



Feature selection based on self-information and entropy measures for incomplete neighborhood decision systems

Meng Yuan^{1,2,3} · Jiucheng Xu^{1,2,3} · Tao Li^{1,2,3} · Yuanhao Sun^{1,2,3}

Received: 7 May 2022 / Accepted: 13 September 2022 / Published online: 11 October 2022
© The Author(s) 2022

Abstract

For incomplete datasets with mixed numerical and symbolic features, feature selection based on neighborhood multi-granulation rough sets (NMRS) is developing rapidly. However, its evaluation function only considers the information contained in the lower approximation of the neighborhood decision, which easily leads to the loss of some information. To solve this problem, we construct a novel NMRS-based uncertain measure for feature selection, named neighborhood multi-granulation self-information-based pessimistic neighborhood multi-granulation tolerance joint entropy (PTSIJE), which can be used to incomplete neighborhood decision systems. First, from the algebra view, four kinds of neighborhood multi-granulation self-information measures of decision variables are proposed by using the upper and lower approximations of NMRS. We discuss the related properties, and find the fourth measure-lenient neighborhood multi-granulation self-information measure (NMSI) has better classification performance. Then, inspired by the algebra and information views simultaneously, a feature selection method based on PTSIJE is proposed. Finally, the Fisher score method is used to delete uncorrelated features to reduce the computational complexity for high-dimensional gene datasets, and a heuristic feature selection algorithm is raised to improve classification performance for mixed and incomplete datasets. Experimental results on 11 datasets show that our method selects fewer features and has higher classification accuracy than related methods.

Keywords Feature selection · Neighborhood multi-granulation rough sets · Self-information · Neighborhood tolerance conditional entropy · Incomplete neighborhood decision system

Introduction

With the rapid development of information technology, the databases expand rapidly. In daily production and life, more and more information is obtained and stored [1–5]. However, these information may contain a great quantity of redundancy, noise, or even missing feature values [6–8]. Nowadays, how to deal with missing values, reduce redundant features, and simplify the complexity of the clas-

sification model, so as to improve the generalization ability of model classification is a huge challenge we are facing [9–15]. As an important step of data preprocessing, feature selection based on granular computing has been widely used in knowledge discovery, data mining, machine learning, and other fields [16–23].

Related work

Neighborhood rough sets [24] and multi-granularity rough sets [25] models, as two commonly used mathematical tools for dealing with uncertainty and incomplete knowledge, are widely employed in feature selection and attribute reduction [26–32]. Hu et al. [33] introduced the neighborhood relationship into different weights, constructed a weighted rough set model based on the weighted neighborhood relationship, and fully tapped the correlation between attributes and decision. Yang et al. [34] introduced fuzzy preference relations to propose a neighborhood rough set model based on the degree of dominance and applied it to the problem of attribute reduction

✉ Meng Yuan
y_m961123@163.com

✉ Jiucheng Xu
xjc@htu.edu.cn

¹ College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China

² Engineering Lab of Intelligence Business & Internet of Things, Xinxiang 453007, Henan Province, China

³ Henan Normal University, 46 Jianshe East Road, Muye District, Xinxiang City, Henan Province, China

in large-scale decision problems. Wang et al. [28] presented a neighborhood discriminant index to characterize the discriminative relationship of neighborhood information, which reflects the distinguishing ability of a feature subset. Qian et al. [35] developed a pessimistic multi-granulation rough sets decision model based on attribute reduction. Lin et al. [36] advanced a neighborhood-based coverage reduction multi-granulation rough set model. Sun et al. [37] combined the fuzzy neighborhood rough sets with the multi-granulation rough set model, and proposed a new fuzzy neighborhood multi-granulation rough set model, which expanded the types of rough set models. Meanwhile, the neighborhood multi-granulation rough sets' (NMRS) model has been widely used. Ma et al. [38] constructed an NMRS model based on the two-particle criterion, which effectively reduced the number of iterations in the attribute reduction calculation. Hu et al. [39] presented a matrix-based incremental method to update the knowledge in NMRS. To obtain more accurate rough approximation and reduce the interference of noise data, Luo et al. [40] developed a neighborhood multi-granular rough set variable precision decision method based on multi-threshold. Although a variety of NMRS-based feature selection methods have been proposed and applied, most of the evaluation functions of them are constructed only based on the lower approximation of decision. This construction method ignores the information contained in the upper approximation, which easily leads to the loss partial information [58].

In various fields datasets, missing values (null or unknown values) often occur [9,41]. Recently, tolerance relation or tolerance rough sets have emerged in the field of processing incomplete datasets [42]. Qian et al. [43] presented an incomplete rough set model based on multiple tolerance relations from the multi-granulation view. Yang et al. [44] proposed supporting feature functions for processing multi-source datasets through the perspective of multi-granulation. Sun et al. [45] developed neighborhood tolerance-dependent joint entropy for feature selection and revealed better classification performance on incomplete datasets. Zhao et al. [42] constructed an extended rough set model based on neighborhood tolerance relation, which was successfully applied to incomplete datasets with mixed categories and numbers. Inspired by these research achievements, this paper is dedicated to developing a heuristic feature selection method based on neighborhood tolerance relation to solve mixed incomplete datasets.

In recent years, the uncertainty measures had developed rapidly from the algebra view or the information view [46,47]. Hu et al. [39] constructed a matrix-based feature selection method to realize the uncertainty of the boundary region in the NMRS model. You et al. [48] proposed the relative reduction for covering information system. Zhang et al. [49] employed local pessimistic multi-granulation rough set to deal with larger size datasets. In general, the above stud-

ies are only from algebra view. Unfortunately, the attribute importance based on algebra view only describes the influence of the features contained in the subset. In the past decades, information entropy and its variants have been extensively used in feature selection as an important method [50,51]. Zeng et al. [52] improved multi-granulation entropy by developing multiple binary relations. Feng et al. [53] studied the reduction problem of multi-granulation fuzzy information system according to merging entropy and conditional entropy. In short, these literatures discussed feature selection only from information view. However, feature significance only from the information view barely reflects the significance on features in uncertainty classification [54,55]. It would be a better topic to integrate the two views to improve the uncertainty measure quality in feature selection for incomplete neighborhood decision systems. Wang et al. [56] studied rough reduction and relative reduction from two views simultaneously. Chen et al. [57] proposed the roughness based on entropy and the approximate roughness measurement of the neighborhood system. Xu et al. [58] presented a feature selection method using fuzzy neighborhood self-information measures and entropy in combination with algebra and information views. Table 1 summarizes and highlights some feature selection methods from the perspective of whether to deal with missing data and uncertainty measure views.

Our work

To solve the problem, some feature evaluation functions only consider the information contained in the lower approximation of the decision, which may lead to loss some information, moreover comprehensively evaluate the uncertainty of the incomplete neighborhood decision systems. This paper will focus on researching a new feature selection method in multi-granulation terms. The main work of this paper is as follows:

- Based on the related definitions of NMRS, the shortcomings of the related neighborhood functions are analyzed.
- Three kinds of uncertain indices are proposed, including decision index, sharp decision index, and blunt decision index, using upper and lower approximations of NMRS. Then, we redefine three types of precision and roughness based on three indices. Next, combining with the concept of self-information, four kinds of neighborhood multi-granulation self-information measures are proposed and their related properties are studied. According to theoretical analysis, the fourth measure, named lenient neighborhood multi-granulation self-information (NMSI), is suitable for select the optimal feature subsets.
- To better study the uncertainty measure for the incomplete neighborhood decision systems from the algebra and information views, the self-information measures

Table 1 Summary of some feature selection methods

Literature	Handling complete/incomplete datasets?	From algebra/ information view?
Qian et al. [35]	Complete	Algebra view
Sun et al. [37]	Complete	Both
Hu et al. [39]	Complete	Algebra view
Qian et al. [43]	Incomplete	Algebra view
Sun et al. [45]	Incomplete	Both
Zeng et al. [52]	Complete	Information view
Feng et al. [53]	Complete	Information view
Xu et al. [58]	Complete	Both

and information entropy are combined to propose a neighborhood multi-granulation self-information-based pessimistic neighborhood multi-granulation tolerance joint entropy (PTSIJE). PTSIJE not only considers the upper and lower approximations of the incomplete decision systems at the same time, but also can measure the uncertainty of the incomplete neighborhood decision systems from the algebra and information views simultaneously.

The structure of the rest of this paper is as follows: some related concepts of self-information and NMRS are reviewed in “Previous knowledge”. “The deficiency of relative function and PTSIJE-based uncertainty measures” illustrates the shortcomings of neighborhood correlation function. Then, to improve the shortcomings, we propose four neighborhood multi-granulation self-information measures and study their related properties. Finally, the fourth measure is combined with neighborhood tolerance joint entropy to construct a feature selection model based on PTSIJE. “PTSIJE-based feature selection method in incomplete neighborhood decision systems” designs a heuristic feature subset selection method. In “Experimental results and analysis” six UCI datasets and five gene expression profile datasets were used to verify the results. “Conclusion” is the conclusion of this paper and our future work.

Previous knowledge

Self-information

Definition 1 [59] Metric $I(x)$ is proposed by Shannon to represent the uncertainty of a signal x , is called the self-information of x if it met the following properties:

- (1) Non-negative: $I(x) \geq 0$.
- (2) If $p(x) \rightarrow 0$, then $I(x) \rightarrow \infty$.
- (3) If $p(x)=0$, then $I(x) = 1$.
- (4) Monotonic: If $p(x) < p(y)$, then $I(x) < I(y)$.

Here, $p(x)$ is the probability of x .

Neighborhood multi-granulation rough sets

Given a neighborhood decision $NDS = \langle U, CA, D, V, f, \Delta, \lambda \rangle$, $U = \{x_1, x_2, \dots, x_m\}$ is an universe, $CA = B_S \cup B_N$ is a conditional attribute set that depicts the samples with mixed data, here B_S is a symbolic attribute set, B_N is the a numerical attribute set; D is the decision attribute set; $V = \cup_{a \in CA \cup DS} V_a$ which V_a is a value of attribute a ; f is the map function; Δ is the distance function; λ is the neighborhood radius and $\lambda \in [0, 1]$. Suppose that $x \in U$ and $f(a, x)$ is equal to a missing value (an unknown value or a null, recorded as “*”), which means that there is at least one attribute $a \in CA$, $f(a, x) = *$, then this decision systems can be called an incomplete neighborhood decision system $INDS = \langle U, CA, D, V, f, \Delta, \lambda \rangle$, it can be abbreviated as $INDS = \langle U, CA, D, \lambda \rangle$.

