Complex & Intelligent Systems (2022) 8:287–305
https://doi.org/10.1007/s40747-021-00356-3

ORIGINAL ARTICLE

Feature selection using self-information and entropy-based uncertainty measure for fuzzy neighborhood rough set

Jiucheng Xu[1,2] · Meng Yuan[1,2] · Yuanyuan Ma[1,2]

Received: 4 January 2021 / Accepted: 25 March 2021 / Published online: 23 April 2021

Abstract

Feature selection based on the fuzzy neighborhood rough set model (FNRS) is highly popular in data mining. However, the dependent function of FNRS only considers the information present in the lower approximation of the decision while ignoring the information present in the upper approximation of the decision. This construction method may lead to the loss of some information. To solve this problem, this paper proposes a fuzzy neighborhood joint entropy model based on fuzzy neighborhood self-information measure (FNSIJE) and applies it to feature selection. First, to construct four uncertain fuzzy neighborhood self-information measures of decision variables, the concept of self-information is introduced into the upper and lower approximations of FNRS from the algebra view. The relationships between these measures and their properties are discussed in detail. It is found that the fourth measure, named tolerance fuzzy neighborhood self-information, has better classification performance. Second, an uncertainty measure based on the fuzzy neighborhood joint entropy has been proposed from the information view. Inspired by both algebra and information views, the FNSIJE is proposed. Third, the K–S test is used to delete features with weak distinguishing performance, which reduces the dimensionality of high-dimensional gene datasets, thereby reducing the complexity of high-dimensional gene datasets, and then, a forward feature selection algorithm is provided. Experimental results show that compared with related methods, the presented model can select less important features and have a higher classification accuracy.

Keywords Fuzzy neighborhood rough set · Feature selection · Self-information · Fuzzy neighborhood joint entropy · Uncertainty measure

Introduction

Feature selection is an important data preprocess in the fields of granular computing and artificial intelligence [1–6]. Its main goal is to reduce redundant features and simplify the complexity of the classification model, thereby improving the generalization ability of classification model [7–12]. So far, feature selection has been widely used in the fields of pattern recognition, data mining, machine learning, and so on [13–17].

Related work

Pawlak proposed the classic rough set model in 1982 [18], which was successfully applied in the field of feature selection [19–27]. However, the Pawlak proposed that rough set is based on the general binary relationship and is only suitable for discrete data [28]. Some information may be lost, while the discretization method is used for continuous data [29]. To solve this problem, many scholars expanded the rough set model. Dubois et al. introduced the concept of fuzzy rough set via combining rough set and fuzzy set, which overcomes the discrete problem and can directly reduce continuous data [30]. Hu et al. [31] proposed a new method of neighborhood granulation and neighborhood rough set for sensitive feature selection. Wang et al. [32] constructed FNRS to use parameterized fuzzy relations to describe fuzzy information granularity, which reduced the possibility of samples being misclassified. Qian et al. [33] studied the pessimistic multi-granularity rough set decision model of attribute reduction,

✉ Yuanyuan Ma
yuanyuanma821@126.com

[1] College of Computer and Information Engineering, Henan Normal University, 46 Jianshe East Road, Muye District, Xinxiang 453007, Henan, China

[2] Engineering Technology Research Center for Computing Intelligence and Data Mining, Xinxiang 453007, China

which overcomes the shortcomings of most models that limit their application due to a single binary relationship.

As one of the most important rough set models, FNRS received extensive attention in the fields of machine learning and data mining. Shreevastava et al. proposed a new intuitionistic fuzzy neighborhood rough set model via combination of intuitionistic fuzzy set and neighborhood rough set, and applied it to heterogeneous datasets [34]. Yue et al. [35] developed fuzzy neighborhoods to be applied to cover data classification. Sun et al. [36] proposed a new fuzzy neighborhood multi-granularity rough set model by combining FNRS with multi-granularity rough set model, which expanded the type of rough set model. Xu et al. [37] redefined the fuzzy neighborhood relationship in FNRS and introduced it into conditional entropy, and proposed a new model—fuzzy neighborhood conditional entropy—which improved the measurement mechanism. In fact, the classification information is not only related to the lower approximation of the decision classification consistency, but also the upper approximation of the decision classification divergence. Wang et al. [38] introduced the concept of self-information into neighborhood rough set, and constructed neighborhood self-information measure using the upper and lower approximations of decision, which is helpful to the select optimal feature subset.

In the past few decades, uncertainty measures for feature selection from the algebra view and information view had been vigorously developed [37,39]. From algebra view, Wang et al. introduced distance measure into fuzzy rough set model and proposed a new attribute reduction method [40]. Liu et al. designed a hash-based algorithm to calculate the positive region of neighborhood rough set, which was applied in attribute reduction [41]. Hu et al. [42] developed a matrix-based dynamic method to calculate the positive, boundary, and negative regions of neighborhood multi-granularity rough set. Fan et al. [43] focused on boundary samples using the largest decision neighborhood rough set model. Liu et al. [58–60] solved some practical problems from the algebra view. The importance of features based on algebra view can only express the influence of features present in feature subsets [40–43]. From information view, information entropy and some of its deformations had been widely used in feature selection in recent years. Xu et al. [37] studied the fuzzy neighborhood conditional entropy to evaluate the feature meaning in FNRS. Zhang et al. [17] proposed information entropy based on fuzzy rough set and applied it in fuzzy information system. The importance of features based on the information view only explains the impact of uncertainty classification on features [17,37,44,45]. If combining the two views for feature selection, it helps to improve the quality of uncertainty measurement in decision system. Wang [45] simultaneously studied rough reduction and relative reduction from two views. Sun et al. [46] constructed an attribute reduction method based on neighborhood multi-granularity rough set, which can handle mixed incomplete datasets from two views at the same time.

Our work

To solve the problem that most feature evaluation functions only consider the information contained in the lower approximation of the decision, which may cause part of the information to be lost, this article focus on studying the feature selection method based on FNSIJE. The main content of this paper is as follows:

- Analysis shows the shortcoming of the evaluation function based on the dependency of FNRS.
- We propose three types of uncertainty indices using upper and lower approximations: decision index, optimistic decision index, and pessimistic decision index. Three definitions of precision and roughness are given based on this basis, and then combined with the concept of self-information, four fuzzy neighborhood self-information measures are proposed, and related properties are studied. Through theoretical analysis, we find the most suitable fuzzy neighborhood self-information for feature selection and apply it in practice.
- To better discuss feature selection methods based on algebra and information views, this paper studies the uncertainty measure method based on fuzzy neighborhood joint entropy. Then, we combine measure and information entropy to propose a fuzzy neighborhood self-information-based fuzzy neighborhood joint entropy (FNSIJE) method for feature selection. FNSIJE not only considers the classification information provided by the upper and lower approximations of the neighborhood decision system at the same time, but also can simultaneously select features from the algebra and information views.

The article is organized as follows: The section “Preliminaries” is the concepts of self-information and FNRS. The section “Insufficiency of neighborhood correlation functions and uncertainty measurement based on FNSIJE” points out the shortcoming of the neighborhood correlation functions; in view of this shortcoming, we propose four fuzzy neighborhood self-information measures, and study their related properties. Then, the feature selection model based on FNSIJE is constructed by combining tolerance decision self-information measure and fuzzy neighborhood joint entropy. In the section “Feature selection method based on FNSIJE model”, we design a heuristic feature subset selection method. In the section “Experimental analysis”, six UCI datasets and four microarray gene expression profile datasets are used for experimental verification. The section “Conclusion” is the conclusion of this article and the outlook for future work.

Preliminaries

In this section, it mainly deals with relevant concepts of self-information and FNRS.

Self-information

Definition 1 [47] The measure $I(x)$ was proposed by Shannon to express the uncertainty of signal x. x is called the self-information if it as the following properties:

(1) Non-negative: $I(x) \geq 0$.
(2) If $p(x) \to 0$, then $I(x) \to \infty$.
(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.

Fuzzy neighborhood rough set

Let NDS $= \langle U, A, D, V, f, \delta\rangle$ be a neighborhood decision system. $U = \{x_1, x_2, \ldots, x_m\}$ is the set of samples. A is the set of conditional attributes and D is the set of decision classes. $V = \bigcup_{a \in A} V_a$. $f : U \times \{A \cup D\} \to V$ is the map function, which $f(a, x)$ is the attribute value of x on attribute a. δ is the parameter of the neighborhood radius, which $0 \leq \delta \leq 1$. The neighborhood decision system is simplified to NDS $= \langle U, A, D\rangle$.

Definition 2 [32,48] Let $B \subseteq A$ be a attribute subset on the universe U, and then, B can induce a fuzzy relation R_B on U; R_B is a fuzzy similarity relation if the following conditions are satisfied:

(1) Reflexivity: $R_B(x, x) = 1$.
(2) Symmetry: $R_B(x, y) = R_B(y, x), \forall x, y \in U$.

Definition 3 Given NDS $= \langle U, A, D\rangle$, fuzzy neighborhood radius parameter $\lambda (0 < \lambda < 1)$ is used to describe the similarity of samples. For any $x, y \in U$, the fuzzy neighborhood similarity relationship between two samples x and y with regard to a is expressed as:

$$R_a = \begin{cases} 0, & |f(a,x) - f(a,y)| > \lambda \\ 1 - |f(a,x) - f(a,y)|, & |f(a,x) - f(a,y)| \leq \lambda. \end{cases} \quad (1)$$

The fuzzy neighborhood similarity matrix is $[x]_a(y) = R_a(x, y)$, for any $B \subseteq A$, $[x]_a(y) = \min_{a \in B}([x]_a y))$ [37].

Definition 4 Given NDS $= \langle U, A, D\rangle$, $B \subseteq A$, for any $x, y \in U$, the parameterized fuzzy neighborhood information granule of x associated with B is expressed as:

$$\alpha_B(x) = [x]_a(y) = \begin{cases} 0, & R_B(x,y) < 1 - \lambda \\ R_B(x,y), & R_B(x,y) \geq 1 - \lambda. \end{cases} \quad (2)$$

Definition 5 Given NDS $= \langle U, A, D\rangle, U = \{x_1, x_2, \ldots, x_m\}$, AT $= \{\mathrm{AT}_1, \mathrm{AT}_2, \ldots, \mathrm{AT}_n\}$ and $U/D = \{D_1, D_2, \ldots, D_t\}$, then the fuzzy decision of the sample derived from D is expressed as:

$$\mathrm{FD} = \{\mathrm{FD}_1^{\mathrm{T}}, \mathrm{FD}_2^{\mathrm{T}}, \ldots, \mathrm{FD}_t^{\mathrm{T}}\}. \quad (3)$$

Here, $\mathrm{FD}_j = \{\mathrm{FD}_j(x_1), \mathrm{FD}_j(x_2), \ldots, \mathrm{FD}_j(x_m)\}$ is the fuzzy rough set of sample decision equivalence class, and $j = 1, 2, \ldots, t$. When $l = 1, 2, \ldots, m$, $\mathrm{FD}_j(x_l)$ is recorded as the degree of membership of $x_l \in U$ on FD_j and denoted [36] by:

$$\mathrm{FD}_j(x_l) = \frac{\left|[x_l]_A(y) \cap D_j\right|}{|[x_l]_A(y)|}, \quad (4)$$

where $[x_l]_A(y)$ is fuzzy neighborhood similarity degree.

