^1	FR-NDM^2
Audio	75.75±2.07	76.21±1.43	74.30±1.24	70.33±1.80	68.17±1.68	61.80±1.29	74.12±1.59	77.27±0.87	78.54±1.40	**79.35±1.28**
Chess	99.58±0.08	99.57±0.08	99.43±0.06	99.09±0.07	96.83±0.13	95.53±0.00	98.66±0.08	98.22±0.09	98.06±0.08	98.92±0.08
Education	26.45±2.89	26.86±2.05	27.40±2.29	28.28±2.47	17.67±2.59	28.91±1.84	28.24±2.37	34.49±2.41	**39.54±1.00**	36.62±1.37
Lym	80.42±2.05	80.81±2.27	77.75±1.75	76.01±1.41	80.29±1.96	72.95±1.23	80.29±1.96	**83.70±1.29**	82.64±1.50	83.45±2.17
Molecular	92.82±0.28	68.43±0.00	62.12±0.58	69.43±0.68	52.41±0.55	62.35±0.40	83.75±0.38	93.15±0.11	**93.34±0.18**	93.13±0.42
Mush	100.00±0.00	**100.00±0.00**	99.70±0.00	99.47±0.05	99.43±0.03	98.82±0.00	99.90±0.00	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**
NPHA	39.34±1.56	49.80±0.44	39.27±1.22	39.57±1.50	39.57±1.50	39.99±1.11	39.57±1.50	**52.51±0.54**	51.48±0.48	52.11±0.01
Risk	94.50±1.11	95.15±1.05	92.25±0.81	94.50±0.84	91.65±0.87	89.95±0.82	95.07±0.87	97.85±0.63	97.00±0.32	**98.70±0.40**
SPECT	74.05±1.63	79.39±0.03	74.44±1.96	75.10±1.18	74.44±1.96	78.75±0.97	74.61±1.35	**79.41±0.03**	79.39±0.04	79.39±0.04
Soybean	86.26±1.37	57.58±1.42	70.25±2.11	78.37±1.53	69.77±1.20	51.06±2.51	70.91±1.67	88.11±0.82	**88.20±0.65**	85.74±0.74
Tic	94.56±0.71	75.42±0.45	72.17±0.80	90.40±0.61	82.23±1.11	76.93±0.36	90.63±0.61	94.56±0.71	94.56±0.71	94.56±0.71
Zoo	95.55±1.36	95.55±1.28	93.68±0.75	94.75±1.15	95.24±0.87	93.68±0.75	95.59±1.04	**99.01±0.03**	98.81±0.60	98.81±0.60
Average	79.94±1.26	75.4±0.87	73.56±1.13	76.28±1.11	72.31±1.2	70.89±0.94	77.61±1.12	83.19±0.63	**83.46±0.58**	83.40±0.65

the CART classifier, their average classification accuracies reach 83.19%, 83.46%, and 83.40% respectively.

Comparative experiments on 12 nominal datasets with six algorithms demonstrate that our three methods exhibit superior performance in classification tasks. Unlike most algorithms that produce limited outputs and fail to provide optimal feature subsets for different classifiers (with the exception of FRC, which allows parameter tuning), FR-NDM, FR-NDM^1, and FR-NDM^2 enable finer-grained mining of fuzzy similarity relations among samples under nominal attributes. By adjusting parameter σ, our methods can generate more diverse feature subsets for classifiers. Furthermore, the improved fitting fuzzy lower approximation derived from (16) enhances the accuracy of the fuzzy dependency function in feature evaluation, thereby strengthening FRS's capability to identify significant features.

Tables 8 and 9 present the sizes of optimal feature subsets obtained by different feature selection algorithms on heterogeneous data and nominal data, respectively. The last row in each table displays the average optimal feature subset size across all datasets for each feature selection algorithm, with the results rounded to integers (decimal places removed). Based on these two tables, we observe the following findings.

1. On the same dataset, the optimal number of feature subsets corresponding to different classifiers is different in most cases. For example, in heterogeneous data, the optimal feature subsets required by CART and GNB differ across 9 datasets when using the feature subsets provided by FR-NDM.
2. Different feature subsets may yield identical classification performance. As demonstrated in nominal data, while FR-NDM and FR-NDM^1 generate optimal feature subsets of size 6 for the Mush dataset under CART classifier, both FRC and FR-NDM^2 produce subsets of size 7—yet all achieve 100.00% classification accuracy.
3. In heterogeneous data, our method achieves an average optimal feature subset size of 9 under both classifiers, which is smaller than those of FScNCE, WAFS, and IFIE. In nominal data, among the fuzzy rough fitting model embedded with different nominal distribution metrics, FR-NDM performs the best, yielding average optimal feature subset sizes of 10 and 11 under the CART and GNB classifiers, respectively. These values are smaller than those obtained by FRC, FScNCE, and IFIE.