Definition 2 Suppose an incomplete neighborhood decision system $INDS = \langle U, CA, D, \lambda \rangle$ with any $B \in CA$ and $B = B_S \cup B_N$, then the neighborhood tolerance relation of B is described as [42]

$$NT_B^\lambda = \{(x, y) \in U \times U | \forall a \in B (f(a, x) = *) \vee (f(a, y) = *) \vee (a \in B_S \rightarrow \Delta_a(x, y) = 0) \wedge (a \in B_N \rightarrow \Delta_a(x, y) \leq \lambda))\} \quad (1)$$

For any $x, y \in U$, the neighborhood tolerance class is expressed as [42]

$$NT_B^\lambda(x) = \{y \in U | (x, y) \in NT_B^\lambda\}. \quad (2)$$

Definition 3 Given $INDS = \langle U, CA, D, \lambda \rangle$ with $X \subseteq U$, $A = \{A_1, A_2, \dots, A_r\}$ and $A \subseteq CA$, the optimistic neighborhood multi-granulation lower and upper approximations of X with regard to $A_1, A_2, \dots, A_r$ are denoted, respectively, as

$$\sum_{i=1}^r A_i^{O, \lambda}(X) = \{x \in U | NT_{A_1}^\lambda(x) \subseteq X \vee$$

$$NT_{A_2}^\lambda(x) \subseteq X \vee \dots \vee NT_{A_r}^\lambda(x) \subseteq X \} \quad (3)$$

$$\overline{\sum_{i=1}^r A_i^{O,\lambda}(X)} = \sim \left(\sum_{i=1}^r A_i^{O,\lambda}(\sim X) \right); \quad (4)$$

here, $NT_{A_i}^\lambda(x)$ represents the neighborhood tolerance class, $A_i \subseteq A$ with $i = 1, 2, \dots, r$. Then, $\left(\sum_{i=1}^r A_i^{O,\lambda}(X), \overline{\sum_{i=1}^r A_i^{O,\lambda}(X)} \right)$ can be called an optimistic NMRS model (ONMRS) in incomplete neighborhood decision systems [60].

Definition 4 Assume that an incomplete neighborhood decision system $INDS = \langle U, CA, D, \lambda \rangle$ with $A_i \subseteq A$, $A = \{A_1, A_2, \dots, A_r\}$, and $U/D = \{D_1, D_2, \dots, D_t\}$, the optimistic positive region and the optimistic dependency degree of D with respect to A based on ONMRS are expressed [60], respectively, as

$$POS_A^{O,\lambda}(D) = \bigcup_{l=1}^t \sum_{i=1}^r A_i^{O,\lambda}(D_l) \quad (5)$$

$$\gamma_A^{O,\lambda}(D) = \frac{|POS_A^{O,\lambda}(D)|}{|U|}, \quad (6)$$

where $A_i \subseteq A$, $i = 1, 2, \dots, r$, $D_l \in U/D$, $l = 1, 2, \dots, t$.

Definition 5 Let $INDS = \langle U, CA, D, \lambda \rangle$ with $X \subseteq U$, $A = \{A_1, A_2, \dots, A_r\}$ and $A \subseteq CA$, then the pessimistic neighborhood multi-granulation lower and upper approximations of X with respect to $A_1, A_2, \dots, A_r$ are denoted [60], respectively, as

$$\sum_{i=1}^r A_i^{P,\lambda}(X) = \{x \in U | NT_{A_1}^\lambda(x) \subseteq X \wedge NT_{A_2}^\lambda(x) \subseteq X \wedge \dots \wedge NT_{A_r}^\lambda(x) \subseteq X\} \quad (7)$$

$$\overline{\sum_{i=1}^r A_i^{P,\lambda}(X)} = \sim \left(\sum_{i=1}^r A_i^{P,\lambda}(\sim X) \right); \quad (8)$$

here, $NT_{A_i}^\lambda(x)$ represents the neighborhood tolerance class, $A_i \subseteq A$ with $i = 1, 2, \dots, r$.

Then, $\left(\sum_{i=1}^r A_i^{P,\lambda}(X), \overline{\sum_{i=1}^r A_i^{P,\lambda}(X)} \right)$ can be called a pessimistic NMRS model (PNMRS) in incomplete neighborhood decision systems [60].

Definition 6 Given $INDS = \langle U, CA, D, \lambda \rangle$ with $A_i \subseteq A$, $A = \{A_1, A_2, \dots, A_r\}$, and $U/D = \{D_1, D_2, \dots, D_t\}$, the pessimistic positive region and the pessimistic dependency

degree of D with respect to A based on PNMRS are expressed [60], respectively, as

$$POS_A^{P,\lambda}(D) = \bigcup_{l=1}^t \sum_{i=1}^r A_i^{P,\lambda}(D_l) \quad (9)$$

$$\gamma_A^{P,\lambda}(D) = \frac{|POS_A^{P,\lambda}(D)|}{|U|}, \quad (10)$$

where $A_i \subseteq A$, $i = 1, 2, \dots, r$, $D_l \in U/D$, $l = 1, 2, \dots, t$.

The deficiency of relative function and PTSIJE-based uncertainty measures

The deficiency of relative function

The classic NMRS model employs Eq. (10) as the evaluation function for feature selection. However, this construction method only considers the positive region, that is to say, only a part of the decision information is taken into account, and the information contained in the upper approximation of the decision is often not negligible, so this construction method is easy to cause the loss of some information. Thus, an ideal evaluation function should take into account the information whether it is consistent with the decision. For this reason, in the next part, we construct PTSIJE to measure the uncertainty in the mixed incomplete neighborhood decision systems, making the feature selection mechanism more comprehensive and reasonable.

PTSIJE-based uncertainty measures

Definition 7 Let $INDS = \langle U, CA, D, \lambda \rangle$, $U/D = \{D_1, D_2, \dots, D_t\}$, $A \subseteq CA$, $A = \{A_1, A_2, \dots, A_r\}$, $A_i \subseteq A$, NT_A^λ is a neighborhood tolerance relation induced by A . The decision index $dec(D_k)$, the sharp decision index $shar_A(D_k)$, and the blunt decision index $blun_A(D_k)$ of D_k are denoted, respectively, by

$$dec(D_k) = |NT_A^\lambda| \quad (11)$$

$$shar_A(D_k) = \left| \sum_{i=1}^r A_i^{P,\lambda}(D_k) \right| \quad (12)$$

$$blun_A(D_k) = \left| \sum_{i=1}^r A_i^{P,\lambda}(D_k) \right|, \quad (13)$$

where the sharp decision index $shar_A(D_k)$ of D_k is illustrated as the cardinal number of its lower approximation, which is expressing the number of samples with consistent neighborhood decision classification. The blunt decision

index $blun_A(D_k)$ of D_k is showed by the cardinal number of its upper approximation, which representing number of samples may belong to D_k . $|\cdot|$ means the cardinality of the set.

Property 1 $shar_A(D_k) \leq dec(D_k) \leq blun_A(D_k)$.

Proof The detailed proof can be found in the supplementary file.

Property 2 Assume $A1 \subseteq A2 \subseteq CA$ and $D_k \in U/D = \{D_1, D_2, \dots, D_t\}$, then

- (1) $shar_{A1}(D_k) \leq shar_{A2}(D_k)$.
- (2) $blun_{A1}(D_k) \geq blun_{A2}(D_k)$.

Proof The detailed proof can be found in the supplementary file.

Property 2 reveals that both the sharp decision and the blunt decision indices are monotonic. The sharp decision index $shar_A(D_k)$ boosts and the consistency of decision is enhanced with the increase of the number of features. A smaller blunt decision index $blun_A(D_k)$ is generated by decrease of decision uncertainty. In other words, attribute reduction is produced by the uncertainty of decision decreases.

Definition 8 For an incomplete neighborhood decision system $INDS = \langle U, CA, D, \lambda \rangle$ with $A \subseteq CA$, $D_k \in U/D$, the precision and roughness of the sharp decision index are defined, respectively, as

$$\theta_A^{(1)}(D_k) = \frac{shar_A(D_k)}{dec(D_k)} \quad (14)$$

$$\sigma_A^{(1)}(D_k) = 1 - \frac{shar_A(D_k)}{dec(D_k)} = 1 - \theta_A^{(1)}(D_k). \quad (15)$$

Evidently, $0 \leq \theta_A^{(1)}(D_k), \sigma_A^{(1)}(D_k) \leq 1$, $\theta_A^{(1)}(D_k)$ shows the degree to which the sample is completely grouped into D_k . $\sigma_A^{(1)}(D_k)$ indicates that the sample is not classified to the degree of correct decision D_k . Both $\theta_A^{(1)}(D_k)$ and $\sigma_A^{(1)}(D_k)$ mean the classification ability of feature subset in different ways.

If $\theta_A^{(1)}(D_k) = 1$, $\sigma_A^{(1)}(D_k) = 0$, then $dec(D_k) = shar_A(D_k)$; it describes that all samples can be correctly divided into corresponding decision by feature subset A ; at this time, feature subset reaches the optimal classification ability. If $\theta_A^{(1)}(D_k) = 0$, $\sigma_A^{(1)}(D_k) = 1$, then $shar_A(D_k) = 0$; it indicates that all samples cannot be classified into correct decision D_k through A ; in this case, feature subset A has the weakest classification ability.

Property 3 Assume that $A1 \subseteq A2 \subseteq CA$ with $D_k \in U/D$, then

$$(1) \theta_{A1}^{(1)}(D_k) \leq \theta_{A2}^{(1)}(D_k).$$

$$(2) \sigma_{A1}^{(1)}(D_k) \geq \sigma_{A2}^{(1)}(D_k).$$

Proof The detailed proof can be found in the supplementary file.

Property 3 explains that the precision and roughness of the sharp decision index are monotonic. As the number of new features increases, higher precision and lower roughness of the sharp decision index will be generated.

Definition 9 Let $A \subseteq CA$, $D_k \in U/D$, the sharp decision self-information of D_k is defined as

$$I_A^1(D_k) = -\sigma_A^{(1)}(D_k) \ln \theta_A^{(1)}(D_k). \quad (16)$$

It is obvious to testify that $I_A^1(D_k)$ meets properties (1), (2), and (3) of the definition of self-information. Then, (4) can be confirmed according to Property 4.

Property 4 Let $A1 \subseteq A2 \subseteq CA$ and $D_k \in U/D$, then $I_{A1}^1(D_k) \geq I_{A2}^1(D_k)$.

Proof The detailed proof can be found in the supplementary file.