Definition 6 Given NDS $= \langle U, A, D\rangle$, $B \subseteq A$, for any $X \subseteq U$, $\alpha_B(x)$ is the parametric fuzzy neighborhood information granule of $x \in U$. Then, the fuzzy neighborhood upper and lower approximations of X with respect to B are, respectively, expressed as:

$$\overline{R_B^{\lambda}}(X) = \{x \in U \,|\alpha_B(x) \cap X \neq \emptyset\} \quad (5)$$

$$\underline{R_B^{\lambda}}(X) = \{x \in U \,|\alpha_B(x) \subseteq X\}. \quad (6)$$

Definition 7 Given NDS $= \langle U, A, D\rangle$, suppose any $D_j \in U/D = \{D_1, D_2, \ldots, D_t\}$, then the fuzzy neighborhood positive region and its dependency degree of D in relation to B are expressed, respectively, by:

$$\mathrm{POS}_B^{\lambda}(D) = \bigcup_{j=1}^{t} \underline{R_B^{\lambda}}(D_j) \quad (7)$$

$$d_B^{\lambda}(D) = \frac{\left|\mathrm{POS}_B^{\lambda}(D)\right|}{|U|}. \quad (8)$$

Insufficiency of neighborhood correlation functions and uncertainty measurement based on FNSIJE

In the first subsection of this section, we will analyze the shortcoming of the evaluation function based on the classical dependency function. To avoid this deficiency, in the second subsection of this section, first, we construct three indices with different meanings using the upper and lower approximations of decision: decision index, optimistic decision index, and pessimistic decision index. Second, we redefine three types of precision and roughness on the basis of the three indices; combining with the concept of self-information, four fuzzy neighborhood self-information measures are proposed and elaborate their properties in detail. Finally, through theoretical analysis, we find that the most suitable measure for feature selection and information entropy is combined to form a new mixed uncertainty measure for feature selection to reduce the noise in the neighborhood decision system.

Insufficiency of neighborhood related functions

Formula (8) is employed as the evaluation function for feature selection in classical FNRS, and a variety of feature selection methods based on formula (8) are proposed and applied. However, this construction method only considers positive samples; in other words, only consistent samples in the lower approximation of the decision are involved, while ignoring the upper approximation that diverges from the decision classification. However, the upper approximation also contains some classification information that cannot be ignored. Therefore, an ideal evaluation function should be a function containing both upper and lower approximations of decision. In the section "Uncertainty measurement based on FNSIJE", we construct the uncertainty measure based on FNSIJE as the evaluation function of feature selection, making the feature selection mechanism more reasonable.

Uncertainty measurement based on FNSIJE

Definition 8 Let $\mathrm{NDS} = \langle U, A, D\rangle$, $U/D = \{D_1, D_2, \ldots, D_t\}$, and $B \subseteq A$. The decision index $\mathrm{dec}(D_r)$, the optimistic decision index $\mathrm{opt}_B(D_r)$, and the pessimistic decision index $\mathrm{pess}_B(D_r)$ of D_r are defined, respectively, as:

$$\mathrm{dec}(D_r) = |D_r| \tag{9}$$

$$\mathrm{opt}_B(D_r) = \left|\underline{R}_B^{\lambda}(D_r)\right| \tag{10}$$

$$\mathrm{pess}_B(D_r) = \left|\overline{R}_B^{\lambda}(D_r)\right|. \tag{11}$$

Here, $\overline{R}_B^{\lambda}(D_r)$ and $\underline{R}_B^{\lambda}(D_r)$ are the fuzzy neighborhood upper and lower approximations, respectively. The optimistic decision index $\mathrm{opt}_B(D_r)$ is represented by the cardinal number of its lower approximation, which denotes the number of samples with consistent classification. The pessimistic decision index $\mathrm{pess}_B(D_r)$ is represented by the cardinal number of its upper approximation, which represents the number of samples that may belong to the decision class. $|\cdot|$ denotes the cardinality of a set:

Property 1 $\mathrm{opt}_B(D_r) \le \mathrm{dec}(D_r) \le \mathrm{pess}_B(D_r)$.

Proof It can be directly inferred from Definition 8.

Property 2 *Let $B_1 \subseteq B_2 \subseteq A$ and $D_r \in U/D$, and then*

(1) $\mathrm{opt}_{B_1}(D_r) \le \mathrm{opt}_{B_2}(D_r)$.
(2) $\mathrm{pess}_{B_1}(D_r) \ge \mathrm{pess}_{B_2}(D_r)$.

Proof (1) Because $B_1 \subseteq B_2 \subseteq A$, thus $R_{B_2}^{\lambda}(D_r) \subseteq R_{B_1}^{\lambda}(D_r)$. It follows from Definition 8 that $\mathrm{opt}_{B_1}(D_r) \le \mathrm{opt}_{B_2}(D_r)$.
(2) It is similar to (1).

Property 2 shows that both optimistic and pessimistic decision indices are monotonous. With the increase of the number of features, the optimistic decision index increases and the decision consistency is enhanced. For pessimistic decision index, attribute reduction is produced with the reduction of decision uncertainty.

Next, we define the precision and roughness of the optimistic decision index to depict the classification ability of a feature subset.

Definition 9 Let $B \subseteq A$, $D_r \in U/D$, and the precision and roughness of the optimistic decision index are, respectively, defined as:

$$\rho_B^{(1)}(D_r) = \frac{\mathrm{opt}_B(D_r)}{\mathrm{dec}(D_r)} \tag{12}$$

$$\omega_B^{(1)}(D_r) = 1 - \frac{\mathrm{opt}_B(D_r)}{\mathrm{dec}(D_r)} = 1 - \rho_B^{(1)}(D_r). \tag{13}$$

Apparently, $0 \le \rho_B^{(1)}(D_r), \omega_B^{(1)}(D_r) \le 1$. $\rho_B^{(1)}(D_r)$ indicates the degree to which the sample is completely divided into D_r. $\omega_B^{(1)}(D_r)$ indicates the degree to which a sample may belong to D_r. Both $\rho_B^{(1)}(D_r)$ and $\omega_B^{(1)}(D_r)$ reflect the classification ability of feature subset B.

When $\rho_B^{(1)}(D_r) = 1$, $\omega_B^{(1)}(D_r) = 0$, then, $\mathrm{dec}(D_r) = \mathrm{opt}_B(D_r)$. In this case, all samples are correctly classified into the corresponding decision, and the feature subset B has the strongest classification ability. When $\rho_B^{(1)}(D_r) = 0$, $\omega_B^{(1)}(D_r) = 1$, then $\mathrm{opt}_B(D_r) = 0$. In this case, all samples have not been assigned to the correct decision. This moment, B, has the weakest classification ability.

Property 3 *Let $B_1 \subseteq B_2 \subseteq A$ and $D_r \in U/D$, and then*

(1) $\rho_{B_1}^{(1)}(D_r) \le \rho_{B_2}^{(1)}(D_r)$.
(2) $\omega_{B_1}^{(1)}(D_r) \ge \omega_{B_2}^{(1)}(D_r)$.

Proof (1) Because $B_1 \subseteq B_2$, it follows from Property 2(1) that $\mathrm{opt}_{B_1}(D_r) \le \mathrm{opt}_{B_2}(D_r)$. Thus, we can have

$\frac{\text{opt}_{B_1}(D_r)}{\text{dec}(D_r)} \leq \frac{\text{opt}_{B_2}(D_r)}{\text{dec}(D_r)}$. By Definition 9, we have $\rho_{B_1}^{(1)}(D_r) \leq \rho_{B_2}^{(1)}(D_r)$.

(2) By Property 3(1), we can obtain that $\rho_{B_1}^{(1)}(D_r) \leq \rho_{B_2}^{(1)}(D_r)$ for $B_1 \subseteq B_2$. Thus, we can have that $1 - \rho_{B_1}^{(1)}(D_r) \geq 1 - \rho_{B_2}^{(1)}(D_r)$. From Definition 9, we can obtain $\omega_{B_1}^{(1)}(D_r) \geq \omega_{B_2}^{(1)}(D_r)$.

Property 3 shows that the precision and roughness of optimistic decision index are monotonous. With the increase of the number of new features, the precision of optimistic decision index gradually increases, while the roughness decreases.

Definition 10 Let $B \subseteq A$, $D_r \in U/D$, and the optimistic decision self-information of D_r can be defined as:

$$I_B^1(D_r) = -\omega_B^{(1)}(D_r) \log \rho_B^{(1)}(D_r). \tag{14}$$

Apparently, Definition 10 satisfies the properties (1), (2), and (3) of Definition 1. Then, Property (4) can be confirmed according to Property 4.

Property 4 *Let* $B_1 \subseteq B_2 \subseteq A$ *and* $D_r \in U/D$, *and then,* $I_{B_1}^1(D_r) \geq I_{B_2}^1(D_r)$.

Proof Because $B_1 \subseteq B_2 \subseteq A$, it follows from Property 3 that $\rho_{B_1}^{(1)}(D_r) \leq \rho_{B_2}^{(1)}(D_r)$. Thus, we can see that $0 \leq -\log \rho_{B_2}^{(1)}(D_r) \leq -\log \rho_{B_1}^{(1)}(D_r)$ and $0 \leq \omega_{B_2}^{(1)}(D_r) \leq \omega_{B_1}^{(1)}(D_r)$. Then, we can obtain $I_{B_1}^1(D_r) \geq I_{B_2}^1(D_r)$.

Definition 11 Let NDS $= \langle U, A, D\rangle$, $U/D = \{D_1, D_2, \ldots, D_t\}$, $B \subseteq A$, and the optimistic decision self-information of *NDS* is defined as:

$$I_B^1(D) = \sum_{r=1}^{t} I_B^1(D_r). \tag{15}$$

Self-information was originally used to describe the instability of signal output. Applying self-information to decision system can describe the uncertainty of decision, which is an effective means to evaluate decision ability.

$I_B^1(D)$ describes the classification information of feature subset B. The smaller $I_B^1(D)$ is, the stronger the classification ability of feature subset B is. $I_B^1(D) = 0$ indicates that all samples in U can be completely classified into their respective classes by feature subset B.

The selected feature subset based on optimistic decision self-information only pays attention to the consistency of the classification objects, ignoring the information contained in the uncertain classification samples. However, these uncertain informations are often not negligible. Therefore, it is necessary to analyze the information contained in uncertain classification objects.

Next, we define the precision and roughness of pessimistic decision to depict the uncertainty of feature subset B.

Definition 12 Let NDS $= \langle U, A, D\rangle$, $U/D = \{D_1, D_2, \ldots, D_t\}$, and $B \subseteq A$, and the precision and roughness of the pessimistic decision index are defined as follows:

$$\rho_B^{(2)}(D_r) = \frac{\text{dec}(D_r)}{\text{pess}_B(D_r)} \tag{16}$$

$$\omega_B^{(2)}(D_r) = 1 - \frac{\text{dec}(D_r)}{\text{pess}_B(D_r)} = 1 - \rho_B^{(2)}(D_r). \tag{17}$$

It is obviously that $0 \leq \rho_B^{(2)}(D_r), \omega_B^{(2)}(D_r) \leq 1$. $\rho_B^{(2)}(D_r)$ can be regarded as the uncertainty degree of the decision equivalence class D_r. $\omega_B^{(2)}(D_r)$ indicates the degree which the sample cannot be correctly classified into decision class.

When $\rho_B^{(2)}(D_r) = 1$, $\omega_B^{(2)}(D_r) = 0$, then $\text{dec}(D_r) = \text{pess}_B(D_r)$. That means all objects which belong toD_r are completely sorted into D_r by B. In this case, feature subset B has the strongest classification ability.

Property 5 *Let* $B_1 \subseteq B_2 \subseteq A$ *and* $D_r \in U/D$, *and then*

(1) $\rho_{B_1}^{(2)}(D_r) \leq \rho_{B_2}^{(2)}(D_r)$.

(2) $\omega_{B_1}^{(2)}(D_r) \geq \omega_{B_2}^{(2)}(D_r)$.