In summary, whether it is heterogeneous data or nominal data, our method can obtain a relatively small number of feature subsets to improve or preserve the classification

Table 7 Classification accuracy of nominal data under GNB classifier

Datasets	Raw data	FRC	LDP	FScNCE	MRNG	MFIGI	IFIE	FR-NDM	FR-NDM[1]	FR-NDM[2]
Audio	70.77±1.11	71.24±0.60	57.48±1.39	61.53±1.26	54.42±1.18	47.46±1.98	64.15±1.54	**71.45±1.86**	69.41±1.03	69.75±1.37
Chess	62.47±0.17	**91.91±0.34**	67.30±0.83	66.74±0.65	66.05±0.00	66.05±0.00	66.19±0.05	70.94±0.14	80.61±0.25	83.62±0.23
Education	31.27±2.39	15.64±1.04	20.06±2.17	31.60±3.00	12.23±1.46	17.02±1.78	20.81±2.14	29.73±2.21	29.74±2.49	**33.71±2.22**
Lym	73.43±2.32	**77.10±1.42**	69.60±0.97	64.46±3.27	70.00±1.01	69.56±1.27	70.00±1.01	75.31±1.14	73.13±1.73	75.31±1.14
Molecular	74.76±0.19	60.55±0.63	54.23±0.89	42.03±0.39	48.64±1.21	53.30±0.81	70.09±0.42	80.72±0.45	**82.83±0.78**	80.33±0.22
Mush	98.73±0.04	99.07±0.02	98.82±0.00	65.19±0.06	98.90±0.01	98.52±0.00	98.87±0.00	99.41±0.00	**99.84±0.02**	99.76±0.01
NPHA	24.54±0.62	43.89±1.14	22.30±0.40	24.46±0.58	24.46±0.58	20.70±1.40	24.46±0.58	48.39±1.32	39.49±2.22	**51.78±0.38**
Risk	98.98±0.24	98.95±0.15	83.50±0.00	97.50±0.00	83.50±0.00	73.10±0.20	90.91±0.19	98.50±0.00	**99.00±0.20**	**99.00±0.00**
SPECT	55.72±0.90	53.19±0.98	78.23±0.57	59.37±0.54	78.23±0.57	78.04±0.97	**78.59±0.80**	62.02±0.77	62.02±0.77	62.02±0.77
Soybean	90.05±0.74	50.24±1.09	71.73±1.32	84.32±0.71	64.27±1.11	50.19±1.38	65.26±1.36	86.76±0.76	**90.64±0.88**	89.12±0.59
Tic	66.85±0.47	68.01±0.78	67.62±0.36	67.51±0.34	67.11±0.30	68.86±0.09	67.46±0.35	**69.94±0.00**	**69.94±0.00**	**69.94±0.00**
Zoo	94.78±0.93	95.17±1.46	87.13±2.22	90.80±1.01	85.67±3.51	87.41±1.94	92.48±1.09	**98.05±0.05**	**98.05±0.05**	**98.05±0.05**
Average	70.2±0.84	68.75±0.80	64.83±0.93	62.96±0.98	62.79±0.91	60.85±0.98	67.44±0.79	74.27±0.72	74.56±0.87	**76.03±0.58**

accuracy of the raw data. Therefore, the proposed fuzzy rough fitting model with nominal distribution metric embedding presents a feasible solution for feature selection in classification tasks.

4.3 Parameter sensitivity analysis

The size of fuzzy information granule can be controlled indirectly by modifying the parameters λ and σ, so the final feature subset may be very different. This section will analyze the relationship between parameters, classification accuracy, and reduction rate.

First, we investigate the impact of parameters σ and λ on the classification accuracy of heterogeneous data under the GNB classifier, as illustrated in Fig. 2. The detailed analysis is presented as follows.

1. The performance of certain datasets exhibits significant sensitivity to parameter σ within specific ranges. For instance, in the Thyroid dataset, classification accuracy decreases sharply as parameter σ increases from 0.3 to 0.6. In contrast, classification accuracy remains relatively stable with variations in parameter λ when σ is held constant. Similar patterns are observed in both the Cancer and Heart datasets.
2. As can be seen in Fig. 2(c), the dataset Cardio is sensitive to a certain range of parameters λ. When the parameter σ is unchanged, the classification accuracy increases sharply with the parameter λ increasing in the range of [0.7, 1.0] and [2.2, 2.5]. However, when λ is fixed, its classification accuracy only fluctuates slightly as σ changes.
3. For the dataset Ger, its classification accuracy fluctuates unordered with the change of parameter σ, while the dataset HCV is affected by parameter λ. These results are attributed to the influence of the distribution and characteristics of the datasets itself.
4. The classification accuracy of datasets Auto, Clinic, Credit, Horse, Obe, and Predict exhibits significant fluctuations in response to variations in parameters σ and λ. Nevertheless, near-optimal classification performance can be achieved across most parameter combinations.

In summary, the classification accuracy of heterogeneous data is indeed influenced by parameters σ and λ, with different datasets exhibiting varying sensitivities to these parameters. In fact, ε_a in (14) serves as the neighborhood radius, and parameter λ controls the size of fuzzy information granules by adjusting this radius. Previous studies [20, 48] have confirmed that this mechanism significantly impacts the results of feature subset selection. Moreover, with the incorporation of parameter σ, the output of FR-NDM

Table 8 Optimal feature subset size for different feature selection algorithms on heterogeneous data

Datasets	Raw data	FScNCE		WAFS		DRFFS		MRNG		MFIGI	IFIE	FR-NDM	
		CART	GNB	CART	GNB	CART	GNB	CART	GNB			CART	GNB
Auto	25	12	13	16	17	5	6	7	7	5	16	11	19
Cancer	9	8	8	8	9	7	6	9	9	4	9	2	4
Cardio	21	19	7	3	3	12	5	18	5	5	12	8	11
Clinic	12	12	5	12	12	11	10	12	5	6	11	6	6
Credit	15	13	13	8	13	10	7	12	13	4	11	4	1
Ger	20	13	13	15	15	10	1	8	9	4	12	6	12
HCV	12	7	7	1	1	9	5	6	6	4	9	5	6
Heart	13	10	12	8	13	1	1	8	12	6	12	8	12
Horse	27	—	—	9	9	8	7	7	6	5	8	15	9
Obe	16	10	15	16	16	3	3	7	10	7	15	4	4
Predict	36	14	18	19	13	1	1	7	7	3	13	25	12
Thyroid	16	7	6	8	8	3	3	6	6	3	8	11	11
Average	18	13	12	10	11	7	5	9	8	5	11	9	9

demonstrates greater diversity, indicating that our method can provide classifiers with a wider range of candidate feature subsets. Therefore, how to efficiently and accurately determine the optimal parameter combination for σ and λ remains an important issue worthy of further investigation.