Definition 10 Given $INDS = \langle U, CA, D, \lambda \rangle$, $U/D = \{D_1, D_2, \dots, D_t\}$, $A \subseteq CA$, then the sharp decision self-information of $INDS$ can be defined as

$$I_A^1(D) = \sum_{k=1}^t I_A^1(D_k). \quad (17)$$

As we know, self-information was originally used to characterize the instability of signal output. Here, the application of self-information in the incomplete neighborhood decision systems can be used to picture the uncertainty of decision, which it is an effective medium to evaluating decision ability.

$I_A^1(D)$ delivers the classification information of the feature subset about the sharp decision. The smaller $I_A^1(D)$ is, the stronger classification ability of feature subset A is. When $I_A^1(D) = 0$, illustrates that all samples in U can be completely classified into the correct categories according to feature subset A .

However, the feature subset selected by sharp decision only focuses on the information contained in the consistent decision, while ignoring the information contained in the uncertain classification object. In addition, these uncertain informations are essential to decision classification, which cannot be ignored. Therefore, it is vital to analyze the information which contained in the uncertain classification objects. Next, we will define the precision and roughness of blunt decision to discuss the classification ability of feature subset.

Definition 11 Letting $INDS = \langle U, CA, D, \lambda \rangle$, $U/D = \{D_1, D_2, \dots, D_t\}$, then precision and roughness of the blunt decision index are denoted, respectively, as

$$\theta_A^{(2)}(D_k) = \frac{dec(D_k)}{blun_A(D_k)} \quad (18)$$

$$\sigma_A^{(2)}(D_k) = 1 - \frac{dec(D_k)}{blun_A(D_k)} = 1 - \theta_A^{(2)}(D_k). \quad (19)$$

Clearly, $0 \leq \theta_A^{(2)}(D_k), \sigma_A^{(2)}(D_k) \leq 1$. $\theta_A^{(2)}(D_k)$ shows the uncertain information contained in D_k . $\sigma_A^{(2)}(D_k)$ expresses the degree that the samples cannot be completely classified into the corresponding decision class. When $\theta_A^{(2)}(D_k) = 1$, $\sigma_A^{(2)}(D_k) = 0$, then $dec(D_k) = blun_A(D_k)$. It explains that all possible decisions are correctly divided into the decision D_k , and the feature subset has the strongest classification ability. On the contrary, feature subset A has no classification ability.

Property 5 Suppose that $A_1 \subseteq A_2 \subseteq CA$ and $D_k \in U/D$, then

$$(1) \theta_{A_1}^{(2)}(D_k) \leq \theta_{A_2}^{(2)}(D_k).$$

$$(2) \sigma_{A_1}^{(2)}(D_k) \geq \sigma_{A_2}^{(2)}(D_k).$$

Proof The detailed proof can be found in the supplementary file.

According to *Property 5*, it explicitly illustrates that the precision and roughness of the blunt decision index are monotonic. When new features are added, the precision of the blunt decision increases with the roughness falls.

Definition 12 Suppose that $A \subseteq CA$ with $D_k \in U/D$, then blunt decision self-information of D_k can be defined as

$$I_A^2(D_k) = -\sigma_A^{(2)}(D_k) \ln \theta_A^{(2)}(D_k). \quad (20)$$

It is obvious to testify that $I_A^2(D_k)$ meets properties (1), (2), and (3) of the definition of self-information, (4) can be confirmed according to *Property 6*.

Property 6 Letting $A_1 \subseteq A_2 \subseteq CA$ and $D_k \in U/D$, then $I_{A_1}^2(D_k) \geq I_{A_2}^2(D_k)$.

Proof The detailed proof can be found in the supplementary file.

Definition 13 Assume an incomplete decision system $INDS = \langle U, CA, D, \lambda \rangle$ with $U/D = \{D_1, D_2, \dots, D_t\}$, and $A \subseteq CA$, then the blunt decision self-information of $INDS$ is defined as

$$I_A^2(D) = \sum_{k=1}^t I_A^2(D_k). \quad (21)$$

Through the above analysis, we can obtain that sharp decision self-information $I_A^1(D)$ relies on samples with consistent decision classification, while blunt decision self-information $I_A^2(D)$ considers samples with inconsistent classification information, but it cannot ensure that all classification information is definitive. Therefore, both $I_A^1(D)$ and $I_A^2(D)$ have insufficient classification capabilities in describing feature subsets, and they are rather one-sided. For this reason, we will propose two other kinds of self-information about classification decision to measure the uncertainty of incomplete neighborhood systems.

Definition 14 Let $A \subseteq CA$ and $D_k \in U/D$, the sharp-blunt decision self-information of D_k can be denoted as

$$I_A^3(D_k) = I_A^1(D_k) + I_A^2(D_k). \quad (22)$$

It is obvious to testify that $I_A^3(D_k)$ meets properties (1), (2), and (3) of the definition of self-information, (4) can be proved by *Property 7*.

Property 7 Assume $A_1 \subseteq A_2 \subseteq CA$ and $D_k \in U/D$, then $I_{A_1}^3(D_k) \geq I_{A_2}^3(D_k)$.

Proof The detailed proof can be found in the supplementary file.

Definition 15 Suppose that $INDS = \langle U, CA, D, \lambda \rangle$, $U/D = \{D_1, D_2, \dots, D_t\}$ and , then the sharp-blunt decision self-information of $INDS$ is defined as

$$I_A^3(D) = \sum_{k=1}^t I_A^3(D_k) \quad (23)$$

Definition 16 Let $A \subseteq CA$ and $D_k \in U/D$, the precision and roughness of the lenient decision index are defined as

$$\theta_A^{(3)}(D_k) = \frac{shar_A(D_k)}{blun_A(D_k)} \quad (24)$$

$$\sigma_A^{(3)}(D_k) = 1 - \frac{shar_A(D_k)}{blun_A(D_k)} = 1 - \theta_A^{(3)}(D_k). \quad (25)$$

Apparently, $0 \leq \theta_A^{(3)}(D_k), \sigma_A^{(3)}(D_k) \leq 1$. $\theta_A^{(3)}(D_k)$ portrays the proportion of the cardinal number of sharp and blunt decision-making samples. In other words, $\theta_A^{(3)}(D_k)$ characterizes the classification ability of feature subset A according to compare sharp decision index with blunt decision index.

When $\theta_A^{(3)}(D_k) = 1$, $\sigma_A^{(3)}(D_r) = 0$. It is the ideal state of the feature subset. At this time, the feature subset A has the strongest classification ability. On the contrary, when $\theta_A^{(3)}(D_r) = 0$, the feature subset A has no effect on classification and has the weakest classification ability.

Property 8 Let $A_1 \subseteq A_2 \subseteq CA$ and $D_k \in U/D$, then

- (1) $\theta_{A1}^{(3)}(D_k) \leq \theta_{A2}^{(3)}(D_k)$, $\sigma_{A1}^{(3)}(D_k) \geq \sigma_{A2}^{(3)}(D_k)$.
 (2) $\theta_A^{(3)}(D_k) = \theta_A^{(1)}(D_k) \cdot \theta_A^{(2)}(D_k)$.
 (3) $\sigma_A^{(3)}(D_k) = \sigma_A^{(1)}(D_k) + \sigma_A^{(2)}(D_k) - \sigma_A^{(1)}(D_k) \cdot \sigma_A^{(2)}(D_k)$.

Proof The detailed proof can be found in the supplementary file.

Definition 17 Assume $A \subseteq CA$ and $D_k \in U/D$, the lenient self-information of D_k can be defined as

$$I_A^4(D_k) = -\sigma_A^{(3)}(D_k) \ln \theta_A^{(3)}(D_k). \quad (26)$$

It is obvious to testify that $I_A^4(D_k)$ meets properties (1), (2), and (3) of the definition of self-information, and (4) can be confirmed according to Property 9.

Property 9 Letting $A1 \subseteq A2 \subseteq CA$ with $D_k \in U/D$, then $I_{A1}^4(D_k) \geq I_{A2}^4(D_k)$.

Proof The detailed proof can be found in the supplementary file.

Property 10 $I_A^4(D_k) \geq I_A^3(D_k)$.

Proof The detailed proof can be found in the supplementary file.

Definition 18 Suppose an incomplete neighborhood decision system $INDS = \langle U, CA, D, \lambda \rangle$ with $U/D = \{D_1, D_2, \dots, D_t\}$, and $A \subseteq CA$, the lenient self-information of $INDS$ is defined as

$$I_A^4(D) = \sum_{k=1}^t I_A^4(D_k). \quad (27)$$

Property 11 Let $A1 \subseteq A2 \subseteq CA$, then $I_{A1}^4(D) \geq I_{A2}^4(D)$.

Proof The detailed proof can be found in the supplementary file.

Remark 1 Through the above theoretical analysis, we receive that the lenient neighborhood multi-granulation self-information (NMSI) not only considers the upper and lower approximations of decision-making, but also can measure the uncertainty of the incomplete neighborhood decision-making system from a more comprehensive perspective. In addition, NMSI is more sensitive for the change of feature subset, and hence, it is more suitable for feature selection.

Definition 19 [45] Assume an incomplete neighborhood decision system $INDS = \langle U, CA, D, \lambda \rangle$, $A \subseteq CA$, $A = \{A_1, A_2, \dots, A_r\}$, $A_i \subseteq A$, $NT_{A_i}^\lambda(x)$ is a neighborhood tolerance class for $x \in U$, the neighborhood tolerance entropy of A is defined as

$$NTE_\lambda(A) = -\frac{1}{|U|} \sum_{i=1}^r \log_2 \frac{|NT_{A_i}^\lambda(x)|}{|U|}. \quad (28)$$

Definition 20 Letting $INDS = \langle U, CA, D, \lambda \rangle$, $U/D = \{D_1, D_2, \dots, D_t\}$, $A = \{A_1, A_2, \dots, A_r\}$, $A \subseteq CA$, $NT_{A_i}^\lambda(x)$ is a neighborhood tolerance class. Then, the neighborhood tolerance joint entropy of A and D is defined as

$$NTE_\lambda(A \cup D) = -\frac{1}{|U|} \sum_{i=1}^r \sum_{k=1}^t \frac{|NT_{A_i}^\lambda(x) \cap D_k|}{|U|} \log_2 \frac{|NT_{A_i}^\lambda(x) \cap D_k|}{|U|}. \quad (29)$$

Definition 21 Assume an incomplete neighborhood decision system $INDS = \langle U, CA, D, \lambda \rangle$, $A = \{A_1, A_2, \dots, A_r\}$, $U/D = \{D_1, D_2, \dots, D_t\}$, $NT_{A_i}^\lambda(x)$ is a neighborhood tolerance class for $x \in U$, the neighborhood multi-granulation self-information-based pessimistic neighborhood multi-granulation tolerance joint entropy (PTSIJE) of A and D is defined as

$$PTSIJE_\lambda^p(A \cup D) = -I_B^4(D) \times \frac{1}{|U|} \sum_{i=1}^r \sum_{k=1}^t \frac{|NT_{A_i}^\lambda(x) \cap D_k|}{|U|} \log_2 \frac{|NT_{A_i}^\lambda(x) \cap D_k|}{|U|}. \quad (30)$$

Here, $I_B^4(D)$ is the lenient NMSI measure of $INDS$.

