Feature selection using Lebesgue and entropy measures for incomplete neighborhood decision systems ☆

Lin Sun [a,*], Lanying Wang [a], Yuhua Qian [b], Jiucheng Xu [a,*], Shiguang Zhang [a,c]

[a] *College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China*
[b] *Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China*
[c] *School of Computer Science and Technology, Tianjin University, Tianjin 300072, China*

ARTICLE INFO

ABSTRACT

Feature selection for mixed and incomplete data in terms of numerical and categorical features with missing values has currently gained considerable attention. The development of the neighborhood rough sets-based feature selection method is an important step in improving classification performance, especially in incomplete data with mixed continuous numerical and categorical features. In this paper, a novel feature selection method based on the neighborhood rough sets using Lebesgue and entropy measures in incomplete neighborhood decision systems is proposed, and the method has the capacity to handle mixed and incomplete datasets; further, it can simultaneously maintain the original classification information. First, a Lebesgue measure based on the neighborhood tolerance class is developed to study the positive region and dependency degree. To thoroughly analyze the uncertainty, noise and incompleteness of incomplete neighborhood decision systems, some neighborhood tolerance entropy-based uncertainty measures are presented based on Lebesgue and entropy measures. Then, by combining an algebraic view with an information view in neighborhood rough sets, the neighborhood tolerance dependency joint entropy is defined in incomplete neighborhood decision systems. Moreover, all the corresponding properties are discussed, and the relationships among these measures are established to meaningfully convey the knowledge essence and investigate the uncertainty of incomplete neighborhood decision systems. Finally, for all high-dimensional datasets, the Fisher score method is used to preliminarily eliminate irrelevant features to significantly reduce the computational complexity, and a heuristic feature selection algorithm is designed to improve the classification performance of mixed and incomplete datasets. Experiments under an instance and fifteen public datasets demonstrate that the proposed feature selection method is effective in selecting the most relevant features, achieving great classification ability for incomplete neighborhood decision systems.

1. Introduction

With the development of big data processing techniques, the amount of data is growing exponentially, and a vast amount of information is collected that may contain many noisy, redundant or missing feature values. Therefore, it is necessary to preprocess the massive data before using it [1,2]. Recently, feature selection has attracted considerable attention of scholars in the fields of pattern recognition, machine learning, and data mining [2,3]. As a significant preprocessing step, the main goal of feature selection is to eliminate redundant and noisy features, handle missing values, classify data, and extract useful information for data applications [4,5]. Roughly speaking, there are three general strategies to categorize feature selection as follows: filter, wrapper and embedded methods [6,7]. The accuracy of the wrapper and embedded methods are not as well as the filter methods [8–10]. Hence, our feature selection method focuses on the filter strategy, in which a heuristic search algorithm is employed to select an optimal feature subset for incomplete datasets.

At present, the traditional rough set model as a popular tool of feature selection can only handle categorical datasets but is not suitable for solving the problem of incomplete and continuous numerical data encountered in mixed and incomplete datasets [11]. The mixed and incomplete datasets refer to data that have continuous numerical and categorical feature values, and some missing values exist in the datasets [11]. As we know, the continuous numerical data should be discretized; however, the process of discretization in rough sets easily ignores the differences among data and affects the data expression of the

---

original feature sets to a certain degree [12]. To solve the problem of information loss in the process of data discretization for the numerical datasets, many scholars introduced neighborhood rough sets to study feature selection models [13], which do not break the neighborhood structure and order structure of datasets in real spaces [14]. Nevertheless, these methods sometimes cannot be enough to regularly find an optimal subset. Chen et al. [15] proposed an attribute reduction algorithm based on the fish swarm algorithm and neighborhood rough sets. To date, most neighborhood rough sets-based reduction methods are used for complete information systems, whereas few studies focus on incomplete information systems, especially mixed and incomplete datasets. For instance, Jing et al. [16] developed a variable precision tolerance neighborhood rough set model for incomplete data. Zhao and Qin [17] presented a heuristic feature selection algorithm based on neighborhood-tolerance conditional entropy in an incomplete decision system. Meng and Shi [18] constructed a positive region-based approach for attribute reduction based on the tolerance relation in an incomplete decision system. In addition, the extended neighborhood rough set model can well address incomplete information systems containing continuous numerical and categorical feature values [11]. Thus, inspired by these contributions, this paper focuses on creating such a solution for mixed and incomplete datasets, and it investigates a heuristic feature selection algorithm based on the neighborhood tolerance relation for incomplete neighborhood decision systems.

As a significant part of the uncertainty analysis based on granular computing, many uncertainty measure methods have been studied in algebraic view or information view [19]. Hu et al. [14] defined dependency degree based on the neighborhood relation to reduce numerical and categorical features. Fan et al. [20] enlarged the positive region by adding samples whose similarity classes have the maximal intersection with some decision classes in neighborhood rough sets. Li et al. [21] studied a positive region-based feature selection algorithm in neighborhood-based decision-theoretic rough sets. In summary, these previously published works on feature selection are all based on neighborhood rough sets from the algebraic view in complete neighborhood systems. Although some of the abovementioned algorithms can achieve feature subsets with criterion preservation, there are still some redundant features that can be further deleted in a sense [1]. In addition, many feature selection algorithms still have higher time-consuming when dealing with high-dimensional datasets in a certain degree [22]. Until now, information entropy as a type of significant uncertainty measure is widely used in feature selection and its variants have been researched extensively [1]. Note that many of feature selection methods based on entropy in rough sets are not fully suitable to measure neighborhood classes of a real-value dataset [23]. Liu et al. [24] investigated neighborhood mutual information and its application on hyper spectral band selection for classification. Wang et al. [25] proposed a feature selection method using conditional discrimination index in neighborhood rough sets. However, the monotonicity of these uncertainty measures does not always hold, and these literatures for feature selection based on neighborhood rough sets just only studied from information view in complete information systems.

In neighborhood rough sets, many existing approaches to feature selection are usually only based on an algebraic view or information view. It is known that the conception of feature significance based on an algebraic view only states the effect of features contained on the classification subset [16,26–29], and this definition of feature significance based on an information view only expounds the influence of features contained in the uncertain subset of the classification in the domain [17,30]. Moreover, the model is applied to some small-scale datasets in most cases [6,

28]. Thus, the two views can be combined to address some issues in real-world applications. Wang et al. [27] researched rough reduction and illustrated the definition of relative reduct in both algebraic view and information view. Chen et al. [31] presented four different uncertainty measurement methods in neighborhood systems. To the best of our knowledge, this research on the combination of algebraic view and information view has not been reported in mixed and incomplete neighborhood decision systems. In general, although these methods have their own merits, they are still inefficient and not appropriate for reducing large-scale and high-dimensional datasets, and their extended algorithms only reduce the computation time to some extent [27,32]. Therefore, this phenomenon inspires us to further investigate incomplete neighborhood decision systems from the two views and obtain the great uncertainty measures, and then this paper will combine algebraic view with information view and develop a heuristic feature selection algorithm for incomplete neighborhood decision systems with mixed data.

Notably, thus far, a large number of existing feature selection algorithms based on rough sets and their variations only concern finite sets. When the traditional feature selection methods deal with infinite sets, they usually produce higher cardinality and lower accuracy [6]. Thus, these limitations may somewhat limit their application. Halmos [33] introduced Lebesgue measure to investigate uncertainty measures. Xu et al. [34] used Lebesgue integral for uncertainty measures over infinite interval and proposed a computation method based on kernel function. Sun et al. [6] proposed an attribute reduction method using Lebesgue and entropy measures in neighborhood rough sets, which has the ability of dealing with infinite sets whilst maintaining the original classification information. Through the analysis of these abovementioned references, we know that the Lebesgue measure can efficiently measure infinite sets. On this basis, the Lebesgue measure is introduced to process infinite sets in incomplete neighborhood decision systems with mixed data. Thus, it is very important to investigate the Lebesgue measure-based uncertainty measure in incomplete neighborhood decision systems and develop a heuristic feature selection algorithm on infinite sets. If this uncertainty measure continues, the Lebesgue measure can be combined with neighborhood tolerance entropy-based uncertainty measures to study the uncertainty in incomplete neighborhood decision systems, an efficient feature selection method using Lebesgue and entropy measures will be proposed, and then a heuristic reduction algorithm can be developed to study the uncertainty and deal with the redundant of mixed datasets in incomplete neighborhood decision systems.

In this paper, to achieve feature selection method in incomplete neighborhood decision systems, our study focuses on three parts and the research motivations are shown as follows:

(1) To solve the problem that most of the neighborhood rough sets-based feature selection methods cannot deal with infinite sets in incomplete information systems, by combining neighborhood rough sets with Lebesgue measure, a new Lebesgue measure-based neighborhood tolerance class is developed, based on which, the concepts of neighborhood upper and lower approximation sets, positive region and dependency degree are redefined in incomplete neighborhood decision systems. Furthermore, the corresponding properties are discussed.

(2) To better research the Lebesgue and entropy-based uncertainty measures from algebraic view and information view in incomplete neighborhood decision systems, neighborhood tolerance entropy is proposed, based on which, neighborhood tolerance joint entropy and neighborhood tolerance dependency joint entropy are studied. Moreover, many corresponding properties are deduced and the relationships among these measures are discussed in incomplete neighborhood decision systems.

(3) To handle mixed and incomplete datasets efficiently, based on neighborhood rough sets, the definitions of a reduct, the internal significance of an attribute, the necessary and unnecessary attribute, the core attribute and the external significance of an attribute are developed in incomplete neighborhood decision systems, and then a new heuristic feature selection algorithm based on the neighborhood tolerance dependency joint entropy for infinite sets is designed in incomplete neighborhood decision systems.

The rest of the paper is constructed as follows. Some related concepts are recalled in Section 2. Section 3 depicts the Lebesgue measure-based dependency degree and neighborhood tolerance entropy-based uncertainty measures. Section 4 proposes a feature selection model based on neighborhood tolerance dependency joint entropy and a comparison analysis with two representatives reducts and reveals the design of a heuristic feature selection algorithm. In Section 5, the experimental results are achieved on seven UCI datasets and eight gene expression datasets. Finally, Section 6 summarizes this study.

2. Previous knowledge

2.1. Information entropy measures

Given a decision system $DS = \langle U, C, D \rangle$, if there exist $a \in C \cup D$ and $x \in U$ so that $f(a, x)$ amounts to a missing value (a null or unknown value, denoted as *), which means for at least one attribute $a \in C \cup D$, $* \in V_a$, then the decision system becomes an incomplete decision system (IDS) [2,35]. Thus, an incomplete decision system can be denoted as $IDS = \langle U, C \cup D, V, f \rangle$, simply written as $IDS = \langle U, C, D \rangle$.

Given an incomplete decision system $IDS = \langle U, C, D \rangle$ with $B \subseteq C$, a binary relation named as a tolerance relation on U is expressed as

$$T_B = \{(x, y) \in U \times U | \forall a \in B, f(a, x)$$
$$= f(a, y) \vee f(a, x) = * \vee f(a, y) = *\}. \tag{1}$$

For any object $x, y \in U$ with respect to $B \subseteq C$, the tolerance class is denoted as

$$T_B(x) = \{y \in U | (x, y) \in T_B\}. \tag{2}$$

Given an incomplete decision system $IDS = \langle U, C, D \rangle$ with $B \subseteq C$, the classification induced by B is $U/T_B = \{T_B(x_1), T_B(x_2), \ldots, T_B(x_h)\}$, and then the information entropy of B is described [36] as

$$H(B) = -\sum_{r=1}^{h} \frac{|T_B(x_r)|}{|U|} \log \frac{|T_B(x_r)|}{|U|}, \tag{3}$$

where $|T_B(x_r)|$ represents the cardinality of tolerance class of x_r with respect to B, $T_B(x_r) \subseteq U/T_B$, and $r = 1, 2, \ldots, h$. Let $U/D = \{d_1, d_2, \ldots, d_l\}$; then, the joint entropy of B and D is represented [35] as

$$H(B \cup D) = -\sum_{r=1}^{h} \sum_{j=1}^{l} \frac{|T_B(x_r) \cap d_j|}{|U|} \log \frac{|T_B(x_r) \cap d_j|}{|U|}, \tag{4}$$

where $d_j \in U/D$ and $j = 1, 2, \ldots, l$. Thus, the conditional entropy of D with respect to B is expressed [36] as

$$H(D|B) = -\sum_{r=1}^{h} \sum_{j=1}^{l} \frac{|T_B(x_r) \cap d_j|}{|U|} \log \frac{|T_B(x_r) \cap d_j|}{|T_B(x_r)|}. \tag{5}$$

2.2. Neighborhood rough sets

Given an incomplete neighborhood decision system $INDS = \langle U, C \cup D, V, f, \Delta, \delta \rangle$, where $U = \{x_1, x_2, \ldots, x_m\}$ is a sample set named as a universe, then $C = B_C \cup B_N$ is a conditional attribute set that describes the samples, in which B_C is the set of categorical attributes and B_N is the set of numerical attributes; D is a decision attribute set; $V = \cup_{a \in \{C \cup D\}} V_a$ and V_a is a value set of attribute a; if there exists $* \in V_a$, it equals $f(a, x) = *$ and f is a map function; $\Delta \to [0, \infty)$ is a distance function, and δ is a neighborhood parameter with $0 \leq \delta \leq 1$. Thus, an incomplete neighborhood decision system can be abbreviated as $INDS = \langle U, C, D, \delta \rangle$.

It is known that the Euclidean distance function can effectively reflect the basic information of the unknown data [37]. Thus, it is employed in this paper as

$$\Delta_B(x, y) = \sqrt{\sum_{k=1}^{N} (f(a_k, x) - f(a_k, y))^2}. \tag{6}$$

where N is the cardinality of the subset B.

Given an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with $B \subseteq C$ and $B = B_C \cup B_N$, the neighborhood tolerance relation with respect to B is expressed [17] as

$$NT_B^\delta = \{(x, y) \in U \times U | f(a, x) = * \vee f(a, y) = * \vee ((a \in B_C \to \Delta_a(x, y) = 0) \wedge (a \in B_N \to \Delta_a(x, y) \leq \delta)), \forall a \in B\}. \tag{7}$$

For any $x \in U$ and $B \subseteq C$, the neighborhood tolerance class is denoted [17] as

$$NT_B^\delta(x) = \{y \in U | (x, y) \in NT_B^\delta\}. \tag{8}$$

Given an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with $B \subseteq C$ and any $x \in X \subseteq U$, the neighborhood upper approximation set and the neighborhood lower approximation set of X with respect to B are described, respectively [17], as

$$\overline{NT_B}(X) = \{x \in U | NT_B^\delta(x) \cap X \neq \emptyset, X \subseteq U\}, \tag{9}$$

$$\underline{NT_B}(X) = \{x \in U | NT_B^\delta(x) \subseteq X, X \subseteq U\}. \tag{10}$$

Let $U/D = \{d_1, d_2, \ldots, d_l\}$; then, the positive region of D with respect to B is represented [17] as

$$POS_B(D) = \sum_{j=1}^{l} \underline{NT_B}(d_j), \tag{11}$$

where $d_j \in U/D$ and $j = 1, 2, \ldots, l$. Thus, the dependency degree of D with respect to B is expressed [17] as

$$\gamma_B(D) = \frac{|POS_B(D)|}{|U|} = \frac{|\sum_{j=1}^{l} \underline{NT_B}(d_j)|}{|U|}. \tag{12}$$

3. Lebesgue measure and entropy-based uncertainty measures in incomplete neighborhood decision systems

3.1. Lebesgue measure-based dependency degree

Aiming at the issue that most existing neighborhood rough set models cannot analyze infinite sets, it is necessary to present a new neighborhood rough set model combined with the Lebesgue measure, such that it takes the neighborhood rough set model and measure theory as the basic theories for infinite sets in incomplete neighborhood decision systems with mixed data.

