Information Sciences 502 (2019) 18–41

Contents lists available at ScienceDirect

Information Sciences

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

Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification

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

[a] *College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China*
[b] *Postdoctoral Mobile Station of Biology, College of Life Science, Henan Normal University, Xinxiang 453007, China*
[c] *Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China*

ARTICLE INFO

Article history:
Received 31 December 2018
Revised 20 May 2019
Accepted 26 May 2019
Available online 11 June 2019

Keywords:
Granular computing
Neighborhood rough sets
Feature selection
Neighborhood entropy
Uncertainty measure
Cancer classification

ABSTRACT

Gene expression data classification is an important technology for cancer diagnosis in bioinformatics and has been widely researched. Due to the large number of genes and the small sample size in gene expression data, feature selection based on neighborhood rough sets is a key step for improving the performance of gene expression data classification. However, some quantitative measures of feature sets may be nonmonotonic in neighborhood rough sets, and many feature selection methods based on evaluation functions yield high cardinality and low predictive accuracy. Therefore, investigating effective and efficient heuristic reduction algorithms is necessary. In this paper, a novel feature selection method based on neighborhood rough sets using neighborhood entropy-based uncertainty measures for cancer classification from gene expression data is proposed. First, some neighborhood entropy-based uncertainty measures are investigated for handling the uncertainty and noise of neighborhood decision systems. Then, to fully reflect the decision-making ability of attributes, the neighborhood credibility and neighborhood coverage degrees are defined and introduced into decision neighborhood entropy and mutual information, which are proven to be nonmonotonic. Moreover, some of the properties and relationships among these measures are derived, which is helpful for understanding the essence of the knowledge content and the uncertainty of neighborhood decision systems. Finally, the Fisher score method is employed to preliminarily eliminate irrelevant genes to significantly reduce complexity, and a heuristic feature selection algorithm with low computational complexity is presented to improve the performance of cancer classification using gene expression data. Experiments on ten gene expression datasets show that our proposed algorithm is indeed efficient and outperforms other related methods in terms of the number of selected genes and the classification accuracy, especially as the size of the genes increases.

1. Introduction

With the rapid development of DNA sequencing technology, researchers can obtain a large amount of gene expression data from various tissue samples, thus providing technical support for the study of tumor pathogenesis at the molecular

[*] Corresponding authors.
E-mail addresses: jinchengqyh@126.com (Y. Qian), jiuchxu@gmail.com (J. Xu).

https://doi.org/10.1016/j.ins.2019.05.072
0020-0255/© 2019 Published by Elsevier Inc.

level [13]. Medical data mining is one of the main research directions of data mining technology and represents a key technology for cancer classification and a bioinformatics research hot spot [35]. When using gene expression data mining technology to find disease genes, protein functions and disease diagnoses are of great significance; therefore, gene selection is the research focus of tumor recognition and classification [21]. Due to the high costs of experiments, the sample sizes of gene expression datasets remain in the hundreds, which is low compared to the tens of thousands of genes involved [39]. Although gene expression datasets are high-dimensional, only a few of the dimensions are beneficial for classification [13]. The resulting dimensionality poses a considerable challenge for classification [31]. Thus, the few beneficial genes must be selected from huge amounts of gene expression data.

Feature selection, as a data mining preprocessing technique, is a dimensionality reduction method that attempts to reserve informative attributes in high-dimensional data, and attribute reduction in rough sets has been recognized as an important feature selection method [4,26]. Feature selection has three main approaches: filter, wrapper and embedded methods [15]. Filter methods are typically employed as preprocessing methods that are independent of the classifier and use feature-ranking techniques as the basis for feature selection. Wrapper methods evaluate the goodness of each feature subset identified by estimating the accuracy percentage of the specific classifier used [15]. However, the wrapper methods not only exhibit sensitivity to the classifier but also tend to present considerable runtimes. Hence, these methods are not extensively used in microarray tasks, and few works in the field have employed them. Compared with wrapper methods, embedded methods integrate feature selection in the training process to reduce the total time required for reclassifying subsets [7]. In this paper, our feature selection method is based on the filter approach, in which a heuristic search algorithm is used to find an optimal feature subset with neighborhood rough sets for gene expression datasets.

Granular computing is an effective technology for uncertainty analyses, and attribute reduction is a fundamental research topic and an important application of granular computing [8,33,37,41]. Traditional rough set-based attribute reduction methods are established based on an equivalence relation, and they are only compatible for categorical datasets and not for continuous numerical datasets [30,43]. To overcome this drawback, Hu et al. [17] established a neighborhood rough set model to process both numerical and categorical datasets via neighborhood relation. Over the last few years, many reduction methods based on neighborhood relation have been investigated [5,10,31]. For instance, Chen et al. [5] studied a gene selection algorithm using neighborhood rough sets and a joint entropy measure. Fan et al. [10] introduced a max-decision neighborhood rough set model to design an attribute reduction algorithm. Sun et al. [31] described a gene selection approach based on Fisher linear discriminant and neighborhood rough sets. Most of the abovementioned feature selection algorithms that use neighborhood rough set models are based on the monotonicity of evaluation functions for heuristic searches [17]. However, there are some issues associated with feature selection based on the monotonicity of the evaluation functions. For example, when the classification performance of the original dataset is poor, the corresponding evaluation functions have low measured values. Therefore, these methods cannot yield good reduction results. To remedy this defect, Li et al. [20] presented a nonmonotonic attribute reduction algorithm for the decision-theoretic rough set model. The ideas of nonmonotonic reduction in [20] inspired us to investigate a new feature selection method based on neighborhood rough sets in this paper. It is known that a gene expression dataset can be granulated by using neighborhood parameters. Thus, some neighborhood entropy measures based on neighborhood rough sets can be further studied and the monotonicity or nonmonotonicity of the neighborhood entropy-based uncertainty measures can be proved. Therefore, a nonmonotonic feature selection algorithm is presented to address the abovementioned problems.