Then we explore the influence of parameter σ on the feature subset of nominal data and the classification accuracy. Figure 3 illustrates the reduction rate of FR-NDM[2] and the classification accuracy of its feature subsets under two classifiers as parameter σ varies. The graphs drawn using FR-NDM as well as FR-NDM[1] are roughly equivalent to FR-NDM[2]. Moreover, the reduction rate is calculated as follows:

$$\text{Reduction rate} = \frac{|C| - |R|}{|C|} \tag{21}$$

where R represents the reduction set R output in Algorithm 1, and C is the conditional attribute set, that is, all the features in the dataset.

As shown in Fig. 3, with the increase of parameter σ, the reduction rates of different datasets show different changing trends. For datasets Chess, NPHA and SPECT, the reduction rate gradually increases as the parameter σ increases. For datasets Education, Lym and Mush, the reduction rate first decreases and then increases. According to Fig. 3, the following analysis results can be obtained.

1. Although parameter σ significantly influences feature selection results, certain datasets (e.g., Audio, Mush, Risk, SPECT, Soybean, and Zoo) can achieve consistently high classification accuracy within specific ranges of σ.
2. By adjusting the parameter σ, a smaller subset of features can be obtained while maintaining or even improving classification performance. For the Education dataset, when σ exceeds 0.5, the classification accuracy initially increases slightly and then stabilizes. Additionally, for the SPECT dataset, when A ranges from 0.3 to 0.6, the reduction rate continues to increase, whereas the CART classification performance remains unchanged.
3. As the reduction rate increases, the classification accuracy of certain datasets (e.g., Chess, Lym, and NPHA) declines but eventually stabilizes. This suggests that excessive attribute elimination leads to information loss, which negatively impacts classification performance.

In general, there is no absolute correlation between the reduction rate and classification performance, which can be attributed to the diversity of data. Additionally, the impact of parameter σ variation is notably significant. Therefore, for most nominal datasets, we recommend narrowing the

Table 9 Optimal feature subset size for different feature selection algorithms on nominal data

Datasets	Raw data	FRC		LDP	FScNCE	MRNG	MFIGI	IFIE	FR-NDM		FR-NDM[1]		FR-NDM[2]	
		CART	GNB						CART	GNB	CART	GNB	CART	GNB
Audio	69	46	63	12	24	10	7	29	28	28	24	35	29	35
Chess	36	35	9	28	21	17	6	19	22	19	20	26	27	18
Education	31	3	3	6	11	6	4	7	7	11	3	23	2	22
Lym	18	17	16	5	8	6	4	6	10	10	6	16	10	10
Molecular	60	2	2	7	15	7	5	8	6	4	21	10	20	5
Mush	22	7	7	3	8	5	2	5	6	4	6	13	7	9
NPHA	14	2	2	12	13	13	6	13	4	3	5	3	1	1
Risk	27	14	17	3	7	4	2	3	4	9	9	14	5	7
SPECT	22	4	20	15	17	15	7	18	1	18	3	18	3	18
Soybean	35	6	6	8	14	7	5	9	12	13	17	17	14	16
Tic	9	4	4	7	8	8	5	8	9	1	9	1	9	1
Zoo	16	15	15	4	7	5	4	5	7	8	8	7	8	7
Average	30	13	14	9	13	9	5	11	10	11	11	15	11	12

tuning range of parameter σ to [0.05, 0.5]. This adjustment can effectively reduce search time while maintaining a high probability of obtaining optimal feature subsets for different classifiers.

4.4 Statistical test

Friedman test [52] and Nemenyi test [53] are used to test whether there are significant differences in the experimental results. Suppose that there are k comparison algorithms and N datasets, the Friedman statistic is defined as follows:

$$\chi_F^2 = \frac{12N}{k\,(k+1)} \left(\sum_{i=1}^{k} R_i^2 - \frac{k\,(k+1)}{4} \right) \tag{22}$$

and

$$F_F = \frac{(N-1)\,\chi_F^2}{N\,(k-1) - \chi_F^2} \tag{23}$$

where R_i is the average rank of the ith algorithm across all datasets, and F_F is the F-distribution with $(k-1)$ and $(k-1)(N-1)$ degrees of freedom. Moreover, the null hypothesis posits that no significant statistical difference exists between the experimental methods.