For any M-dimensional Euclidean space R^M, let E be a point set in R^M, and for an open interval I_i of each column covered by E, $\bigcup_{i=1}^{\infty} I_i \supset E$ holds. The sum of its volume is $\mu = \sum_{i=1}^{\infty} |I_i|$, and

all of μ constitute a set of numbers that are bounded from below. The infimum is called the Lebesgue outer measure of E, which can be represented as $m^*(E)$ [6], i.e.,

$$m^*(E) = \inf_{E \subset \bigcup_{i=1}^{\infty} I_i} \sum_{i=1}^{\infty} |I_i|. \tag{13}$$

The Lebesgue inner measure can be denoted as $m_*(E) = |I| - m^*(I - E)$. If $m_*(E) = m^*(E)$; then, we can say E is measurable and written as $m(E)$. In this paper, $m(X)$ is uniformly regarded as the Lebesgue measure of a set X, i.e., $|X|$.

Definition 1. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U, and any $B \subseteq C$, NT_B^{δ} is the neighborhood tolerance relation with respect to B. Then, for any $x, y \in U$, the Lebesgue measure-based neighborhood tolerance class with respect to B is defined as

$$m(NT_B^{\delta}(x)) = m(\{y \in U | (x, y) \in NT_B^{\delta}\}). \tag{14}$$

Property 1. *Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with a non-empty infinite set U, for any P, $Q \subseteq C$, and $x \in U$ exists; then, the following properties hold*
   (1) $m(U) = |U|$.
   (2) $m(NT_P^{\delta}(x)) = m(\cap_{p \in P} NT_p^{\delta}(x))$.
   (3) If $Q \subseteq P$, then $m(NT_P^{\delta}(x)) \leq m(NT_Q^{\delta}(x))$.

Proof. (1) This proof is straightforward.
   (2) Suppose that any $p \in P \subseteq C$; it follows from Theorem 1 in [17] that $NT_P^{\delta}(x) = \cap_{p \in P} NT_p^{\delta}(x)$. Thus, $m(NT_P^{\delta}(x)) = m(\cap_{p \in P} NT_p^{\delta}(x))$ can be easily obtained.
   (3) Suppose that any $Q \subseteq P$ and $x \in U$; it follows immediately from (2) and Theorem 2 in [17] that $NT_P^{\delta}(x) = \cap_{p \in P} NT_p^{\delta}(x) = (\cap_{p \in Q} NT_p^{\delta}(x)) \cap (\cap_{p \in P-Q} NT_p^{\delta}(x)) = NT_Q^{\delta}(x) \cap NT_{P-Q}^{\delta}(x) \subseteq NT_Q^{\delta}(x)$, i.e., $NT_P^{\delta}(x) \subseteq NT_Q^{\delta}(x)$. Therefore, $m(NT_P^{\delta}(x)) \leq m(NT_Q^{\delta}(x))$ holds.

Definition 2. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U, for any $B \subseteq C$, $x, y \in U$; then, $X \subseteq U$, the neighborhood upper approximation set and the neighborhood lower approximation set of X with respect to B based on the Lebesgue measure are defined, respectively, as

$$m(\overline{NT_B}(X)) = m(\{x \in U | NT_B^{\delta}(x) \cap X \neq \emptyset, X \subseteq U\}), \tag{15}$$

$$m(\underline{NT_B}(X)) = m(\{x \in U | NT_B^{\delta}(x) \subseteq X, X \subseteq U\}). \tag{16}$$

Definition 3. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with a non-empty infinite set U, for any $B \subseteq C$, and $U/D = \{d_1, d_2, \ldots, d_l, \ldots\}$, the positive region of D with respect to B based on Lebesgue measure is defined as

$$m(POS_B(D)) = \sum_{j=1}^{\infty} m(\underline{NT_B}(d_j)), \tag{17}$$

where $d_j \in U/D$ and $j = 1, 2, \ldots, l, \ldots$.

Proposition 1. *Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U, and any $Q \subseteq P \subseteq C$. Then, $POS_Q(D) \subseteq POS_P(D)$ and $m(POS_Q(D)) \leq m(POS_P(D))$.*

Proof. For any $Q \subseteq P$ and $x \in U$, it follows from Property 1 that $NT_P^{\delta}(x) \subseteq NT_Q^{\delta}(x)$ and $m(NT_P^{\delta}(x)) \leq m(NT_Q^{\delta}(x))$. For any $d_j \in U/D$, one has $NT_P^{\delta}(d_j) \subseteq NT_Q^{\delta}(d_j)$, where $j = 1, 2, \ldots, l, \ldots$. Hence, it can be obtained from Eq. (17) that $POS_Q(D) \subseteq POS_P(D)$ and $m(POS_Q(D)) \leq m(POS_P(D))$.

Table 1
An incomplete neighborhood decision system with mixed data.

U	a_1	a_2	a_3	a_4	D
x_1	0.15	1	1	0.2	1
x_2	0.7	0	0	*	1
x_3	0.2	*	*	0.5	1
x_4	0.3	0	0	0.7	2
x_5	0.8	0	0	0.8	0
x_6	0.85	0	*	*	0

Definition 4. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U, for any $B \subseteq C$, and $U/D = \{d_1, d_2, \ldots, d_l, \ldots\}$; then, a dependency degree of D with respect to B based on the Lebesgue measure is defined as

$$\gamma_B(D) = \frac{m(POS_B(D))}{m(U)} = \frac{\sum_{j=1}^{\infty} m(\underline{NT_B}(d_j))}{m(U)}, \tag{18}$$

where $d_j \in U/D$ and $j = 1, 2, \ldots, l, \ldots$.

Remark 1. The above described ideas can also be interpreted as an ability to classify objects. More precisely, if $\gamma_B(D) = 1$, then all elements of the universe can be classified to elementary categories of U/D by the conditional attribute subset B. If $\gamma_B(D) \neq 1$, only those elements of the universe that belong to the positive region can be classified into categories of the decision attribute D through employing B. In particular, if $\gamma_B(D) = 0$, then none of the elements of the universe can be classified by B with respect to elementary categories of D.

Example 1. Consider an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$, where $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, $C = B_N \cup B_C = \{a_1, a_4\} \cup \{a_2, a_3\}$, $D = \{d\}$ with the value $\{0, 1, 2\}$, and $\delta = 0.15$, the incomplete neighborhood decision system with mixed data is shown in Table 1.
   In Table 1, on the neighborhood tolerance relation, the universe U is divided into three equivalence classes by D, i.e., $U/D = \{d_1, d_2, d_3\}$. The neighborhood tolerance class with respect to D can be obtained by $d_1 = \{x_1, x_2, x_3\}$, $d_2 = \{x_4\}$, and $d_3 = \{x_5, x_6\}$.
   Let $C = \{a_1, a_2, a_3, a_4\}$; then, according to Eq. (8), the neighborhood tolerance classes with respect to C are computed by $NT_C^{\delta}(x_1) = \{x_1\}$, $NT_C^{\delta}(x_2) = \{x_2, x_5, x_6\}$, $NT_C^{\delta}(x_3) = \{x_3\}$, $NT_C^{\delta}(x_4) = \{x_4\}$, $NT_C^{\delta}(x_5) = \{x_2, x_5, x_6\}$, and $NT_C^{\delta}(x_6) = \{x_2, x_5, x_6\}$.
   Thus, the positive region and the dependency degree of D with respect to C are calculated, respectively, as
   $POS_C(D) = \{x_1, x_3, x_4\}$ and $\gamma_C(D) = \frac{m(POS_C(D))}{|U|} = \frac{3}{6} = \frac{1}{2}$.

Proposition 2. *Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with a non-empty infinite set U and any $Q \subseteq P \subseteq C$, $\gamma_Q(D) \leq \gamma_P(D)$ holds.*

Proof. For any $Q \subseteq P \subseteq C$, it follows from Proposition 1 that $m(POS_Q(D)) \leq m(POS_P(D))$. Hence, from Definition 4, $\gamma_Q(D) \leq \gamma_P(D)$ can be obtained.

3.2. Lebesgue measure and neighborhood tolerance entropy-based uncertainty measures

In recent years, to evaluate the real-value datasets on the neighborhood tolerance relation, the definition of the neighborhood has been introduced into information entropy to extend Shannon entropy [12,31]. However, the neighborhood tolerance relation-based measures and their variations are only based on finite sets in most cases, which limits their practical applications

to a certain extent. Fortunately, the Lebesgue measure can efficiently measure the uncertainty of infinite sets [6]. Therefore, the Lebesgue measure is introduced to solve the problem of uncertainty measures on infinite sets in incomplete neighborhood decision systems with mixed data.

Definition 5. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with any $B \subseteq C$ exists, and $NT_B^\delta(x_i)$ is a neighborhood tolerance class of $x_i \in U$. Then, the neighborhood tolerance entropy of B is defined as

$$NTE_\delta(B) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|NT_B^\delta(x_i)|}{|U|}. \tag{19}$$

Definition 6. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with a non-empty infinite set U, for any $B \subseteq C$ exists and $NT_B^\delta(x_i)$ is a neighborhood tolerance class of $x_i \in U$. Then, the neighborhood tolerance entropy of B based on the Lebesgue measure is defined as

$$NTE_\delta(B) = -\frac{1}{m(U)} \sum_{i=1}^{\infty} \int_{x_i \in U} \log_2 \frac{m(NT_B^\delta(x_i))}{m(U)} dx. \tag{20}$$

Proposition 3. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with a non-empty infinite set U, for any $Q \subseteq P \subseteq C$, $NTE_\delta(Q) \leq NTE_\delta(P)$ holds.

Proof. Suppose that for any $Q \subseteq P \subseteq C$ and $x_i \in U$, from Property 1, $m(NT_P^\delta(x_i)) \leq m(NT_Q^\delta(x_i))$ holds. Then, $-\frac{1}{m(U)} \sum_{i=1}^{|U|} \log_2 \frac{m(NT_P^\delta(x_i))}{m(U)} \leq -\frac{1}{m(U)} \sum_{i=1}^{|U|} \log_2 \frac{m(NT_Q^\delta(x_i))}{m(U)}$. Hence, $NTE_\delta(Q) \leq NTE_\delta(P)$ can be easily obtained.

Proposition 3 shows that in an incomplete neighborhood decision system, the neighborhood tolerance entropy increases monotonically with the increase of the features. Obviously, $NTE_\delta(B) \geq 0$, because of $m(NT_B^\delta(x)) \leq m(U)$ coming into existence. In fact, $NTE_\delta(B) = 0$ if and only if $m(NT_B^\delta(x)) = m(U)$ for any $x \in U$.

Definition 7. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with a non-empty infinite set U, for any $B \subseteq C$, $NT_B^\delta(x_i)$ is a neighborhood tolerance class of $x_i \in U$, and $NT_D(x_i)$ is the neighborhood tolerance class of x_i formed by D, with $U/D = \{d_1, d_2, \ldots, d_l, \ldots\}$, where $NT_D^\delta(x_i) \in U/D$ because $* \in V_D$ does not usually exist. Then, the neighborhood tolerance joint entropy of B and D based on the Lebesgue measure is defined as

$$NTE_\delta(B \cup D) = -\frac{1}{m(U)} \sum_{j=1}^{\infty} \int_{x_i \in U} \log_2 \left(\frac{m(NT_B^\delta(x_i) \cap d_j)}{m(U)} \right) dx, \tag{21}$$

where $d_j \in U/D$ and $j = 1, 2, \ldots, l, \ldots$.

Proposition 4. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with a non-empty infinite set U, for any $Q \subseteq P \subseteq C$, $NTE_\delta(Q \cup D) \leq NTE_\delta(P \cup D)$ holds.

Proof. For any $Q \subseteq P \subseteq C$ and $x_i \in U$, according to Property 1, one has $NT_P^\delta(x_i) \subseteq NT_Q^\delta(x_i)$ and $m(NT_P^\delta(x_i)) \leq m(NT_Q^\delta(x_i))$. Since the values of D are usually complete in an incomplete neighborhood decision system, the neighborhood tolerance class $NT_D^\delta(x_i)$ is equal to the equivalence class $[x_i]_d$ in rough sets. Then, $NT_D^\delta(x_i) \subseteq U/D = \{d_1, d_2, \ldots, d_l, \ldots\}$, and $\{x_i\} \subseteq NT_P^\delta(x_i) \cap d_j \subseteq NT_Q^\delta(x_i) \cap d_j \subseteq U$ holds. It follows that $1 = |[x_i]_d| \leq m(NT_P^\delta(x_i) \cap d_j) \leq m(NT_Q^\delta(x_i) \cap d_j) \leq m(U)$. Obviously, $\frac{1}{m(U)} \leq \frac{m(NT_P^\delta(x_i) \cap d_j)}{m(U)} \leq \frac{m(NT_Q^\delta(x_i) \cap d_j)}{m(U)} \leq$

1. Thus, $\log_2 \frac{1}{m(U)} \leq \log_2 \frac{m(NT_P^\delta(x_i) \cap d_j)}{m(U)} \leq \log_2 \frac{m(NT_Q^\delta(x_i) \cap d_j)}{m(U)} \leq$ 0. Hence, it is clear that $0 \leq -\frac{1}{m(U)} \sum_{i=1}^{|U|} \log_2(\frac{m(NT_Q^\delta(x_i) \cap d_j)}{m(U)}) \leq -\frac{1}{m(U)} \sum_{i=1}^{|U|} \log_2(\frac{m(NT_P^\delta(x_i) \cap d_j)}{m(U)}) \leq \frac{1}{m(U)} \log_2 m(U)$. Namely, it can be easily proved that $NTE_\delta(Q \cup D) \leq NTE_\delta (P \cup D)$.

Proposition 5. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U, for any $B \subseteq C$ and $x_i \in U$, $d_j \in U/D = \{d_1, d_2, \ldots, d_l, \ldots\}$, where $j = 1, 2, \ldots, l, \ldots$; then, one has $NTE_\delta(B \cup D) \geq NTE_\delta(B)$.

Proof. It follows immediately from Definitions 6 and 7 that

$$NTE_\delta(B \cup D) - NTE_\delta(B)$$
$$= -\frac{1}{m(U)} \left(\sum_{j=1}^{\infty} \int_{x_i \in U} \log_2 \left(\frac{m(NT_B^\delta(x_i) \cap d_j)}{m(U)} \right) dx \right.$$
$$\left. - \sum_{j=1}^{\infty} \int_{x_i \in U} \log_2 \frac{m(NT_B^\delta(x_i))}{m(U)} dx \right)$$
$$= -\frac{1}{m(U)} \sum_{j=1}^{\infty} \int_{x_i \in U} \left(\log_2 \left(\frac{m(NT_B^\delta(x_i) \cap d_j)}{m(U)} \right) - \log_2 \frac{m(NT_B^\delta(x_i))}{m(U)} \right) dx$$
$$= -\frac{1}{m(U)} \sum_{j=1}^{\infty} \int_{x_i \in U} \left(\log_2 \frac{m(NT_B^\delta(x_i) \cap d_j)}{m(NT_B^\delta(x_i))} \right) dx.$$

Since $n_B^\delta(x_i) \cap [x_i]_D \subseteq n_B^\delta(x_i) \subseteq U$, where $[x_i]_D \in U/D$, then $|n_B^\delta(x_i) \cap [x_i]_D| \leq |n_B^\delta(x_i)| \leq |U|$ in neighborhood class. Similarly, under the neighborhood tolerance class, $|NT_B^\delta(x_i) \cap d_j| \leq |NT_B^\delta(x_i)| \leq |U|$. It follows that $m(NT_B^\delta(x_i) \cap d_j) \leq m(NT_B^\delta(x_i)) \leq m(U)$. Thus, $\frac{m(NT_B^\delta(x_i) \cap d_j)}{m(NT_B^\delta(x_i))} \leq 1$. Therefore, $NTE_\delta(B \cup D) - NTE_\delta(B) \geq 0$ holds, i.e., $NTE_\delta(B \cup D) \geq NTE_\delta(B)$.

From Proposition 5, the value of the neighborhood tolerance joint entropy is larger than that of the neighborhood tolerance entropy of a feature subset. Thus, it can be concluded that the neighborhood tolerance joint entropy has a stronger distinguishing capacity when adding new features.