Note that the reduction calculation of neighborhood decision systems is a key problem in neighborhood rough sets. In addition, the reduct sets of an information system needs to be achieved to further extract rule-like knowledge from an information system [42]. In practical decision-making applications, the certainty factor and the object coverage factor of rules are two important standards for evaluating the decision-making ability of decision systems [43]. However, some of these existing reduction methods cannot objectively reflect the change of the decision-making ability of classification. The credibility and coverage degrees are known efficiently reflect the classification ability of conditional attributes with respect to the decision attribute [35]. Therefore, the conditional attributes with higher credibility and coverage degrees are important with respect to the decision attribute. Until now, the literature has not considered neighborhood rough sets, which inspires our investigation of new measures to fully reflect the classification performance and decision-making ability of neighborhood decision systems. Consequently, new uncertainty measures and an effective heuristic search algorithm must be investigated. Moreover, the concepts of coverage degree and credibility degree should be introduced into neighborhood decision systems as measures to reflect the classification ability of conditional attributes with respect to the decision attribute, and then the credibility degree and the coverage degree based on neighborhood relation should be integrated into neighborhood entropy measures to demonstrate the decision-making ability of attributes in neighborhood decision systems.

Available pretreatment methods for dimensionality reduction include the principal component analysis, which is the most common linear dimension-reduction method. However, in many real-world datasets, the low-dimensional structure hidden in high-dimensional data is nonlinear, and this reduction is not effective for mapping such high-dimensional data. Locally linear embedding (LLE) approximates the input data with a low-dimensional surface and reduces its dimensionality by learning a mapping of the surface [45]. Unfortunately, LLE has a drawback in that its computation cost is high [24]. The Fisher score, which is a common attribute relevance criterion, is a supervised learning technique with many advantages, such as few calculations, high accuracy, and strong operability, and it can efficiently reduce computational complexity [47]. However, the Fisher score method occasionally selects redundant attributes, which affects the classification result [14]. This phenomenon inspires us to combine the Fisher score with neighborhood rough sets to reduce the initial dimensions and improve the classification performance of high-dimensional gene expression datasets. Then, the appropriate genes are selected to form a candidate gene subset, and some neighborhood entropy-based uncertainty measures are studied to address the uncertainty and noise of gene expression datasets. To fully reflect the decision-making ability of attributes, a neighborhood credibility degree and a neighborhood coverage degree are introduced into decision neighborhood entropy and mutual information with nonmonotonicity. Thus, a heuristic nonmonotonic feature selection algorithm with Fisher score in neighborhood decision systems is designed to improve the classification performance of gene expression datasets. The experimental results for several gene expression datasets show that our proposed method can find optimal reduct sets with few genes and high classification accuracy.

The remainder of this paper is organized as follows. Section 2 reviews some basic concepts. Section 3 investigates some neighborhood entropy-based uncertainty measures and develops a nonmonotonic feature selection approach with Fisher score for gene expression data classification. Section 4 shows and analyzes the experimental results. Finally, Section 5 summarizes this study.

2. Previous knowledge

In this section, we briefly review several basic concepts of decision systems, information entropy measures and neighborhood rough sets described in previous studies [17,22,25].

2.1. Information entropy measure

Consider a decision system $DS =< U, C, D, V, f >$, which is usually written more simply as $DS =< U, C, D >$, where $U = \{x_1, x_2, \cdots, x_n\}$ is a sample set named universe, $C = \{a_1, a_2, \cdots, a_m\}$ is a conditional attribute set that describes the samples, D is a classification attribute set, f: $U \times \{C \cup D\} \to V$ is an information function that associates a unique value of each attribute with every object belonging to U, and $f(x, a)$ represents the value of sample $x \in U$ on any attribute $a \in C$. Then, for any attribute subset $B \subseteq C$, an equivalence relation from B is described as

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

Given a decision system $DS =< U, C, D >$ with $B \subseteq C$, and $U/B = \{X_1, X_2, \cdots, X_{n'}\}$, the information entropy of B is expressed as

$$H(B) = -\sum_{i=1}^{n'} p(X_i) \log(p(X_i)), \tag{2}$$

where $p(X_i) = \frac{|X_i|}{|U|}$ is the probability of $X_i \subseteq U/B$, and $|X_i|$ denotes the cardinality of the equivalence class (block) X_i.