Proof (1) Because $B_1 \subseteq B_2$, it follows from Property 2(2) that $\text{pess}_{B_1}(D_r) \geq \text{pess}_{B_2}(D_r)$. Thus, we can have $\frac{\text{dec}(D_r)}{\text{pess}_{B_1}(D_r)} \leq \frac{\text{dec}(D_r)}{\text{pess}_{B_2}(D_r)}$. By Definition 12, we have $\rho_{B_1}^{(2)}(D_r) \leq \rho_{B_2}^{(2)}(D_r)$.

(2) By Property 5(1), we can obtain that $\rho_{B_1}^{(2)}(D_r) \leq \rho_{B_2}^{(2)}(D_r)$ for $B_1 \subseteq B_2$. Thus, we can have that $1 - \rho_{B_1}^{(2)}(D_r) \geq 1 - \rho_{B_2}^{(2)}(D_r)$. From Definition 12, we can obtain $\omega_{B_1}^{(2)}(D_r) \geq \omega_{B_2}^{(2)}(D_r)$.

Property 5 shows that the precision and roughness of the pessimistic decision index are monotonous. As the number of new features increases, the precision of the pessimistic decision index gradually increases and the roughness gradually decreases.

Definition 13 Let $B \subseteq A$ and $D_r \in U/D$, and the pessimistic decision self-information of D_r can be denoted as:

$$I_B^2(D_r) = -\omega_B^{(2)}(D_r) \log \rho_B^{(2)}(D_r). \tag{18}$$

Apparently, Definition 13 satisfies the properties (1), (2), and (3) of Definition 1. Then, Property (4) can be confirmed according to Property 6.

Property 6 *Let* $B_1 \subseteq B_2 \subseteq A$ *and* $D_r \in U/D$, *and then,* $I_{B_1}^2(D_r) \geq I_{B_2}^2(D_r)$.

Proof Because $B_1 \subseteq B_2$, it follows from Property 5(1) that $\rho_{B_1}^{(2)}(D_r) \le \rho_{B_2}^{(2)}(D_r)$. Thus, we can see that $0 \le -\log \rho_{B_2}^{(2)}(D_r) \le -\log \rho_{B_1}^{(2)}(D_r)$ and $0 \le \omega_{B_2}^{(2)}(D_r) \le \omega_{B_1}^{(2)}(D_r) \le 1$. By Definition 13, we can obtain $I_{B_1}^2(D_r) \ge I_{B_2}^2(D_r)$.

Definition 14 Let $NDS = \langle U, A, D\rangle$, $U/D = \{D_1, D_2, \ldots, D_t\}$, and $B \subseteq A$, and the pessimistic decision self-information of *NDS* is defined as:

$$I_B^2(D) = \sum_{r=1}^{t} I_B^2(D_r). \tag{19}$$

Through the above analysis, the optimistic decision self-information $I_B^1(D)$ focuses on the ability to completely divide the sample into correct decision D_r through feature subset B. Although the pessimistic decision self-information $I_B^2(D)$ can obtain samples that may belong to decision D_r, it cannot ensure that all classification information is consistent. Hence, both $I_B^1(D)$ and $I_B^2(D)$ describe the classification ability of feature subset B from one-sided perspective, which cannot reflect the comprehensive information included in the decision. Therefore, we will propose two other self-information to measure the uncertainty of classification information. They can not only avoid considering the uncertainty of decision information from a one-sided perspective to a greater extent, but they are also more in line with actual decision in theory.

Definition 15 Let $B \subseteq A$ and $D_r \in U/D$, and the optimistic-pessimistic decision self-information of D_r is defined as:

$$I_B^3(D_r) = I_B^1(D_r) + I_B^2(D_r). \tag{20}$$

Property 7 *Let $B_1 \subseteq B_2 \subseteq A$ and $D_r \in U/D$, and then, $I_{B_1}^3(D_r) \ge I_{B_2}^3(D_r)$.*

Proof From Properties 4 and 6, we can know that $I_{B_1}^1(D_r) \ge I_{B_2}^1(D_r)$ and $I_{B_1}^2(D_r) \ge I_{B_2}^2(D_r)$ for $B_1 \subseteq B_2$. Thus, we can obtain that $I_{B_1}^1(D_r) + I_{B_1}^2(D_r) \ge I_{B_2}^1(D_r) + I_{B_2}^2(D_r)$. By Definition 15, we have $I_{B_1}^3(D_r) \ge I_{B_2}^3(D_r)$.

Definition 16 Let $NDS = \langle U, A, D\rangle$, $U/D = \{D_1, D_2, \ldots, D_t\}$ and $B \subseteq A$, the optimistic-pessimistic decision self-information of *NDS*:

$$I_B^3(D) = \sum_{r=1}^{t} I_B^3(D_r). \tag{21}$$

Definition 17 Let $B \subseteq A$ and $D_r \in U/D$, and the precision and roughness of the tolerance decision index are defined as follows:

$$\rho_B^{(3)}(D_r) = \frac{\mathrm{opt}_B(D_r)}{\mathrm{pess}_B(D_r)} \tag{22}$$

$$\omega_B^{(3)}(D_r) = 1 - \frac{\mathrm{opt}_B(D_r)}{\mathrm{pess}_B(D_r)} = 1 - \rho_B^{(3)}(D_r). \tag{23}$$

Apparently, $0 \le \rho_B^{(3)}(D_r), \omega_B^{(3)}(D_r) \le 1$. $\rho_B^{(3)}(D_r)$ depicts the ratio of optimistic decision and pessimistic decision samples, which characterizes the classification ability of feature subset B. When $\rho_B^{(3)}(D_r) = 1, \omega_B^{(3)}(D_r) = 0$, it is the ideal state of the feature subset B, and the feature subset B has the optimal classification ability at this time. On the contrary, when $\rho_B^{(3)}(D_r) = 0$, the feature subset B has no effect on classification and has the weakest classification ability.

Property 8 *Let $B_1 \subseteq B_2 \subseteq A$, $D_r \in U/D$ and $B \subseteq A$, and then:*

(1) $\rho_{B_1}^{(3)}(D_r) \le \rho_{B_2}^{(3)}(D_r)$, $\omega_{B_1}^{(3)}(D_r) \ge \omega_{B_2}^{(3)}(D_r)$.
(2) $\rho_B^{(3)}(D_r) = \rho_B^{(1)}(D_r) \cdot \rho_B^{(2)}(D_r)$.
(3) $\omega_B^{(3)}(D_r) = \omega_B^{(1)}(D_r) + \omega_B^{(2)}(D_r) - \omega_B^{(1)}(D_r) \cdot \omega_B^{(2)}(D_r)$.

Proof (1) Because $B_1 \subseteq B_2$, it follows from Property 2 that $\mathrm{opt}_{B_1}(D_r) \le \mathrm{opt}_{B_2}(D_r)$ and $\mathrm{pess}_{B_1}(D_r) \ge \mathrm{pess}_{B_2}(D_r)$. Then, we can have $\frac{\mathrm{opt}_{B_1}(D_r)}{\mathrm{pess}_{B_1}(D_r)} \le \frac{\mathrm{opt}_{B_2}(D_r)}{\mathrm{pess}_{B_2}(D_r)}$. By Definition 17, we have $\rho_{B_1}^{(3)}(D_r) \le \rho_{B_2}^{(3)}(D_r)$. Thus, we can have that $1 - \rho_{B_1}^{(3)}(D_r) \ge 1 - \rho_{B_2}^{(3)}(D_r)$. Therefore, we can obtain $\omega_{B_1}^{(3)}(D_r) \ge \omega_{B_2}^{(3)}(D_r)$.

(2) From Definitions 9 and 12, we have $\rho_B^{(1)}(D_r) = \frac{\mathrm{opt}_B(D_r)}{\mathrm{dec}(D_r)}$ and $\rho_B^{(2)}(D_r) = \frac{\mathrm{dec}(D_r)}{\mathrm{pess}_B(D_r)}$, then $\rho_B^{(1)}(D_r) \cdot \rho_B^{(2)}(D_r) = \frac{\mathrm{opt}_B(Dr)}{\mathrm{dec}_B(Dr)} \cdot \frac{\mathrm{dec}_B(Dr)}{\mathrm{pess}_B(Dr)} = \frac{\mathrm{opt}_B(Dr)}{\mathrm{pess}_B(Dr)}$. From Definition 17, we can obtain $\rho_B^{(3)}(D_r) = \frac{\mathrm{opt}_B(D_r)}{\mathrm{pess}_B(D_r)}$, and thus, $\rho_B^{(3)}(D_r) = \rho_B^{(1)}(D_r) \cdot \rho_B^{(2)}(D_r)$.

(3) From Definition 17 and Property 8(2), we have:

$$\begin{aligned}\omega_B^{(3)}(D_r) &= 1 - \rho_B^{(3)}(D_r)\\ &= 1 - \rho_B^{(1)}(D_r) \cdot \rho_B^{(2)}(D_r)\\ &= 1 - \left[1 - \omega_B^{(1)}(D_r)\right] \cdot \left[1 - \omega_B^{(2)}(D_r)\right]\\ &= \omega_B^{(1)}(D_r) + \omega_B^{(2)}(D_r) - \omega_B^{(1)}(D_r) \cdot \omega_B^{(2)}(D_r).\end{aligned}$$

Definition 18 Let $B \subseteq A$ and $D_r \in U/D$, and the tolerance decision self-information of D_r is defined as:

$$I_B^4(D_r) = -\omega_B^{(3)}(D_r) \log \rho_B^{(3)}(D_r). \tag{24}$$

Clearly, Definition 18 satisfies (1), (2), and (3) of Definition 1. Property (4) is verified by Property 9.

Property 9 *Let $B_1 \subseteq B_2 \subseteq A$ and $D_r \in U/D$, and then, $I^4_{B_1}(D_r) \geq I^4_{B_2}(D_r)$.*

Proof Because $B_1 \subseteq B_2$, it follows from Property 8(1) that $\rho^{(3)}_{B_1}(D_r) \leq \rho^{(3)}_{B_2}(D_r)$. Thus, we can see that $0 \leq -\log \rho^{(3)}_{B_2}(D_r) \leq -\log \rho^{(3)}_{B_1}(D_r)$ and $0 \leq \omega^{(3)}_{B_2}(D_r) \leq \omega^{(3)}_{B_1}(D_r) \leq 1$. Then, from Definition 18, we can obtain $I^4_{B_1}(D_r) \geq I^4_{B_2}(D_r)$.

Property 10 $I^4_B(D_r) \geq I^3_B(D_r)$.