Definition 8. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U, for any $B \subseteq C$, and $NT_B^\delta(x_i)$ and $NT_D(x_i)$ are two neighborhood tolerance classes of $x_i \in U$ formed by B and D, respectively, with $U/D = \{d_1, d_2, \ldots, d_l, \ldots\}$, and $m(POS_B(D))$ is the positive region of D with respect to B based on the Lebesgue measure. Then, the neighborhood tolerance dependency joint entropy of D and B based on the Lebesgue measure is defined as

$$NTDE(B, D) = -\frac{m(POS_B(D))}{m(U)^2} \sum_{j=1}^{\infty} \int_{x_i \in U} \log_2 \left(\frac{m(NT_B^\delta(x_i) \cap d_j)}{m(U)} \right) dx, \tag{22}$$

where $d_j \in U/D$ and $j = 1, 2, \ldots, l, \ldots$.

Property 2. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U, and for any $B \subseteq C$, the dependency degree of D in respect to B is $\gamma_B(D)$, and the neighborhood tolerance joint entropy of B and D is $NTE_\delta(B \cup D)$. Then, $NTDE(B, D) = \gamma_B(D) \cdot NTE_\delta(B \cup D) \geq 0$.

Proof. Suppose that for any $B \subseteq C$, $NT_B^\delta(x_i)$ is the neighborhood tolerance class of $x_i \in U$, and $d_j \in U/D = \{d_1, d_2, \ldots, d_l, \ldots\}$,

where $j = 1, 2, \ldots, l, \ldots$. It follows immediately from Definition 8, that

$$
NTDE(B, D) = -\frac{1}{m(U)^2} m(POS_B(D))
$$

$$
\cdot \sum_{j=1}^{\infty} \int_{x_i \in U} \log_2(\frac{m(NT_B^\delta(x_i) \cap d_j)}{m(U)}) dx
$$

$$
= -\frac{1}{|U|^2} |POS_B(D)| \cdot \sum_{j=1}^{\infty} \int_{x_i \in U} \log_2(\frac{|NT_B^\delta(x_i) \cap d_j|}{|U|}) dx
$$

$$
= \frac{|POS_B(D)|}{|U|} \cdot (-\frac{1}{|U|} \sum_{j=1}^{\infty} \int_{x_i \in U} \log_2(\frac{|NT_B^\delta(x_i) \cap d_j|}{|U|}) dx)
$$

$$
= \gamma_B(D) \cdot NTE_\delta(B \cup D).
$$

From Property 1, one has $m(U) = |U|$. Then, it is known from Proposition 5 that $m(NT_B^\delta(x_i) \cap d_j) \le m(U)$. Thus, it is obvious that $\log_2(\frac{m(NT_B^\delta(x_i) \cap d_j)}{|U|}) \le 0$ and $m(POS_B(D)) \ge 0$. Hence, $NTDE(B, D) \ge 0$ holds.

Note that Wang et al. [27] stated that all definitions and calculations in rough sets based on the upper and lower approximation sets are denoted from algebraic view of rough set theory, and the concepts of information entropy and its extensions are described from an information view in rough sets. It can be seen from Definition 8 and Property 2 that $\gamma_B(D)$ is the dependency degree of D with respect to B in the algebraic view, and $NTE_\delta(B \cup D)$ is the neighborhood tolerance joint entropy of B and D in the information view. Therefore, Definition 8 can analyze and measure the uncertainty of incomplete neighborhood decision systems from both an algebraic view and an information view based on Lebesgue and entropy measures.

4. Feature selection method in incomplete neighborhood decision systems

4.1. Feature selection based on neighborhood tolerance dependency joint entropy

Proposition 6. *Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U, for any $Q \subseteq P \subseteq C$, and $U/D = \{d_1, d_2, \ldots, d_l, \ldots\}$, where $d_j \in U/D$ and $j = 1, 2, \ldots, l, \ldots$. $NTDE(Q, D) \le NTDE(P, D)$.*

Proof. Let any $Q \subseteq P \subseteq C$ and $x_i \in U$. It follows from Proposition 2 that $\gamma_Q(D) \le \gamma_P(D)$. According to Proposition 4, $NTE_\delta(Q \cup D) \le NTE_\delta(P \cup D)$ obviously holds. Then, $-\frac{1}{m(U)} \sum_{i=1}^{|U|} \log_2$ $(\frac{m(NT_Q^\delta(x_i) \cap d_j)}{m(U)}) \le -\frac{1}{m(U)} \sum_{i=1}^{|U|} \log_2(\frac{m(NT_P^\delta(x_i) \cap d_j)}{m(U)})$ can be obtained. For any $x_i \in U$, when $NT_Q^\delta(x_i) = NT_P^\delta(x_i)$, $\gamma_Q(D) = \gamma_P(D)$ and $NT_Q^\delta(x_i) \cap d_j = NT_P^\delta(x_i) \cap d_j$. Thus, $NTDE(Q, D) = NTDE(P, D)$ can be obtained. Therefore, $NTDE(Q, D) \le NTDE(P, D)$ holds.

Remark 2. The monotonicity is one of the most important properties of uncertainty measures for feature selection. From Proposition 6, the neighborhood tolerance dependency joint entropy increases monotonically with the number of elements in the attribute subset. Moreover, the monotonic increase of the neighborhood tolerance dependency joint entropy contributes to the selection of the greedy method for feature selection in incomplete neighborhood decision systems.

Definition 9. Suppose there is an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U and $B \subseteq C$; if $NTDE(B, D) = NTDE(C, D)$, and for any $a \in B$, there exists $NTDE(B, D) > NTDE(B - \{a\}, D)$. One can say B is a reduct of C with respect to D.

Definition 10. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U, for any $B \subseteq C$ and $a \in B$; then, the internal significance of the attribute a in B with respect to D is defined as

$$
Sig^{inner}(a, B, D) = NTDE(B, D) - NTDE(B - \{a\}, D). \quad (23)
$$

Definition 11. Suppose there is an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U and $B \subseteq C$; if $Sig^{inner}(a, B, D) > 0$ for any $a \in B$, then the attribute a in B is necessary; otherwise, the a is unnecessary. If each a in B is necessary, then B is independent.

Definition 12. Suppose there is an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U and $B \subseteq C$; if $NTDE(C, D) > NTDE(C - \{a\}, D)$ for any $a \in C$, i.e., $Sig^{inner}(a, C, D) > 0$, then the a is called a core attribute of C relative to D.

Definition 13. Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U, for any $B \subseteq C$ and $b \in C - B$; then, the external significance of attribute b with respect to D is defined as

$$
Sig^{outer}(b, B, D) = NTDE(B \cup \{b\}, D) - NTDE(B, D). \quad (24)
$$

Remark 3. From Definition 13, the significance measure $Sig^{outer}(b, B, D)$ indicates the significance of attribute b, which is added to B with respect to D in an incomplete neighborhood decision system, and it provides a powerful reference for decision making. The larger the value of Sig^{outer} is, the more important the attribute is in incomplete neighborhood decision systems; otherwise, the lower its significance is. Therefore, when the significance of each attribute is calculated, the attribute of significance with 0 is removed, and the reduced subset of attributes can be obtained finally.

Property 3. *Suppose that an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U and any $B \subseteq C$; then, the following properties can be immediately obtained*

    (1) $0 \le Sig^{inner}(a, C, D) \le 1$ for any $a \in C$.
    (2) $0 \le Sig^{outer}(b, B, D) \le 1$ for any $b \in C - B$.
    (3) When $B = C$, $Sig^{outer}(, C, D) = 0$.
    (4) Any $b \in C - B$ is not necessary if and only if $Sig^{outer}(b, B, D) = 0$.

Note that in an $INDS = \langle U, C, D, \delta \rangle$ with a non-empty infinite set U and $B \subseteq C$, for any $a \in B$, when $Sig^{inner}(a, B, D)$ is calculated, $NTDE(B - \{a\}, D)$ is only calculated because $NTDE(B, D)$ is a constant. Similarly, for any $b \in C - B$, when calculating $Sig^{outer}(b, B, D)$, $NTDE(B \cup \{b\}, D)$ only needs to be calculated.

Example 2. Consider the incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ in Table 1, where $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, $C = B_N \cup B_C = \{a_1, a_4\} \cup \{a_2, a_3\}$, $D = \{d\}$ with the value $\{0, 1, 2\}$ and $\delta = 0.15$.

From Definitions 7, 8 and 10, the following formulas can be computed by

$$
NT_C^\delta(x_1) \cap d_1 = \{x_1\}, NT_C^\delta(x_1) \cap d_2 = \emptyset, NT_C^\delta(x_1) \cap d_3 = \emptyset;
$$

$$
NT_C^\delta(x_2) \cap d_1 = \{x_2\}, NT_C^\delta(x_2) \cap d_2 = \emptyset, NT_C^\delta(x_2) \cap d_3 = \{x_5, x_6\};
$$

$$
NT_C^\delta(x_3) \cap d_1 = \{x_3\}, NT_C^\delta(x_3) \cap d_2 = \emptyset, NT_C^\delta(x_3) \cap d_3 = \emptyset;
$$

$$
NT_C^\delta(x_4) \cap d_1 = \emptyset, NT_C^\delta(x_4) \cap d_2 = \{x_4\}, NT_C^\delta(x_4) \cap d_3 = \emptyset;
$$

$NT_C^\delta(x_5) \cap d_1 = \{x_2\}, NT_C^\delta(x_5) \cap d_2 = \emptyset, NT_C^\delta(x_5) \cap d_3 = \{x_5, x_6\};$

$NT_C^\delta(x_6) \cap d_1 = \{x_2\}, NT_C^\delta(x_6) \cap d_2 = \emptyset, NT_C^\delta(x_6) \cap d_3 = \{x_5, x_6\}.$

Then, from Eq. (21), the neighborhood tolerance joint entropy of D with respect to C is calculated by

$$NTE_\delta(C \cup D) = -\frac{1}{6} \sum_{j=1}^{3} \sum_{i=1}^{6} \log_2 \frac{m(NT_C^\delta(x_i) \cap d_j)}{m(U)} = 2.252.$$

It follows immediately from Definition 8 that

$$NTDE(C, D) = -\frac{1}{36} m(POS_B(D)) \sum_{j=1}^{3} \sum_{i=1}^{6} \log_2 \frac{m(NT_B^\delta(x_i) \cap d_j)}{6}$$

$$= 1.126.$$

In the same way, for all attributes in C, the internal significance of attributes is calculated by $Sig^{inner}(a_1, C, D) = NTDE(C, D) - NTDE(C - \{a_1\}, D) = 0.806$. Similarly, $Sig^{inner}(a_2, C, D) = 0$, $Sig^{inner}(a_3, C, D) = 0$ and $Sig^{inner}(a_4, C, D) = 0.806$.

4.2. Comparison analysis with two representative reducts

As we know, Wang et al. [27] stated that the relative reduct of a decision system from an information view must be its relative reduct from an algebraic view, and then heuristic reduction algorithms may be developed in decision systems. Sun et al. [26] concluded that the definition of the conditional rough entropy reduct from the information view can include that of the positive region reduct from the algebraic view in incomplete decision systems. On the basis of the ideas in [26,27], the definition of the reduct based on neighborhood tolerance dependency joint entropy in an incomplete neighborhood decision system can be extended from an algebraic view and information view in neighborhood rough sets. For convenience, the reduct in Definition 9 is named as a neighborhood tolerance reduct. Meng and Shi [18] proposed a positive region-based reduct in incomplete decision systems, and the positive region reduct as a representative reduct is called a reduct in the algebraic view of neighborhood rough sets. Zhao and Qin [17] designed a feature selection algorithm based on conditional entropy under the neighborhood tolerance relation for incomplete data, which is a representative reduct from the information view in neighborhood rough sets.

Given an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with a non-empty infinite set U, for any $B \subseteq C$ and $D = \{d\}$, a positive region reduct of the incomplete neighborhood decision system is proposed [18] as follows. For any $a \in B$, if $|POS_B(D)| = |POS_C(D)|$ and $|POS_{B-\{a\}}(D)| < |POS_B(D)|$, B is a relative reduct of the incomplete neighborhood decision system, where $POS_B(D) = \cup\{\underline{B}(X)_\delta | X \in U/T_B\}$ is the positive region of D with respect to B.

Proposition 7. *Suppose there is an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with non-empty infinite set U and any $B \subseteq C$; if B is a neighborhood tolerance reduct of C with respect to D in the incomplete neighborhood decision system; then, B is a positive region reduct of C with respect to D in the incomplete neighborhood decision system.*

Proof. Let $U = \{x_1, x_2, \ldots, x_m, \ldots\}$ and $U/D = \{d_1, d_2, \ldots, d_l, \ldots\}$. Suppose that for any $B \subseteq C$, from Definition 9, if $NTDE(B, D) = NTDE(C, D)$ and for any $a \in B$, there exists $NTDE(B, D) > NTDE(B - \{a\}, D)$, and B is a neighborhood tolerance reduct of C relative to D in the incomplete neighborhood decision

system. When $NTDE(B, D) = NTDE(C, D)$, it can be obtained from Proposition 6 that $NT_B^\delta(x) = NT_C^\delta(x)$, $\gamma_B(D) = \gamma_C(D)$ and $NT_B^\delta(x_i) \cap d_j = NT_C^\delta(x_i) \cap d_j$, where $x_i \in U$, $d_j \in U/D$ and $j = 1, 2, \ldots, l, \ldots$. By Eq. (10), one has that $NT_B(d_j) = NT_C(d_j)$. Thus, it is obvious that $POS_B(D) = POS_C(D)$ with a non-empty infinite set U, i.e., $|POS_B(D)| = |POS_C(D)|$. For any $a \in B$, $B - \{a\} \subseteq B$ and from Property 15 in [18], $B - \{a\}(D)_\delta \subseteq \underline{B}(D)_\delta$ can be obtained. Therefore, $POS_{B-\{a\}}(D) \subseteq POS_B(D)$ holds. Because for any $a \in B$, there is $NTDE(B, D) > NTDE(B - \{a\}, D)$, one has $B - \{a\}(D)_\delta \subseteq \underline{B}(D)_\delta$. From Proposition 1, $POS_{B-\{a\}}(D) \subseteq POS_B(D)$ holds. Obviously, $|POS_{B-\{a\}}(D)| < |POS_B(D)|$ for any $a \in B$. Hence, B is a positive region reduct of C with respect to D in the incomplete neighborhood decision system.

It is noted that the inverse relation of Proposition 7 is generally unavailable. Proposition 7 shows that the definition of the neighborhood tolerance reduct in the information view contains that of the positive region reduct in the algebraic view for the incomplete neighborhood decision system.

Given an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ with a non-empty infinite set U, any $B \subseteq C$ and $D = \{d\}$, for any $a \in B$, a reduct of the incomplete neighborhood decision system called a neighborhood conditional entropy reduct is proposed [17] as follows: If $NTE(D|B) = NTE(D|C)$ and $NTE(D|B - \{a\}) < NTE(D|B)$, B is a neighborhood tolerance conditional entropy reduct of C with respect to D in the incomplete neighborhood decision system, where $NTE(D|B) = -\frac{1}{|U|} \sum_{i=1}^{|U|} (\frac{NT_B(x_i)}{|U|} - \frac{NT_B(x_i) \cap NT_D(x_i)}{|U|})$ denotes the neighborhood tolerance conditional entropy of B and D, $NT_B(x_i)$ and $NT_D(x_i)$ are the neighborhood tolerance classes of x_i with respect to B and D, respectively, and $NT_D(x_i) \in U/D$.