In this study, each experimental group involves 6 comparison algorithms and 12 datasets, yielding $k = 7$ and $N = 12$. With the significance level set at 0.1, the critical value is calculated as 1.87. Furthermore, the chi-square distribution table gives $\chi_F^2\,(6) = 10.65$. Based on these values, we conduct Friedman tests for the proposed methods, where FR-NDM, FR-NDM[1], and FR-NDM[2] are separately compared with other algorithms in nominal data. All test results are presented in Tables 10, 11, 12 and 13. Table 10 demonstrates that for heterogeneous data, FR-NDM achieves the highest ranking under both CART and GNB classifiers, with χ_F^2 exceeding $\chi_F^2\,(6)$ and F_F surpassing the critical value. Tables 11 to 13 present the Friedman test results of FR-NDM, FR-NDM[1], and FR-NDM[2] compared with other algorithms on nominal data, respectively. It can be observed that our three methods achieve the highest classification accuracy rankings under both classifiers. Similarly, χ_F^2 and F_F are larger than $\chi_F^2\,(6)$ and critical value. Consequently, we reject the null hypothesis, confirming that the proposed methods exhibit statistically significant differences compared to other algorithms in both heterogeneous and nominal data scenarios.

Further, we use the Nemenyi test to explore the differences between our approaches and other methods. In the Nemenyi test, the critical distance (CD) is defined as follows:

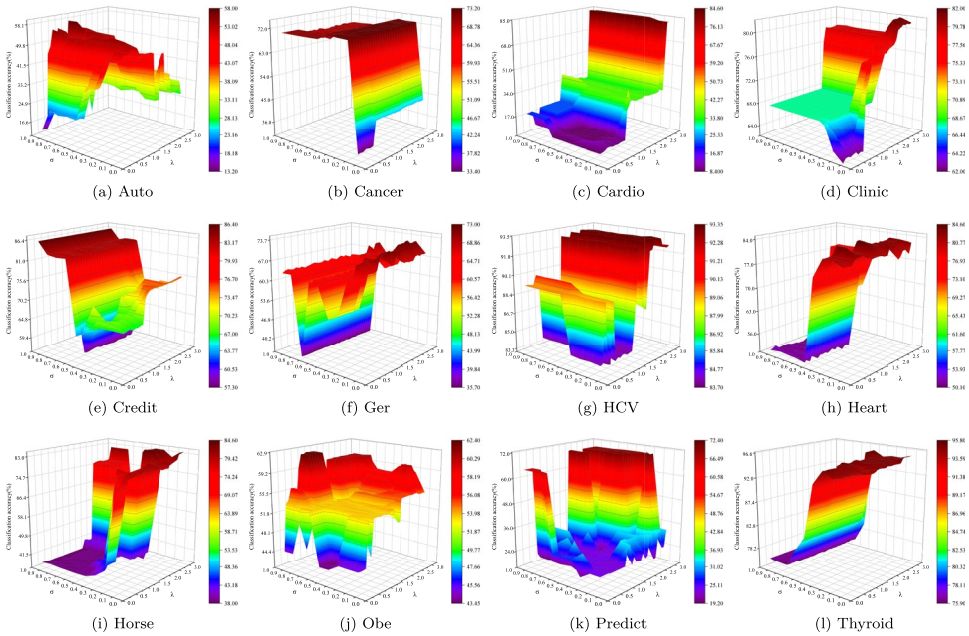


Fig. 2 Variation trend of classification accuracy of heterogeneous data with parameters σ and λ under GNB classifier