Proof

$$\begin{aligned}
I^{(4)}_B(D_r) &= -\omega^{(3)}_B(D_r)\log\rho^{(3)}_B(D_r)\\
&= -\left[\omega^{(1)}_B(D_r) + \omega^{(2)}_B(D_r) - \omega^{(1)}_B(D_r)\cdot\omega^{(2)}_B(D_r)\right]\\
&\quad\cdot\log\left[\rho^{(1)}_B(D_r) + \rho^{(2)}_B(D_r)\right]\\
&= -\left[\omega^{(1)}_B(D_r) + \omega^{(2)}_B(D_r) - \omega^{(1)}_B(D_r)\cdot\omega^{(2)}_B(D_r)\right]\\
&\quad\cdot\left[\log\rho^{(1)}_B(D_r) + \log\rho^{(2)}_B(D_r)\right]\\
&= -\omega^{(1)}_B(D_r)\cdot\log\rho^{(1)}_B(D_r) - \omega^{(2)}_B(D_r)\cdot\log\rho^{(2)}_B(D_r)\\
&\quad+\left[\omega^{(1)}_B(D_r) - 1\right]\cdot\omega^{(2)}_B(D_r)\cdot\log\rho^{(1)}_B(D_r)\\
&\quad+\left[\omega^{(2)}_B(D_r) - 1\right]\cdot\omega^{(1)}_B(D_r)\cdot\log\rho^{(2)}_B(D_r)\\
&= I^1_B(D_r) + I^2_B(D_r) + \left[\omega^{(1)}_B(D_r) - 1\right]\cdot\omega^{(2)}_B(D_r)\\
&\quad\cdot\log\rho^{(1)}_B(D_r) + \left[\omega^{(2)}_B(D_r) - 1\right]\cdot\omega^{(1)}_B(D_r)\cdot\log\rho^{(2)}_B(D_r).
\end{aligned}$$

It is clear that $0 \leq \rho^{(1)}_B(D_r) \leq 1$, $0 \leq \omega^{(1)}_B(D_r) \leq 1$ from Definition 9, and we have $\log\rho^{(1)}_B(D_r) \leq 0$, $\omega^{(1)}_B(D_r) - 1 \leq 0$. Similarly, we can see $0 \leq \rho^{(2)}_B(D_r) \leq 1$, $0 \leq \omega^{(2)}_B(D_r) \leq 1$ from Definition 12, and we have $\log\rho^{(1)}_B(D_r) \leq 0$, $\omega^{(2)}_B(D_r) - 1 \leq 0$. It can be proved that:

$$\begin{cases}\left[\omega^{(1)}_B(D_r) - 1\right]\cdot\omega^{(2)}_B(D_r)\cdot\log\rho^{(1)}_B(D_r) \geq 0 & ①\\ \left[\omega^{(2)}_B(D_r) - 1\right]\cdot\omega^{(1)}_B(D_r)\cdot\log\rho^{(2)}_B(D_r) \geq 0 & ②.\end{cases}$$

Thus, $\Delta =$ ① + ② ≥ 0:

$$\begin{aligned}
I^4_B(D_r) &= I^1_B(D_r) + I^2_B(D_r) + \Delta\\
&= I^3_B(D_r) + \Delta\\
&\geq I^3_B(D_r).
\end{aligned}$$

Definition 19 Let NDS $= \langle U, A, D\rangle$, $U/D = \{D_1, D_2, \ldots, D_t\}$, and $B \subseteq A$, and the tolerance decision self-information of *NDS* is defines as:

$$I^4_B(D) = \sum_{r=1}^{t} I^4_B(D_r). \tag{25}$$

Property 11 *Let $B_1 \subseteq B_2 \subseteq A$, and then, $I^4_{B_1}(D) \geq I^4_{B_2}(D)$.*

Proof It is straightforward from Property 9.

Remark 1 According to the above theoretical analysis, tolerance fuzzy neighborhood decision self-information (FNSI) can not only consider the decision uncertainty from a more comprehensive perspective, but also is sensitive to the change of feature subset size, and thus, it is more suitable for feature selection.

Definition 20 [36] Let NDS $= \langle U, A, D\rangle$ and $U = \{x_1, x_2, \ldots, x_m\}$, and then, the fuzzy neighborhood entropy of A is denoted by:

$$\mathrm{FNE}_\alpha(A) = -\sum_{l=1}^{m}\frac{|\alpha_A(x_l)|}{|U|}\log_2\frac{|\alpha_A(x_l)|}{|U|}. \tag{26}$$

Here, $\alpha_A(x_l)$ is the parameterized fuzzy neighborhood granule, $x_l \in U$ and $l = 1, 2, \ldots, m$.

Definition 21 Let NDS $= \langle U, A, D\rangle$, $U/D = \{D_1, D_2, \ldots, D_t\}$ and AT $\subseteq A$, and the fuzzy neighborhood joint entropy of AT and D is defined as:

$$\begin{aligned}\mathrm{FNE}_\alpha(\mathrm{AT}, D) = &-\frac{1}{|U|}\sum_{l=1}^{m}\sum_{j=1}^{t}\frac{\left|\alpha_A(x_t)\cap\mathrm{FD}_j\right|}{|U|}\\ &\times\log_2\frac{\left|\alpha_A(x_l)\cap\mathrm{FD}_j\right|}{|U|}.\end{aligned} \tag{27}$$

Here, $x_l \in U$, $l = 1, 2, \ldots, m$, and $j = 1, 2, \ldots, t$, and $\left|\alpha_A(x_l)\cap\mathrm{FD}_j\right|$ indicates that the degree of membership of $\alpha_A(x_l)$ is not greater than the number of non-zero values of the samples of FD_j [37].

Definition 22 Let NDS $= \langle U, A, D\rangle$, $U/D = \{D_1, D_2, \ldots, D_t\}$, and AT $\subseteq A$, and then, the FNSI-based fuzzy neighborhood joint entropy (FNSIJE) of AT and D is defined as:

$$\begin{aligned}&\mathrm{FNSIJE}_\alpha(\mathrm{AT}_i, D) = -I^4_B(D)\\ &\times\frac{1}{|U|}\sum_{l=1}^{m}\sum_{j=1}^{t}\frac{\left|\alpha_{A_i}(x_l)\cap\mathrm{FD}_j\right|}{|U|}\log_2\frac{\left|\alpha_{A_i}(x_l)\cap\mathrm{FD}_j\right|}{|U|}.\end{aligned} \tag{28}$$

Here, $I^4_B(D)$ is the tolerance decision self-information of NDS, $\alpha_A(x_l)$ is the parameterized fuzzy neighborhood granule, $x_l \in U$ and $l = 1, 2, \ldots, m$, $\left|\alpha_A(x_l)\cap\mathrm{FD}_j\right|$ indicates that the degree of membership of $\alpha_A(x_l)$ is not greater than the number of non-zero values of the samples of FD_j, $j = 1, 2, \ldots, t$.

Property 12 *Let NDS* $= \langle U, A, D\rangle$, $U/D = \{D_1, D_2, \ldots, D_t\}$, $AT \subseteq A$, $AT = \{AT_1, AT_2, \ldots, AT_n\}$ *and* $U = \{x_1, x_2, \ldots, x_m\}$, *and then:*

$$FNSIJE_\alpha(AT_i, D) = I_B^4(D) \times FNE_\alpha(AT_i, D) \geq 0.$$

Remark 2 From Definition 22 and Property 12, $I_B^4(D)$ is the fuzzy neighborhood self-information measure from algebra view, and $\mathrm{FNE}_\alpha(\mathrm{AT}_i, D)$ denotes the fuzzy neighborhood joint entropy from the information view. Therefore, FNSIJE can simultaneously measure the uncertainty of neighborhood decision from both the algebra and information views.

Feature selection method based on FNSIJE model

In this section, we will propose a feature selection method based on the FNSIJE model and apply it to the neighborhood decision system.

FNSIJE-based feature selection

Definition 23 Let NDS $= \langle U, A, D\rangle$, $\mathrm{AT}' \subseteq \mathrm{AT} \subseteq A$, $\mathrm{AT} = \{\mathrm{AT}_1, \mathrm{AT}_2, \ldots, \mathrm{AT}_n\}$, AT' is the reduction of AT with regard to D, if it is satisfied:

(1) $\mathrm{FNSIJE}_\alpha(\mathrm{AT}', D) = \mathrm{FNSIJE}_\alpha(\mathrm{AT}, D)$.
(2) $\mathrm{FNSIJE}_\alpha(\mathrm{AT}', D) > \mathrm{FNSIJE}_\alpha(\mathrm{AT}' - \mathrm{AT}_i, D)$, for any $\mathrm{AT}_i \subseteq \mathrm{AT}'$.

Equation (1) shows that the reduced subset has the same classification ability as the entire dataset, and Eq. (2) ensures that the reduced subset has no redundant feature.

Definition 24 Let NDS $= \langle U, A, D\rangle$, $\mathrm{AT}' \subseteq \mathrm{AT} \subseteq A$, $\mathrm{AT} = \{\mathrm{AT}_1, \mathrm{AT}_2, \ldots, \mathrm{AT}_n\}$, $\mathrm{AT}_i \subseteq \mathrm{AT}'$ and $i = 1, 2, \ldots, n$, and then, the attribute significance of subset AT_i with respect to D is defined as:

$$\begin{aligned}\mathrm{SIG}(\mathrm{AT}_i, \mathrm{AT}', D) = {} & \mathrm{FNSIJE}_\alpha(\mathrm{AT}' - \mathrm{AT}_i, D) \\ & -\mathrm{FNSIJE}_\alpha(\mathrm{AT}', D).\end{aligned} \tag{29}$$

Feature selection algorithm

According to the above definition, a FNSIJE-based feature selection (FNSIJE-KS) method is demonstrated in Algorithm 1.

Algorithm 1 FNSIJE-KS

```
Require: $NDS =< U, A, D >$, a fuzzy neighborhood radius parameters $\lambda$.
Ensure: An optimal feature subset $red$.
1: $\forall a \in A$, compute the relation matrix $\mathrm{R}_\mathrm{a}^\lambda$
2: Initialize $red = \emptyset$, $B = A - red$, start=1.
3: for each $D_r$ in $U/D$ do
4:     Compute pessimistic decision index $pess_B(D_r)$
5:     Compute optimistic decision index $opt_B(D_r)$
6:     Compute precision function $\rho_B(D_r)$
7:     Compute $I_B^4(D_r) = -(1 - \rho_B(D_r)) \log(\rho_B(D_r))$
8: end for
9: while $FNSIJE_\alpha(red, D) \neq FNSIJE_\alpha(A, D)$ do
10:     for j =1 to |B| do
11:         Compute $FNSIJE_\alpha(red \cup a_j, D)$
12:         Select$a_j$ that satisfies max
                $\{a_j \mid FNSIJE_\alpha(red \cup a_j, D)\}$
13:     end for
14:     Let $red = red \cup a_j$, $B = A - red$, and compute $FNSIJE_\alpha(red, D)$
15: end while
16: for $k = 1$ to $red$ do
17:     Select $B_k \in red$
18:     Compute $SIG(B_k, red, D)$
19:     if $SIG(B_k, red, D) > 0$ then
20:         Let $red = red - \{B_k\}$
21:     end if
22: end for
23: return $red$
```

For this FNSIJE-KS method, the most significant impact on complexity is the calculation of parameterized fuzzy neighborhood granule. The time complexity is about $O(mn)$, where m is the size of objects and n is the size of features. This method is a loop in steps 9–15, and its time complexity is $O(n^3m)$ at worst. Suppose the size of the selected feature subset is r, the time complexity of fuzzy neighborhood granule is about $O(rm)$. As n is much greater than r for the most part. Thus, the total time complexity is close to $O(mn)$.

Experimental analysis

In this section, we conduct a series of experiments to test the feasibility and stability of FNSIJE-KS. The experiment is divided into five parts: the section "Datasets and experimental

Table 1 Ten datasets in experiments

No.	Datasets	Samples	Features	Classes
1	Glass	214	10	6
2	Ionosphere	351	33	2
3	Sonar	208	60	2
4	Wdbc	569	31	2
5	Wine	178	13	3
6	Wpbc	198	34	2
7	Colon	62	2000	2
8	DLBCL	77	5469	2
9	Leukemia	72	7129	2
10	Prostate	136	12,600	2

design” is the datasets and experimental environment used in the experiments; the section “Performance on different fuzzy neighborhood para-meters” is the impact of different fuzzy neighborhood parameters on the classification performance of our method; the section “Classification results of the UCI datasets” is the classification results of the UCI datasets; the section “Classification results of gene datasets” is the classification results of the gene datasets; the section “Statistical analysis” is the statistical analysis.

Datasets and experimental design

The experiment performs feature selection on ten public datasets including six low-dimensional UCI datasets and four high-dimensional microarray gene expression profile datasets (hereinafter referred to as gene datasets). The six UCI datasets can be downloaded at http://archive.ics.uci.edu/ml/datasets.php, and the four gene datasets can be downloaded at http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi.