Property 12 Letting $INDS = \langle U, CA, D, \lambda \rangle$, $U/D = \{D_1, D_2, \dots, D_t\}$, $A \subseteq CA$, $A = \{A_1, A_2, \dots, A_r\}$, $NT_{A_i}^\lambda(x)$ is a neighborhood tolerance class for $x \in U$, then

$$PTSIJE_\lambda^p(A \cup D) = I_A^4(D) \times NTE_\lambda(A \cup D) \geq 0.$$

Remark 2 From Definition 20 and Property 12, we can vividly realize that $I_A^4(D)$ represents the lenient NMSI measure in algebraic view, and $NTE_\lambda(A \cup D)$ is the neighborhood tolerance joint entropy of A and D in information view. Therefore, PTSIJE can measure the uncertainty of incomplete neighborhood decision systems from both algebraic and information views based on self-information measures and entropy.

PTSIJE-based feature selection method in incomplete neighborhood decision systems

Feature selection based on PTSIJE

Definition 22 Letting $INDS = \langle U, CA, D, \lambda \rangle$, $A = \{A_1, A_2, \dots, A_r\}$, $A' \subseteq A \subseteq CA$, then A' is the reduction of A and D if:

$$(1) \quad PTSIJE_\lambda^p(A', D) = PTSIJE_\lambda^p(CA, D).$$

Algorithm 1 PTSIJE-FS

Require: An incomplete neighborhood decision system $INDS = \langle U, CA, D, \lambda \rangle$ with a neighborhood parameter λ and $A = \{A_1, A_2, \dots, A_r\}$.

Ensure: An optimal feature subset red .

- 1: Initialize: $red = \emptyset, B = A - red, start=1$;
- 2: **for** each D_k in U/D **do**
- 3: Compute sharp decision index $shar_A(D_k)$
- 4: Compute blunt decision index $blun_A(D_k)$
- 5: Compute presion function $\theta_A(D_k)$
- 6: Compute $I_A^4(D_k) = -(1 - \theta_A(D_k)) \ln \theta_A(D_k)$
- 7: **end for**
- 8: **while** $PTSIJE_\lambda^p(red, D) \neq PTSIJE_\lambda^p(A, D)$ **do**
- 9: **for** $j = 1$ to $|B|$ **do**
- 10: Compute $PTSIJE_\lambda^p(red \cup a_j, D)$
- 11: Select a_j that satisfies $\max \{a_j \mid PTSIJE_\lambda^p(red \cup a_j, D)\}$
- 12: **end for**
- 13: Let $red = red \cup a_j, B = A - red$, and compute $PTSIJE_\lambda^p(red, D)$
- 14: **end while**
- 15: **for** $k = 1$ to red **do**
- 16: Select $B_k \in red$
- 17: Compute $SIG(B_k, red, D)$
- 18: **if** $SIG(a_k, red, D) > 0$ **then**
- 19: Let $red = red - \{a_k\}$
- 20: **end if**
- 21: **end for**
- 22: **return** red

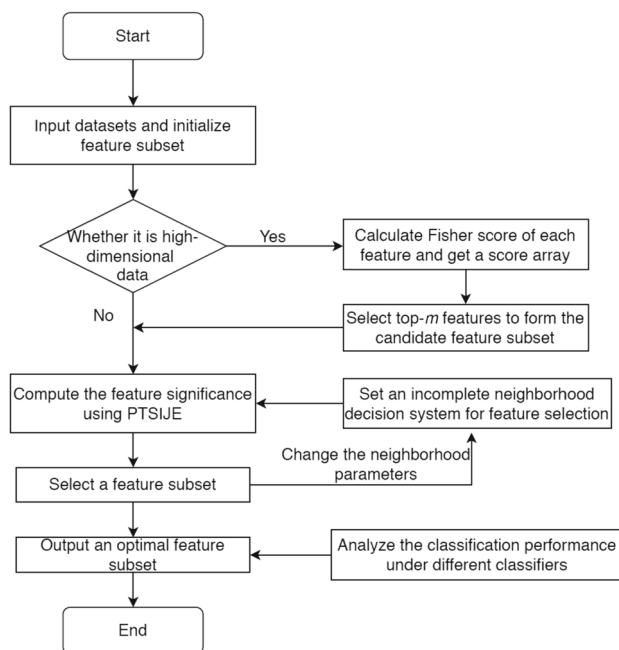


Fig. 1 Process flow of the feature selection method for data classification

- (2) For any $A_i \subseteq A'$, $PTSIJE_\lambda^p(A', D) < PTSIJE_\lambda^p(A' - A_i, D)$.

Here, the formula (1) illustrates that the reduced subset and the entire dataset have the same classification ability, and the formula (2) guarantees that the reduced subset has no redundant attributes.

Definition 23 Let $INDS = \langle U, CA, D, \lambda \rangle$, $A = \{A_1, A_2, \dots, A_r\}$, $A' \subseteq A \subseteq CA$, for any $A_i \subseteq A'$, $i = 1, 2, \dots, r$, the attribute significance of attribute subset A' with respect to D in granularity set A_i is defined as

$$SIG(A_i, A', D) = PTSIJE_\lambda^p(A' - A_i, D) - PTSIJE_\lambda^p(A', D) \quad (31)$$

Feature selection algorithm

To show the feature selection method more clearly, the process of data classification for feature selection is expressed in Fig. 1, and the algorithm description is shown in Algorithm 1.

For the PTSIJE-FS algorithm, it mainly involves two aspects: getting neighborhood tolerance classes and counting PTSIJE. The calculation of neighborhood tolerance class has a great impact on the time complexity. To reduce the computational complexity of the neighborhood tolerance classes, the bucket sorting algorithm [37] is used here, and the time of complexity of neighborhood tolerance classes is cut back to $O(mn)$. Here, m is the number of samples and n is the number of features. Meanwhile, the computational time complexity of the PTSIJE is $O(n)$. The PTSIJE-FS algorithm is a loop in steps 8–14. In the worst case, its time complexity is $O(n^3m)$. Assuming that the number of selected granularities is n_R , since only candidate granularities need to be considered but not all granularities, the time complexity of calculating the neighborhood tolerance classes is $O(n_Rm)$. In most cases $n_R \ll n$, so the time complexity of PTSIJE-FS is about $O(mn)$.

Experimental results and analysis

Experiment preparation

To demonstrate the effectiveness and robustness of our proposed feature selection method, we conducted experiments on 11 public datasets, including 6 UCI datasets and 5 high-dimensional microarray gene expression data-sets. The datasets used are listed in Table 1. Among them, six UCI datasets can be downloaded from <http://archive.ics.uci.edu/ml/datasets.php>, and five gene expression datasets can be downloaded from <http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi>. It should be noted here that the Wine, Wdbc, Sonar, Heart datasets and five gene expression datasets are usually complete. Therefore, to convenience, we

Table 2 Eleven datasets in experiments

No.	Datasets	Samples	Features	Classes
1	Credit [60]	690	15	9
2	Heart [60]	270	13	2
3	Sonar [69]	208	60	2
4	Wdbc [70]	569	31	2
5	Wine [71]	178	13	3
6	Wdbc [72]	198	34	2
7	DLBCL [60]	77	5469	2
8	Leukemia [45]	72	7129	2
9	Lung [45]	181	12533	2
10	MLL [60]	72	12582	3
11	Prostate [73]	136	12600	2

randomly modify some known feature values to missing values to achieve incomplete neighborhood decision systems.

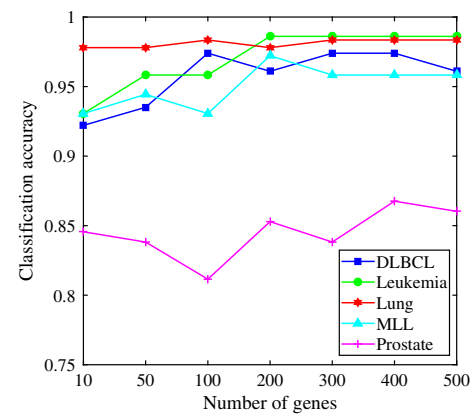
Table 2 lists all datasets.

All simulation experiments are running on MATLAB 2016a, under Window 10 operating system with an Intel(R) i5 CPU at 3.20 GH and 4.0 GB RAM. The classification accuracy is verified by three classifiers KNN, CART, and C4.5 whose default values of all parameters are selected under Weka 3.8 software. To ensure the consistency of experiments, we use the tenfold cross-validation method in all our experiments.

Effect of different neighborhood parameters

This subsection will focus on the analysis of the impact of different neighborhood parameters λ on the classification performance of our method, and find the best parameters for each dataset. For five high-dimensional gene expression datasets (hereinafter referred to as gene datasets), to effectively reduce the time cost, we use Fisher score (FS) [45] for preliminary dimensionality reduction. There are some advances of FS: less calculation, strong operability, and can effectively reduce the computational complexity. Figure 2 illustrates the variation trend of the accuracy of the number of selected genes under the classifier KNN in different dimensions (10, 50, 100, 200, 300, 400, 500) on the five gene datasets.

As shown in Fig. 2, the accuracy in most cases changes with the size of the gene subset. The optimal balance between the size and accuracy of the selected feature gene subset is found to obtain the appropriate dimension of each gene subset for the next feature selection. Therefore, the values of two datasets DLBCL and Lung are set as 100 dimensions, the 200 dimensions are favorable for datasets Leukemia and MLL, and the 400 dimensions are appropriate for dataset Prostate.

**Fig. 2** Classification accuracy via the size of gene subsets on five gene datasets

The classification accuracy of selected feature subsets using PTSIJE-FS on 11 datasets is obtained with different neighborhood parameters. For six UCI datasets, the classification performance is evaluated under two classifiers KNN ($k=10$) and CART, which is displayed in Fig. 3a–f, and the classification performance for five gene datasets under three classifiers KNN ($k=10$), CART, and C4.5 is illustrated in Fig. 3g–k. The classification accuracy under different parameters λ is shown in Fig. 3, where the abscissa represents different neighborhood parameters $\lambda \in [0.05, 1]$, and the ordinate is the classification accuracy.