Given a decision system $DS =< U, C, D >$ with $B \subseteq C$, $U/B = \{X_1, X_2, \cdots, X_{n'}\}$, and $U/D = \{Y_1, Y_2, \cdots, Y_{m'}\}$, the joint entropy of B and D is characterized as

$$H(D \cup B) = -\sum_{i=1}^{n'} \sum_{j=1}^{m'} p(X_i \cap Y_j) \log(p(X_i \cap Y_j)), \tag{3}$$

where $p(X_i \cap Y_j) = \frac{|X_i \cap Y_j|}{|U|}$, $i = 1, 2, \cdots, n'$, and $j = 1, 2, \cdots, m'$.

Given a decision system $DS =< U, C, D >$ with $B \subseteq C$, $U/B = \{X_1, X_2, \cdots, X_{n'}\}$, and $U/D = \{Y_1, Y_2, \cdots, Y_{m'}\}$, the conditional information entropy of D with respect to B is defined as

$$H(D|B) = -\sum_{i=1}^{n'} p(X_i) \sum_{j=1}^{m'} p(Y_j|X_i) \log(p(Y_j|X_i)), \tag{4}$$

where $p(Y_j|X_i) = \frac{|Y_j \cap X_i|}{|X_i|}$, $i = 1, 2, \cdots, n'$, and $j = 1, 2, \cdots, m'$.

Given a decision system $DS =< U, C, D >$ with $B \subseteq C$, there exists $H(D|B) = H(D \cup B) - H(B)$.

Given a decision system $DS =< U, C, D >$ with $B \subseteq C$, the mutual information between B and D is defined as

$$I(B; D) = H(D) - H(D|B). \tag{5}$$

Given a decision system $DS =< U, C, D >$ with $B \subseteq C$, if B is a reduct of C with respect to D, then for any $a \in B$, there are $I(B; D) = I(C; D)$ and $I(B - \{a\}; D) < I(B; D)$.

2.2. Neighborhood rough sets

Consider a neighborhood decision system $NS =< U, C, D, \delta >$, where $U = \{x_1, x_2, \cdots, x_n\}$ is a sample set named universe, $C = \{a_1, a_2, \cdots, a_m\}$ is a conditional attribute set that describes the samples, $D = \{d\}$ is a decision attribute set that contains only one attribute, and δ is a neighborhood parameter with $0 \le \delta \le 1$. Then, for any conditional attribute subset $B \subseteq C$ and any neighborhood parameter δ, the similarity relation from B is defined as

$$NR_\delta(B) = \{(x, y) \in U \times U | \Delta_B(x, y) \le \delta\}, \tag{6}$$

where $\Delta_B(x, y)$ is a distance metric function used to determine the shape of the neighborhood, and δ is a threshold used to control the size of the neighborhood.

For any sample $x \in U$ and any subset of attributes $B \subseteq C$, the δ neighborhood of x induced by B is defined as

$$n_B^\delta(x) = \{y | x, y \in U, \Delta_B(x, y) \le \delta\}. \tag{7}$$

Currently, there are three types of widely employed classical distance metrics, including Manhattan distance, Euclidean distance and Chebyshev distance. Since the Euclidean distance function effectively reflects the basic information of the unknown data [5], it is used in this paper and expressed as

$$\Delta_B(x, y) = \sqrt{\sum_{k=1}^{|B|} |f(x, a_k) - f(y, a_k)|^2}. \tag{8}$$

Given a neighborhood decision system $NS =< U, C, D, \delta >$ with $B \subseteq C$ and $X \subseteq U$, the neighborhood lower approximation set $\underline{B}_\delta(X)$ and the neighborhood upper approximation set $\overline{B}_\delta(X)$ of X with respect to B are described, respectively, as

$$\underline{B}_\delta(X) = \{x_i | n_B^\delta(x_i) \subseteq X, x_i \in U, i = 1, 2, \cdots, |U|\}, \tag{9}$$

$$\overline{B}_\delta(X) = \{x_i | n_B^\delta(x_i) \cap X \ne \emptyset, x_i \in U, i = 1, 2, \cdots, |U|\}. \tag{10}$$

3. Neighborhood entropy-based uncertainty measures

In recent years, some correlative concepts of neighborhood entropy are defined to measure the uncertainty of numerical data [16]. Considering information entropy and its variants, several feature selection algorithms with monotonicity are proposed to address the analysis of real-valued data [5,16,31]. However, when the classification performance of the original dataset is poor, the corresponding evaluation functions have lower measured values; thus, monotonic attribute reduction methods cannot obtain great reduction results [19]. To address this issue, some concepts of neighborhood entropy-based uncertainty measures are proposed to investigate the uncertainty of knowledge in neighborhood decision systems. In addition, important properties and the relationships of these measures are deduced.