Table 1 lists all datasets.

All experiments are implemented using MATLAB R2016a under Windows 10 with an Intel Core i5-3470 CPU at 3.20 GHz and 4.0 GB RAM. The classification accuracy is verified under the three classifiers KNN, CART and C4.5 in WEKA 3.8, and all parameters under the three classifiers are set to default values. To ensure the consistency of experiments, we use tenfold cross-validation in the following subsections.

Performance on different fuzzy neighborhood para-meters

This subsection analyzes the influence of different fuzzy neighborhood parameters on the classification performance of our method, and finds the most appropriate parameter for each dataset.

To reduce the time complexity originate from the four gene datasets with high dimensions, we use the Kolmogorov–Smirnov test (K–S test) for preliminary dimensionality reduction.

K–S test is a common non-parametric method used to compare whether two types of samples belong to the same distribution. It is extremely sensitive in identifying the difference of distribution morphology of two types of samples [49]. It has made development and breakthrough in the fields of cancer gene data analysis [50], emotional recognition, etc. [51]. Not only that, it also has many advantages: high calculation speed and strong operability, which can effectively reduce time complexity.

Assume that the original dataset covers two types of independent samples, denoted as positive (A) and negative (B). The total number of samples in the gene dataset is recorded as n, take a gene X in the original gene dataset as an example, the observed value of gene X is $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, refer to its eigenvalues to sort in descending order, and the corresponding order observation value can be obtained as $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. Therefore, the cumulative distribution function of the gene is [52]:

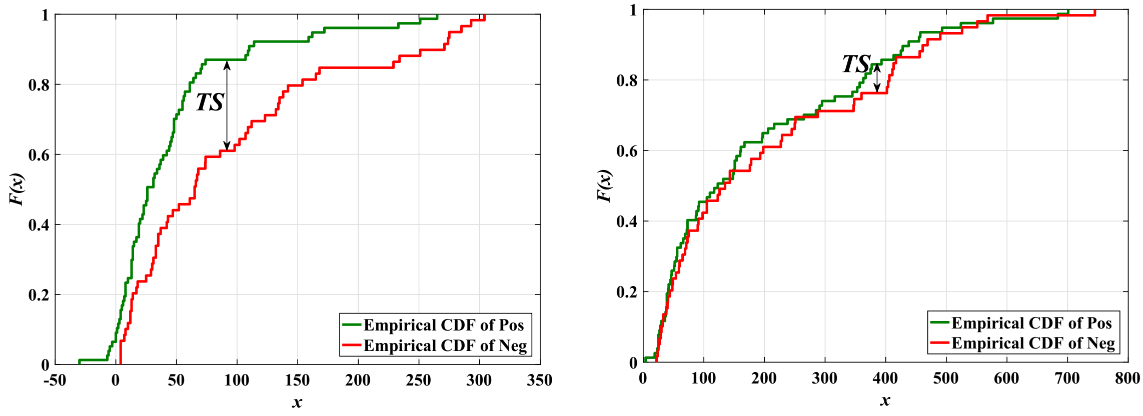
$$F(x) = \begin{cases} 0, & x < x_{(1)} \\ \dfrac{k}{n}, x(k) \le x \le x(k+1); & k = 1, 2, \ldots, n-1 \\ n, & x \le x(n). \end{cases} \tag{30}$$

On the basis of formula (30), the cumulative distribution functions $F_{A(x)}$ and $F_{B(x)}$ of the two types of samples are calculated, respectively, and then, the K–S test statistic TS is recorded as:

$$TS = \max \left| F_{A(x)} - F_{B(x)} \right| . \tag{31}$$

According to the K–S test principle, under the significance level β, if TS $<$ $\mathrm{TS}_{\mathrm{crit}}$ ($\mathrm{TS}_{\mathrm{crit}}$ is the critical value under the significance level β), it is believed that the gene is not significantly different between the positive sample and the negative sample. If TS $\ge$ $\mathrm{TS}_{\mathrm{crit}}$, it is believed that there is a significant difference between the positive sample and the negative sample of the gene at the significance level of $1-\beta$ [53]. β and the corresponding critical value $\mathrm{TS}_{\mathrm{crit}}$ (here, $s(n) = \sqrt{\dfrac{n_1 + n_2}{n_1 n_2}}$, n_1 and n_2 are the number of positive and negative genes, respectively) are shown in Table 2 [54].

It can be seen from formula (31) that the larger the value TS, the greater the difference between the positive and negative of the gene, which means that the gene has a stronger ability to distinguish between the positive and negative samples. Taking the gene dataset Prostate as an example, which contains 77 normal samples and 59 tumor samples, and



(a) The 78th gene

(b) The 3474th gene

Fig. 1 Examples of K–S test statistics

Table 2 Significance level of K–S test and corresponding critical value

β	$\mathrm{TS}_{\mathrm{crit}}/s(n)$
0.10	1.22
0.005	1.36
0.01	1.63
0.05	1.73
0.001	1.95

each sample contains 12,600 genes. As shown in Fig. 1a, b, comparing the cumulative distribution probability and K–S test statistics TS of the 78th gene and the 3474th gene (the abscissa x represents the gene value, the ordinate $F(x)$ represents the cumulative probability corresponding to x), we can see the difference of the 3474th gene in the two types of samples is significantly greater than that of the 78th gene, so 3474th gene has better distinguishing ability.

To reduce the time complexity of gene datasets, this paper uses K–S test method to preprocess the data, to reduce the dimensionality reduction. The algorithm description of K–S test method is shown in Algorithm 2.

Algorithm 2 K-S Test

```
Require: Original gene data set $OD$, significance level $\beta$.
Ensure: A selected gene subset $S$.
Initialize S =∅.
for $i = 1$ to |OD| do
    Compute $F^i_{A(x)}$ and $F^i_{B(x)}$
    Find $TS_i$ that satisfies $TS_i = \max\left|F^i_{A(x)} - F^i_{B(x)}\right|$
    if $TS_i \geq TS_{crit}(\beta)$ then
        Let $S = S \cup i$
    end if
end for
return S
```

Table 3 The number of genes in the four gene datasets after dimensionality reduction with K–S test under different significance levels

Datasets	Original	Different significance levels				
		0.1	0.005	0.01	0.05	0.001
Colon	2000	477	105	146	324	44
DLBCL	5469	2116	802	941	1640	459
Leukemia	7129	2637	903	1141	2011	288
Prostate	12,600	4240	1163	1536	3024	685

Since the K–S test can sensitively detect the difference in the distribution shape of different gene samples, it is helpful to eliminate irrelevant genes and achieve the purpose of reducing the dimensionality of the dataset. Table 3 shows the number of genes in the four gene datasets under different significance levels (0.10, 0.005, 0.01, 0.05, and 0.001) after dimension reduction by K–S test. Classification accuracy changes with the size of gene subsets in most cases. Figure 2 illustrates the change trend of the accuracy via the gene selected feature subsets under different significance levels for four gene datasets under the classifier KNN. Here, the abscissa represents five different significance levels, and the ordinate represents classification accuracy.

Adhering to the principle that the gene selected feature subset has a smaller dimension and higher classification accuracy, combined with Table 3 and Fig. 2, it can be seen that significance level β sets to 0.005 is more suitable for dataset Colon, β sets to 0.001 is more suitable for dataset DLBCL and Leukemia, and β sets to 0.1 is more suitable for dataset Prostate.

In the following portion, we will discuss the impact of the fuzzy neighborhood parameters λ on FNSIJE-KS. The change curve of classification accuracy on different λ for 10 datasets is shown in Fig. 3 (abscissa represents parameters λ

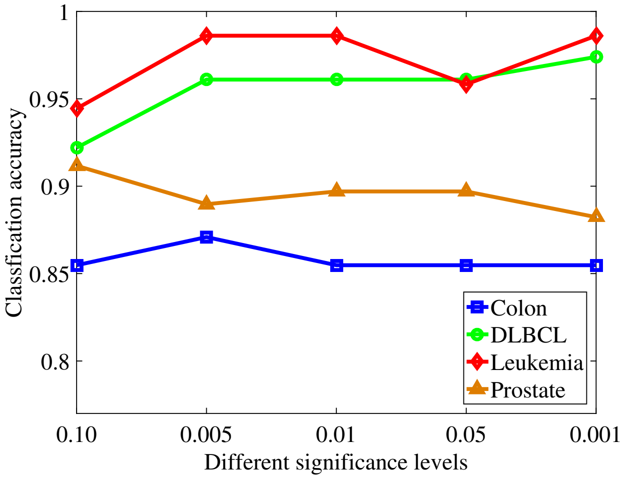


Fig. 2 Classification accuracy of gene subsets under different significance levels

Table 4 Optimal fuzzy neighborhood parameters for six UCI datasets

Datasets	KNN	CART
Glass	0.95	0.95
Ionosphere	0.15	0.3
Sonar	0.65	0.45
Wdbc	0.4	0.05
Wine	0.5	0.65
Wpbc	0.65	0.65

Table 5 Optimal fuzzy neighborhood parameters for four gene datasets

Datasets	KNN	CART	C4.5
GDLBCL	0.35	0.3	0.35
Colon	0.35	0.2	0.65
Prostate	0.7	0.75	0.7
Leukemia	0.25	0.1	0.1

and ordinate represents classification accuracy). We design the value of λ to vary from 0 to 1, with a step size of 0.05. The classification accuracy of the six UCI datasets under the classifiers KNN and CART is shown in Fig. 3a–f; the accuracy of the four gene datasets under the classifiers KNN, CART, and C4.5 is shown in Fig. 3g–j.

It can be seen from Fig. 3 that the change of λ does have a certain impact on the classification performance of FNSIJE-KS, but all datasets can achieve a high classification accuracy in a wider range of λ. Tables 4 and 5 list the optimal fuzzy neighborhood parameters selected for each UCI dataset and gene dataset under different classifiers, respectively.

Classification results of the UCI datasets

In this subsection, to verify the classification effectiveness of the FNSIJE-KS method on UCI datasets, we compare the size and classification accuracy of selected feature subsets by FNSIJE-KS and three existing methods. The three methods include FNRS (method based on fuzzy neighborhood rough set) [32], FNCE (method based on fuzzy neighborhood conditional entropy) [37], and FNPME-FS (method based on fuzzy neighborhood pessimistic multigranulation entropy) [36].