Figure 3 reveals that the neighborhood parameters increase from 0.05 to 1, and the classification accuracy of features selected by PTSIJE-FS is changing. Different parameters have a certain impact on the classification performance of PTSIJE-FS. Fortunately, every dataset can reach the high result in the wider range of λ . Figure 3a displays that for Credit dataset, the neighborhood parameter will be 0.4 under KNN and 0.1 under CART. From Fig. 3b, for Heart dataset, the neighborhood parameter set 1.0 under classifier KNN and 0.75 under classifier CART, the classification accuracy achieves maximum. In Fig. 3c, for Sonar dataset, the classification performance is at best level when neighborhood parameters are 0.5 and 0.6 under classifiers KNN and CART, respectively. As shown in Fig. 3d, when the neighborhood parameter is set to 0.15 under classifiers KNN and CART, the classification performance is optimal. It can be seen from Fig. 3e that for the Wine dataset, the classification performance is best when the neighborhood parameter is set to 0.8 under the classifier KNN, and the neighborhood parameter is set to 0.2 under the classifier CART. From Fig. 3f that on the Wdbc dataset, when the neighborhood parameters are set to 0.15 and 0.3, the classification performance of the selected feature subset under the classifiers KNN and CART reaches the high level at the same time. Figure 3g shows the accuracy of DLBCL dataset and the neighborhood parameters will be set 0.45, 1.0, and 0.4 under three classifiers

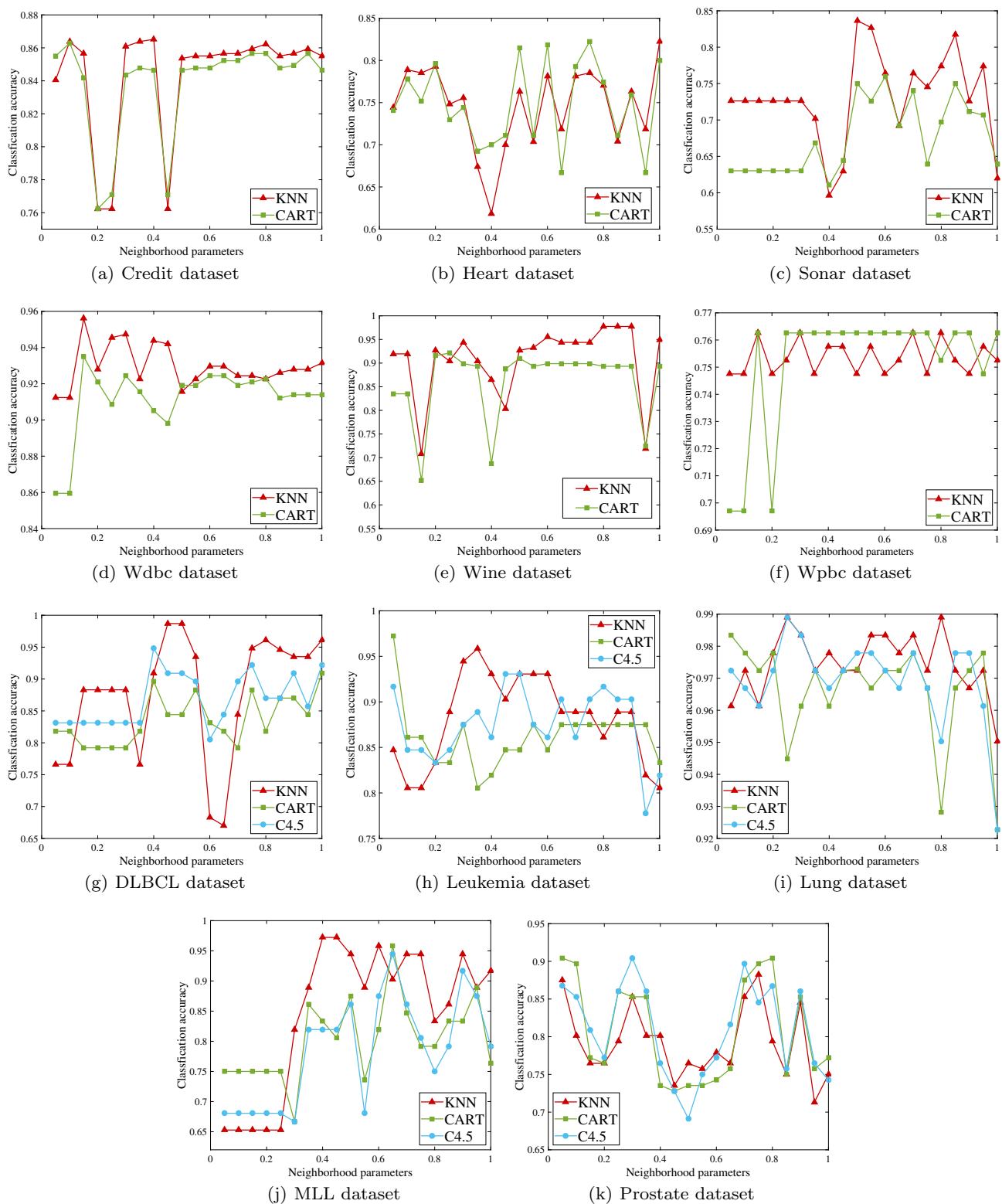


Fig. 3 Classification accuracy for 11 datasets with different neighborhood parameter values

KNN, CART, and C4.5, respectively. In Fig. 3h, for the gene dataset Leukemia, when the neighborhood parameter is set to 0.35 under KNN, 0.05 under CART, and 0.45 under C4.5,

the gene subset with the best classification performance can be obtained. As shown in Fig. 3i, when the neighborhood parameter under KNN is set to 0.8, CART is set to 0.05,

Table 3 The number of selected features with the seven methods on the six UCI datasets

Datasets	Original	FSNTDJE	DMRA	FPRA	FRFS	IFGAS	PDJE-AR	PTSIJE-FS
Credit	15	9.5	10.2	9.6	10.0	8.4	9.1	7.2
Heart	13	10.9	11.8	11.4	10.7	8.4	9.6	9.4
Sonar	60	19.7	35.3	17.1	19.5	13.0	18.4	17.7
Wdbc	31	16.9	27.9	20.8	24.4	17.1	15.7	18.4
Wine	13	8.8	12.1	10.9	11.4	6.2	5.7	5.5
Wdbc	34	7.6	25.5	19.0	22.1	10.4	12.5	5.0
Mean	27.67	12.23	20.47	14.8	16.35	10.58	11.83	10.53

Table 4 Optimal feature subset with PTSIJE-FS for six UCI datasets

Datasets	KNN	CART
Credit	{3 5 8 9 11 13 14 15}	{4 5 9 11 15}
Heart	{1 3 4 5 8 10 11 12}	{1 2 3 4 6 7 9 11 12 13}
Sonar	{3 10 11 16 27 37 46 47 60}	{3 11 12 13 14 16 18 24 27 32 35 38 45 46 48 49 52 55}
Wdbc	{1 3 5 8 9 12 14 15 16 17 18 19 20 21 25 27 28 30}	{1 3 5 8 9 12 14 15 16 17 18 19 20 21 25 27 28 30}
Wine	{1 7 10 11 13}	{1 2 7 9 11}
Wdbc	{15 16 17 20 23 25 26}	{17 18}

and C4.5 is set to 0.7, the classification performance of the gene subset selected from Lung dataset achieves the best level. Figure 3j demonstrates that the optimal neighborhood parameter values should be set to 0.4 under KNN, and set to 0.65 under classifiers CART and C4.5 for MLL. Similarly, for the Prostate dataset, the parameter values are 0.75, 0.8, and 0.3 under classifiers KNN, CART, and C4.5, respectively.

Classification results of the UCI datasets

In this subsection, to illustrate the classification performance of the method PTSIJE-FS on the low-dimensional UCI datasets, the PTSIJE-FS is compared with six existing feature selection methods: (1) the neighborhood tolerance dependency joint entropy-based feature selection method (FSNTDJE) [45], (2) the discernibility matrix-based reduction algorithm (DMRA) [27], (3) the fuzzy positive region-based accelerator algorithm (FPRA) [61], (4) the fuzzy boundary region-based feature selection algorithm (FRFS) [62], (5) the intuitionistic fuzzy granule-based attribute selection algorithm (IFGAS) [63], and (6) the pessimistic neighborhood multi-granulation dependency joint entropy (PDJE-AR) [60].

The first part of this subsection focuses on the size of feature subsets selected by all comparative feature selection methods. Using neighborhood parameters λ , the tenfold cross-validation method is used on six UCI datasets. The original data and average size of the selected feature subset by the above seven feature selection methods are shown in Table 3. Under two classifiers KNN and CART, PTSIJE-FS selects the optimal feature subset for six UCI datasets as dis-

played in Table 4 where “original” represents the original dataset.

Table 3 shows the average number of the selected feature subset by seven feature selection methods. Compared with method IFGAS, PTSIJE-FS selects more features for datasets Heart and Sonar. The average size of the selected feature subsets by PTSIJE-FS is 7.2, 5.5, and 5.0, respectively, and reaches the minimum on datasets Credit, Wine, and Wdbc. In a word, the mean number of the selected feature subset by PTSIJE-FS is the minimum compare with six related methods for six UCI datasets.

The second part in this subsection is to exhibit the classification effectiveness for PTSIJE-FS. There are six feature selection methods, including FSNTDJE, DMRA, FPRA, FRFS, IFGAS, and PDJE-AR, are used to prove the accuracy on the selected feature subsets under two classifiers KNN and CART for six UCI datasets. Tables 5 and 6, respectively, list the average classification accuracy using the seven methods of selected feature subsets under two classifiers KNN and CART. In Tables 3, 7, the bold font shows that the size of the reduced datasets is least with respect to other methods. In Tables 5, 6, 9, 10, the bold numbers indicate that the classification accuracy of the selected feature subsets is the highest with respect to other methods.

Combined with Tables 3 and 5, we can clearly see the differences between the seven methods. For almost all UCI datasets, the classification accuracy of the selected feature subsets by PTSIJE-FS apparently outperforms the other six methods under the KNN classifier. In addition, PTSIJE-FS not only selects the fewer features, but also has the highest classification accuracy on datasets Credit, Wine, and Wdbc.