3.1. Neighborhood entropy in neighborhood decision systems

Given a neighborhood decision system $NS =< U, C, D, \delta >$ with $B \subseteq C$, and $n_B^\delta(x_i)$ as a neighborhood class of $x_i \in U$, Hu et al. [16] defined the neighborhood entropy of x_i as

$$H_\delta^{x_i}(B) = -\log\left(\frac{|n_B^\delta(x_i)|}{|U|}\right). \tag{11}$$

Then, the average neighborhood entropy of the sample set is computed by

$$H_\delta(B) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n_B^\delta(x_i)|}{|U|}\right). \tag{12}$$

Given a neighborhood decision system $NS =< U, C, D, \delta >$ with $B \subseteq C$ and $[x_i]_D \in U/D$, Hu et al. [16] described the joint entropy of B and D as

$$H_\delta(D \cup B) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n_B^\delta(x_i) \cap [x_i]_D|}{|U|}\right). \tag{13}$$

Property 1. *Suppose that a neighborhood decision system is represented by $NS = \ < U, C, D, \delta >$ and $n_C^\delta(x_i) \subseteq U$ for any $x_i \in U$. Then, $0 \le H_\delta(C) \le \log |U|$.*

Proof. Since there exists $n_C^\delta(x_i) \subseteq U$ for any $x_i \in U$ according to Eq. (7), it follows that $\frac{1}{|U|} \le \frac{|n_C^\delta(x_i)|}{|U|} \le 1$. Thus, $0 \le H_\delta(C) \le \log |U|$ holds. □

Proposition 1. *Suppose that a neighborhood decision system is represented by $NS =< U, C, D, \delta >$ exists. For any $x_i \in U$, if $B_1 \subseteq B_2 \subseteq C$, then $H_\delta(B_2) \ge H_\delta(B_1)$.*

Proof. Let $B_1 \subseteq B_2 \subseteq C$; it follows from Proposition 1 in [5] that $n_{B_1}^\delta(x_i) \supseteq n_{B_2}^\delta(x_i)$. Obviously, $|n_{B_1}^\delta(x_i)| \ge |n_{B_2}^\delta(x_i)|$. Hence, according to Eq. (12), $H_\delta(B_2) \ge H_\delta(B_1)$ holds.

The joint entropy introduced by Hu et al. [16] was used to solve the inability of traditional rough set-based feature selection methods to directly process continuous datasets. As shown in from Eq. (13), $\frac{|n_B^\delta(x_i) \cap [x_i]_D|}{|U|}$ represents the necessity degree of a decision rule in a decision system [43]. However, the joint entropy fails to fully reflect its decision-making ability of a decision system. □

3.2. Decision neighborhood entropy and mutual information in neighborhood decision systems

Notably, the credibility and coverage degrees can reflect the decision-making ability and the classification ability of conditional attributes with respect to the decision attribute [40]. In general, the credibility degree indicates the adequacy of the proposition, and the coverage degree describes the necessity of the proposition [40,43]. Thus, to fully reflect the decision-making ability and the classification ability of a neighborhood decision system, this paper investigates some neighborhood entropy-based uncertainty measures by combining the credibility degree with the coverage degree.

Given a decision system $DS = < U, C \cup D >$ with $B \subseteq C$, $U/B = \{X_1, X_2, \cdots, X_{n'}\}$, and $U/D = \{Y_1, Y_2, \cdots, Y_{m'}\}$, the credibility degree $\alpha_{ij} = \frac{|X_i \cap Y_j|}{|X_i|}$ and the coverage degree $\kappa_{ij} = \frac{|X_i \cap Y_j|}{|Y_j|}$ based on the partition are described respectively in [40,43], where $i = 1, 2, \cdots, n'$ and $j = 1, 2, \cdots, m'$.

Definition 1. Suppose that a neighborhood decision system $NS = < U, C, D, \delta >$ with $B \subseteq C$ exists. $n_B^\delta(x_i)$ is a neighborhood class of $x_i \in U$ generated by the neighborhood relation $NR_\delta(B)$, and $[x_i]_D$ is an equivalence class of $x_i \in U$ generated by the equivalence relation $R(D)$. A neighborhood credibility degree α_i and a neighborhood coverage degree κ_i for $x_i \in U$ are defined, respectively, as

$$\alpha_i = \frac{|n_B^\delta(x_i) \cap [x_i]_D|}{|n_B^\delta(x_i)|}, \tag{14}$$

$$\kappa_i = \frac{|n_B^\delta(x_i) \cap [x_i]_D|}{|[x_i]_D|}, \tag{15}$$

where α_i describes a classification accuracy of B for a classification of D, and κ_i denotes the true positive rate of B with respect to D.