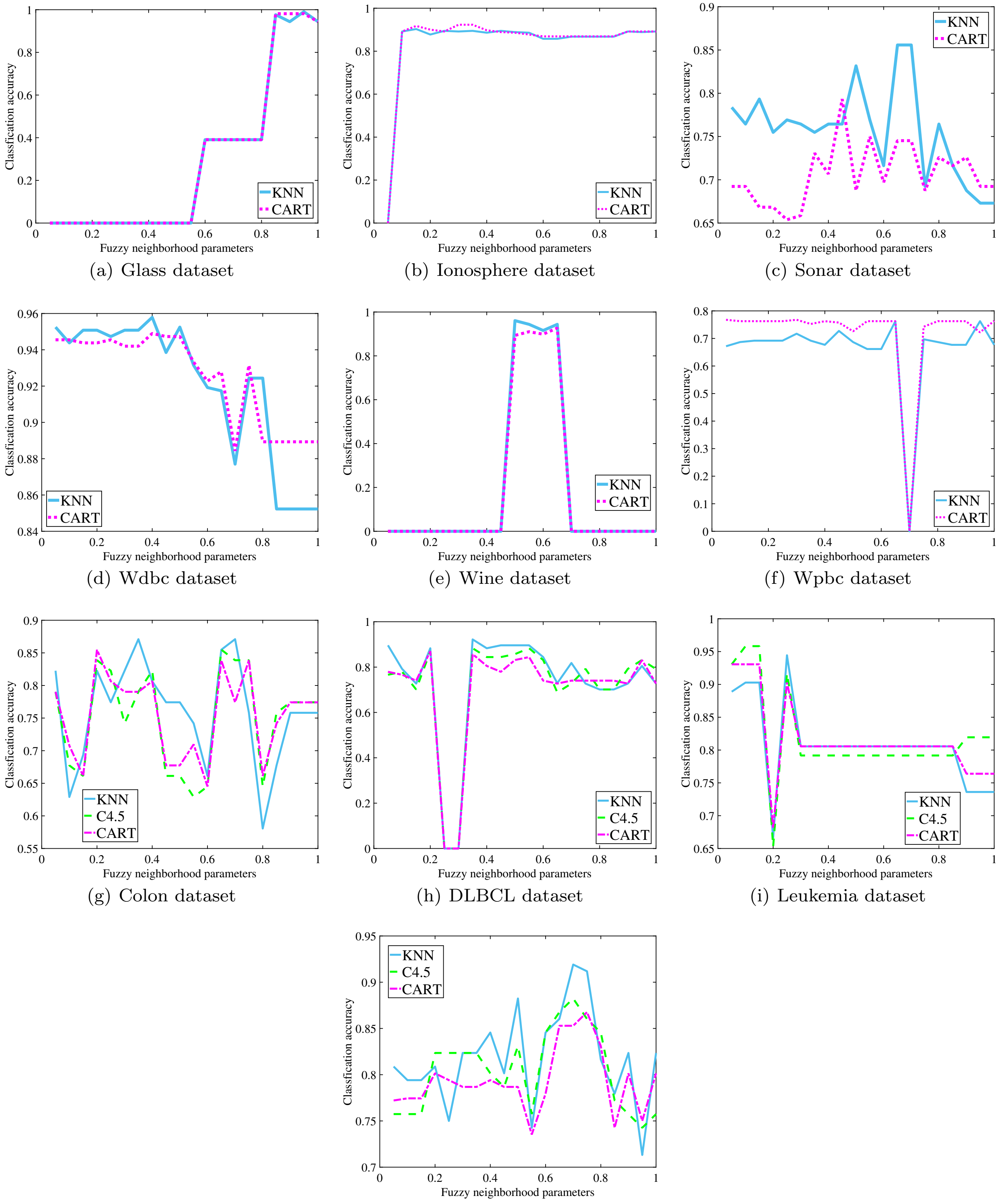
In the first portion of the subsection, we use the obtained fuzzy neighborhood parameters λ, let the four feature selection methods use ten-fold cross-validation on six UCI datasets, and the average size of the original datasets and the selected feature subsets by different methods are exhibited in Table 6. The optimal selected feature subsets by FNSIJE-KS for the six UCI datasets under the classifiers KNN and CART are shown in Table 7.

Table 6 describes the average size of the selected feature subsets by the four methods. In Table 6, the bolded numbers show that the size of the reduced datasets is the least with respect to other methods. The selected feature subsets by FNSIJE-KS on the datasets Glass, Ionosphere, and Wine reach the minimum size, which are 2.1, 2.2, and 4.75, respectively. In short, comparing with FNRS, FNCE, and FNPME-FS, the mean size of selected feature subsets using FNSIJE-KS is the least for six UCI datasets.

The second portion pays attention to the effectiveness of classification for FNSIJE-KS. Tables 8 and 9 list classification accuracy of selected feature subsets by FNSIJE-KS and other three feature selection methods (FNRS [32], FNCE [37], and FNPME-FS [36]) under classifiers KNN and CART, respectively. In Tables 8 and 9, the bolded numbers indicate that the classification accuracy of the selected feature subsets is the best with respect to other methods.

No method always outperforms other methods under different classifiers and learning tasks. As observed from the results of four methods demonstrate in Tables 6, 8, and 9 illustrate the differences among the four methods. From Tables 6 and 8, it can be seen that FNSIJE-KS has the highest classification accuracy of the selected feature subsets on datasets Glass and Wpbc, 99.07 and 77.27%, respectively, under the classifier KNN. Furthermore, the average size of the selected feature subset by FNSIJE-KS is the 2.8, 10.2, and 3.6 smaller than those by FNRS, FNCE, and FNPME-FS on Glass dataset, respectively. The selected feature subsets by FNSIJE-KS and FNPME-FS have the same classification accuracy of 90.31% on dataset Ionosphere, and the selected feature subset by FNSIJE-KS is 9.4 less than those selected by FNPME-FS.

Similarly, from Tables 6 and 9, it is clear that the average classification accuracy of FNSIJE-KS is larger than FNRS,



(a) Glass dataset

(b) Ionosphere dataset

(c) Sonar dataset

(d) Wdbc dataset

(e) Wine dataset

(f) Wpbc dataset

(g) Colon dataset

(h) DLBCL dataset

(i) Leukemia dataset

(j) Prosate dataset

Fig. 3 Classification accuracy under different fuzzy neighborhood parameter values on ten datasets

Table 6 Number of selected features with four feature selection methods

Datasets	Original	FNRS	FNCE	FNPME-FS	FNSIJE-KS
Glass	15	4.9	12.3	5.7	**2.1**
Ionosphere	33	3.2	16	11.6	**2.2**
Sonar	60	**1.5**	21.7	17.5	7.95
Wdpc	31	11.8	**4.2**	7	11.5
Wine	13	9.1	5.9	6	**4.75**
Wpbc	34	9	8.1	**5.7**	7.9
Mean	31	6.5	11.37	8.91	**6.1**

Table 7 Optimal subset with FNSIJE-KS for six UCI datasets

Datasets	KNN	CART
Glass	{1}	{1}
Ionosphere	{30 4 9}	{10 25 22}
Sonar	{11 22 31 36 39 48 54 17 35 37 38 13 19}	{11 13 26 33 34 50}
Wdpc	{1 2 5 7 16 22 25 28 9 21}	{1 2 5 9 12 16 18 22 25 27 28 21}
Wine	{1 2 10 12 13 4}	{1 8 10 7}
Wpbc	{3 6 13 21 26 1}	{3 6 13 21 26 1}

Table 8 Classification accuracy with four methods on classifier KNN

Datasets	Original	FNRS	FNCE	FNPME-FS	FNSIJE-KS
Glass	0.9112	0.9393	0.8627	0.9393	**0.9907**
Ionosphere	0.8245	0.8664	0.8823	**0.9031**	**0.9031**
Sonar	0.8653	0.5385	0.8001	**0.8846**	0.8558
Wdpc	0.9124	0.9174	**0.9825**	0.9279	0.9579
Wine	0.9195	**0.9720**	0.9073	0.9719	0.9607
Wpbc	0.6969	0.7273	0.6872	0.7222	**0.7727**
Mean	0.8550	0.8268	0.8537	0.8915	**0.9068**

Table 9 Classification accuracy with four methods on classifier CATR

Datasets	Original	FNRS	FNCE	FNPME-FS	FNSIJE-KS
Glass	0.9672	**0.9813**	0.8475	**0.9813**	**0.9813**
Ionosphere	0.8641	0.8832	0.8053	0.9117	**0.9231**
Sonar	0.7115	0.6298	0.7616	0.7789	**0.7933**
Wdpc	0.8595	0.9438	0.8807	0.9314	**0.9455**
Wine	0.8932	0.8932	0.8249	**0.9326**	0.9270
Wpbc	0.7121	0.7275	0.6143	0.7374	**0.7626**
Mean	0.8346	0.8431	0.7891	0.8789	**0.8888**

FNCE, and FNPME-FS on almost UCI datasets under classifier CART , except only one dataset Wine. In addition, FNPME-FS selects fewer features than FNRS and FNPME-FS when its classification accuracy is equal to FNRS and FNPME-FS on dataset Glass. According to Tables 8 and 9, comparing FNSIJE-KS with FNRS, FNCE, and FNPME-FS, the mean classification accuracy of FNSIJE-KS has been improved under two classifiers KNN and CART on the UCI datasets.

In summary, the FNSIJE-KS method is easy to eliminate redundant features as a whole and displays better classification performance than other compared methods on UCI datasets.

Classification results of gene datasets

This subsection is dealing with the classification effectiveness of the FNSIJE-KS method on gene datasets.

In the first portion of this subsection, we compare the size of selected gene subsets by the FNSIJE-KS method and three existing methods (FNRS [32], FNCE [37], and FNPME-FS [36]).

We use the obtained λ, let the four feature selection methods use tenfold cross-validation on four gene datasets, and the average size of the original datasets and the selected gene subsets by different methods are described in Table 10. In Table 10, the bolded numbers show that the size of the reduced gene datasets is the least with respect to other methods. Then, the optimal selected gene subsets by FNSIJE-KS for the four gene datasets under three classifiers KNN, C4.5, and CART are demonstrated in Table 11.

As is clearly shown from Table 10, for three gene datasets Colon, DLBCL, and Leukemia, FNSIJE-KS is significantly superior to the other three methods. Moreover, the mean size of the selected gene subset by FNSIJE-KS is less than those by FNRS, FNCE, and FNPME-FS. Briefly, FNSIJE-KS can select less genes for gene datasets.

In the second portion of this subsection, we discuss the classification performance of the selected gene subsets by FNSIJE-KS and other four feature selection methods under classifiers KNN and C4.5. Among them, four feature selection methods include FNRS [32], FNCE [37], FNPME-FS [36], and EGGS (entropy gain-based gene selection method) [55].

Let five feature selection methods use tenfold cross-validation on four gene datasets, and the classification accuracy of selected gene subsets under the classifiers KNN and C4.5 are illustrated in Tables 12 and 13, respectively. In Tables 12 and 13, the bolded numbers indicate that the classification accuracy of the selected gene subsets is the best with respect to other four methods.

From Tables 10 and 12, the classification accuracy of selected gene subsets by FNSIJE-KS is 13.23, 6.75, 2.88, and 27.97%, respectively, which is higher than those selected by FNRS, FNCE, FNPME-FS, and EGGS under classifier KNN on dataset Prostate. Both FNSIJE-KS and FNPME-FS have the highest classification accuracy of selected gene subsets on dataset Leukemia, which is 94.44%; furthermore, the selected gene subset by FNSIJE-KS is less than those of FNPME-FS. However, FNSIJE-KS selects 2.5 and 2.8 fewer genes than FNCE and FNPME-FS on datasets Colon and DLBCL, respectively, although its classification accuracy is slightly lower than FNCE and FNPME-FS.

In the same way, from Tables 10 and 13, the classification accuracy of selected gene subset by FNSIJE-KS achieves highest value of 85.48% under classifier C4.5 on dataset Colon, which is 3.22, 11.29, 4.83, and 20.84% higher than those by the other four methods, respectively. The classification accuracy of selected gene subsets by FNSIJE-KS and FNPME-FS reaches the highest value at the same time on dataset Prostate, which is 88.23%, and the selected gene subset by FNSIJE-KS is 4.45 fewer than those by FNPME-FS.

In short, from Tables 12 and 13, comparing with FNRS, FNCE, FNPME-FS, and EGGS, the mean classification accuracy of selected gene subsets by the FNSIJE-KS method has been improved under classifiers KNN and C4.5, both reaching the highest level.

This final portion is verify the classification performance of the FNSIJE-KS method under the classifier CART for four gene datasets. Three feature selection methods compared with FNSIJE-KS include FNRS [32], FNCE [37], and FNPME-FS [36].