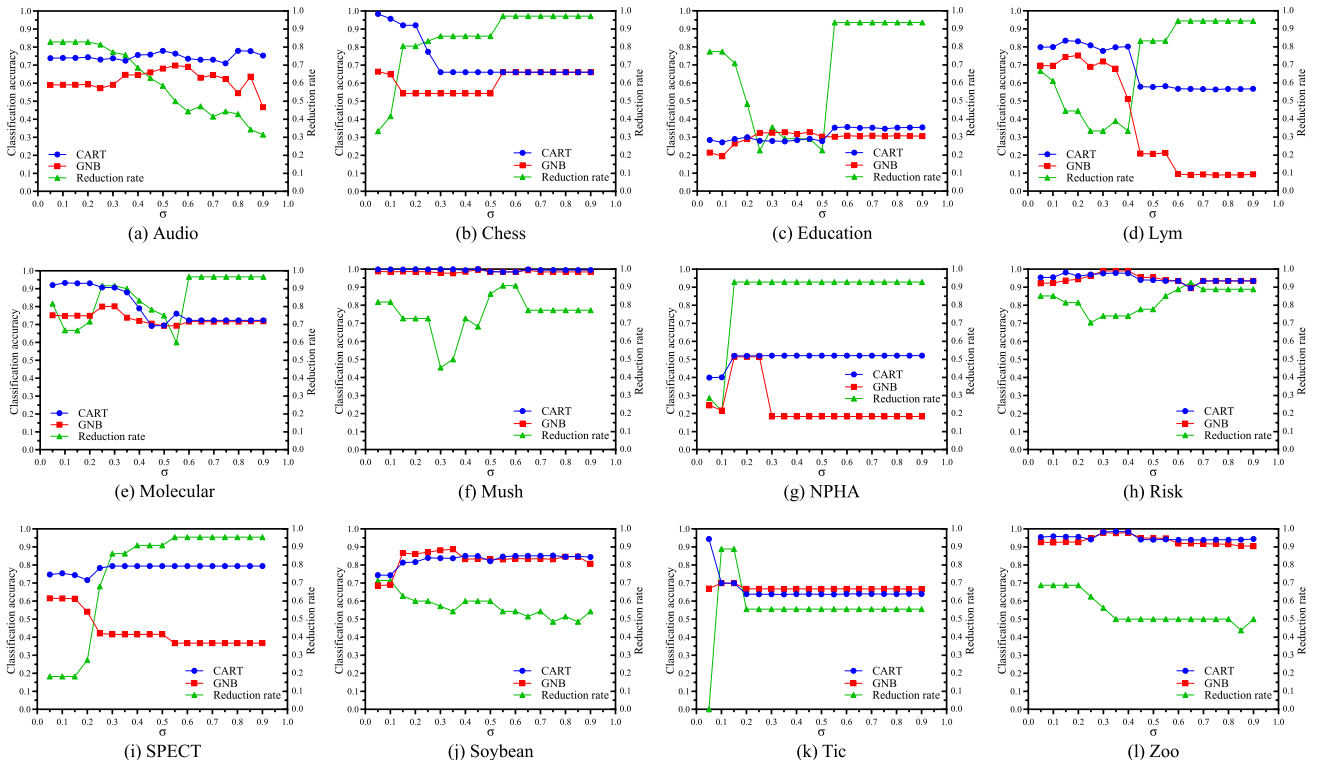


Fig. 3 Trend of classification accuracy and reduction rate with parameter σ in nominal data of FR-NDM[2]

Table 10 Friedman test of different feature selection algorithms on heterogeneous data

Algorithms	Average ranks							χ_F^2	F_F	χ_F^2 (6)	Critical value
	FScNCE	WAFS	DRFFS	MRNG	MFIGI	IFIE	FR-NDM				
CART	3.63	4.71	3.42	4.51	5.38	5.13	1.21	31.44	8.53	10.65	1.87
GNB	4.29	5.00	3.67	4.63	3.92	5.17	1.33	25.88	6.17		

Table 11 Friedman test of FR-NDM versus other feature selection algorithms on nominal data

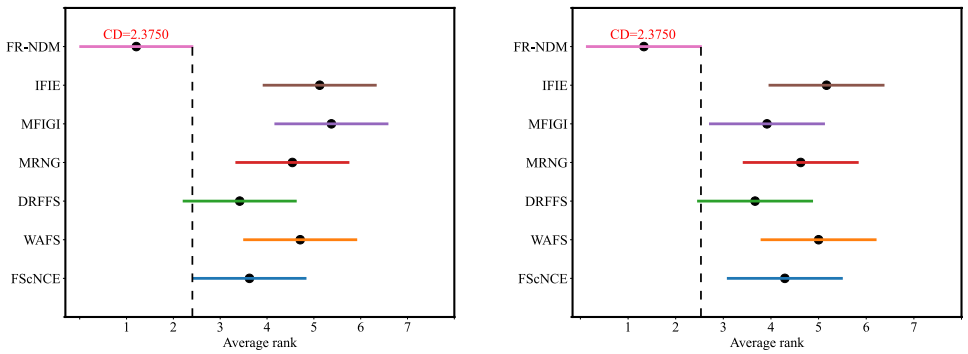
Algorithms	Average ranks							χ_F^2	F_F	$\chi_F^2(6)$	Critical value
	FRC	LDP	FScNCE	MRNG	MFIGI	IFIE	FR-NDM				
CART	3.13	5.08	4.00	5.50	5.54	3.38	1.38	35.61	10.76	10.65	1.87
GNB	3.00	4.42	4.50	5.25	5.63	3.54	1.67	29.01	7.42		

Table 12 Friedman test of FR-NDM[1] versus other feature selection algorithms on nominal data

Algorithms	Average ranks							χ_F^2	F_F	$\chi_F^2(6)$	Critical value
	FRC	LDP	FScNCE	MRNG	MFIGI	IFIE	FR-NDM[1]				
CART	3.08	5.08	4.00	5.50	5.54	3.38	1.42	35.24	10.55	10.65	1.87
GNB	2.92	4.42	4.50	5.25	5.63	3.54	1.75	28.47	7.20		

Table 13 Friedman test of FR-NDM[2] versus other feature selection algorithms on nominal data

Algorithms	Average ranks							χ_F^2	F_F	$\chi_F^2(6)$	Critical value
	FRC	LDP	FScNCE	MRNG	MFIGI	IFIE	FR-NDM[2]				
CART	3.08	5.08	4.00	5.50	5.54	3.46	1.33	36.12	11.07	10.65	1.87
GNB	3.00	4.42	4.58	5.25	5.63	3.54	1.58	30.26	7.97		



(a) CART: Heterogeneous Data (b) GNB: Heterogeneous Data

Fig. 4 Nemenyi test for different feature selection methods on heterogeneous data

$$CD_a = \rho_a \sqrt{\frac{k(k+1)}{6N}} \tag{24}$$

where ρ_a is the critical value of the two-tailed Nemenyi test and is equal to 2.693 when the significance level a is 0.1 [53]. Therefore, we calculate the CD as 2.3750.

Based on the average ranks presented in Tables 10 to 13, we construct Nemenyi test diagrams comparing different algorithms under two classifiers, as shown in Figs. 4 and 5. In these diagrams, when the vertical dashed line at the end of FR-NDM's critical distance intersects with other algorithms, it indicates no statistically significant superiority. Figure 4(a) demonstrates that in the CART classification experiments on heterogeneous data, FR-NDM significantly outperforms WAFS, MRNG, MFIGI, and IFIE. Figure 4(b) shows that under the GNB classifier, FR-NDM achieves significant superiority over FScNCE, WAFS, MRNG, MFIGI, and IFIE. Figure 5 presents the Nemenyi test results comparing the proposed method with other algorithms on nominal data. Specifically, Figs. 5(a) to 5(c) display the Nemenyi test results of FR-NDM, FR-NDM[1], and FR-NDM[2], respectively, against other algorithms using the CART classifier. Similarly, Figs. 5(d) to 5(f) show the Nemenyi test results of FR-NDM, FR-NDM[1], and FR-NDM[2], respectively, against other algorithms using the GNB classifier. It can be visually observed that regardless of which nominal distribution metric is embedded in the FRS model, its performance significantly outperforms LDP, FScNCE, MRNG, and MFIGI. In summary, the proposed method exhibits nice performance and superiority across both heterogeneous and nominal data scenarios.