Proposition 8. *Suppose there is an incomplete neighborhood decision system $INDS = \langle U, C, D, \delta \rangle$ exists with a non-empty infinite set U and $B \subseteq C$; B is a neighborhood tolerance reduct of C with respect to D in the incomplete neighborhood decision system if and only if B is a neighborhood conditional entropy reduct of C with respect to D in the incomplete neighborhood decision system.*

Proof. $\Rightarrow$ Let $U = \{x_1, x_2, \ldots, x_m, \ldots\}$, and $U/D = \{d_1, d_2, \ldots, d_l, \ldots\}$. Suppose that for any $B \subseteq C$, it follows from Definition 9 that if $NTDE(B, D) = NTDE(C, D)$ and for any $a \in B$, there exists $NTDE(B, D) > NTDE(B - \{a\}, D)$; then, B is a neighborhood tolerance reduct of C relative to D. Similar to the proof of Proposition 7, when $NTDE(B, D) = NTDE(C, D)$, for any $x_i \in U$, $d_j \in U/D$ and $j = 1, 2, \ldots, l, \ldots$, $NT_B^\delta(x_i) = NT_C^\delta(x_i)$ and $NT_B^\delta(x_i) \cap d_j = NT_C^\delta(x_i) \cap d_j$. It follows that $NTE(D|B) = NTE(D|C)$ with a non-empty infinite set U. Because any $B - \{a\} \subseteq B$, from Theorem 6 in [23], we have $NTE(D|B - \{a\}) < NTE(D|B)$. For any $a \in B$, there exists $NTDE(B, D) > NTDE(B - \{a\}, D)$; then, $NTE(D|B - \{a\}) < NTE(D|B)$ holds. Thus, B is the neighborhood conditional entropy reduct of C with respect to D in the incomplete neighborhood decision system.

$\Leftarrow$ Suppose that for the non-empty infinite set U, for any $B \subseteq C$ and $a \in B$, if $NTE(D|B) = NTE(D|C)$ and $NTE(D|B - \{a\}) < NTE(D|B)$, then B is a neighborhood conditional entropy reduct of C relative to D. Similar to the proof of Proposition 7, from Definitions 2, 3 and 4, when $NT_B^\delta(x) = NT_C^\delta(x)$, $\gamma_B(D) = \gamma_C(D)$ and $NT_B^\delta(x_i) \cap d_j = NT_C^\delta(x_i) \cap d_j$, where for any $x_i \in U$, $d_j \in U/D$ and $j = 1, 2, \ldots, l, \ldots$. Therefore, it can be obtained from Definition 8 that $NTDE(B, D) = NTDE(C, D)$. Since any $B - \{a\} \subseteq B$, from Proposition 6, $NTDE(B, D) > NTDE(B - \{a\}, D)$ holds. For any $a \in B$, there exists $NTE(D|B - \{a\}) < NTE(D|B)$ with a non-empty infinite set U; then, $NTDE(B, D) > NTDE(B - \{a\}, D)$ holds. Therefore, B is the neighborhood tolerance reduct of C with respect to D in the incomplete neighborhood decision system.

Remark 4. Proposition 8 demonstrates that in an incomplete neighborhood decision system, the neighborhood tolerance reduct is equivalent to the neighborhood conditional entropy reduct in the information view. According to Propositions 7 and 8, it can be concluded that the definition of neighborhood tolerance reduct contains two representative reducts proposed in the algebraic view and information view, respectively. Therefore, the definition of the neighborhood tolerance reduct represents a mathematical quantitative measure to evaluate the knowledge uncertainty of different attribute sets in incomplete neighborhood decision systems.

4.3. Feature selection algorithm

To facilitate understanding the feature selection method more clearly, the process of data classification for feature selection is shown in Fig. 1, where the Fisher score method [6] is introduced for preliminary dimension reduction of the high-dimensional datasets. To support efficient feature selection, a feature selection algorithm based on the neighborhood tolerance dependency joint entropy (FSNTDJE) is designed as Algorithm 1.

Algorithm 1. FSNTDJE
Input: An incomplete neighborhood decision system $INDS = < U, C, D, \delta >$, and neighborhood parameter δ.
Output: An optimal attribute subset B.
(1) Initialize $B = \emptyset$ and $R = \emptyset$.
(2) Calculate $NTDE(C, D)$.
(3) **FOR** $i = 1$ to $|C|$ **do**
(4)   Calculate $Sig^{inner}(c_i, C, D)$.
(5)   **IF** $Sig^{inner}(c_i, C, D) > 0$
(6)   then $B = B \cup \{c_i\}$.
(7)   **ENDIF**
(8) **ENDFOR**
(9) Let $R = C - B$.
(10) **WHILE** $NTDE(B, D) \neq NTDE(C, D)$
(11)   **FOR** $j = 1$ to $|R|$ **do**
(12)   Calculate $NTDE(B \cup \{a_j\}, D)$.
(13)   Select a_j to make it satisfy
$\max\{a_j \in R | NTDE(B \cup \{a_j\}, D)\}$, and if multiple attributes satisfy the maximum, then the front should be selected.
(14)   **ENDFOR**
(15) Let $B = B \cup \{a_j\}$ and $R = R - \{a_j\}$, and calculate $NTDE(B, D)$.
(16) **ENDWHILE**
(17) **FOR** $k = 1$ to $|B|$ **do**
(18) Select $b_k \in B$.
(19) Compute $NTDE(B - \{b_k\}, D)$.
(20)   **IF** $NTDE(B - \{b_k\}, D) > NTDE(B, D)$
(21)   then $B = B - \{b_k\}$.
(22)   **EDNIF**
(23) **ENDFOR**
(24) **RETURN** An optimal attribute subset B.

In the FSNTDJE algorithm, the computation of neighborhood tolerance classes is frequent in incomplete neighborhood decision systems. The process of computing neighborhood tolerance classes exerts a great influence on the time complexity of feature selection. It should be noted that the main calculation of FSNTDJE involves two important aspects: obtaining the neighborhood tolerance classes and calculating the neighborhood tolerance dependency joint entropy. First, in order to further reduce the computational time complexity of neighborhood tolerance classes, the bucket sorting algorithm [38] is employed. Then, the time complexity of neighborhood tolerance classes is $O(mn)$, where m is the number of samples, and n is the number of features.

At the same time, the computational time complexity of the neighborhood tolerance dependency joint entropy is $O(n)$. Since $O(n) < O(mn)$, the complexity of reaching the neighborhood tolerance dependency joint entropy is $O(mn)$. Thus, in the worst case, the time complexity of FSNTDJE is $O(n^3m)$, since there are two loops at Steps 3–8 and 10–16. In the process of dimension reduction, suppose that the number of selected features is n_R, and in the calculation of neighborhood tolerance classes, we only need to consider the candidate features without traversing the whole feature subset. Hence, the time complexity of calculating neighborhood tolerance classes is $O(n_R m)$. For the FSNTDJE algorithm, the times of the outer loop and the inner loop are m and $n - n_R$, respectively. The total time complexity of FSNTDJE is $O(n_R m(n - n_R)n)$. It is known that $n_R \ll n$ in most cases. Therefore, the time complexity of FSNTDJE is nearly $O(mn)$. Moreover, the FSNTDJE algorithm achieves better performance than some of the existing algorithms for feature selection [18,26,38–43] in terms of the computational complexity for decision systems. Moreover, the space complexity of FSNTDJE is $O(mn)$.

5. Experimental results and analysis

5.1. Experiment preparation

The primary task of a feature selection method often includes two aspects: one is to choose a small number of features, and the other is to retain high classification accuracy. To demonstrate the classification performance of our feature selection method proposed in Section 4.3, the comprehensive results of all contrasted algorithms are obtained and analyzed on fifteen public datasets (seven UCI datasets and eight DNA microarray gene expression datasets). The seven selected UCI datasets with low-dimensions include Nursery, Credit, Mushroom, Wpbc, Soybean, Annealing and Ozone level, downloaded from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/index.php). The eight high-dimensional DNA microarray gene expression datasets include Colon, DLBCL, Brain, Leukemia, Breast, Lung, MLL and Prostate, downloaded from the Cancer Program Datasets (http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi). Notably, a gene expression dataset can be described by a neighborhood decision system, where an object corresponds to a sample, a conditional attribute represents a gene, and a decision attribute expresses a subclass of cancer. It should be noted that the Credit and Wpbc datasets in UCI and the eight gene expression datasets are usually complete; therefore, to create incomplete neighborhood decision systems, for convenience, we randomly change some known values of features into missing values. All of the datasets are described in detail in Table 2, where the two datasets contain mixed features, the three datasets include categorical features, and the remaining ten datasets have numerical features.

The experiments are performed on a personal computer running Windows 10 with an Intel(R) i5 CPU at 3.20 GHz with 4.0 GB memory. All the simulation experiments are implemented in MATLAB 2016a, and the four classifiers, including Naive Bayes, C4.5, KNN and CART, are selected to illustrate classification results in WEKA 3.8. The following experimental comparisons for classification on the selected features are implemented using 10-fold cross-validation with all test datasets, where every dataset is first randomly divided into ten portions, which are the same size; then, one data subset is used as the testing dataset and the other nine data subsets are used as the training datasets; each of the ten data subsets is employed exactly once as the testing dataset. Cross-validation is repeated ten times, and the average of the ten test results is the obtained number of selected features and classification accuracy [3,6,40,44].
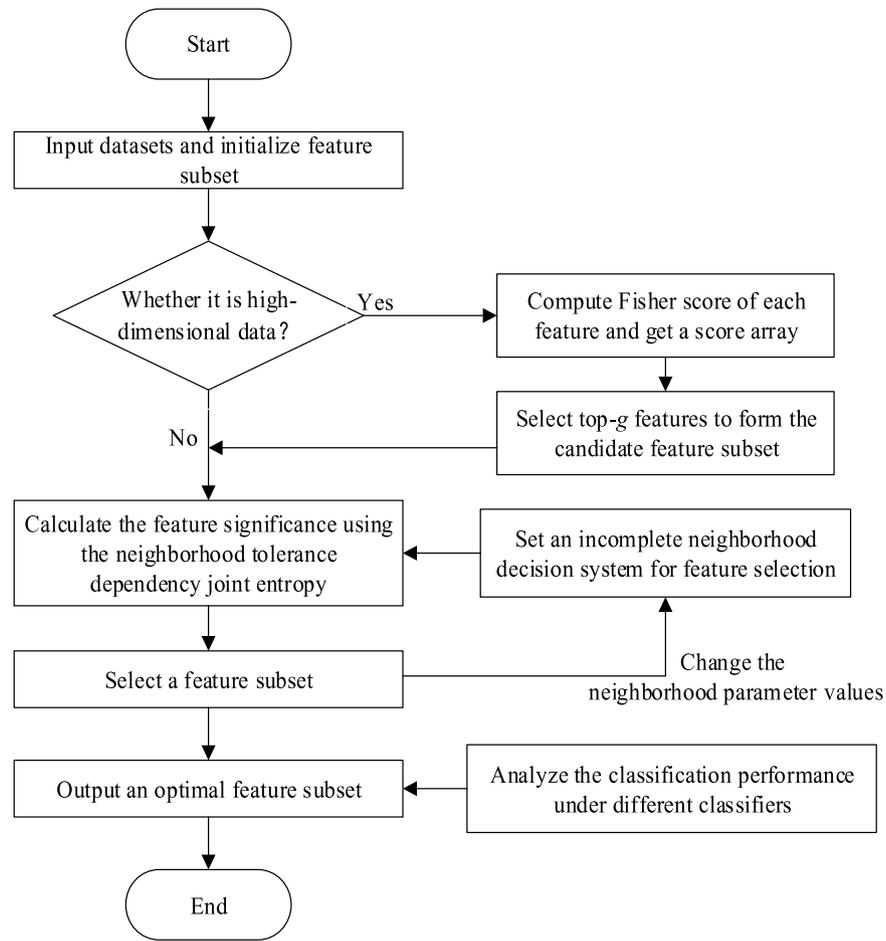
Fig. 1. Process flow of the feature selection method for data classification.

Table 2
The description of the fifteen public datasets.

No.	Datasets	Samples	Features			Classes
			Total	Numerical features	Categorical features	
1	Nursery	12 960	8	0	8	5
2	Credit	690	15	6	9	2
3	Mushroom	8 124	22	0	22	2
4	Wpbc	198	34	34	0	2
5	Soybean	307	35	0	53	19
6	Annealing	798	38	6	32	6
7	Ozone level	2 534	73	73	0	2
8	Colon	62	2 000	2 000	0	2
9	DLBCL	77	5 469	5 469	0	2
10	Brain	90	5 920	5 920	0	5
11	Leukemia	72	7 129	7 129	0	2
12	Breast	84	9 216	9 216	0	5
13	Lung	181	12 533	12 533	0	2
14	MLL	72	12 582	12 582	0	3
15	Prostate	136	12 600	12 600	0	2

5.2. Effect of different neighborhood parameter values

The second portion of our experiments pays more attention to the classification accuracy and the reduction rate under the different neighborhood parameters. The classification accuracy versus the reduction rate on a feature subset for the different neighborhood parameters is performed to obtain an appropriate neighborhood parameter value. To explain the classification accuracy and the reduction performance with the values of different neighborhood parameters, a reduction rate needs to be presented to evaluate the performance of feature redundancy for our proposed feature selection method.

Definition 14. A reduction rate for datasets is defined as

$$Rate = 1 - \frac{\min(|R|)}{|C|}, \tag{25}$$

where $|C|$ describes the number of conditional attributes, and $|R|$ denotes the number of selected features generated under the given neighborhood parameter. Since a higher reduction rate shows that the method has a stronger reduction capacity for datasets, a higher reduction means the redundancy degree will be lower.

For the eight high-dimensional gene expression datasets, the Fisher score method [38] is employed to compute the value of the Fisher score and sort it based on all the genes from the eight datasets, and g genes are selected to form a candidate gene subset. The classification accuracy under the following seven dimensions (10, 50, 100, 200, 300, 400 and 500) is acquired so that the appropriate dimension can be selected for feature selection. Fig. 2 demonstrates the variation trend between the classification accuracy and the number of genes on the eight gene expression datasets.

According to Fig. 2, it is obvious that when the number of genes increases, the accuracy generally changes. As we know, the classification accuracy and the cardinality of selected features are two important aspects for evaluating the classification performance of the feature selection methods. It is necessary to select the appropriate number of genes from Fig. 2. Hence, the
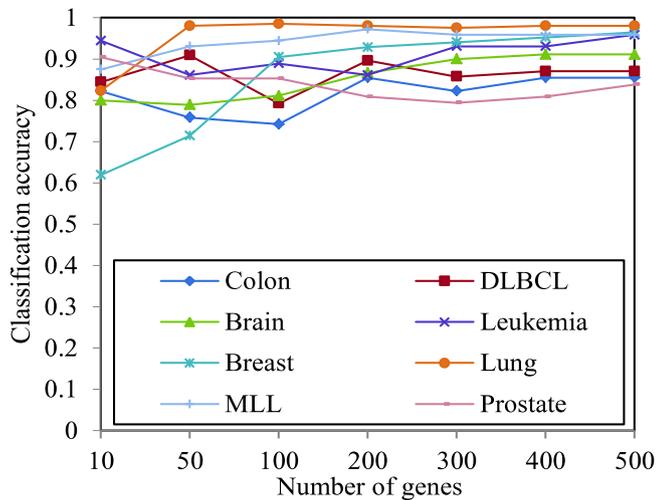
Fig. 2. Classification accuracy versus the number of genes on the eight gene expression datasets.

dimensions of genes can be set to 200-dimension features for the Colon and MLL datasets, and 50-dimension features for the DLBCL, Lung and Prostate datasets. For the Brain dataset, the number of appropriate genes can be set as 400-dimension features. Both the Leukemia and Breast datasets can employ 100-dimension features for feature selection.

Note that if all of the feature values are categorical, the neighborhood tolerance relation will degenerate to the tolerance relation [17]. Thus, the neighborhood parameter will be set as 0 for three datasets (Nursery, Mushroom and Soybean) in our experiments. The classification accuracy of selected features on the twelve datasets is obtained by using the FSNTDJE algorithm with the different neighborhood parameter values. After obtaining the results of feature selection with different parameters, the classification accuracy under the Naive Bayes and C4.5 classifiers is achieved for the four UCI datasets. For the eight gene expression datasets, the classification accuracy and the reduction rate are evaluated under the KNN ($k = 10$) and C4.5 classifiers. The results of different parameter values are shown in Fig. 3, where the horizontal coordinate represents the different neighborhood parameter values with $\delta \in [0.05, 1]$ at an interval of 0.05, and the left and right vertical coordinates state the accuracy and the reduction rate, respectively.