Definition 2. Suppose that a neighborhood decision system $NS = < U, C, D, \delta >$ with $B \subseteq C$ exists. $n_B^\delta(x_i)$ is a neighborhood class of $x_i \in U$ generated by $NR_\delta(B)$, and $[x_i]_D$ is an equivalence class of $x_i \in U$ generated by $R(D)$. A decision neighborhood entropy of D with respect to B is defined as

$$H_\delta(D|B) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log\left(\frac{|n_B^\delta(x_i) \cap [x_i]_D|^2}{|[x_i]_D||n_B^\delta(x_i)|}\right). \tag{16}$$

Proposition 2. *Suppose that a neighborhood decision system $NS = < U, C, D, \delta >$ with $B \subseteq C$ exists. α_i is the neighborhood credibility degree for $x_i \in U$, and κ_i is the neighborhood coverage degree for $x_i \in U$; thus, $H_\delta(D|B) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log(\alpha_i \kappa_i)$.*

Proof. It follows immediately from Definitions 1 and 2 that

$$\begin{aligned} H_\delta(D|B) &= -\frac{1}{|U|}\sum_{i=1}^{|U|} \log\left(\frac{|n_B^\delta(x_i) \cap [x_i]_D|^2}{|[x_i]_D||n_B^\delta(x_i)|}\right) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|} \log\left(\frac{|n_B^\delta(x_i) \cap [x_i]_D|}{|[x_i]_D|} \cdot \frac{|n_B^\delta(x_i) \cap [x_i]_D|}{|n_B^\delta(x_i)|}\right) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|} \log(\alpha_i \kappa_i). \end{aligned}$$

Proposition 2 establishes the relationships among the decision neighborhood entropy, the neighborhood credibility degree and the neighborhood coverage degree in a neighborhood decision system. These relationships help characterize that the decision neighborhood entropy can reflect the decision-making ability of attributes in the neighborhood decision system. Then, in a neighborhood decision system $NS = < U, C, D, \delta >$ with any subset $B \subseteq C$, the decision neighborhood entropy $H_\delta(D|B)$ can measure the uncertainty and reflect the decision-making ability of B with respect to D. In what follows, to discuss the monotonicity or the nonmonotonic of entropy measures in neighborhood rough sets, some relevant definitions and propositions of neighborhood entropy measures in neighborhood decision systems need to be further investigated. □

Definition 3. Suppose that a neighborhood decision system $NS = < U, C, D, \delta >$ with $B \subseteq C$ exists. $n_B^\delta(x_i)$ is a neighborhood class of $x_i \in U$, and $[x_i]_D$ is an equivalence class of $x_i \in U$. Then, a new neighborhood joint entropy of B and D is defined as

$$H_\delta(D, B) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log\left(\frac{|n_B^\delta(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|}\right). \tag{17}$$

Proposition 3. *Suppose that a neighborhood decision system $NS = < U, C, D, \delta >$ with $B_1 \subseteq B_2 \subseteq C$ exists. Then $H_\delta(D, B_1) \le H_\delta(D, B_2)$, where the equality holds if and only if $n_{B_1}^\delta(x_i) = n_{B_2}^\delta(x_i)$ for any $x_i \in U$.*

Proof. Suppose that $B_1 \subseteq B_2 \subseteq C$, and it follows from Proposition 1 that $n^{\delta}_{B_1}(x_i) \supseteq n^{\delta}_{B_2}(x_i)$. Then, $U \supseteq n^{\delta}_{B_1}(x_i) \cap [x_i]_D \supseteq n^{\delta}_{B_2}(x_i) \cap [x_i]_D \supseteq \{x_i\}$. It is easily obtained that $|U| \geq |n^{\delta}_{B_1}(x_i) \cap [x_i]_D| \geq |n^{\delta}_{B_2}(x_i) \cap [x_i]_D| \geq |\{x_i\}| = 1$. Thus, $\frac{|U|^2}{|U||[x_i]_D|} \geq \frac{|n^{\delta}_{B_1}(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|} \geq \frac{|n^{\delta}_{B_2}(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|} \geq \frac{1}{|U||[x_i]_D|}$, and then $\log(\frac{|U|}{|[x_i]_D|}) \geq \log(\frac{|n^{\delta}_{B_1}(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|}) \geq \log(\frac{|n^{\delta}_{B_2}(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|}) \geq \log(\frac{1}{|U||[x_i]_D|})$. Thus, $-\frac{1}{|U|}\sum_{i=1}^{|U|}\log(\frac{|U|}{|[x_i]_D|}) \leq -\frac{1}{|U|}\sum_{i=1}^{|U|}\log(\frac{|n^{\delta}_{B_1}(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|}) \leq -\frac{1}{|U|}\sum_{i=1}^{|U|}\log(\frac{|n^{\delta}_{B_2}(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|}) \leq -\frac{1}{|U|}\sum_{i=1}^{|U|}\log(\frac{1}{|U||[x_i]_D|})$. It can be concluded from Definition 3 that $H_{\delta}(D, B_1) \leq H_{\delta}(D, B_2)$. When $n^{\delta}_{B_1}(x_i) = n^{\delta}_{B_2}(x_i)$, $\frac{|n^{\delta}_{B_1}(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|} = \frac{|n^{\delta}_{B_2}(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|}$. Therefore, $H_{\delta}(D, B_1) \leq H_{\delta}(D, B_2)$ holds.