5 Conclusion

In this paper, a fuzzy rough fitting model with nominal distribution metric embedding is proposed. The proposed model considers both the distribution characteristics and decision tendencies of samples when calculating fuzzy similarity for nominal data, while ensuring that each sample attains maximum membership degree within its corresponding decision category. Then, a feature selection algorithm is designed based on the new model and compared with other 8 feature

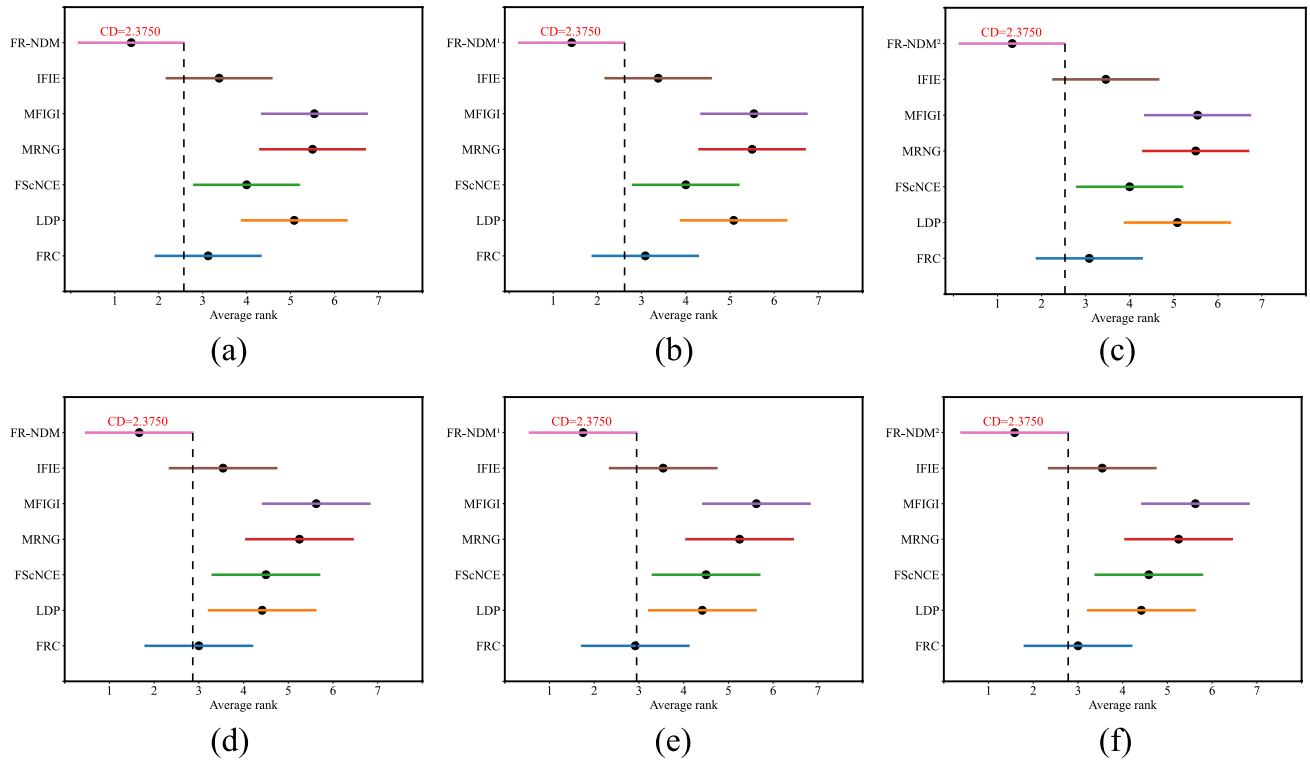


Fig. 5 Nemenyi test for different feature selection methods on nominal data

selection algorithms on 24 UCI datasets. The experimental results demonstrate that our proposed model improves classification performance and effectively removes redundant features. Furthermore, this work systematically examines parameter sensitivity on experimental outcomes and provides specific parameter tuning guidelines for nominal data processing. In future research, different distance metrics for nominal data could be further explored to enhance the feature selection performance of FRS in heterogeneous and nominal datasets. Additionally, under scenarios with missing labels, investigating how to mine similarity or dissimilarity relationships among nominal data would constitute meaningful work.

Data Availability The UCI public datasets used in the experiments are available at https://archive.ics.uci.edu/.

Declarations

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.