Fig. 3 shows that when the neighborhood parameter values change from 0.05 to 1, the classification accuracy of selected features by FSNTDJE is increasing and the reduction rate is decreasing in most cases. It can be observed that the different parameters produce a great impact on the classification performance of FSNTDJE. Thus, for each dataset, the optimal neighborhood parameter can be chosen in Fig. 3. It can be seen from Fig. 3(a) that the classification accuracy of selected features reaches the best performance when the parameter equals 0.5 on the Credit dataset under the Naive Bayes and C4.5 classifiers. For the Wpbc dataset in Fig. 3(b), the neighborhood parameter can be set to 0.15 under the Naive Bayes classifier and 0.35 under the C4.5 classifier. Fig. 3(c) demonstrates the classification accuracy of the Annealing dataset with different neighborhood parameters, and the neighborhood parameter can be set to 0.3 under both the Naive Bayes and C4.5 classifiers. Fig. 3(d) shows the classification accuracy of the Ozone level dataset with the different neighborhood parameter values, and its neighborhood parameter can be set as 0.1. Fig. 3(e) reveals that the classification accuracy of the selected feature subset reaches the best value

when the parameter value is 0.2 for the Colon dataset under the KNN and C4.5 classifiers. Similar to Fig. 3(e),(f) demonstrates that the reduction rate decreases as the neighborhood parameter values increase, and the classification accuracy of selected features reaches the relative maximum when the neighborhood parameter is 0.1 under the KNN and C4.5 classifiers. Fig. 3(g) indicates that for the Brain dataset, under the two different classifiers, the neighborhood parameter can be set as 0.45. For the Leukemia dataset in Fig. 3(h), the neighborhood parameters can be set to 0.15 and 0.05 under the KNN and C4.5 classifiers, respectively. In Fig. 3(i), on the KNN and C4.5 classifiers, the parameter can be set as 0.2 on the Breast dataset. For the Lung dataset in Fig. 3(j), we can see the slight difference of classification accuracy between the two classifiers. It follows that the neighborhood parameters can be set to 0.3 under the KNN classifier and 0.35 under the C4.5 classifier. The parameters can be set as 0.15 for the MLL dataset and 0.25 for the Prostate dataset, as observed from Fig. 3(k) and (l) under the KNN and C4.5 classifiers, respectively.

5.3. Classification results on low-dimensional UCI datasets

This part of our experiments evaluates the performance of our proposed FSNTDJE algorithm in terms of the number of selected features and the classification performance on the low-dimensional UCI datasets. The FSNTDJE algorithm is compared with four state-of-the-art feature selection algorithms, described as follows: (1) the conditional entropy-based feature selection algorithm in rough sets (FSCE) [26], (2) the positive approximation-based incomplete feature selection algorithm in rough sets (IF-SPA) [43], (3) the positive region-based feature selection algorithm using the rough set model (FSPR) [18], and (4) the rough set theory-based heuristic SetCover feature selection algorithm (SetCover) [26]. By using the different neighborhood parameters obtained in Section 5.2 and following the techniques proposed in [18,26,39,40], the average number of selected features and the appropriate neighborhood parameters on the seven UCI datasets from Table 2 are obtained and shown in Table 3, where the average sizes of selected feature subsets of the five methods are obtained from the 10-fold cross-validation method. Note that the bold font indicates the best result in all the following subsections.

Table 3 lists the average number of features selected by the five different algorithms using 10-fold cross-validation. As seen from Table 3, with the Naive Bayes and C4.5 classifiers, the average number of features selected by FSNTDJE is less than that of FSCE, IFSPA, FSPR and SetCover in most cases. For the Nursery and Soybean datasets, the SetCover and FSNTDJE algorithms have obtained nearly the same average number of features under the two different classifiers; however, both of them have approximately one more than the other three algorithms. On the Credit and Annealing datasets, the average number of features selected by FSNTDJE is slightly lower than that of FSPR, and they achieve the minimum among the five algorithms. For the Mushroom dataset, the number of features selected by FSNTDJE is 4.2, which is 0.2–3.4 lower than that of the other four algorithms. For the Wpbc dataset, the number of selected features of FSNTDJE is 4.9 under the Naive Bayes classifier, which reaches the lowest value. Nevertheless, under the C4.5 classifier, the average number of features selected by FSNTDJE is nearly the same as that by the SetCover algorithm for Wpbc. Under the two classifiers, FSNTDJE reaches the minimal number of features on the Ozone level dataset. Furthermore, the Mean index denotes the mean value of all the results in the following subsections. It is obvious that the mean result of FSNTDJE under the Naive Bayes classifier in Table 3 is the minimal. On the whole, the proposed FSNTDJE algorithm is efficient in terms of the average number of selected features for all datasets.

(a) Credit dataset

(b) Wpbc dataset

(c) Annealing dataset

(d) Ozone level dataset

(e) Colon dataset

(f) DLBCL dataset

(g) Brain dataset

(h) Leukemia dataset

(i) Breast dataset

(j) Lung dataset
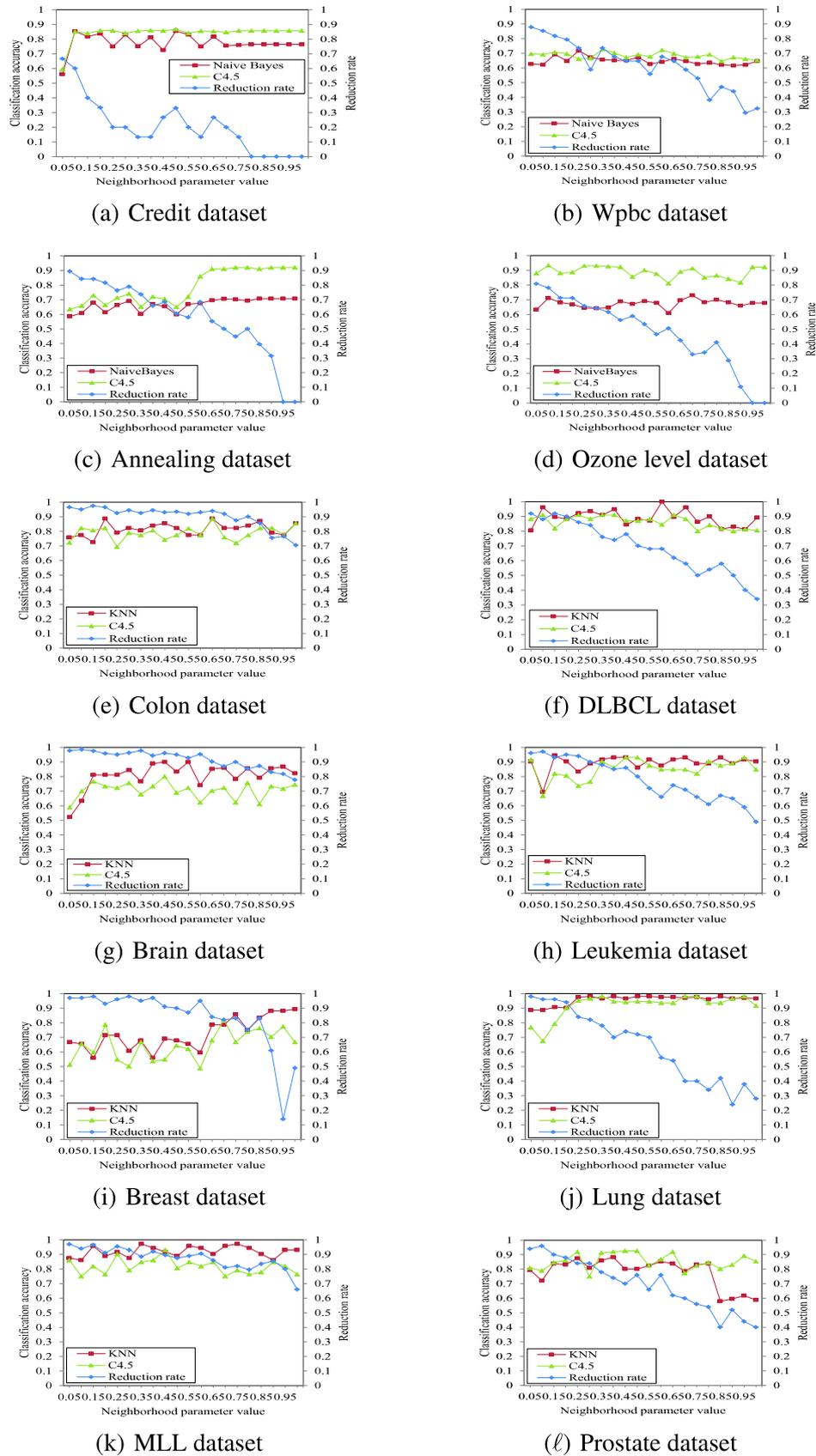
(k) MLL dataset

(ℓ) Prostate dataset

Fig. 3. The classification accuracy and the reduction rate for the twelve datasets with different neighborhood parameter values.

Table 3
The number of selected features with the five algorithms.

Datasets	FSCE	IFSPA	FSPR	SetCover	FSNTDJE	
					Naive Bayes(δ)	C4.5(δ)
Nursery	7.9	**6.7**	7	8.4	8.2(0)	8.2(0)
Credit	11.9	11.2	9.7	10.9	**9.5**(0.5)	**9.5**(0.5)
Mushroom	4.4	7.6	4.5	5.5	**4.2**(0)	**4.2**(0)
Wpbc	11.6	12.1	10.4	8.9	**4.9**(0.15)	7.6(0.35)
Soybean	9	9.3	**8.9**	9.5	9.8(0)	9.8(0)
Annealing	12.1	10.2	7.6	9.8	**7.5**(0.3)	**7.5**(0.3)
Ozone level	31.8	30.4	21.5	21.1	**16.4**(0.1)	**16.4**(0.1)
Mean	12.7	12.5	9.9	10.6	**8.6**	9.0

Table 4
The classification accuracy of the six methods under the Naive Bayes classifier.

Datasets	ODP	FSCE	IFSPA	FSPR	SetCover	FSNTDJE
Nursery	0.9211	0.8994	0.8929	0.8991	0.9032	**0.9451**
Credit	0.7638	0.8124	0.8191	0.8086	0.8065	**0.8536**
Mushroom	0.9201	0.9219	0.9271	0.9019	0.9169	**0.9583**
Wpbc	0.6111	0.6569	0.6625	0.6706	**0.6718**	0.6649
Soybean	0.9055	0.8828	0.8801	0.8792	**0.9159**	0.8156
Annealing	0.5877	0.6166	0.6072	0.6002	0.6118	**0.9136**
Ozone level	0.6784	0.6595	0.6819	0.7004	0.6827	**0.7808**
Mean	0.7751	0.7785	0.7851	0.7800	0.7870	**0.8747**

Table 5
The classification accuracy of the six methods under the C4.5 classifier.

Datasets	ODP	FSCE	IFSPA	FSPR	SetCover	FSNTDJE
Nursery	**0.9705**	0.9501	0.9337	0.9435	**0.9705**	0.9219
Credit	0.8565	0.8245	0.8461	0.8352	0.8568	**0.8765**
Mushroom	**1**	0.9766	0.9816	**1**	**1**	0.9951
Wpbc	0.6919	0.7071	0.7262	0.7065	0.7124	**0.7374**
Soybean	0.8338	0.8217	0.8338	0.8470	0.8369	**0.8769**
Annealing	0.9048	**0.9235**	0.9063	0.9044	0.9125	0.9201
Ozone level	0.9215	0.9214	0.9258	0.9305	0.9254	**0.9522**
Mean	0.8793	0.8750	0.8791	0.8810	0.8878	**0.8972**

In what follows, the average classification accuracy of our proposed method is illustrated using the six methods to evaluate the classification performance on the selected features, i.e., the feature subsets are selected with 10-fold cross-validation; moreover, to obtain objective classification results and reduce random errors, all the compared methods are performed 10 times, and the results are the average value of the 10 evaluations of the classification accuracy. The FSNTDJE algorithm is compared with the above four feature selection methods (FSCE, IFSPA, FSPR and SetCover) and the original data processing method (ODP). The two classifiers (Naive Bayes and C4.5) are employed to test the classification performance. Following the designed techniques in [18,26,39,40], the average classification accuracy of optimal features selected by the six methods under the Naive Bayes and C4.5 classifiers is shown in Tables 4 and 5, respectively.

As seen from Table 3, there is little differences in the average number of features selected by the five algorithms, on the basis of which, Tables 4 and 5 show the differences among the six methods. It is obvious that the classification accuracy of the FSNTDJE algorithm outperforms that of the other five methods on most of the datasets, except for the Wpbc and Soybean datasets under the Naive Bayes classifier, and the Nursery, Mushroom and Annealing datasets under the C4.5 classifier. Moreover, the mean accuracy of FSNTDJE in Tables 4 and 5 has been improved and is the highest on the two different classifiers. From Tables 3 and 4, under the Naive Bayes classifier, although FSNTDJE does not perform as well as FSCE, IFSPA and FSPR in the average sizes of selected features, the classification accuracy of FSNTDJE is 2.4–5.22% higher than that of ODP, FSCE, IFSPA and FSPR for the Nursery dataset; however, FSNTDJE is inferior to SetCover on the Soybean dataset. Though there are some differences in the average sizes of features

selected by the five algorithms, the average accuracy of FSNTDJE is higher than that of other methods on all datasets, except for the Wpbc and Soybean datasets. Similarly, as seen from Tables 3 and 5, under the C4.5 classifier, the mean accuracy of FSNTDJE is 0.94–2.22% higher than that of the other five methods, and the accuracy of FSNTDJE is nearly the same as that of the SetCover in all datasets. For the Mushroom dataset, the classification accuracy of FSNTDJE is 0.05% lower than that of ODP, FSPR and SetCover; however, FSNTDJE selects fewer features than the other methods, and it displays better classification performance than FSCE and IFSPA. In summary, in terms of the mean accuracy, our FSNTDJE algorithm demonstrates great stability with Naive Bayes and C4.5, whereas the accuracy of the ODP, FSPR and SetCover algorithms is slightly unstable. It can be ascertained from the results in Tables 4 and 5 that for the Nursery dataset under Naive Bayes, and the Mushroom and Annealing datasets under C4.5, it is possible that FSNTDJE reduces some important features in the process of feature selection, which results in decreasing the classification accuracy of the selected feature subsets.

In the process of conducting the experiments, the rough sorting of the five feature selection methods on the time complexity is obtained as follows: $O(FSNTDJE) < O(FSCE) = O(SetCover) < O(FSPR) < O(IFSPA)$, where $O(A)$ represents the time complexity of the A algorithm. Suppose that there are m samples and n features, and the time complexity of FSCE is $O(mn^2)$ [26]. For the low-dimensional UCI datasets, the time complexity of FSCE is low. Notably, these UCI datasets usually have a large number of samples and a small number of features in most instances; however, the microarray gene expression datasets with the small sample size are high-dimensional. Since $m \ll n$ on the large-scale and high-dimensional datasets, the time complexity of FSNTDJE is $O(mn)$, which is much less than that of FSCE. For the SetCover algorithm, its time complexity is $O(mn^2)$ [40]. Since the time complexity of the FSPR algorithm is mainly spent on the calculation of tolerance classes of each sample for the different decision classes, FSPR runs slowly, and its time complexity is $O(mn^2 \log m)$ [18]. In addition, the time complexity of IFSPA is $O(mn^3)$ [39], which is very time-consuming. For computing tolerance classes, IFSPA consumes much time. Therefore, one can conclude that the FSNTDJE algorithm achieves lower time complexity and can efficiently eliminate the redundant features and optimize the classification performance of incomplete datasets.

The third part of this subsection is to further illustrate the classification results of the FSNTDJE algorithm on some selected datasets in terms of the number of selected features and their classification accuracy. Here, four state-of-the-art feature selection methods for incomplete datasets used in comparison include (1) the heuristic discernibility matrix-based feature selection algorithm using fuzzy rough sets (DMFS) [45], (2) the rough sets-based backward attribute reduction algorithm for incomplete ordered information systems with fuzzy decision (BKAR) [46], (3) the mutual information-based feature selection algorithm with the forward greedy strategy from incomplete decision systems in rough sets (MIFS) [36], and (4) the discernibility matrix-based knowledge reduction algorithm using dominance rough sets (DMKR) [47]. It should be noted that to compare the FSNTDJE algorithm with the abovementioned four feature selection methods and the ODP method, the Credit and Annealing datasets are selected from Table 2 for convenience. Following the experimental techniques and parameters designed in [36,45–47], Tables 6 and 7 show the experimental results of the six different methods, where the average number of selected features with 10-fold cross-validation and the average accuracy of the 10 evaluations can be implemented under the Naive Bayes and C4.5 classifiers.