Table 5 Classification accuracy of the seven methods under the KNN classifier

Datasets	Original	FSNTDJE	DMRA	FPRA	FRFS	IFGAS	PDJE-AR	PTSIJE-FS
Credit	0.7046	0.7046	0.7927	0.8023	0.7827	0.8130	0.8551	0.8652
Heart	0.7652	0.7741	0.7697	0.7444	0.7545	0.7889	0.8037	0.8222
Sonar	0.7262	0.6975	0.7214	0.7407	0.7357	0.7055	0.7240	0.8365
Wdbc	0.9124	0.9466	0.9404	0.9337	0.9432	0.9311	0.9561	0.9561
Wine	0.9195	0.9057	0.9108	0.9474	0.9330	0.9429	0.9494	0.9775
Wpbc	0.7475	0.7374	0.7021	0.7471	0.7368	0.6840	0.7576	0.7626
Mean	0.7959	0.7943	0.8062	0.8193	0.8143	0.8109	0.8410	0.8700

Table 6 Classification accuracy of the seven methods under the CART classifier

Datasets	Original	FSNTDJE	DMRA	FPRA	FRFS	IFGAS	PDJE-AR	PTSIJE-FS
Credit	0.7766	0.7766	0.8724	0.8173	0.8245	0.8317	0.8420	0.8629
Heart	0.7852	0.7653	0.7444	0.7407	0.7370	0.7481	0.8000	0.8222
Sonar	0.6302	0.6707	0.7362	0.7224	0.7260	0.6871	0.6904	0.7596
Wdbc	0.8595	0.9124	0.9208	0.9175	0.8990	0.9296	0.9402	0.9349
Wine	0.8351	0.8027	0.8492	0.8592	0.8544	0.8727	0.9101	0.9157
Wpbc	0.6970	0.6869	0.6241	0.6439	0.6818	0.6658	0.7071	0.7626
Mean	0.7639	0.7691	0.7912	0.7835	0.7871	0.7892	0.815	0.8430

In brief, the PTSIJE-FS method deletes redundant features to the greatest extent and still shows better classification performance than six compared methods on UCI datasets.

Similarly, from Tables 3 and 6, the differences among seven feature selection methods are illustrated under the classifier CART. The average accuracy of PTSIJE-FS is larger than other six comparison methods on the datasets Heart, Sonar, Wine, and Wpbc, which are 82.22%, 75.96%, 91.57%, and 76.26%, respectively. Although the average accuracy of the feature subset selected by PTSIJE-FS on the dataset Credit is 0.95% lower than that of the feature subset selected by DMRA, PTSIJE-FS selects the fewer features for the dataset Credit.

In terms of time complexity, the time complexity of DMRA and IFGAS is $O(m^2n)$ [27,63], the time complexity of FPRA is $O(m\log n)$ [61], the time complexity of FRFS is $O(n^2)$ [62], and the time complexity of PDJE-AR and FSNTDJE is $O(mn)$ [45,60]. Therefore, the rough ranking of the seven methods on times complexity is as follows: $O(FRFS) < O(FPRA) < O(PDJE - AR) = O(FSNTDJE) = O(PTSIJE - FS) < O(DMRA) = O(IFGAS)$.

Under different classifiers and learning tasks, no one method always performs better than other methods. PTSIJE-FS shows superior classification performance and stability under the classifiers KNN and CART.

In summary, PTSIJE-FS can eliminate redundant features as a whole, and shows outstanding classification performance on UCI datasets.

Classification results of the gene expression datasets

This subsection illustrates the classification performance of the method PTSIJE-FS on high-dimensional gene expression datasets. PTSIJE-FS is compared with six existing feature selection methods: (1) the neighborhood tolerance dependency joint entropy-based feature selection method (FSNTDJE) [45], (2) the mutual information-based attribute reduction algorithm for knowledge reduction (MIBARK) [64], (3) the decision neighborhood entropy-based heuristic attribute reduction algorithm (DNEAR) [21], (4) the entropy gain-based gene selection algorithm (EGGS) [65], (5) EGGS [65] algorithm combined with Fisher Score (EGGS-FS) [66], and (6) the pessimistic neighborhood multi-granulation dependency joint entropy (PDJE-AR) [60].

The performance of seven feature selection methods is verified on the gene datasets DLBCL, Leukemia, Lung, MLL, and Prostate under tenfold cross-validation. Table 7 lists the average size of gene subsets selected by each feature selection method, where “Original” represents the original dataset and “-” denotes that we have not obtained the result of this method. Under the three classifiers KNN, CART, and C4.5, The optimal gene subsets selected by PTSIJE-FS method for five gene datasets under three classifiers KNN, CART, and C4.5 are demonstrated in Table 8.

Table 7 reveals the average size of gene subsets selected by seven feature selection methods. PTSIJE-FS do select more genes for five gene datasets, and the average size of the gene subset selected by PTSIJE-FS is smaller than EGGS.

In the next part, according to the results of Tables 7 and 8, PTSIJE-FS and related six feature selection methods are used

Table 7 Optimal gene subset with PSINTE-FS for five gene datasets

Datasets	Original	FSNTDJE	MIBARK	DNEAR	EGGS	EGGS-FS	PDJE-AR	PTSIJE-FS
DLBCL	5469	6.4	2.8	10	19.5	3.7	6	7.4
Leukemia	7129	7.0	4.4	7.8	8.1	5.0	7.6	7.6
Lung	12533	9.0	5.9	6.0	11.4	6.4	8	11.0
MLL	12582	7.0	–	–	–	–	7.8	14.8
Prostate	12600	8.0	4.4	5.1	7.7	14.0	4.0	5.5
Mean	10062.6	7.48	4.38	7.23	11.68	7.28	6.68	9.26

Table 8 Optimal feature subset with PTSIJE-FS for five gene datasets

DLBCL	{1071 3942 689 856 584 874 952 930 5184}	{3942 689 3724 4552}	{4767 52 584 823 1640 952}
Leukemia	{1745 2020 2111 6005 461 4328 4377 6281 2354 1260 }	{2288 4196 2288}	{461 4328 4377 6281 2354 1260}
Lung	{2882 3202 7249 11957 12298 4815 1673 7625 6405 8564 2421}	{7200 111958 8564 2421}	{2882 3202 7249 11957 12298 4815 1673 7625 6405 8564 2421}
MLL	{12418 8518 8428 3768 11229 4341 1316 9929 9845 3675 3882 9084 10856}	{12418 8518 3634 11297 3768 9586 9882 6278 11718 9005 8212 3316 9741 10457 8165 11325 11603}	{12418 8518 3634 11297 3768 9586 9882 6278 11718 9005 8212 3316 9741 10457 8165 11325 11603}
Prostate	{6185 8058}	{6185 5486 8058}	{6185 8958 5979 3333 1322 5486 9230 8236 9357}

to verify the average classification accuracy of gene subsets under the two classifiers KNN and C4.5, and the results are exhibited in Tables 9 and 10.

From Tables 9 and 10, PTSIJE-FS displays the highest classification accuracy, except for dataset Prostate under the C4.5 classifier. Especially on the datasets leukemia and MLL, the average classification accuracy of the gene subset selected by PTSIJE-FS is significantly higher than the other six feature selection methods under classifier KNN, and the average accuracy is increased by about 8%-42% and 10%-32%, respectively. As we have seen, the gene subsets selected by PTSIJE-FS show better classification performance on the gene datasets DLBCL, Leukemia, Lung, and MLL under classifier C4.5. There are 4%-39% gaps which obviously exist between PTSIJE-FS and FSNTDJE, MIBARK, DNEAR, EGGS, and EGGS-FS, only lower 0.76% than PDJE-AR. In general, under the classifiers KNN and C4.5, the mean classification accuracy of PTSIJE-FS is higher than the other six feature selection methods, and reach the highest level for five gene datasets.

In terms of time complexity, the time complexity of FSNTDJE, DNEAR, and PDJE-AR is $O(mn)$ [21,45,60], the time complexity of MIBARK is $O(m^2)$ [64], and the time complexity of EGG and EGG-FS is $O(m^3n)$ [65]. Therefore, the rough ranking of the seven methods on times complexity is as follows: $O(MIBARK) < O(PDJE-AR) = O(DENAR) = O(FSNTDJE) = O(PTSIJE-FS) < O(EGG) = O(EGG-FS)$.

In summary, PTSIJE-FS can eliminate redundant features as a whole on five gene datasets, and shows better classification performance against the six related methods.

Statistical analysis

To systematically compare the statistical performance of classification accuracy of all methods, Friedman test and corresponding post-test will be employed in this subsection. The Friedman statistic [67] is expressed as

$$\chi^2 = \frac{12n}{k(k+1)} \sum_{i=1}^k r_i^2 - \frac{k(k+1)^2}{4} \quad (32)$$

Table 9 Classification accuracy of the seven methods under the KNN classifier

Datasets	Original	FSNTDJE	MIBARK	DNEAR	EGGS	EGGS-FS	PDJE-AR	PTSIJE-FS
DLBCL	0.896	0.805	0.765	0.698	0.854	0.87	0.935	0.9870
Leukemia	0.734	0.833	0.828	0.533	0.629	0.801	0.875	0.9583
Lung	0.931	0.892	0.958	0.82	0.859	0.979	0.941	0.9890
MLL	0.6528	0.9167	–	–	–	–	0.8722	0.9722
Prostate	0.782	0.787	0.512	0.611	0.639	0.849	0.868	0.8824
Mean	0.7992	0.8467	0.7658	0.6655	0.7453	0.8748	0.8982	0.9578

Table 10 Classification accuracy of the seven methods under the C4.5 classifier

Datasets	Original	FSNTDJE	MIBARK	DNEAR	EGGS	EGGS-FS	PDJE-AR	PTSIJE-FS
DLBCL	0.809	0.831	0.778	0.718	0.826	0.801	0.877	0.9481
Leukemia	0.814	0.917	0.834	0.671	0.733	0.813	0.806	0.9306
Lung	0.926	0.916	0.964	0.819	0.966	0.955	0.951	0.9890
MLL	0.7500	0.8889	–	–	–	–	0.9028	0.9444
Prostate	0.640	0.772	0.566	0.570	0.591	0.863	0.912	0.9044
Mean	0.7878	0.8650	0.7855	0.6945	0.7790	0.8580	0.8898	0.9433

$$F = \frac{(n-1)\chi^2}{n(k-1) - \chi^2}. \quad (33)$$

rankings, which is defined as

$$CD_\alpha = q_\alpha \sqrt{\frac{k(k+1)}{6n}}, \quad (34)$$

Here, r_i is the mean rank, and n and k represent the number of datasets and methods, respectively. F obeys the distribution with $(k-1)$ and $(k-1)(n-1)$ degrees of freedom. For the six low-dimensional UCI datasets in Tables 5 and 6, PTSIJE-FS, FSNTDJE, DMRA, FPRA, FRFS, IFGAS, and PDJE-AR are employed to conduct the Friedman statistic. According to the classification accuracy obtained in Tables 5 and 6, the rankings of the seven feature selection methods under the classifiers KNN and CART are shown in Tables 11 and 12.