References

1. Qin J, Martínez L, Pedrycz W, et al (2023) An overview of granular computing in decision-making: Extensions, applications, and challenges. Inf Fusion p 101833. https://doi.org/10.1016/j.inffus.2023.101833
2. Bargiela A, Pedrycz W (2008) Toward a theory of granular computing for human-centered information processing. IEEE Trans Fuzzy Syst 16(2):320–330. https://doi.org/10.1109/TFUZZ.2007.905912
3. Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11:341–356. https://doi.org/10.1007/BF01001956
4. Li Z, Yang T, Li J (2023) Semi-supervised attribute reduction for partially labelled multiset-valued data via a prediction label strategy. Inf Sci 634:477–504. https://doi.org/10.1016/j.ins.2023.03.127
5. Dai J, Hu Q, Zhang J et al (2017) Attribute selection for partially labeled categorical data by rough set approach. IEEE Trans Cybern 47(9):2460–2471. https://doi.org/10.1109/TCYB.2016.2636339
6. Xu J, Zhou C, Xu S et al (2024) Feature selection based on multi-perspective entropy of mixing uncertainty measure in variable-granularity rough set. Appl Intell 54(1):147–168. https://doi.org/10.1007/s10489-023-05194-z
7. Radzikowska AM, Kerre EE (2002) A comparative study of fuzzy rough sets. Fuzzy Sets Syst 126(2):137–155. https://doi.org/10.1016/S0165-0114(01)00032-X
8. Dubois D, Prade H (1990) Rough fuzzy sets and fuzzy rough sets. Int J Gen Syst 17(2–3):191–209. https://doi.org/10.1080/03081079008935107
9. Shi Z, Li L, Xie S et al (2024) The variable precision fuzzy rough set based on overlap and grouping functions with double weight method to madm. Appl Intell 54(17):7696–7715. https://doi.org/10.1007/s10489-024-05554-3
10. Yu B, Zheng Z, Cai M et al (2024) Frcm: a fuzzy rough c-means clustering method. Fuzzy Sets Syst 480:108860. https://doi.org/10.1016/j.fss.2024.108860
11. Yuan Z, Chen H, Li T, et al (2021a) Outlier detection based on fuzzy rough granules in mixed attribute data. IEEE Trans Cybern 52(8):8399–8412. https://doi.org/10.1109/TCYB.2021.3058780