Proposition 3 states that the neighborhood joint entropy of knowledge in neighborhood rough sets monotonically increases as the knowledge granularity formed by the neighborhood relation becomes smaller at a finer classification. Then, the neighborhood joint entropy has monotonicity in decision neighborhood systems. □

Proposition 4. *Suppose that a neighborhood decision system $NS = < U, C, D, \delta >$ with $B \subseteq C$ exits. Then $H_{\delta}(D, B) \geq H_{\delta}(B)$.*

Proof. It follows immediately from Definition 3 and Eq. (12) that

$$\begin{aligned} H_{\delta}(D,B) - H_{\delta}(B) &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n^{\delta}_B(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|}\right) + \frac{1}{|U|}\sum_{i=1}^{|U|}\log(\frac{|n^{\delta}_B(x_i)|}{|U|}) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n^{\delta}_B(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|} \cdot \frac{|U|}{|n^{\delta}_B(x_i)|}\right) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n^{\delta}_B(x_i) \cap [x_i]_D|}{|[x_i]_D|} \cdot \frac{|n^{\delta}_B(x_i) \cap [x_i]_D|}{|n^{\delta}_B(x_i)|}\right). \end{aligned}$$

Since there exist $n^{\delta}_B(x_i) \cap [x_i]_D \subseteq n^{\delta}_B(x_i)$ and $n^{\delta}_B(x_i) \cap [x_i]_D \subseteq [x_i]_D$, it is easily obtained that $|n^{\delta}_B(x_i) \cap [x_i]_D| \leq |n^{\delta}_B(x_i)|$ and $|n^{\delta}_B(x_i) \cap [x_i]_D| \leq |[x_i]_D|$. Then, $\frac{|n^{\delta}_B(x_i) \cap [x_i]_D|}{|n^{\delta}_B(x_i)|} \leq 1$ and $\frac{|n^{\delta}_B(x_i) \cap [x_i]_D|}{|[x_i]_D|} \leq 1$. Hence, $H_{\delta}(D, B) - H_{\delta}(B) \geq 0$ holds, i.e., $H_{\delta}(D, B) \geq H_{\delta}(B)$. □

Proposition 5. *Suppose that a neighborhood decision system $NS = < U, C, D, \delta >$ with $B \subseteq C$ exists. Then, $H_{\delta}(D|B) = H_{\delta}(D, B) - H_{\delta}(B)$.*

Proof. It follows immediately from Definitions 2 and 3 and the Eq. (12) that

$$\begin{aligned} H_{\delta}(D,B) - H_{\delta}(B) &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n^{\delta}_B(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|}\right) + \frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n^{\delta}_B(x_i)|}{|U|}\right) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n^{\delta}_B(x_i) \cap [x_i]_D|^2}{|U||[x_i]_D|} \cdot \frac{|U|}{|n^{\delta}_B(x_i)|}\right) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n^{\delta}_B(x_i) \cap [x_i]_D|}{|[x_i]_D|} \cdot \frac{|n^{\delta}_B(x_i) \cap [x_i]_D|}{|n^{\delta}_B(x_i)|}\right) \\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n^{\delta}_B(x_i) \cap [x_i]_D|^2}{|[x_i]_D||n^{\delta}_B(x_i)|}\right) \\ &= H_{\delta}(D|B). \end{aligned}$$

Thus, $H_{\delta}(D|B) = H_{\delta}(D, B) - H_{\delta}(B)$. □

Definition 4. Suppose that a neighborhood decision system $NS = < U, C, D, \delta >$ with $B \subseteq C$ exists. Then, the mutual information of B and D is defined as

$$MI_{\delta}(D; B) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|n^{\delta}_B(x_i)||[x_i]_D|^2}{|U||n^{\delta}_B(x_i) \cap [x_i]_D|^2}. \quad (18)$$

Proposition 6. *Suppose that a neighborhood decision system $NS = < U, C, D, \delta >$ with $B \subseteq C$ exists. If $n^{\delta}_B(x_i) = [x_i]_D$ for any $x_i \in U$, then $MI_{\delta}(D; B) = H_{\delta}(D)$.*

Proof. Let $n_B^\delta(x_i) = [x_i]_D$ for any $x_i \in U$. It follows from Definition 4 and Eq. (12) that $MI_\delta(D;B) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|n_B^\delta(x_i)||[x_i]_D|^2}{|U||n_B^\delta(x_i)\cap[x_i]_D|^2} = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|[x_i]_D|}{|U|} = H_\delta(D)$. Hence, when $n_B^\delta(x_i) = [x_i]_D$ for any $x_i \in U$, $MI_\delta(D;B) = H_\delta(D)$. □

Proposition 7. *Suppose that a neighborhood decision system* $NS = < U, C, D, \delta >$ *with* $B \subseteq C$ *exists. Then, the following properties hold:*

(1) $MI_\delta(D;\ B) \geq\ 0$,
(2) $MI_\delta(D;B) = H_\delta(D) + H_\delta(B) - H_\delta(D,B)$,
(3) $MI_\delta(D;B) = H_\delta(D) - H_\delta(D|B)$.

Proof. (1) This proof is straightforward.