Calling *icdf* calculation in MATLAB 2016a, when $\alpha=0.1$, $F(6,30)=1.9803$. Assuming that these seven methods are equivalent in classification performance, then the value of Friedman statistics will not exceed the critical value $F(6,30)$. Otherwise, it is said that these seven methods have significant differences in feature selection performance. According to Friedman statistics, $F=7.4605$ for the classifier KNN and $F=5.3738$ for the classifier CART. Obviously, the F values under the classifiers KNN and CART are far greater than the critical value $F(6,30)$, indicating that these seven methods are significantly different under the classifiers KNN and CART on six UCI datasets.

Subsequently, we need to perform post-testing on the differences between the seven methods. The post-hoc test used here is the Nemenyi test [68]. The statistics needs to determine the critical value of the mean distance between

where q_α means a critical value. It is can be obtained that $q_{0.1}=2.693$ when the methods number is 7 and $\alpha=0.1$. Then, from formula (33), we can know that $CD_{0.1}=3.3588$ ($k=7$, $n=6$). According to Table 10, the distance between mean rankings of PTSIJE-FS to FSNTDJE, DMRA, FRFS, and IFGAS are 4.25, 4.25, 3.75, and 3.92, respectively, which are greater than 3.3588 for classifier KNN. In this case, Nemenyi test reveals that PSITE-FS is far superior to FSNTDJE, DMRA, FRFS, and IFGAS under the classifier KNN at $\alpha=0.1$. In the same way, from Table 11, the distances between mean rankings of PTSIJE-FS to FSNTDJE, FPRA, FRFS, and FSNTDJE are 4.17, 3.84, and 3.84, respectively, indicating that PSINTE-FS is significantly better than FSNTDJE, FPRA, and FRFS.

In the second part of this subsection, Friedman test is performed on the classification accuracy of the seven feature selection methods in Tables 9 and 10 under the classifiers KNN and C4.5. Tables 13 and 14, respectively, list the mean rankings of the seven methods under the classifiers KNN and C4.5.

After calculation, when $\alpha=0.1$, the critical value $F(6,24)=2.0351$, $F=8.2908$ for the classifier KNN and $F=3.4692$ for the classifier C4.5. It is clear that above two values are both greater than the critical value $F(6,24)$. It explains that these seven methods have significant differences on the classifiers KNN and C4.5 on the five gene datasets.

Table 11 Rank of the seven methods with KNN

Datasets	FSNTDJE	DMRA	FPRA	FRFS	IFGAS	PDJE-AR	PTSIJE-FS
Credit	7	5	4	6	3	2	1
Heart	4	5	7	6	3	2	1
Sonar	7	5	2	3	6	4	1
Wdbc	3	5	6	4	7	1.5	1.5
Wine	7	6	3	5	4	2	1
Wpbc	4	6	3	5	7	2	1
Mean	5.33	5.33	4.17	4.83	5	2.25	1.08

Table 12 Rank of the seven methods with CART

Datasets	FSNTDJE	DMRA	FPRA	FRFS	IFGAS	PDJE-AR	PTSIJE-FS
Credit	7	1	6	5	4	3	2
Heart	3	5	6	7	4	2	1
Sonar	7	2	4	3	6	5	1
Wdbc	6	4	5	7	3	1	2
Wine	7	6	4	5	3	2	1
Wpbc	3	7	6	4	5	2	1
Mean	5.5	4.17	5.17	5.17	4.17	2.5	1.33

Table 13 Rank of the seven methods with KNN

Datasets	FSNTDJE	MIBARK	DNEAR	EGGS	EGGS-FS	PDJE-AR	PTSIJE-FS
DLBCL	5	6	7	4	3	2	1
Leukemia	3	4	7	6	5	2	1
Lung	5	3	7	6	2	4	1
MLL	2	–	–	–	–	3	1
Prostate	4	7	6	5	3	2	1
Mean	3.8	5	6.75	5.25	3.25	2.6	1

Table 14 Rank of the seven methods with C4.5

Datasets	FSNTDJE	MIBARK	DNEAR	EGGS	EGGS-FS	PDJE-AR	PTSIJE-FS
DLBCL	4	6	7	3	5	2	1
Leukemia	2	3	7	6	4	5	1
Lung	6	3	7	2	4	5	1
MLL	3	–	–	–	–	2	1
Prostate	4	7	6	5	3	1	2
Mean	3.8	4.75	6.75	4	4	3	1.2

Next, Nemenyi test is performed on the classification accuracy of the seven feature selection methods in Tables 9 and 10 under the classifiers KNN and C4.5. It can be computed that $CD_{0.1}=3.6793(k=7, n=5)$ when the number of methods is 7, and $q_{0.1}=2.693$. The distances between mean rankings of PTSIJE-FS to IBARK, DNEAR, EGGS under the classifier KNN are 4, 5.75, 4.25. This result demonstrates PTSIJE-FS is apparently better than IBARK, DNEAR, EGGS. The distance between mean rankings of PTSIJE-FS to DNEAR is 5.55 greater than the critical value,

showing that PTSIJE-FS is far better than DNEAR under the classifier C4.5.

In a word, the PTSIJE-FS method is superior than the other corresponding methods through the Friedman statistic test.

Conclusion

NMRS model is an effectively tool to improve the classification performance in incomplete neighborhood decision systems. However, most feature evaluation functions based on NMRS only consider the information contained in the

lower approximation. Such a construction method is likely to lead to the loss of some information. In fact, the upper approximation also contains some classification information that cannot be ignored. To solve this problem, we propose a feature selection model based on PTSIJE. First, from the algebra view, using the upper and lower approximations in NMRS and combining the concept of self-information, four types of neighborhood multi-granulation self-information measures are defined, and related properties are discussed in detail. The proof proves that the fourth neighborhood multi-granulation self-information measure is more sensitive and helps to select the optimal feature subset. Second, from the information views, the neighborhood tolerance joint entropy is given to analyze the redundancy and noise in the incomplete decision systems. Then, inspiring by both algebra and information view, combined with self-information measure and information entropy, a PTSIJE model is proposed to analyze the uncertainty of incomplete neighborhood decision systems. Finally, a heuristic forward feature selection method is designed and compared with other relevant methods. The experimental results show that our proposed method can select fewer features and have higher classification accuracy than related methods. In the future, we will not only focus on more efficient search strategies based on NMRS and self-information measures to achieve the best balance between classification accuracy and feature subset size, but also on constructing more general feature selection methods.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40747-022-00882-8>.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 61976082, 62076089, 62002103, and in part by the Key Scientific Research Project of Henan Provincial Higher Education under Grant 22B520013, and in part by the Key Scientific and Technological Project of Henan Province under Grant 222102210169.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Asunción J, Juan MM, Salvador P (2021) A novel embedded min-max approach for feature selection in nonlinear Support Vector Machine classification. *Eur J Oper Res* 293(1):24–35
- Miao JY, Ping Y, Chen ZS, Jin XB, Li PJ, Niu LF (2021) Unsupervised feature selection by non-convex regularized self-representation. *Expert Syst Appl*. <https://doi.org/10.1016/j.eswa.2021.114643>
- Lang GM, Li QG, Cai MJ, Yang T, Xiao QM (2017) Incremental approaches to knowledge reduction based on characteristic matrices. *Int J Mach Learn Cybern* 8:203–222
- Zhang XH, Bo CX, Smarandache F, Dai J (2018) New inclusion relation of neutrosophic sets with application and related lattice structure. *Int J Mach Learn Cybern* 9:1783–1763
- Xu JC, Shen KL, Sun L (2022) Multi-Label feature selection based on fuzzy neighborhood rough sets. *Complex Intell Syst*. <https://doi.org/10.1007/s40747-021-00636-y>
- Gao C, Lai ZH, Zhou J, Zhao CR, Miao DQ (2018) Maximum decision entropy-based attribute reduction in decision-theoretic rough set model. *Knowl-Based Syst* 143:179–191
- Huang QQ, Li T, Huang YY, Yang X (2020) Incremental three-way neighborhood approach for dynamic incomplete hybrid data. *Inf Sci* 541:98–122
- Shen HT, Zhu Y, Zheng W, Zhu X (2020) Half-Quadratic minimization for unsupervised feature selection on incomplete data. *IEEE Trans Neural Netw Learn Syst*. <https://doi.org/10.1109/TNNLS.2020.3009632>
- Xie XJ, Qin XL (2017) A novel incremental attribute reduction approach for dynamic incomplete decision systems. *Int J Approx Reason* 93:443–462
- Dong HB, Li T, Ding R, Sun J (2018) A novel hybrid genetic algorithm with granular information for feature selection and optimization. *Appl Soft Comput* 65:33–46
- Thabtah F, Kamalov F, Hammoud S, Shahamiri SR (2020) Least Loss: A simplified filter method for feature selection. *Inf Sci* 534:1–15
- Zhang CC, Dai JH, Chen JL (2020) Knowledge granularity based incremental attribute reduction for incomplete decision systems. *Int J Mach Learn Cybern* 11:1141–1157
- Fu J, Dong J, Zhao F (2020) A deep learning reconstruction framework for differential phase-contrast computed tomography with incomplete data. *IEEE Trans Image Process* 29:2190–2202
- Tran CT, Zhang MJ, Andreae P, Xue B, Bui LT (2018) Improving performance of classification on incomplete data using feature selection and clustering. *Appl Soft Comput* 73:848–861
- Yang W, Shi Y, Gao Y, Wang L, Yang M (2018) Incomplete-data oriented multiview dimension reduction via sparse low-rank representation. *IEEE Transactions on Neural Networks and Learning Systems* 29(12):6276–6291
- Ding WP, Lin CT, Cao ZH (2019) Deep neuro-cognitive co-evolution for fuzzy attribute reduction by quantum leaping PSO with nearest-neighbor memplexes. *IEEE Trans Cybern* 49(7):2744–2757
- Wu CH, Li WJ (2021) Enhancing intrusion detection with feature selection and neural network. *Int J Intell Syst*. <https://doi.org/10.1002/int.22397>
- Ding WP, Lin CT, Cao ZH (2019) Shared nearest neighbor quantum game-based attribute reduction with hierarchical co-evolutionary spark and its consistent segmentation application in neonatal cerebral cortical surfaces. *IEEE Trans Neural Netw Learn Syst* 30(7):2013–2027
- Wang CZ, Huang Y, Shao MW, Fan XD (2019) Fuzzy rough set-based attribute reduction using distance measures. *Knowl-Based Syst* 164:205–212