12. Yuan Z, Chen H, Xie P, et al (2021b) Attribute reduction methods in fuzzy rough set theory: An overview, comparative experiments, and new directions. Appl Soft Comput 107:107353. https://doi.org/10.1016/j.asoc.2021.107353
13. Dhal P, Azad C (2022) A comprehensive survey on feature selection in the various fields of machine learning. Appl Intell 52(4):4543–4581. https://doi.org/10.1007/s10489-021-02550-9
14. Jensen R, Shen Q (2004) Fuzzy-rough attribute reduction with application to web categorization. Fuzzy Sets Syst 141(3):469–485. https://doi.org/10.1016/S0165-0114(03)00021-6
15. Jensen R, Shen Q (2007) Fuzzy-rough sets assisted attribute selection. IEEE Trans Fuzzy Syst 15(1):73–89. https://doi.org/10.1109/TFUZZ.2006.889761
16. Jensen R, Shen Q (2008) New approaches to fuzzy-rough feature selection. IEEE Trans Fuzzy Syst 17(4):824–838. https://doi.org/10.1109/TFUZZ.2008.924209
17. Qian W, Xu F, Huang J et al (2023) A novel granular ball computing-based fuzzy rough set for feature selection in label distribution learning. Knowl-Based Syst 278:110898. https://doi.org/10.1016/j.knosys.2023.110898
18. Wang C, Wang C, An S, et al (2024a) Feature selection and classification based on directed fuzzy rough sets. IEEE Trans Syst Man Cybern: Syst 55(1):699–711. https://doi.org/10.1109/TSMC.2024.3492337
19. Wang C, Wang C, Qian Y, et al (2024b) Feature selection based on weighted fuzzy rough sets. IEEE Trans Fuzzy Syst 32(7):4027–4037. https://doi.org/10.1109/TFUZZ.2024.3387571
20. Yuan Z, Chen H, Li T et al (2021) Unsupervised attribute reduction for mixed data based on fuzzy rough sets. Inf Sci 572:67–87. https://doi.org/10.1016/j.ins.2021.04.083
21. Wang C, Qi Y, Shao M et al (2016) A fitting model for feature selection with fuzzy rough sets. IEEE Trans Fuzzy Syst 25(4):741–753. https://doi.org/10.1109/TFUZZ.2016.2574918
22. Qiu Z, Zhao H (2022) A fuzzy rough set approach to hierarchical feature selection based on hausdorff distance. Appl Intell 52(10):11089–11102. https://doi.org/10.1007/s10489-021-03028-4
23. Cornelis C, Cock M, Kerre EE (2003) Intuitionistic fuzzy rough sets: at the crossroads of imperfect knowledge. Expert Syst 20(5):260–270. https://doi.org/10.1111/1468-0394.00250
24. Tan A, Wu WZ, Qian Y et al (2018) Intuitionistic fuzzy rough set-based granular structures and attribute subset selection. IEEE Trans Fuzzy Syst 27(3):527–539. https://doi.org/10.1109/TFUZZ.2018.2862870
25. Tan A, Shi S, Wu WZ et al (2020) Granularity and entropy of intuitionistic fuzzy information and their applications. IEEE Trans Cybern 52(1):192–204. https://doi.org/10.1109/TCYB.2020.2973379
26. An S, Zhang M, Wang C et al (2023) Robust fuzzy rough approximations with knn granules for semi-supervised feature selection. Fuzzy Sets Syst 461:108476. https://doi.org/10.1016/j.fss.2023.01.011
27. Huang Z, Li J, Wang C (2024) Robust feature selection using multigranulation variable-precision distinguishing indicators for fuzzy covering decision systems. IEEE Trans Syst Man Cybern Syst 54(2):903–914. https://doi.org/10.1109/TSMC.2023.3321315
28. Sang B, Xu W, Chen H et al (2023) Active antinoise fuzzy dominance rough feature selection using adaptive k-nearest neighbors. IEEE Trans Fuzzy Syst 31(11):3944–3958. https://doi.org/10.1109/TFUZZ.2023.3272316
29. Kong L, Qu W, Yu J et al (2019) Distributed feature selection for big data using fuzzy rough sets. IEEE Trans Fuzzy Syst 28(5):846–857. https://doi.org/10.1109/TFUZZ.2019.2955894
30. Bai S, Lin Y, Lv Y et al (2021) Kernelized fuzzy rough sets based online streaming feature selection for large-scale hierarchical classification. Appl Intell 51:1602–1615. https://doi.org/10.1007/s10489-020-01863-5
31. Hu Q, Zhang L, Zhou Y et al (2017) Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets. IEEE Trans Fuzzy Syst 26(1):226–238. https://doi.org/10.1109/TFUZZ.2017.2647966
32. Wan J, Chen H, Li T et al (2021) Interactive and complementary feature selection via fuzzy multigranularity uncertainty measures. IEEE Trans Cybern 53(2):1208–1221. https://doi.org/10.1109/TCYB.2021.3112203
33. Yu D, An S, Hu Q (2011) Fuzzy mutual information based min-redundancy and max-relevance heterogeneous feature selection. Int J Comput Intell Syst 4(4):619–633. https://doi.org/10.2991/ijcis.2011.4.4.18
34. Wang C, Wang Y, Shao M et al (2019) Fuzzy rough attribute reduction for categorical data. IEEE Trans Fuzzy Syst 28(5):818–830. https://doi.org/10.1109/TFUZZ.2019.2949765
35. Li Z, Guo R, Lin N et al (2025) Local fuzzy rough attribute reduction for large-scale mixed data with limited missing labels based on local fuzzy self information. Inf Sci 691:121613 https://doi.org/10.1016/j.ins.2024.121613
36. Luo S, Miao D, Zhang Z et al (2020) A neighborhood rough set model with nominal metric embedding. Inf Sci 520:373–388. https://doi.org/10.1016/j.ins.2020.02.015
37. Zhang Y, Cheung YM (2020) A new distance metric exploiting heterogeneous interattribute relationship for ordinal-and-nominal-attribute data clustering. IEEE Trans Cybern 52(2):758–771. https://doi.org/10.1109/TCYB.2020.2983073
38. Zhang Y, Ym C (2021) Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes. IEEE Trans Pattern Anal Mach Intell 44(7):3560–3576. https://doi.org/10.1109/TPAMI.2021.3056510
39. Cheng V, Li CH, Kwok JT et al (2004) Dissimilarity learning for nominal data. Pattern Recogn 37(7):1471–1477. https://doi.org/10.1016/j.patcog.2003.12.015
40. Yuan F, Yang Y, Yuan T (2020) A dissimilarity measure for mixed nominal and ordinal attribute data in k-modes algorithm. Appl Intell 50(5):1498–1509. https://doi.org/10.1007/s10489-019-01583-5
41. Stanfill C, Waltz D (1986) Toward memory-based reasoning. Commun ACM 29(12):1213–1228. https://doi.org/10.1145/7902.7906
42. Hindi K (2013) Specific-class distance measures for nominal attributes. AI Commun 26(3):261–279. https://doi.org/10.3233/AIC-130565
43. Ahmad A, Dey L (2007) A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. Pattern Recogn Lett 28(1):110–118. https://doi.org/10.1016/j.patrec.2006.06.006
44. Jia H, Ym C, Liu J (2015) A new distance metric for unsupervised learning of categorical data. IEEE Trans Neural Netw Learn Syst 27(5):1065–1079. https://doi.org/10.1109/TNNLS.2015.2436432
45. Wang C, Cao L, Wang M, et al (2011) Coupled nominal similarity in unsupervised learning. In: Proceedings of the 20th ACM international conference on Information and knowledge management, pp 973–978, https://doi.org/10.1145/2063576.2063715
46. Zhu P, Hu Q, Zuo W et al (2014) Multi-granularity distance metric learning via neighborhood granule margin maximization. Inf Sci 282:321–331. https://doi.org/10.1016/j.ins.2014.06.017
47. Jian S, Cao L, Pang G, et al (2017) Embedding-based representation of categorical data by hierarchical value coupling learning. In: IJCAI International Joint Conference on Artificial Intelligence, http://hdl.handle.net/10453/126349
48. Zhang P, Li T, Yuan Z et al (2022) Heterogeneous feature selection based on neighborhood combination entropy. IEEE Trans Neural Netw Learn Syst 35(3):3514–3527. https://doi.org/10.1109/TNNLS.2022.3193929

49. Dai J, Zhu Z, Li M, et al (2024a) Attribute reduction for heterogeneous data based on monotonic relative neighborhood granularity. Int J Approx Reason 170:109210. https://doi.org/10.1016/j.ijar.2024.109210
50. Dai J, Zhu Z, Zou X (2024) Fuzzy rough attribute reduction based on fuzzy implication granularity information. IEEE Trans Fuzzy Syst 32(6):3741–3752. https://doi.org/10.1109/TFUZZ.2024.3381993
51. Yan S, Qian J, Yu Y et al (2025) A feature selection method driven by fuzzy implication granularity. Eng Appl Artif Intell 158:111298 https://doi.org/10.1016/j.engappai.2025.111298
52. Friedman M (1940) A comparison of alternative tests of significance for the problem of m rankings. Ann Math Statist 11(1):86–92. https://doi.org/10.1214/aoms/1177731944
53. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. The J Mach Learn Res 7:1–30. https://doi.org/10.1007/s10846-005-9016-2

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.