(2) It follows immediately from Definitions 3 and 4 and the Eq. (12) that

$$\begin{aligned}
&H_\delta(D) + H_\delta(B) - H_\delta(D,B)\\
&= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|[x_i]_D|}{|U|}\right) - \frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n_B^\delta(x_i)|}{|U|}\right) + \frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n_B^\delta(x_i)\cap[x_i]_D|^2}{|U||[x_i]_D|}\right)\\
&= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|[x_i]_D|}{|U|}\cdot\frac{|n_B^\delta(x_i)|}{|U|}\cdot\frac{|U||[x_i]_D|}{|n_B^\delta(x_i)\cap[x_i]_D|^2}\right)\\
&= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|[x_i]_D|^2|n_B^\delta(x_i)|}{|U||n_B^\delta(x_i)\cap[x_i]_D|^2}\right)\\
&= MI_\delta(D;B).
\end{aligned}$$

Thus, $MI_\delta(D;B) = H_\delta(D) + H_\delta(B) - H_\delta(D,B)$ holds.

(3) It follows immediately from Definitions 2 and 4 and the Eq. (12) that

$$\begin{aligned}
H_\delta(D) - H_\delta(D|B) &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|[x_i]_D|}{|U|}\right) + \frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n_B^\delta(x_i)\cap[x_i]_D|^2}{|[x_i]_D||n_B^\delta(x_i)|}\right)\\
&= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|[x_i]_D|}{|U|}\cdot\frac{|[x_i]_D||n_B^\delta(x_i)|}{|n_B^\delta(x_i)\cap[x_i]_D|^2}\right)\\
&= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|[x_i]_D|^2|n_B^\delta(x_i)|}{|U||n_B^\delta(x_i)\cap[x_i]_D|^2}\right)\\
&= MI_\delta(D;B).
\end{aligned}$$

Therefore, $MI_\delta(D;B) = H_\delta(D) - H_\delta(D|B)$ holds.

Proposition 7 establishes the relationships among the neighborhood entropy, the decision neighborhood entropy, the neighborhood joint entropy and the mutual information. These relationships are helpful for understanding the essence of the knowledge content and the uncertainty of neighborhood decision systems. Thus far, the abovementioned measures and their relationships have not been reported in neighborhood decision systems. Therefore, the above methods for measuring uncertainty can characterize the uncertainty of knowledge in neighborhood decision systems. □

4. Nonmonotonic feature selection for gene expression data classification

4.1. Nonmonotonic feature selection in neighborhood decision systems

In neighborhood rough sets, many evaluation functions for feature subsets of feature selection methods are developed, and then, heuristic reduction algorithms based on the monotonicity of evaluation functions are established [10,17]. However, in the rough set model of decision theory, the positive region of a decision attribute does not satisfy monotonicity, i.e., as the conditional attribute increases, the positive region of the decision attribute may decrease. Thus, these existing feature selection methods cannot be constructed based on the consistency of the positive region. Meanwhile, when the classification performance of the original dataset is poor and the positive domains of the decision attribute are relatively small, the selected feature subset should include a large positive reduction set. To solve this problem, Li et al. [19] proved the nonmonotonicity of the positive region in decision rough set models and constructed an attribute reduction algorithm in which the positive region of the reduction set is not less than that of the original dataset. Therefore, in this case, the classification performance of datasets can be improved efficiently. On the inspiring ideas in [19], the nonmonotonicity of decision neighborhood entropy and mutual information is investigated, and a nonmonotonic feature selection method is designed for neighborhood decision systems.

Proposition 8. *Suppose that a neighborhood decision system* $NS = < U, C, D, \delta >$ *with* $B \subseteq C$ *exists. Then, the monotonicity of* $H_\delta(D|B)$ *is uncertain, that is, the decision neighborhood entropy does not satisfy monotonicity.*

Table 1
A neighborhood decision system.

U	a	b	c	d
x_1	0.12	0.41	0.61	Y
x_2	0.21	0.15	0.14	Y
x_3	0.31	0.11	0.26	N
x_4	0.61	0.13	0.23	N

Proof. Suppose $B_1 \subseteq B_2 \subseteq C$, and it follows immediately from Definition 3 that

$$\begin{aligned}\Delta &= H_\delta(D|B_2) - H_\delta(D|B_1)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n^\delta_{B_2}(x_i)\cap[x_i]_D|}{|[x_i]_D|}\cdot\frac{|n^\delta_{B_2}(x_i)\cap[x_i]_D|}{|n^\delta_{B_2}(x_i)|}\right)+\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n^\delta_{B_1}(x_i)\cap[x_i]_D|}{|[x_i]_D|}\cdot\frac{|n^\delta_{B_1}(x_i)\cap[x_i]_D|}{|n^\delta_{B_1}(x_i)|}\right)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n^\delta_{B_2}(x_i)\cap[x_i]_D|}{|[x_i]_D|}\cdot\frac{|n^\delta_{B_2}(x_i)\cap[x_i]_D|}{|n^\delta_{B_2}(x_i)|}\cdot\frac{1}{\frac{|n^\delta_{B_1}(x_i)\cap[x_i]_D|}{|[x_i]_D|}\cdot\frac{|n^\delta_{B_1}(x_i)\cap[x_i]_D|}{|n^\delta_{B_1}(x_i)|}}\right)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\left(\frac{|n^\delta_{B_2}(x_i)\cap[x_i]_D|}{|n^\delta_{B_1}(x_i)\cap[x_i]_D|}\right)^2\cdot\frac{|n^\delta_{B_1}(x_i)|}{|n^\delta_{B_2}(x_i)|}\right)\\ &= \frac{1}{|U|}\sum_{i=1}^{|U|}\left(\log\left(\frac{|n^\delta_{B_1}(x_i)\cap[x_i]_D|}{|n^\delta_{B_2}(x_i)\cap[x_i]_D|}\right)^2-\log\frac{|n^\delta_{B_1}(x_i)|}{|n^\delta_{B_2}(x_i)|}\right).\end{aligned}$$