20. Chen HM, Li TR, Fan X, Luo C (2019) Feature selection for imbalanced data based on neighborhood rough sets. *Inf Sci* 483:1–20
21. Sun L, Zhang XY, Qian YH, Xu JC, Zhang SG (2019) Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. *Inf Sci* 502:18–41
22. Wang CZ, Huang Y, Ding WP, Cao ZH (2021) Attribute reduction with fuzzy rough self-information measures. *Inf Sci* 549:68–86
23. Hudec M, Mináriková E, Mesiari R, Saranti A, Holzinger A (2022) Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions. *Knowl-Based Syst* 2:106916
24. Zhang J, Lu G, Li JQ, Li CW (2021) An ensemble classification method for high-dimensional data using neighborhood rough set. *Complexity*. <https://doi.org/10.1155/2021/8358921>
25. Chu XL, Sun BZ, Chu XD, Wu JQ, Han KY, Zhang Y, Huang QC (2022) Multi-granularity dominance rough concept attribute reduction over hybrid information systems and its application in clinical decision-making. *Inf Sci* 597:274–299
26. Hu QH, Zhao H, Yu DR (2008) Efficient symbolic and numerical attribute reduction with neighborhood rough sets. *Moshi Shibie Yu Rengong Zhineng/pattern Recogn Artif Intell* 21(6):732–738
27. Huang YY, Guo KJ, Yi XW, Li Z, Li TR (2022) Matrix representation of the conditional entropy for incremental feature selection on multi-source data. *Inf Sci* 591:263–286
28. Wang CZ, Hu QH, Wang XZ, Chen DG, Qian YH, Dong Z (2018) Feature selection based on neighborhood discrimination index. *IEEE Trans Neural Netw Learn Syst* 29(7):2986–2999
29. Mohamed AE, Abu-Donia Hassan M, Rodyna AH, Saeed LH, Rehab AL (2022) Improved evolutionary-based feature selection technique using extension of knowledge based on the rough approximations. *Inf Sci* 594:76–94
30. Yang J, Wang GY, Zhang QH, Wang HM (2020) Knowledge distance measure for the multigranularity rough approximations of a fuzzy concept. *IEEE Trans Fuzzy Syst* 28(4):706–717
31. Gao Y, Chen XJ, Yang XB, Mi JS (2019) Ensemble-based neighborhood attribute reduction: a multigranularity view. *Complexity* 2019:1–17
32. Li JH, Huang CC, Qi JJ, Qian YH, Liu WQ (2017) Three-way cognitive concept learning via multi-granularity. *Inf Sci* 378:244–263
33. Hu M, Tsang ECC, Guo YT, Chen DG, Xu WH (2021) A novel approach to attribute reduction based on weighted neighborhood rough sets. *Knowl-Based Syst*. <https://doi.org/10.1016/j.knsys.2021.106908>
34. Yang S, Zhang H, Baets BD, Jah M, Shi G (2021) Quantitative dominance-based neighborhood rough sets via fuzzy preference relations. *IEEE Trans Fuzzy Syst* 29(3):515–529
35. Qian YH, Li SY, Liang JY, Shi ZZ, Wang F (2014) Pessimistic rough set-based decisions: a multi-granulation fusion strategy. *Inf Sci* 264(20):196–210
36. Lin GP, Qian YH, Li JJ (2012) NMGRS: neighborhood-based multi-granulation rough sets. *Int J Approx Reason* 53(7):1080–1093
37. Sun L, Wang LY, Ding WP, Qian YH, Xu JC (2021) Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy neighborhood multigranulation rough sets. *IEEE Trans Fuzzy Syst* 29(1):19–33
38. Ma FM, Chen JW, Zhang TF (2017) Quick attribute reduction algorithm for neighborhood multi-granulation rough set based on double granulate criterion. *Kongzhi yu Juece/Control Decis* 32(6):1121–1127
39. Hu CX, Zhang L, Wang BJ, Zhang Z, Li FZ (2018) Incremental updating knowledge in neighborhood multigranulation rough sets under dynamic granular structures. *Knowl-Based Syst* 163:811–829
40. Luo GZ, Qian JL (2019) Neighborhood multi-granulation rough set based on multi-threshold for variable precision decisions. *Appl Res Comput* 2:2
41. Lang GM, Cai MJ, Fujita H, Xiao QM (2018) Related families-based attribute reduction of dynamic covering decision information systems. *Knowl-Based Syst* 163:161–173
42. Zhao H, Qin KY (2014) Mixed feature selection in incomplete decision table. *Knowl-Based Syst* 57:181–190
43. Qian YH, Liang JY, Dang CY (2010) Incomplete multigranulation rough set. *IEEE Trans Syst Man Cybern Part A Syst Hum* 40(2):420–431
44. Yang L, Zhang XY, Xu WH, Sang BB (2019) Multi-granulation rough sets and uncertainty measurement for multi-source fuzzy information system. *Int J Fuzzy Syst* 21:1919–1937
45. Sun L, Wang LY, Qian YH, Xu JC, Zhang SG (2019) Feature selection using Lebesgue and entropy measures for incomplete neighborhood decision systems. *Knowl Based Syst* 2:104942
46. Zhang XY, Yang JL, Tang LY (2020) Three-way class-specific attribute reducts from the information viewpoint. *Inf Sci* 507:840–872
47. Wang CZ, Wang Y, Shao MW, Qian YH, Chen DG (2020) Fuzzy rough attribute reduction for categorical data. *IEEE Trans Fuzzy Syst* 28(5):818–830
48. You XY, Li JJ, Wang HK (2019) Relative reduction of neighborhood-covering pessimistic multigranulation rough set based on evidence theory. *Information (Switzerland)*. <https://doi.org/10.3390/info10110334>
49. Zhang J, Zhang XY, Xu WH, Wu YX (2019) Local multi-granulation decision-theoretic rough set in ordered information systems. *Soft Comput* 23:13247–13291
50. Fan XD, Zhao WD, Wang CZ, Huang Y (2018) Attribute reduction based on max-decision neighborhood rough set model. *Knowl-Based Syst* 151:16–23
51. Xu JC, Wang Y, Xu KQ, Zhang TL (2019) Feature genes selection using fuzzy rough uncertainty metric for tumor diagnosis. *Comput Math Methods Med*. <https://doi.org/10.1155/2019/6705648>
52. Zeng K, She K, Xiu XZ (2013) Multi-granulation entropy and its applications. *Entropy* 15(6):2288–2302
53. Feng T, Fan HT, Mi JS (2017) Uncertainty and reduction of variable precision multi-granulation fuzzy rough sets based on three-way decisions. *Int J Approx Reason* 85:36–58
54. Wang GY (2003) Rough reduction in algebra view and information view. *Int J Intell Syst* 18(6):679–688
55. Xu JC, Wang Y, Mu HY, Huang FZ (2019) Feature genes selection based on fuzzy neighborhood conditional entropy. *J Intell Fuzzy Syst* 36(1):117–126
56. Zhang X, Mei CL, Chen DG, Yang YY, Li JH (2020) Active incremental feature selection using a fuzzy-rough-set-based information entropy. *IEEE Trans Fuzzy Syst* 28(5):901–915
57. Chen YM, Wu KS, Chen XH, Tang CH, Zhu QX (2014) An entropy-based uncertainty measurement approach in neighborhood systems. *Inf Sci* 279:239–250
58. Xu JC, Yuan M, Ma YY (2022) Feature selection using self-information and entropy-based uncertainty measure for fuzzy neighborhood rough set. *Complex Intell Syst*. <https://doi.org/10.1007/s40747-021-00356-3>
59. Shannon CE (2001) A mathematical theory of communication. *Bell Syst Tech J* 5(3):3–55
60. Sun L, Wang LY, Ding WP, Qian YH, Xu JC (2020) Neighborhood multi-granulation rough sets-based attribute reduction using Lebesgue and entropy measures in incomplete neighborhood decision systems. *Knowl-Based Syst* 2:105373
61. Qian YH, Wang Q, Cheng HH, Liang JY, Dang CY (2015) Fuzzy-rough feature selection accelerator. *Fuzzy Sets Syst* 258:61–78
62. Jensen R, Shen Q (2009) New approaches to fuzzy-rough feature selection. *IEEE Trans Fuzzy Syst* 17(4):824–838

63. Tan AH, Wu WZ, Qian YH, Liang JY, Chen JK, Li JJ (2019) Intuitionistic fuzzy rough set-based granular structures and attribute subset selection. *IEEE Trans Fuzzy Syst* 27(3):527–539
64. Xu FF, Miao DQ, Wei L (2009) Fuzzy-rough attribute reduction via mutual information with an application to cancer classification. *Comput Math Appl* 57(6):1010–1017
65. Chen YM, Zhang ZJ, Zheng JZ, Ma Y, Xue Y (2017) Gene selection for tumor classification using neighborhood rough sets and entropy measures. *J Biomed Inform* 67:59–68
66. Saqlain SM, Sher M, Shah FA, Khan I, Ashraf MU, Awais M, Ghani A (2019) Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. *Knowl Inf Syst* 58:139–167
67. Friedman M (1940) A comparison of alternative tests of significance for the problem of rankings. *Ann Math Stat* 11:86–92
68. Demsar J, Schuurmans D (2006) Statistical comparison of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30
69. Wang TH (2018) Kernel learning and optimization with Hilbert-Schmidt independence criterion. *Int J Mach Learn Cybern* 9:1707–1717
70. Li WW, Huang ZQ, Jia XY, Cai XY (2016) Neighborhood based decision-theoretic rough set models. *Int J Approx Reason* 69:1–17
71. Faris H, Mafarja MM, Heidari AA, Alijarah I, Zoubi AMA, Mirjalili S, Fujita H (2018) An efficient binary Salp Swarm Algorithm with crossover scheme for feature selection problems. *Knowl-Based Syst* 154:43–67
72. Chen YM, Zeng ZQ, Lu JW (2017) Neighborhood rough set reduction with fish swarm algorithm. *Soft Comput* 21:6907–6918
73. Mu HY, Xu JC, Wang Y, Sun L (2018) Feature genes selection using Fisher transformation method. *J Intell Fuzzy Syst* 34(6):4291–4300

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