As shown in Table 6, FSNTDJE achieves the lowest number of the selected Credit features and the highest classification accuracy. Compared with the other five methods, the average number

Table 6

The classification accuracy of the selected Credit features with the six methods.

Methods	Features	Naive Bayes	C4.5	Mean
ODP	15	0.7638	0.8565	0.8101
DMFS	11.9	0.7891	0.7873	0.7882
BKAR	11.5	0.7849	0.8059	0.7954
MIFS	12.4	0.8014	0.8373	0.8194
DMKR	13.7	0.7699	0.7924	0.7812
FSNTDJE	**9.5**	**0.8536**	**0.8765**	**0.8651**

Table 7

The classification accuracy of the selected Annealing features with the six methods.

Methods	Features	Naive Bayes	C4.5	Mean
ODP	38	0.5877	0.9048	0.7463
DMFS	10.7	0.9170	0.8823	0.8997
BKAR	9.2	**0.9224**	0.9107	0.9166
MIFS	9.6	0.8351	**0.9241**	0.8796
DMKR	10.0	0.9146	0.9033	0.9089
FSNTDJE	**7.5**	0.9136	0.9201	**0.9169**

Table 8

The number of selected genes with the five algorithms.

Datasets	CFS	FCBF	INT	IG	FSNTDJE	
					Naive Bayes(δ)	C4.5(δ)
Colon	23.6	14.4	14	10.1	**9.5**(0.4)	11.5(0.2)
DLBCL	65.1	37.2	50.5	**4.8**	7.3(0.25)	6(0.1)
Brain	35.7	**1.4**	49.1	13.4	14.4(0.45)	16.1(0.45)
Breast	130.1	99.0	102.0	11.9	**7**(0.2)	**7**(0.2)
Prostate	89.3	76.5	72.5	9.2	**5.9**(0.2)	7.2(0.25)
Mean	68.76	45.7	57.62	9.88	**9.04**	9.56

of features selected by our algorithm is 2–4.5 lower than that achieved by five methods. Moreover, with the Naive Bayes and C4.5 classifiers, the accuracy of features selected by FSNTDJE is higher than that of the other five methods, i.e., under the Naive Bayes classifier, the accuracy of our algorithm is 5.22–8.98% higher than the other methods, and under the C4.5 classifier, it is 2–8.92% higher than them. Hence, the FSNTDJE algorithm can achieve great classification performance for the Credit dataset.

According to the classification results in Table 7, the FSNT-DJE algorithm obtains the lowest number of selected Annealing features and the highest mean accuracy, based on which, the accuracy of features selected by our algorithm under the Naive Bayes classifier is as good as that of the DMFS, BKAR and DMKR algorithms and is 32.59% and 7.85% higher than that of ODP and MIFS, respectively. Moreover, the accuracy of FSNTDJE is similar to MIFS and higher than that of the other four methods under the C4.5 classifier. Thus, our proposed FSNTDJE algorithm can remove redundant features from the original Annealing dataset.

From all the above results and analysis, it is obvious that there is no algorithm consistently better than the others for different learning tasks and classifiers. In general, it can be observed from Tables 4 to 7 that our FSNTDJE algorithm, compared with these other feature selection methods, can reflect the decision-making ability of features, avoid the loss of useful information caused by discretization, and tackle the uncertainty and efficiently improve the classification performance in incomplete neighborhood decision systems. Therefore, the proposed FSNTDJE algorithm outperforms the other related feature selection methods on incomplete low-dimensional UCI datasets.

5.4. Classification results on high-dimensional gene expression datasets

This section of our experiments demonstrates the classification performance of our proposed method on high-dimensional gene expression datasets. The FSNTDJE algorithm is compared with four state-of-the-art feature selection methods, which include (1) the rough sets-based correlation feature selection algorithm (CFS) [44], (2) the fast correlation-based filter feature selection algorithm in rough sets (FCBF) [42], (3) the interacting features selection algorithm, which can handle feature interaction and efficiently selects relevant features (INT) [42], and (4) the information gain and divergence-based feature selection algorithm for statistical machine learning (IG) [43]. Similar to Section 5.3, the average sizes of feature subsets selected by the

above methods using 10-fold cross-validation and the appropriate neighborhood parameters on the gene expression datasets are obtained and shown in Table 8. Following the designed experiments in [41–44], the five gene expression datasets (Colon, DLBCL, Brain, Breast and Prostate) are selected from Table 2 to conveniently compare the abovementioned four algorithms.

Table 8 displays the average number of genes selected by the five feature selection algorithms using 10-fold cross-validation under the Naive Bayes and C4.5 classifiers. In Table 8, it is clear that our proposed FSNTDJE algorithm is superior to the CFS, FCBF and INT algorithms in most cases; however, for the Brain dataset, FCBF provides the best result, and IG selects the optimal number of genes on the DLBCL dataset. The FSNTDJE algorithm selects the minimum average genes for the Colon and Prostate datasets under the Naive Bayes classifier. For the Breast dataset, the number of genes selected by FSNTDJE is 7 under the two different classifiers and reaches the minimal value. Furthermore, the mean number of genes selected by FSNTDJE is the best, which is 0.84 and 0.32 lower than that of IG on the two different classifiers. In summary, our proposed approach can select the fewest genes for the high-dimensional gene expression datasets.

Based on the results in Table 8, the Naive Bayes and C4.5 classifiers are employed to evaluate the classification results for the five gene expression datasets (Colon, DLBCL, Brain, Breast and Prostate). It is well known that three indices including accuracy (*Acc*), the true positive rate (*TPR*) and the false positive rate (*FPR*), are usually used to evaluate the classification performance of feature selection. The higher the *TPR* is, the lower the *FPR* is, and the better the method is [44]. The formulas of three indices are denoted, respectively [44,48,49], as

$$Acc = \frac{TN + TP}{TN + TP + FN + FP}, \tag{26}$$

$$TPR = \frac{TP}{TP + FN}, \tag{27}$$

$$FPR = \frac{FP}{TN + FP}, \tag{28}$$

where True Positive (*TP*) represents the number of positive samples detected as correctly, False Positive (*FP*) describes the number of positive samples detected as falsely, True Negative (*TN*) denotes the number of negative instances diagnosed as true, and False Negative (*FN*) states the negative instances diagnosed as false. Tables 9 and 10 show the values of *Acc*, *TPR* and *FPR* of selected genes with the six methods under the Naive Bayes and C4.5 classifiers, respectively. Similar to the previous subsections, all the compared methods are executed 10 times, and the values of *Acc* are evaluated as the average of 10 classification operations.

It can be easily observed from Table 8 that the five algorithms differ greatly in the average number of selected genes. According to Tables 9 and 10, the classification accuracy of FSNTDJE outperforms that of the other five methods, except for the Colon and Brain datasets under the Naive Bayes classifier and the Brain

Table 9
Three indices with the six methods under the Naive Bayes classifier.

Methods	Indices	Colon	DLBCL	Brain	Breast	Prostate	Mean
ODP	*Acc*	0.7742	0.9091	**0.9111**	0.7380	0.5882	0.7841
	TPR	0.7740	0.0900	0.9000	0.7380	0.9150	0.6834
	FPR	0.1820	0.1030	0.0130	0.4040	0.3350	0.2074
CFS	*Acc*	**0.8500**	0.9000	0.8100	0.3700	0.2600	0.6380
	TPR	0.7600	0.9600	0.5000	1	**1**	0.8440
	FPR	**0.1000**	0.1600	**0**	1	1.0000	0.4520
FCBF	*Acc*	0.8000	0.9000	0.6100	0.3700	0.2600	0.5880
	TPR	0.7600	0.9600	1	1	**1**	0.9440
	FPR	0.1800	0.1600	0.6000	1	1	0.5880
INT	*Acc*	0.7700	0.9000	0.8100	0.3700	0.2600	0.6220
	TPR	0.7600	0.9600	0.5000	1	**1**	0.8440
	FPR	0.2300	0.1700	**0**	1	1.0000	0.4800
IG	*Acc*	0.7900	0.9400	0.8600	0.3200	0.2600	0.6340
	TPR	0.7200	0.9600	0.7000	0.3200	0.2600	0.5920
	FPR	0.1800	0.0800	0.0700	1	0.9600	0.4580
FSNTDJE	*Acc*	0.8312	**0.9659**	0.8952	**0.7857**	**0.6938**	**0.8344**
	TPR	**0.8550**	1	0.9170	1	0.9490	**0.9442**
	FPR	0.1650	**0.0260**	**0**	**0.0140**	0.3050	0.1020

Table 10
Three indices with the six methods under the C4.5 classifier.

Methods	Indices	Colon	DLBCL	Brain	Breast	Prostate	Mean
ODP	*Acc*	0.8548	0.8312	0.7444	0.7500	0.8529	0.8066
	TPR	0.8550	0.6320	0.8000	0.7500	0.8700	0.7814
	FPR	**0.0750**	0.1030	0.0480	0.0960	0.1690	**0.0982**
CFS	*Acc*	0.7900	0.7400	**0.9700**	0.6800	0.2600	0.6880
	TPR	0.6800	0.6600	**0.9600**	0.7100	1	0.8020
	FPR	0.1500	0.1600	**0**	0.3400	1.0000	0.3300
FCBF	*Acc*	0.7900	0.7200	0.8600	0.5800	0.2600	0.6420
	TPR	0.6400	0.6200	0.8700	0.2800	1	0.6820
	FPR	0.1300	0.1600	0.0600	0.2500	1	0.3200
INT	*Acc*	0.7900	0.7000	**0.9700**	0.7900	0.2600	0.7020
	TPR	0.7200	0.6200	**0.9600**	0.7100	1	0.8020
	FPR	0.1800	0.2000	**0**	0.1700	1	0.3100
IG	*Acc*	0.7200	0.7600	0.7100	0.4700	0.2600	0.5840
	TPR	0.7800	0.6600	0.7400	0.2800	1	0.6920
	FPR	0.3000	0.1200	0.1600	0.4700	1	0.4100
FSNTDJE	*Acc*	**0.9119**	**0.8957**	0.8742	**0.8214**	**0.9367**	**0.8880**
	TPR	**0.8870**	**0.7890**	0.9170	**0.8950**	0.9480	**0.8872**
	FPR	0.1250	**0.0520**	0.2000	**0.0270**	**0.1190**	0.1046

Table 11
The classification accuracy of the six methods under the KNN classifier.

Datasets	ODP	DMRA	FPRA	FRFS	IFPR	FSNTDJE
Colon	0.7258	0.7625	0.7958	0.6991	0.6574	**0.8763**
Leukemia	0.7344	0.7776	0.7477	0.7619	0.8042	**0.9019**
Breast	0.6909	**0.8158**	0.6196	0.7321	0.6725	0.8027
MLL	0.6528	0.8859	0.9112	0.8160	0.9405	**0.9615**
Mean	0.7010	0.8105	0.7686	0.7523	0.7687	**0.8856**

Table 12
The classification accuracy of the six methods under the CART classifier.

Datasets	ODP	DMRA	FPRA	FRFS	IFPR	FSNTDJE
Colon	0.5967	0.7875	0.8125	0.7252	0.7595	**0.8910**
Leukemia	0.5903	0.5492	0.5492	0.6800	0.7556	**0.7628**
Breast	0.5714	0.7131	0.6104	0.7000	0.7312	**0.7642**
MLL	0.7500	0.8018	0.7537	0.8227	**0.9112**	0.8492
Mean	0.6271	0.7129	0.6815	0.7320	0.7894	**0.8168**

dataset under the C4.5 classifier. In addition, the values of *TPR* and *FPR* of FSNTDJE achieve better results on most of the five datasets. From Tables 8 and 9, under the Naive Bayes classifier, the differences among the six methods can be clearly identified. Although FSNTDJE does not perform as well as IG in terms of the average genes selected from the DLBCL and Brain datasets, the mean values of *Acc*, *TPR* and *FPR* are the best with the Naive Bayes classifier. The *Acc* of FSNTDJE is nearly 1.8% and 1.59% lower than that of CFS and ODP for the Colon and Brain datasets, respectively, because the FSNTDJE algorithm lose some important genes of the Colon and Brain datasets during reduction, resulting in the reduction of classification accuracy. It is known that the lower the *FPR* is, the better the performance of the method is [44]. For the Breast dataset, FSNTDJE achieves the minimum in the average number of genes and obtains the best results in the three indices. Although FSNTDJE is approximately 6% lower than that of CFS, FCBF and INT for the Prostate dataset, and 9% lower than that of FCBF for the Brain dataset in the *TPR*, the values of *FPR* are minimal for the Prostate and Brain datasets. Similarly, as seen from Tables 8 and 10, under the C4.5 classifier, the mean values of *Acc* of FSNTDJE are nearly 8.2–30.4% higher than those of the other five methods. Compared with the results of ODP, the *TPR* of our method has been significantly improved with the exception of the Prostate dataset. Furthermore, regarding the *FPR*, for the DLBCL and Prostate datasets, FSNTDJE has the lowest *FPR*; however, its mean value is nearly 6.4% higher than that of the ODP. Based on the results in Table 8, 9 and 10, although FSNTDJE does not select the fewest genes on the DLBCL and Brain datasets, FSNTDJE reaches the relative best results in most of the gene expression datasets. Overall, the experimental results show that our approach is efficient in eliminating redundant genes and improves *Acc* and *TPR* on the high-dimensional gene expression datasets.

Similar to the previous analysis of time complexity for the low-dimensional UCI datasets, the comparison of the above five methods illustrates a rough order in terms of time complexity as follows: $O(\text{FSNTDJE}) < O(\text{FCBF}) < O(\text{INT}) < O(\text{CFS}) < O(\text{IG})$. The time complexity of CFS is $O(n^2)$ [44,50]. For gene expression datasets, the number of genes is much larger than the number of samples, and the time complexity of CFS is greater than that of FSNTDJE. For the FCBF algorithm, the time complexity is not more than $O(m^2 n)$ [41]. The time complexity of INT is $O(mn^2)$ [42]. For the IG algorithm in [43], the time complexity is not more than $O(mn^2 \log n)$. It can be concluded that FSNTDJE achieves a lower time complexity than the other four algorithms for high-dimensional gene expression datasets.

The following portion of this experiment continues testing the classification performance of the FSNTDJE method on the four gene expression datasets (Colon, Leukemia, Breast and MLL), selected from Table 2. Four state-of-the-art methods compared with FSNTDJE are described as follows: (1) the discernibility matrix-based reduction algorithm in fuzzy rough sets (DMRA) [51], (2) the fuzzy positive region-based accelerator algorithm in rough sets (FPRA) [52], (3) the fuzzy rough sets-based feature selection algorithm by using the boundary region (FRFS) [53], and (4) the intuitionistic fuzzy positive region-based gene selection algorithm in fuzzy rough sets (IFPR) [54]. Here, by using different neighborhood parameters obtained in Section 5.2, similar to the previous evaluated results, all the compared methods are run 10 times, and the average classification accuracy of the four gene expression datasets under the KNN ($k = 10$) and CART classifiers is the mean of the 10 evaluations. Following the experimental techniques presented in [51–54], Tables 11 and 12 demonstrate the classification accuracy of the six methods under the KNN and CART classifiers, respectively.

As seen from Tables 11 and 12, the average classification accuracy of the FSNTDJE algorithm outperforms that of the other five methods on almost all datasets, except for the Breast dataset under the KNN classifier and the MLL dataset under the CART classifier, and the mean accuracy of FSNTDJE is the highest. According to Table 11, under the KNN classifier, for the Colon, Leukemia and MLL datasets, the classification accuracy of FSNTDJE reaches the highest values, i.e., 87.63%, 90.19% and 96.15%, respectively. However, the accuracy of FSNTDJE is slightly inferior to that of

Table 13

The classification accuracy of the five methods under the Naive Bayes classifier.