Let the four feature selection methods use the tenfold cross-validation method on four gene datasets, and classification accuracy of selected gene subsets under the classifier CART is shown in Table 14. In Table 14, the bolded numbers indicate that the classification accuracy of the selected gene subsets is the best with respect to other three methods.

As seen from the average classification result illustrated in Table 14, and combining Table 10, our method FNSIJE-KS achieves the largest value for the gene dataset Colon under the classifier CART. Meanwhile FNSIJE-KS selects fewest genes than three other feature selection methods. FNSIJE-KS and FNRS have the highest classification performance on selected gene subset of dataset Prostate, which is 86.03%. Comparing FNSIJE-KS with FNRS, FNCE, and FNPME-FS, the mean classification accuracy of the selected gene subset by FNSIJE-KS reached the highest level under the classifier CART.

In summary, the FNSIJE-KS method can eliminate redundant features in the mass and exhibits better classification performance than other compared methods on the gene datasets.

Statistical analysis

To systematically study the statistical performance on classification accuracy of all compared methods in this paper, Friedman test and corresponding post hoc will be carried out in this subsection. Friedman statistic [56] is conveyed as:

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

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

where r_i is the mean ranking of method on all datasets, n and k represent the number of datasets and methods, respectively, and F is a F-distribution under $(k-1)$ and $(k-1)\,(n-1)$ freedom degrees.

Table 10 Number of selected genes with four feature selection methods

Datasets	Original	FNRS	FNCE	FNPME-FS	FNSIJE-KS
Colon	2000	2.5	4.9	7	**2.4**
DLBCL	5469	5.2	16.3	6.4	**3.6**
Leukemia	7129	4.7	8.7	6	**2.2**
Prostate	12,600	**3.5**	13.7	11.5	7.05
Mean	6799.5	3.98	10.9	7.73	**3.81**

Table 11 Gene subset with FNSIJE-KS for four gene datasets

Datasets	KNN	C4.5	CART
Colon	{377 493 765 249}	{377 265 765}	{765 897}
DLBCL	{1939 3769 920}	{1939 3769 920}	{1939 3769 398}
Leukemia	{1882 760 2258}	{1882 255}	{1882 255}
Prostate	{8554 9850 3474 4173 3793 7756 48 }	{8554 9850 3474 4173 3793 7756 48}	{8554 9850 3474 4173 3794 7756}

Table 12 Classification accuracy with five methods on classifier KNN

Datasets	Original	FNRS	FNCE	FNPME-FS	EGGS	FNSIJE-KS
Colon	0.7903	0.8548	**0.9231**	0.8548	0.6493	0.8710
DLBCL	0.8701	0.8831	0.8917	**0.9611**	0.8540	0.9221
Leukemia	0.7344	0.9306	0.8970	**0.9444**	0.6292	**0.9444**
Prostate	0.7647	0.7868	0.8516	0.8903	0.6394	**0.9191**
Mean	0.7899	0.8638	0.8909	0.9127	0.6929	**0.9142**

In this first portion, the classification accuracy of the four feature selection methods in Tables 8 and 9 under the classifiers KNN and CART is statistically analyzed.

First, for the classification accuracy on six UCI datasets in Tables 8 and 9, the Friedman tests are achieved by the comparison of FNSIJE-KS with FNRS, FNCE, and FNPME-FS. Tables 15 and 16 list the mean rankings of the four methods under the classifiers KNN and CART, respectively. Calling $icdf$ in MATLAB 2016a calculates critical value $F(3, 15) = 2.4898$, when $\alpha = 0.1$.

Assuming that the four methods are equivalent in classification performance, the value of Friedman statistics should not exceed the critical value $F(3, 15)$. On the contrary, the four methods differ significantly in feature selection performance.

Table 13 Classification accuracy with five methods on classifier C4.5

Datasets	Original	FNRS	FNCE	FNPME-FS	EGGS	FNSIJE-KS
Colon	0.7419	0.8226	0.7419	0.8065	0.6464	**0.8548**
DLBCL	0.7922	**0.9351**	0.8181	0.8832	0.8264	0.8832
Leukemia	0.8143	0.9306	0.8754	**0.9583**	0.7333	0.9306
Prostate	0.6400	0.8603	0.7945	**0.8823**	0.5913	**0.8823**
Mean	0.7471	0.8872	0.7945	0.8825	0.6993	**0.8877**

Table 14 Classification accuracy with four methods on classifier CART

Datasets	Original	FNRS	FNCE	FNPME-FS	FNSIJE-KS
Colon	0.5967	0.8387	0.7964	0.8064	**0.8548**
DLBCL	0.8181	0.8442	0.7698	**0.9091**	0.8961
Leukemia	0.7556	0.9306	0.8342	**0.9722**	0.9306
Prostate	0.6917	**0.8603**	0.8046	0.8224	**0.8603**
Mean	0.7155	0.8685	0.8013	0.8775	**0.8855**

Table 15 Rank of four methods with classifier KNN on UCI datasets

Datasets	FNRS	FNCE	FNPME-FS	FNSIJE-KS
Glass	2.5	4	2.5	1
Ionosphere	4	3	1.5	1.5
Sonar	4	3	1	2
Wdpc	4	1	3	2
Wine	1	4	2	3
Wpbc	2	4	3	1
Mean	2.92	3.17	2.17	1.75

Table 16 Rank of four methods with classifier CART on UCI datasets

Datasets	FNRS	FNCE	FNPME-FS	FNSIJE-KS
Glass	2	4	2	2
Ionosphere	3	4	2	1
Sonar	4	3	2	1
Wdpc	2	4	3	1
Wine	3	4	1	2
Wpbc	3	4	2	1
Mean	2.83	3.83	2	1.33

According to Friedman statistics, we can get that $F = 1.8431$ under the classifier KNN and $F = 11.0948$ under the classifier CART. Obviously, the value under the classifier CART is greater than the critical value $F(3, 15)$. This phenomenon indicates that the FNRS, FNCE, FNPME-FS, and FNSIJE-KS are significantly different.

Second, we need to do a post hoc test for the difference between the four methods. The post hoc test used here is Nemenyi test [57]. The statistics first needs to determine the critical value of the distance between the mean ranking values, which is denoted as:

$$\mathrm{CD}_\alpha = q_a \sqrt{\frac{k\,(k+1)}{6n}}. \tag{34}$$

Here, q_a represents the critical tabulated value of this test. It can be obtained that $q_{0.1} = 2.291$ when $\alpha = 0.1$ and the number of methods is 4. Thus, $\mathrm{CD}_{0.1} = 1.7076$.

If the distance of the corresponding mean rankings between the two methods is larger than the critical distance $\mathrm{CD}_{0.1}$, the two methods are considered to be significantly different. From Table 15, the distance between mean rankings of FNSIJE-KS to FNCE is 1.95, which is greater than 1.7076 under the classifier KNN. Therefore, the Nemenyi tests show that FNSIJE-KS is significantly greater than FNCE at $\alpha = 0.1$. However, the distances between mean rankings of FNSIJE-KS to FNRS and FNPME-FS are less than 1.7076. This phenomenon repeals that there is no significant difference between FNSIJE-KS to FNRS and FNPME-FS. For the classifier CART, the distance between mean rankings of FNSIJE-KS to FNCE is 2.5, which is greater than 1.7076; it is obviously obtained that FNSIJE-KS is significantly greater than FNCE at $\alpha = 0.1$. However, the distances between mean rankings of FNSIJE-KS to FNRS and FNPME-FS are less than 1.7076. This reveals that there is no significant difference between FNSIJE-KS to FNRS and FNPME-FS.

Table 17 Rank of five methods with classifier KNN on gene datasets

Datasets	FNRS	FNCE	FNPME-FS	EGGS	FNSIJE-KS
Colon	3.5	1	3.5	5	2
DLBCL	4	3	1	5	2
Leukemia	3	4	1.5	5	1.5
Prostate	4	3	2	5	1
Mean	3.63	2.75	2	5	1.63

Table 18 Rank of five methods with classifier C4.5 on gene datasets

Datasets	FNRS	FNCE	FNPME-FS	EGGS	FNSIJE-KS
Colon	2	4	3	5	1
DLBCL	1	5	2.5	4	2.5
Leukemia	2.5	4	1	5	2.5
Prostate	3	4	1.5	5	1.5
Mean	2.13	4.25	2	4.75	1.88

The following portion is executed on five different methods FNRS, FNCE, FNPME-FS, and FNSIJE-KS for four gene datasets from Tables 12 and 13 under two classifiers KNN and C4.5. Tables 17 and 18 list the mean rankings of the five methods under the classifiers KNN and C4.5, respectively.

After calculation, $F(4, 12) = 2.4801$ when $\alpha = 0.1$, and then, we can compute that $F = 8.5221$ for Classifier KNN and $F = 10.0225$ for Classifier C4.5. It is apparent to obtain that the two values are greater than the critical value $F(4, 12)$. This result exhibits that five methods are significantly different.

Then, we perform Nemenyi tests on the five feature selection methods in Tables 12 and 13 under the classifiers KNN and C4.5. It is easily obtained that $q_{0.1} = 2.459$ when the number of methods is 5 and $\alpha = 0.1$. Thus, it can calculate that $\mathrm{CD}_{0.1} = 2.7492$ $(k = 5, n = 4)$ according to formula (34). For classifier KNN, the distances between mean rankings of FNSIJE-KS to other four methods are less than 2.7492, and this result shows that five methods are no significant difference. For classifier C4.5, the distance between mean rankings of FNSIJE-KS to EGGS is greater than 2.7492. Thus, the Nemenyi tests show that FNSIJE-KS is significantly better than EGGS at $\alpha = 0.1$.

Table 19 Rank of five methods with classifier CART on gene datasets

Datasets	FNRS	FNCE	FNPME-FS	FNSIJE-KS
Colon	2	4	3	1
DLBCL	3	4	1	2
Leukemia	2.5	4	1	2.5
Prostate	1.5	4	3	1.5
Mean	2.25	4	2	1.75

In the final portion of this subsection, we discuss the statistical results of the four compared methods for four gene datasets from Table 14 under the classifier CART. The mean rankings of four methods under classifier CART are computed and demonstrated in Table 19.

It is apparently obtained that $F(3, 9) = 2.8129$ when $\alpha = 0.1$, and then, we can compute that 5.00 for classifier CART. This exhibits that the values is greater than the critical value $F(3, 9)$. Thus, four methods are significantly different under classifier CART.

Then, we perform Nemenyi test on the four feature selection methods in Table 14 under the classifier CART. It is easily obtained that $q_{0.1} = 2.291$ when the number of methods is 4 and $\alpha = 0.1$. Thus, it can calculate that 2.0914 ($k = 4, n = 4$) according to formula (34). For the classifier CART, the distance between mean rankings of FNSIJE-KS to FNCE is greater than 2.0914. Thus, the Nemenyi test shows that FNSIJE-KS is significantly better than FNCE at $\alpha = 0.1$.

To sum up, FNSIJE-KS is better than the other compared methods in the Friedman statistic test.

Conclusion

In neighborhood decision system, the traditional dependency function based on FNRS only considers the classification information contained in the lower approximation while ignoring the classification information contained in the upper approximation. This construction method may lead to the loss of some information. To solve this problem, we propose an improved model based on self-information and information entropy. First, from the algebra view, the four types of fuzzy neighborhood self-information measures are defined using the upper and lower approximations in FNRS and combining with the concept of self-information, and the related properties are discussed in detail. It is proved that the fourth measure changes more with the change of feature subset, which is helpful to select the optimal feature subset. The significance of features based on algebra view can interpret the influence of the features included in feature subset. Second, the fuzzy neighborhood joint entropy is given from the information view. Information view-based feature significance can illustrate the importance on features from uncertainty classification. Then, from both algebra and information views, a model based on FNSIJE is proposed to analyze the noise, uncertainty, and ambiguity of the neighborhood decision system. Third, a new forward search method for feature selection is designed, which makes that the selected feature subsets have higher classification performance. All the designed experiments demonstrate that our FNSIJE-KS method can select fewer features for some low-dimensional UCI datasets and high-dimensional gene datasets, which has the optimal classification performance. In the future work, we will further study more effective search strategies based on FNSIJE to balance the size of selected feature subsets and classification accuracy as much as possible, and explore on incomplete information systems.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant (61976082, 61976120, and 62002103), and in part by the Key Scientific and Technological Projects of Henan Province under Grant 202102210165.