Datasets	ODP	DBAGEL	DGAFS	DGAFS-MI	FSNTDJE
Colon	0.7742	0.7702	0.7370	0.7698	**0.8312**
DLBCL	0.9091	0.7795	0.7481	0.8536	**0.9659**
Breast	0.7380	0.7616	0.6293	0.6837	**0.7857**
Lung	0.8226	0.8877	0.7914	0.8780	**0.9803**
Prostate	0.5882	**0.7766**	0.5231	0.6884	0.6938
Mean	0.7664	0.7951	0.6858	0.7747	**0.8514**

Table 14

The classification accuracy of the five methods under the KNN classifier.

Datasets	ODP	DBAGEL	DGAFS	DGAFS-MI	FSNTDJE
Colon	0.7258	0.8298	0.7350	0.7649	**0.8763**
DLBCL	0.8052	0.9205	0.7434	0.8565	**0.9610**
Breast	0.6909	**0.8384**	0.7362	0.7818	0.8027
Lung	0.9605	0.9228	0.8911	0.8868	**0.9803**
Prostate	0.8309	0.8234	0.7332	0.6842	**0.8750**
Mean	0.8027	0.8670	0.7678	0.7948	**0.8991**

DMRA for the Breast dataset. Similarly, as shown in Table 12, under the CART classifier, FSNTDJE achieves the highest accuracy on almost all datasets; however, for the MLL dataset, the average accuracy of FSNTDJE is 6.2% lower than that of IFPR, and 2.65–9.92% higher than that of the other four methods. Overall, in terms of the average accuracy, our FSNTDJE algorithm demonstrates stronger stability for the four high-dimensional gene expression datasets under the KNN and CART classifiers, whereas the classification performance of the DMRA, FPRA, FRFS and IFPR algorithms is slightly unstable. Hence, it can be proven that the FSNTDJE algorithm can eliminate the redundant genes, significantly improve the classification performance, and outperform the other five related feature selection methods for high-dimensional gene expression datasets.

The third part of this subsection further evaluates the classification performance of the FSNTDJE algorithm compared with three state-of-the-art feature selection methods in terms of the classification accuracy of selected gene subsets. These feature selection methods for contrasts include (1) the dynamic Bayesian genetic feature selection algorithm, which is designed by enhancing the principles of the Bayesian genetic algorithm in rough sets (DBAGEL) [55], (2) the dynamic genetic algorithm-based feature selection method for selecting the significant features (DGAFS) [56], and (3) the rough sets-based feature selection algorithm by selecting significant features and imputing missing values (DGAFS-MI) [56]. Following the experimental techniques designed in [55,56], the five gene expression datasets (Colon, DLBCL, Breast, Lung and Prostate) are selected from Table 2, the Naive Bayes and KNN ($k = 10$) classifier are employed, and then Tables 13 and 14 show the experimental results of the five different feature selection methods in detail under the two different classifiers, respectively.

According to Tables 13 and 14, the differences among the five feature selection methods can be clearly identified. As shown in Table 13 under the Naive Bayes classifier, FSNTDJE performs markedly better than the other four methods in classification accuracy except for the Prostate dataset, and the accuracy of the ODP, DBAGEL, DGAFS and DGAFS-MI algorithms is similar on the five gene expression datasets. On the Prostate dataset, the accuracy of FSNTDJE is 8.28% lower than that of DBAGEL, and it is 0.54–17.07% higher than that of the other three methods. This reason is that when FSNTDJE processes the gene dataset, some noises of the Prostate dataset still exist, so that this case decreases the accuracy. However, the mean accuracy of FSNTDJE is nearly 5.63–16.54% higher than that of the other four methods

Table 15

The optimal features selected by FSNTDJE on the seven UCI datasets under the Naive Bayes classifier.

Datasets	Selected feature subsets	Number of features
Nursery	{1, 2, 3, 4, 5, 6, 7, 8}	8
Credit	{1, 2, 3, 4, 6, 9, 10, 12, 13, 14}	10
Mushroom	{1, 4, 5, 22}	4
Wpbc	{1, 2, 6, 14, 16, 32}	6
Soybean	{1, 4, 6, 7, 8, 12, 15, 19, 22, 35}	10
Annealing	{1, 6, 8, 10, 16, 32, 33, 34}	8
Ozone level	{1, 3, 4, 7, 9, 12, 25, 33, 37, 39, 46, 47, 52, 55, 62, 63}	16

Table 16

The optimal features selected by FSNTDJE on the seven UCI datasets under the C4.5 classifier.

Datasets	Selected feature subsets	Number of features
Nursery	{1, 2, 3, 4, 5, 6, 7, 8}	8
Credit	{1, 2, 3, 4, 6, 9, 10, 12, 13, 14}	10
Mushroom	{1, 4, 5, 22}	4
Wpbc	{1, 3, 5, 12, 13, 16, 23, 24, 32}	9
Soybean	{1, 4, 6, 7, 8, 12, 15, 19, 22, 35}	10
Annealing	{1, 6, 8, 10, 16, 32, 33, 34}	8
Ozone level	{1, 3, 4, 7, 9, 12, 25, 33, 37, 39, 46, 47, 52, 55, 62, 63}	16

and reaches the highest value. As seen from Table 14, under the KNN classifier, the average classification accuracy of the gene subsets selected by FSNTDJE is the best on the Colon, DLBCL, Lung and Prostate datasets. However, for the Breast dataset, the accuracy of FSNTDJE is 3.57% lower than that of DBAGEL, and it is 2.09–11.18% higher than that of the other three methods. The reason is that FSNTDJE cannot sufficiently eliminate the noise genes, which results in weakening the classification performance of the selected Breast genes. In summary, the FSNTDJE model can efficiently reduce the dimensions of high-dimensional gene expression datasets and achieve great classification performance on these large-scale and high-dimensional datasets.

5.5. Optimal results of feature selection with FSNTDJE

In this subsection of our experiments, by using the FSNTDJE algorithm, the above neighborhood parameter values are set as in Section 5.2. Following the experimental techniques designed by Wang et al. [25], the optimal results of feature selection on the seven UCI datasets are shown in Tables 15 and 16, and the eight gene expression datasets are illustrated in Tables 17 and 18. It can be seen from Tables 15 and 16 that all the UCI datasets achieve the same feature subset under the Naive Bayes and C4.5 classifiers, expect for the Wpbc dataset. From Tables 17 and 18, the selected gene subsets for the Colon, DLBCL, Brain, Breast, MLL and Prostate datasets under the KNN classifier are the same as that under the C4.5 classifier.

5.6. Statistical analysis

To demonstrate the statistical performance of the results of feature selection, the Friedman test [57] and Bonferroni–Dunn test [58] are employed to further study the classification accuracy of each classifier with the several different methods. The Friedman statistic is represented as

$$\chi_F^2 = \frac{12T}{s(s+1)}\left(\sum_{a=1}^{s} R_a^2 - \frac{s(s+1)^2}{4}\right), \tag{29}$$

$$F_F = \frac{(T-1)\chi_F^2}{T(s-1) - \chi_F^2}, \tag{30}$$

Table 17
The optimal features selected by FSNTDJE on the eight gene expression datasets under the KNN classifier.

Datasets	Selected feature subsets	Number of features
Colon	{493, 1770, 590, 384, 765, 822, 1423, 1541, 1060, 581, 1247}	11
DLBCL	{453, 4809, 3371, 1156, 1656, 4767}	6
Brain	{1879, 2095, 2459, 3019, 2295, 4151, 5175, 5413, 5604, 820, 5281, 4560, 4578, 1602, 633, 4801}	16
Leukemia	{6696, 2010, 4925, 4211, 5300, 6801, 4609}	7
Breast	{6425, 744, 6802, 2904, 6024, 5644, 8700}	7
Lung	{8457, 11150, 7934, 6597, 6200, 7905, 10547, 12431}	8
MLL	{7930, 3054, 8020, 8992, 2933, 10581, 7811}	7
Prostate	{7710, 8850, 4483, 5155, 6185, 5314, 8768, 5757}	8

Table 18
The optimal features selected by FSNTDJE on the eight gene expression datasets under the C4.5 classifier.

Datasets	Selected feature subsets	Number of features
Colon	{493, 1770, 590, 384, 765, 822, 1423, 1541, 1060, 581, 1247}	11
DLBCL	{453, 4809, 3371, 1156, 1656, 4767}	6
Brain	{1879, 2095, 2459, 3019, 2295, 4151, 5175, 5413, 5604, 820, 5281, 4560, 4578, 1602, 633, 4801}	16
Leukemia	{758, 3108, 4267, 6575}	4
Breast	{6425, 744, 6802, 2904, 6024, 5644, 8700}	7
Lung	{7642, 8457, 11150, 7934, 6597, 6200, 7905, 10547, 4584, 8396}	10
MLL	{7930, 3054, 8020, 8992, 2933, 10581, 7811}	7
Prostate	{7710, 8850, 4483, 5155, 6185, 5314, 8768, 5757}	8

Table 19
The ranking of the six methods under the Naive Bayes classifier.

Datasets	ODP	FSCE	IFSPA	FSPR	SetCover	FSNTDJE
Nursery	1.5	3	5	4	1.5	6
Credit	3	6	4	5	2	1
Mushroom	2	6	5	2	2	4
Wpbc	6	4	2	5	3	1
Soybean	4.5	6	4.5	2	3	1
Annealing	4	1	3	5	2	6
Ozone level	5	6	3	2	4	1
Mean	3.71	4.57	3.79	3.57	2.5	2.86

Table 20
The ranking of the six methods under the C4.5 classifier.

Datasets	ODP	FSCE	IFSPA	FSPR	SetCover	FSNTDJE
Nursery	2	4	6	5	3	1
Credit	6	3	2	4	5	1
Mushroom	4	3	2	6	5	1
Wpbc	6	5	4	2	1	3
Soybean	2	3	4	5	1	6
Annealing	6	2	4	5	3	1
Ozone level	5	6	4	2	3	1
Mean	4.43	3.71	3.71	4.14	3	2

Table 21
The ranking of the six methods under the two classifiers.

Methods	Naive Bayes		Mean	C4.5		Mean
	Credit	Annealing		Credit	Annealing	
ODP	6	6	6	2	4	3
DMFS	3	2	2.5	6	6	6
BKAR	4	1	2.5	4	3	3.5
MIFS	2	5	3.5	3	1	2
DKMR	5	3	4	5	5	5
FSNTDJE	1	4	2.5	1	2	1.5

where s is the number of methods, T is the number of datasets, and R_a is the mean ranking of method A over all the datasets. F_F follows a Fisher distribution with $s-1$ and $(s-1)(T-1)$ degrees of freedom. If the null hypothesis is rejected after the Friedman test, the Bonferroni–Dunn test can be introduced to further detect which algorithms are different in statistical terms [59]. Following the experiments designed in [59], on the basis of the test results of the above subsections, if the average level of the distance exceeds the critical distance, the two algorithms will be significantly different [59,60]. The critical distance [61] is denoted as

$$CD_\alpha = q_\alpha \sqrt{\frac{s(s+1)}{6T}}, \tag{31}$$

where q_α is the critical tabulated value for the test and α is the significance level of the Bonferroni–Dunn test.

For the seven low-dimensional UCI datasets in Tables 4 and 5, the FSNTDJE algorithm is compared with the five methods (ODP, FSCE, IFSPA, FSPR and SetCover) to conduct the Friedman statistic. We develop two Friedman tests to investigate whether the classification performance of the six feature selection algorithms is significantly different. From the classification accuracy obtained in Tables 4 and 5, the ranking results of the six algorithms under the Naive Bayes and C4.5 classifiers are shown in Tables 19 and 20, respectively.

The values of the different evaluation indices under the Naive Bayes and C4.5 classifiers can be computed according to Eqs. (29) and (30). According to Tables 19 and 20, the Bonferroni–Dunn tests on the two different classifiers denote that the FSNTDJE algorithm is significantly superior to the other five algorithms on

the whole. Since $s=6$ and $T=7$, the values of the two evaluation indices (Friedman statistics χ_F^2 and Bonferroni–Dunn test F_F) can be obtained as follows: $\chi_F^2 = 7.59$ and $F_F = 1.66$ under the Naive Bayes classifier, and $\chi_F^2 = 5.38$ and $F_F = 1.09$ under the C4.5 classifier. When the significance level $\alpha = 0.1$, the critical value of $F(5, 30)$ is 2.05. Thus, one can carry out the two Bonferroni–Dunn tests. Namely, $CD = 2.24$ can be easily obtained by Eq. (31), where the critical value $q_{0.1} = 2.24$ can be easily found in [61]. As seen from Tables 19 and 20, under the Naive Bayes classifier, the Bonferroni–Dunn tests indicate that FSNTDJE has a statistical advantage over the other five algorithms; nevertheless, under the C4.5 classifier, SetCover achieves the best results, and FSNTDJE follows as second-best.

From Tables 6 and 7, it follows that $s=6$ and $T=2$, and the ranking values of the six feature selection methods (ODP, DMFS, BKAR, MIFS, DMKR and FSNTDJE) for the Credit and Annealing datasets under the Naive Bayes and C4.5 classifiers are shown in Table 21. As seen from Table 21, under the Naive Bayes classifier, the values of χ_F^2 and F_F equal 5.43 and 1.19, respectively. In the same way, under the C4.5 classifier, $\chi_F^2 = 8.57$ and $F_F = 6$. Thus, if $\alpha = 0.1$ and the critical value of $F(5, 5)$ equals 3.45, the two Bonferroni–Dunn tests should be executed and $CD = 4.19$, where the critical value $q_{0.1} = 2.24$. From Table 21, the results of the Bonferroni–Dunn test demonstrates that our FSNTDJE method achieves better statistical performance under the Naive Bayes and C4.5 classifiers.

The following portion is devoted to statistical analysis of all the high-dimensional gene expression datasets. First, for the five gene expression datasets (Colon, DLBCL, Brain, Breast and Prostate) in Tables 9 and 10, we use the five feature selection methods (ODP, CFS, FCBF, INT and IG) to compare with FSNTDJE, and then the Friedman statistic is evaluated. According to the

Table 22
The ranking of the six methods under the Naive Bayes classifier.

Dataset	ODP	CFS	FCBF	INT	IG	FSNTDJE
Colon	5	1	3	6	4	2
DLBCL	3	5	5	5	2	1
Brain	1	5	6	5	3	2
Breast	2	3	4.5	4.5	6	1
Prostate	2	4.5	4.5	4.5	4.5	1
Mean	2.6	3.7	4.6	5	3.9	1.4

Table 23
The ranking of the six methods under the C4.5 classifier.

Dataset	ODP	FSCE	IFSPA	PR	SetCover	FSNTDJE
Colon	2	4	4	4	6	1
DLBCL	2	4	5	6	3	1
Brain	5	1.5	4	1.5	6	3
Breast	3	4	5	2	6	1
Prostate	2	4.5	4.5	4.5	4.5	1
Mean	2.8	3.6	4.5	3.6	5.1	1.4

Table 24
The ranking of the six methods under the KNN classifier.

Datasets	ODP	DMRA	FPRA	FRFS	IFPR	FSNTDJE
Colon	4	3	2	5	6	1
Leukemia	6	3	5	4	2	1
Breast	4	1	6	3	5	2
MLL	6	4	3	5	2	1
Mean	5	2.75	4	4.25	3.75	1.25

Table 25
The ranking of the six methods under the CART classifier.

Datasets	ODP	DMRA	FPRA	FRFS	IFPR	FSNTDJE
Colon	6	3	2	5	4	1
Leukemia	4	5.5	5.5	3	2	1
Breast	6	3	5	4	2	1
MLL	6	4	5	3	2	1
Mean	5.5	3.875	4.375	3.75	2.25	1.25

Table 26
The ranking of the five methods under the Naive Bayes classifier.

Datasets	ODP	DBAGEL	DGAFS	DGAFS-MI	FSNTDJE
Colon	2	3	5	4	1
DLBCL	2	4	5	3	1
Breast	3	2	5	4	1
Lung	4	2	5	3	1
Prostate	4	1	5	3	2
Mean	3	2.4	5	3.4	1.2

Table 27
The ranking of the five methods under the KNN classifier.