References

1. Hoque N, Singh M, Bhattacharyya DK (2018) EFS-MI: an ensemble feature selection method for classification. Complex Intell Syst 4:105–118
2. Chen J-K, Mi K-S, Lin Y-J (2020) A graph approach for fuzzy-rough feature selection. Fuzzy Sets Syst 391:96–116
3. Yu N, Wu M-J, Liu J-X, Zheng C-H, Xu Y (2020) Correntropy-based hypergraph regularized NMF for clustering and feature selection on multi-cancer integrated data. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2020.3000799
4. Capo M, Perez A, Lozano JA (2020) A cheap feature selection approach for the K-means algorithm. IEEE Trans Netw. https://doi.org/10.1109/TNNLS.2020.3002576
5. Ding W, Lin C-T, Prasad M, Cao Z, Wang J-D (2017) A layered-coevolution-based attribute-boosted reduction using adaptive quantum behavior PSO and its consistent segmentation for neonates brain tissue. IEEE Trans Fuzzy Syst 26(3):1177–1191
6. Haq AU, Zhang D, Peng H, Rahman SU (2019) Combining multiple feature-ranking techniques and clustering of variables for feature selection. IEEE Access 7:151482–151492
7. Dikshit-Ratnaparkhi A, Bormane D, Ghongade R (2020) A novel entropy-based weighted attribute selection in enhanced multicriteria decision-making using fuzzy TOPSIS model for hesitant fuzzy

rough environment. Complex Intell Syst. https://doi.org/10.1007/s40747-020-00187-8

8. Sun L, Zhang X-Y, Qian Y-H, Xu J-C, Zhang S-G (2019) Feature selection using neighbor-hood entropy-based uncertainty measures for gene expression data classification. Inf Sci 502:18–41
9. Pudaruth S, Soyjaudah KMS, Gunputh RP (2018) An innovative multi-segment strategy for the classification of legal judgments using the k-nearest neighbour classifier. Complex Intell Syst 4:1–10. https://doi.org/10.1007/s40747-017-0042-z
10. Wang C-Z, Hu Q-H, Wang X-Z, Chen D-G, Qian Y-H, Dong Z (2017) Feature selection based on neighborhood discrimination index. IEEE Trans Netw 29(8):2986–2999
11. Sun L, Zhang X-Y, Xu J-C, Wang W, Liu R-N (2017) A gene selection approach based on the fisher linear discriminant and the neighborhood rough set. Bioengineered 9:144–151
12. Lang G-M, Li Q-G, Cai M-J, Yang T, Xiao Q-M (2017) Incremental approaches to know-ledge reduction based on characteristic matrices. Int J Mach Learn Cybern 8(1):203–222
13. Dong L-J, Chen D-G, Wang N-L, Lu Z-H (2020) Key energy-consumption feature selection of thermal power systems based on robust attribute reduction with rough sets. Inf Sci 532:61–71
14. Wang C-Z, Shi Y-P, Fan X-D, Shao M-W (2018) Attribute reduction based on k-nearest neighborhood rough sets. Int J Approx Reason 106:18–31
15. Dong H-B, 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
16. Yenny VR (2019) Maximal similarity granular rough sets for mixed and incomplete information systems. Soft Comput 23(13):4617–4631
17. Zhang X, Mei C-L, Chen D, Yang Y-Y, Li J-H (2019) Active incremental feature selection using a fuzzy rough set-based information entropy. IEEE Trans Fuzzy Syst. https://doi.org/10.1109/TFUZZ.2019.2959995
18. Pawlak Z (1982) Rough sets. Int Comput Inf Sci 11(5):341–356
19. Deng Z-X, Zheng Z-L, Deng D-Y, Wang T-X, He Y-R, Zhang D-W (2020) Feature selection for multi-label learning based on f-neighborhood rough sets. IEEE Access 8:39678–39688
20. Zhan J-M, Jiang H-B, Yao Y-Y (2020) Covering-based variable precision fuzzy rough sets with PROMETHEE-EDAS methods. Inf Sci 538:314–336
21. Che X-Y, Chen D-G, Mi J-S (2020) A novel approach for learning label correlation with application to feature selection of multi-label data. Inf Sci 512:795–812
22. Zhang Q-H, Zhao F, Yang J, Wang G-Y (2020) Three-way decisions of rough vague sets from the perspective of fuzziness. Inf Sci 523:111–132
23. Liang J-Y, Wang F, Dang C-Y, Qian Y-H (2014) A group incremental approach to feature selection applying rough set technique. IEEE Trans Knowl Data Eng 26(2):294–308
24. Xie X-J, Qian X-L (2018) A novel incremental attribute reduction approach for dynamic incomplete decision systems. Int J Approx Reason 93:443–462
25. Sang S-S, Liu L-Z, Wang S-W (2020) An incremental attribute reduction algorithm for the dominant relationship rough set. Comput Sci 47(08):137–143
26. Chen H-M, Li T-R, Fan X, Luo C (2019) Feature selection for imbalanced data based on neighborhood rough sets. Inf Sci 483:1–20
27. Wu W-Z, Shao M-W, Wang X (2017) Using single axioms to characterize (S, T)-intuitionistic fuzzy rough approximation operators. Int J Mach Learn Cybern 10:27–42
28. Wang C-Z, Qi Y-L, Shao M-W, Hu Q-H, Chen D-G, Qian Y-H, Lin Y-J (2017) A fitting model for feature selection with fuzzy rough sets. IEEE Trans Fuzzy Syst 25(4):741–753
29. Chen L-L, Chen D-G, Wang H (2019) Fuzzy kernel alignment with application to attribute reduction of heterogeneous data. IEEE Trans Fuzzy Syst 27(7):1469–1478
30. Dubois D, Prade H (1990) Rough fuzzy sets and fuzzy rough sets. Int J Gen Syst 17:191–209
31. Hu Q-H, Liu J-F, Yu D-R (2008) Mixed feature selection based on granulation and approximation. Knowl Based Syst 21(4):294–304
32. Wang C-Z, Shao M-W, He Q, Qian Y-H, Qi Y-L (2016) Feature subset selection based on fuzzy neighborhood rough sets. Knowl Based Syst 111(1):173–179
33. Qian Y-H, Li S-Y, Liang J-Y, Shi Z-Z, Wang F (2014) Pessimistic rough set-based decisions: a multi-granulation fusion strategy. Inf Sci 264:196–210
34. Shreevastava S, Tiwari AK, Som T (2018) Intuitionistic fuzzy neighborhood rough set model for feature selection. Int J Fuzzy Syst Appl 7(2):75–84
35. Yue X-D, Chen Y-F, Miao D-Q, Fujita H (2020) Fuzzy neighborhood covering for three-way classification. Inf Sci 507:795–808
36. Sun L, Wang L-Y, Ding W-P, Qian Y-H, Xu J-C (2021) Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy neighborhood multigranulation rough sets. IEEE Trans Fuzzy Syst 29(1):19–33
37. Xu J-C, Wang Y, Mu H-Y, Huang F-Z (2018) Feature genes selection based on fuzzy neighborhood conditional entropy. J Intell Fuzzy Syst 36(1):117–126
38. Wang C-Z, Huang Y, Shao M-W, Hu Q-H, Chen D-G (2020) Feature selection based on neighborhood self-information. IEEE Trans Cybern 50(9):4031–4042
39. Sun L, Wang L-Y, Xu J-C, Zhang S-G (2019) A neighborhood rough sets-based attribute reduction method using Lebesgue and entropy measures. Entropy. https://doi.org/10.3390/e21020138
40. Wang C-Z, Huang Y, Shao M-W, Fan X-D (2019) Fuzzy rough set-based attribute reduction using distance measures. Knowl Based Syst 164:205–212
41. Liu Y, Huang W-L, Jiang Y-L, Zeng Z-Y (2014) Quick attribute reduct algorithm for neighborhood rough set model. Inf Sci 271:65–81
42. Hu C-X, Zhang L, Wang B-J, Zhang Z, Li F-Z (2019) Incremental updating knowledge in neighborhood multi-granulation rough sets under dynamic granular structures. Knowl Based Syst 163:811–829
43. Fan X-D, Zhao W-D, Wang C-Z, Huang Y (2018) Attribute reduction based on max-decision neighborhood rough set model. Knowl Based Syst 151:16–23
44. Zeng K, She K, Niu X-Z (2013) Multi-granulation entropy and its applications. Entropy 15(6):2288–2302
45. Wang G-Y (2003) Rough reduction in algebra view and information view. Int J Intell Syst 18(6):679–688
46. Sun L, Wang L-Y, Ding W-P, Qian Y-H, Xu J-C (2020) Neighborhood multigranulation rough sets-based attribute reduction using Lebesgue and entropy measures in incomplete neighborhood decision systems. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2019.105373
47. Shannon C-E (2001) A mathematical theory of communication. Bell Syst Tech J 5(3):3–55
48. Zadeh LA (1968) Probability measures of fuzzy events. J Math Anal Appl 23(2):421–427
49. Al-Labadi L, Zarepour M (2017) Two-sample Kolmogorov–Smirnov test using a Bayesian nonparametric approach. Math Methods Stat 26(3):212–225
50. Xie J-Y, Hu Q-F, Dong Y-F (2016) Gene selection algorithm combined with K–S test and mRMR. Appl Res Comput 33(4):1013–1018 **(in Chinese)**
51. Zhang L-J, Li Z-J (2009) Gene selection in cancer classification problems with microarray data. J Comput Res Dev 46:784–802

52. Huang S-G, Yeo AA, Li S-D (2007) Modification of Kolmogorov–Smirnov test for DNA content data analysis through distribution alignment. Assay Drug Dev Technol 5(5):663–672
53. Young IT (1977) Proof without prejudice: use of Kolmogorov–Smirnov test for the analysis of histograms from flow systems and other sources. J Histochem Cytochem 25(7):935–941
54. Buckland WR (1969) Handbook of tables for probability and statistics. J R Stat Soci Ser A (Gen) 132(3):452–452
55. Chen Y-M, Zhang Z-J, Zheng J-Z, Ma Y, Xue Y (2017) Gene selection for tumor classification using neighborhood rough sets and entropy measures. J Biomed Inform 67:59–68
56. Friedman M (1940) A comparison of alternative tests of significance for the problem of *m* rankings. Ann Math Stat 11(1):86–92
57. Demsar J, Schuurmans D (2006) Statistical comparison of classifiers over multiple data sets. J Mach Learn Res 7:1–30
58. Liu L, Liu Y-J, Chen A-J, Tong S-C, Philip Chen CL (2020) Integral barrier Lyapunov function-based adaptive control for switched nonlinear systems. Sci China Inf Sci 63(3):212–225
59. Liu L, Li X-S, Liu Y-J, Tong S-C (2021) Neural network based adaptive event trigger control for a class of electromagnetic suspension systems. Control Eng Pract. https://doi.org/10.1016/j.conengprac.2020.104675
60. Liu L, Li X-S (2020) Event-triggered tracking control for active seat suspension systems with time-varying full-state constraints. IEEE Trans Syst Man Cybern. https://doi.org/10.1109/TSMC.2020.3003368

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