Datasets	ODP	DBAGEL	DGAFS	DGAFS-MI	FSNTDJE
Colon	5	2	4	3	1
DLBCL	4	2	5	3	1
Breast	5	1	4	3	2
Lung	2	3	4	5	1
Prostate	2	3	4	5	1
Mean	3.6	2.2	4.2	3.8	1.2

(ODP, DBAGEL, DGAFS, DGAFS-MI and FSNTDJE) under the Naive Bayes and KNN classifiers, respectively. One has that $\chi_F^2 = 15.52$ and $F_F = 13.86$ under Naive Bayes, and $\chi_F^2 = 12.64$ and $F_F = 6.87$ under KNN. Similarly, $F(4, 16)$ is 2.33 and $CD = 2.24$. Therefore, Tables 26 and 27 illustrate that the FSNTDJE method excels the other four methods in the Friedman statistic test.

6. Conclusion

Feature selection is one of the important parts of classification learning and can improve the classification performance and decrease the cost of classification in most cases. Uncertainty measures for calculating the distinguishing ability of feature subsets play an important role in the process of feature selection. The neighborhood rough set model can effectively solve the reduction problem of mixed and incomplete-valued information systems. In this paper, a novel feature selection method based on Lebesgue and entropy measures in neighborhood rough sets is proposed to improve the classification performance of mixed and incomplete datasets. Based on Lebesgue and entropy measures, some neighborhood tolerance entropy-based uncertainty measures are investigated in incomplete neighborhood decision systems. Then, the neighborhood tolerance dependency joint entropy is presented for dealing with the uncertainty, noise and incompleteness of incomplete neighborhood decision systems, which integrates an algebraic view with an information view in neighborhood rough sets. Moreover, some of their properties and relationships are established. Thus, a heuristic feature selection algorithm is designed to optimize the classification efficiency of selected features in incomplete neighborhood decision systems with mixed data. The experimental results demonstrate that the proposed method can select a small, effective feature subset with great classification ability in mixed and incomplete datasets. However, our proposed feature selection model cannot optimally balance the size of the selected feature subset and the classification accuracy in all large-scale and high-dimensional datasets. In our future work, to improve the classification performance and computational efficiency of the proposed algorithm for classification tasks with large-scale mixed and incomplete datasets, more efficient search strategies and uncertainty measures based on neighborhood rough sets should be explored.

Acknowledgments

classification results of Tables 9 and 10, the ranking results of the six feature selection methods under the Naive Bayes and C4.5 classifiers are shown in Tables 22 and 23.

It can be easily ascertained from Tables 9 and 10 that $s = 6$ and $T = 5$, and then under the Naive Bayes classifier, the values of χ_F^2 and F_F are 14.69 and 5.69, respectively. Similarly, one has that $\chi_F^2 = 12.11$ and $F_F = 3.76$ under the C4.5 classifier. When the significance level $\alpha = 0.1$, the critical value of $F(5, 50)$ is 2.16. Thus, two Bonferroni–Dunn tests can be performed. The critical value $q_{0.1} = 2.24$ and $CD = 2.65$ are calculated. From Tables 22 and 23, under the Naive Bayes and the C4.5 classifiers, the FSNTDJE algorithm outperforms the other five algorithms in terms of the Bonferroni–Dunn test in general.

Second, according to Tables 11 and 12, another validation is conducted for the Colon, Leukemia, Breast and MLL datasets under the KNN and CART classifiers, and the FSNTDJE algorithm is compared with the ODP, DMRA, FPRA, FRFS and IFPR methods. When $s = 6$ and $T = 4$, the ranking results of the six feature selection methods under the KNN and CART classifiers are shown in Tables 24 and 25, respectively. It follows that $\chi_F^2 = 10$ and $F_F = 3$ under the KNN classifier, and $\chi_F^2 = 13.25$ and $F_F = 5.89$ under the C4.5 classifier. Therefore, $CD = 1.48$ can be obtained. In summary, it can be easily proven that FSNTDJE is statistically superior to the other five algorithms under the two different classifiers.

Third, this experiment further continues the statistical analysis on the five gene expression datasets (Colon, DLBCL, Breast, Lung and Prostate) in Tables 13 and 14. Since $s = 5$ and $T = 5$, Tables 26 and 27 show the ranking values of the five methods

References

[1] C. Gao, Z.H. Lai, J. Zhou, C.R. Zhao, D.Q. Miao, Maximum decision entropy-based attribute reduction in decision-theoretic rough set model, Knowl.-Based Syst. 143 (2018) 179–191.

[2] X.J. Xie, X.L. Qin, A novel incremental attribute reduction approach for dynamic incomplete decision systems, Internat. J. Approx. Reason. 93 (2017) 443–462.

[3] L. Sun, X.Y. Zhang, Y.H. Qian, J.C. Xu, S.G. Zhang, Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification, Inform. Sci. 502 (2019) 18–41.

[4] H.B. Dong, T. Li, R. Ding, J. Sun, A novel hybrid genetic algorithm with granular information for feature selection and optimization, Appl. Soft Comput. 65 (2018) 33–46.

[5] L. Sun, X.C. Meng, J.C. Xu, S.G. Zhang, An image segmentation method based on improved regularized level set model, Appl. Sci. 8 (12) (2018) 2393.

[6] L. Sun, L.Y. Wang, J.C. Xu, S.G. Zhang, A neighborhood rough sets-based attribute reduction method using Lebesgue and entropy measures, Entropy 21 (2) (2019) 138.

[7] L. Sun, X.Y. Zhang, J.C. Xu, W. Wang, R.N. Liu, A gene selection approach based on the fisher linear discriminant and the neighborhood rough set, Bioengineered 9 (1) (2018) 144–151.

[8] A.K. Das, S. Sengupta, S. Bhattacharyya, A group incremental feature selection for classification using rough set theory based genetic algorithm, Appl. Soft Comput. 65 (2018) 400–411.

[9] L. Sun, S.S. Chen, J.C. Xu, Y. Tian, Improved Monarch Butterfly Optimization algorithm based on opposition-based learning and random local perturbation, Complexity 2019 (2019) 4182148.

[10] L. Sun, X.L. Kong, J.C. Xu, Z.A. Xue, R.B. Zhai, S.G. Zhang, A hybrid gene selection method based on ReliefF and Ant Colony Optimization algorithm for tumor classification, Sci. Rep. 9 (2019) 8978.

[11] V.R. Yenny, Maximal similarity granular rough sets for mixed and incomplete information systems, Soft Comput. 23 (2019) 4617–4631.

[12] Q.H. Hu, L. Zhang, D. Zhang, W. Pan, S. An, W. Pedrycz, Measuring relevance between discrete and continuous features based on neighborhood mutual information, Expert Syst. Appl. 38 (2011) 10737–10750.

[13] C.Z. Wang, Y.P. Shi, X.D. Fan, M.W. Shao, Attribute reduction based on k-nearest neighborhood rough sets, Internat. J. Approx. Reason. 106 (2019) 18–31.

[14] Q.H. Hu, D. Yu, J.F. Liu, C.X. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inform. Sci. 178 (2018) 3577–3594.

[15] Y.M. Chen, Z.Q. Zeng, J.W. Lu, Neighborhood rough set reduction with fish swarm algorithm, Soft Comput. 21 (2016) 6907–6918.

[16] S.Y. Jing, K. She, S. Ali, A universal neighborhood rough sets model for knowledge discovering from incomplete heterogeneous data, Expert Syst. 30 (1) (2013) 89–96.

[17] H. Zhao, K.Y. Qin, Mixed feature selection in incomplete decision table, Knowl.-Based Syst. 57 (2014) 181–190.

[18] Z.Q. Meng, Z.Z. Shi, A fast approach to attribute reduction in incomplete decision systems with tolerance relation-based rough sets, Inform. Sci. 179 (16) (2009) 2774–2793.

[19] C.Z. Wang, Y. Huang, M.W. Shao, D.G. Chen, Uncertainty measures for general fuzzy relations, Fuzzy Sets and Systems 360 (2019) 82–96.

[20] X.D. Fan, W.D. Zhao, C.Z. Wang, Y. Huang, Attribute reduction based on max-decision neighborhood rough set model, Knowl.-Based Syst. 151 (2018) 16–23.

[21] W.W. Li, Z.Q. Huang, X.Y. Jia, X.Y. Cai, Neighborhood based decision-theoretic rough set models, Internat. J. Approx. Reason. 69 (2016) 1–17.

[22] L. Sun, W. Wang, J.C. Xu, S.G. Zhang, Improved LLE and neighborhood rough sets-based gene selection using Lebesgue measure for cancer classification on gene expression data, J. Intell. Fuzzy Systems (2019) http://dx.doi.org/10.3233/JIFS-181904.

[23] Y.M. Chen, Z.J. Zhang, J.Z. Zheng, Y. Ma, Y. Xue, Gene selection for tumor classification using neighborhood rough sets and entropy measures, J. Biomed. Inform. 67 (2017) 59–68.

[24] Y. Liu, H. Xie, Y.H. Chen, K.Z. Tan, X. Wu, Neighborhood mutual information and its application on hyperspectral band selection for classification, Chemometr. Intell. Lab. Syst. 157 (2016) 140–151.

[25] C.Z. Wang, Q.H. Hu, X.Z. Wang, D.G. Chen, Y.H. Qian, Z. Dong, Feature selection based on neighborhood discrimination index, IEEE Trans. Neural Netw. Learn. Syst. 29 (2017) 2986–2999.

[26] L. Sun, J.C. Xu, Y. Tian, Feature selection using rough entropy-based uncertainty measures in incomplete decision systems, Knowl.-Based Syst. 36 (2012) 206–216.

[27] G.Y. Wang, Rough reduction in algebra view and information view, Int. J. Intell. Syst. 18 (2003) 679–688.

[28] S.H. Teng, M. Lu, A.F. Yang, J. Zhang, Y.J. Nian, M. He, Efficient attribute reduction from the viewpoint of discernibility, Inform. Sci. 326 (2016) 297–314.

[29] B. Qin, F.P. Zeng, K.S. Yan, Knowledge structures in a tolerance knowledge base and their uncertainty measures, Knowl.-Based Syst. 151 (2018) 198–215.

[30] L. Sun, J.C. Xu, Information entropy and mutual information-based uncertainty measures in rough set theory, Appl. Math. Inf. Sci. 8 (4) (2014) 1973–1985.

[31] Y.M. Chen, K.S. Wu, X.H. Chen, C.H. Tang, Q.X. Zhu, An entropy-based uncertainty measurement approach in neighborhood systems, Inform. Sci. 279 (2014) 239–250.

[32] H. Ge, C.J. Yang, L.S. Li, Positive region reduct based on relative discernibility and acceleration strategy, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 26 (4) (2018) 521–551.

[33] P.R. Halmos, Measure Theory, World Publishing Corporation, 2007.

[34] X. Xu, Z.Z. Lu, X.P. Luo, A kernel estimate method for characteristic function-based uncertainty importance measure, Appl. Math. Model. 42 (2016) 58–70.

[35] J.H. Dai, Rough set approach to incomplete numerical data, Inform. Sci. 241 (2013) 43–57.

[36] W.B. Qian, W.H. Shu, Mutual information criterion for feature selection from incomplete data, Neurocomputing 168 (30) (2015) 210–220.

[37] L. Sun, X.Y. Zhang, J.C. Xu, S.G. Zhang, An attribute reduction method using neighborhood entropy measures in neighborhood rough sets, Entropy 21 (2) (2019) 155.

[38] L. Sun, X.Y. Zhang, Y.H. Qian, J.C. Xu, S.G. Zhang, Y. Tian, Joint neighborhood entropy-based gene selection method with fisher score for tumor classification, Appl. Intell. 49 (4) (2019) 1245–1259.

[39] Y.H. Qian, J.Y. Liang, W. Pedrycz, C.Y. Dang, An efficient accelerator for attribute reduction from incomplete data in rough set framework, Pattern Recognit. 44 (8) (2011) 1658–1670.

[40] J.H. Dai, W.T. Wang, H.W. Tian, L. Liu, Attribute selection based on a new conditional entropy for incomplete decision systems, Knowl.-Based Syst. 39 (2013) 207–213.

[41] J. Biesiada, W. Duch, Feature selection for high-dimensional data: A kolmogorov-smirnov correlation-based filter, Soft Comput. 30 (2005) 95–103.

[42] Z. Zhao, H. Liu, Searching for interacting features in subset selection, Intell. Data Anal. 13 (2) (2009) 207–2281.

[43] C. Lee, G.G. Lee, Information gain and divergence based feature selection for machine learning based text categorization, Inf. Process. Manage. 42 (1) (2006) 155–165.

[44] A. Paul, J. Sil, C.D. Mukhopadhyay, Gene selection for designing optimal fuzzy rule base classifier by estimating missing value, Appl. Soft Comput. 55 (2017) 276–288.

[45] Q. He, C.X. Wu, D.G. Chen, S.Y. Zhao, Fuzzy rough set based attribute reduction for information systems with fuzzy decisions, Knowl.-Based Syst. 24 (2011) 689–696.

[46] W.B. Qian, W.H. Shu, Attribute reduction in incomplete ordered information systems with fuzzy decision, Appl. Soft Comput. 73 (2018) 242–253.

[47] M.W. Shao, W.X. Zhang, Dominance relation and rules in an incomplete ordered information system, Int. J. Intell. Syst. 20 (2005) 13–27.

[48] L. Sun, R.N. Liu, J.C. Xu, S.G. Zhang, Y. Tian, An affinity propagation clustering method using hybrid kernel function with LLE, IEEE Access 6 (2018) 68892–68909.

[49] L. Sun, R.N. Liu, J.C. Xu, S.G. Zhang, An adaptive density peaks clustering method with Fisher linear discriminant, IEEE Access 7 (2019) 72936–72955.

[50] N. Gopika, A.E.A. Meena Kowshalaya, Correlation based feature selection algorithm for machine learning, in: Proceedings of the 3rd International Conference on Communication and Electronics Systems, 2018, pp. 692–695.

[51] D.G. Chen, L. Zhang, S.Y. Zhao, Q.H. Hu, P.F. Zhu, A novel algorithm for finding reducts with fuzzy rough sets, IEEE Trans. Fuzzy Syst. 20 (2) (2012) 385–389.

[52] Y.H. Qian, Q. Wang, H.H. Cheng, J.Y. Liang, C.Y. Dang, Fuzzy-rough feature selection accelerator, Fuzzy Sets and Systems 258 (1) (2015) 61–78.

[53] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Trans. Fuzzy Syst. 17 (4) (2009) 824–838.

[54] A.H. Tan, W.Z. Wu, Y.H. Qian, J.Y. Liang, J.K. Chen, J.J. Li, Intuitionistic fuzzy rough set-based granular structures and attribute subset selection, IEEE Trans. Fuzzy Syst. 27 (3) (2019) 527–539.

[55] R.D. Priya, S. Kuppuswami, A genetic algorithm-based approach for imputing missing discrete values in databases, Trans. Inf. Sci. Appl. 9 (2012) 169–178.

[56] R.D. Priya, R. Sivaraj, Dynamic genetic algorithm-based feature selection and incomplete value imputation for microarray classification, Current Sci. 112 (1) (2017) 126–131.

[57] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1) (1940) 86–92.

[58] O.J. Dunn, Multiple comparisons among means, J. Amer. Statist. Assoc. 56 (1961) 52–64.

[59] C.Z. Wang, Y.L. Qi, M.W. Shao, Q.H. Hu, D.G. Chen, Y.H. Qian, Y.J. Lin, A fitting model for feature selection with fuzzy rough sets, IEEE Trans. Fuzzy Syst. 25 (4) (2017) 741–753.

[60] L. Sun, J.C. Xu, S.W. Liu, S.G. Zhang, Y. Li, C.A. Shen, A robust image watermarking scheme using arnold transform and BP neural network, Neural Comput. Appl. 30 (8) (2018) 2425–2440.

[61] J. Demsar, D. Schuurmans, Statistical comparison of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.