Since $B_1 \subseteq B_2 \subseteq C$, it can be obtained from Proposition 1 that $n^\delta_{B_1}(x_i) \supseteq n^\delta_{B_2}(x_i)$, and then $\frac{|n^\delta_{B_1}(x_i)|}{|n^\delta_{B_2}(x_i)|} \geq 1$. Let $\frac{|n^\delta_{B_1}(x_i)|}{|n^\delta_{B_2}(x_i)|} = f_1(x_i)$; then, $f_1(x_i) \geq$ 1. Let $\frac{|n^\delta_{B1}(x_i)\cap[x_i]_d|}{|n^\delta_{B_2}(x_i)\cap[x_i]_d|} = f_2(x_i)$; then, $f_2(x_i) \geq$ 1. Thus, $\Delta = \frac{1}{|U|}\sum_{i=1}^{|U|}(\log f_2(x_i)^2 - \log f_1(x_i)) = \frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{(f_2(x_i))^2}{f_1(x_i))}$. Since $f_1(x_i) \geq 1$ and $f_2(x_i) \geq 1$, the size of $\frac{(f_2(x_i))^2}{f_1(x_i)}$ is uncertain, and the plus or minus sign of Δ cannot be determined. That is, the monotonicity of $H_\delta(D|B)$ is uncertain.

In the following example, the performance of Proposition 8 in a neighborhood decision system is shown. □

Example 1. Consider the neighborhood decision system $NS =< U, C, D, \delta >$ in Table 1, where $U = \{x_1, x_2, x_3, x_4\}$, $C = \{a, b, c\}$, $D = \{d\}$, and $\delta = 0.3$.

In Table 1, the neighborhood class of each attribute is calculated using the Euclidean distance function. For an attribute subset $\{a\}$, $\Delta_{\{a\}}(x_1, x_2) = 0.09$, $\Delta_{\{a\}}(x_1, x_3) = 0.19$, $\Delta_{\{a\}}(x_1, x_4) = 0.49$, $\Delta_{\{a\}}(x_2, x_3) = 0.1$, $\Delta_{\{a\}}(x_2, x_4) = 0.4$, and $\Delta_{\{a\}}(x_3, x_4) = 0.3$. Then, the neighborhood classes of any $x_i \in U$ can be computed by

$n^\delta_{\{a\}}(x_1) = \{x_1, x_2, x_3\}$, $n^\delta_{\{a\}}(x_2) = \{x_1, x_2, x_3\}$, $n^\delta_{\{a\}}(x_3) = \{x_1, x_2, x_3, x_4\}$, and $n^\delta_{\{a\}}(x_4) = \{x_3, x_4\}$.

According to $D = \{d\}$ in Table 1, it follows that $U/\{d\} = \{X_1, X_2\} = \{\{x_1, x_2\}, \{x_3, x_4\}\}$. Thus,

$H_\delta(D, \{a\}) = -\frac{1}{4}(\log(\frac{2^2}{2\times3})+\log(\frac{2^2}{2\times3})+\log(\frac{2^2}{2\times4})+\log(\frac{2^2}{2\times2})) = 0.1633$.

Similarly, $H_\delta(D, \{b\}) = 0.301$, $H_\delta(D, \{c\}) = 0.3578$, $H_\delta(D, \{a, b\}) = 0.2698$, $H_\delta(D, \{a, c\}) = 0.4515$, $H_\delta(D, \{b, c\}) = 0.3578$, and $H_\delta(D, \{a, b, c\}) = 0.301$.

From the above calculated results, it can be observed that $H_\delta(D, \{c\}) > H_\delta(D, \{b\}) > H_\delta(D, \{a\})$. Since $H_\delta(D, \{c\})$ is the maximum, the attribute c should be added to the candidate set, i.e., $R = \{c\}$. When the additional attributes are added, $H_\delta(D, \{a, c\}) > H_\delta(D, \{c\})$ and $H_\delta(D, \{b, c\}) = H_\delta(D, \{c\})$, which illustrates that the decision neighborhood entropy proposed in Section 3.4 does not strictly increase with the increase of the number of attributes. Furthermore, through computation, $H_\delta(D, \{b\}) > H_\delta(D, \{a, b\})$, and the decision neighborhood entropy decreases with the increase of the number of attributes. In summary, the decision neighborhood entropy is nonmonotonic.

Proposition 8 shows that $H_\delta(D|B)$, similar to conditional information entropy, can be used to characterize the ability of B to distinguish samples with different decision, and it is clear that the nonmonotonicity of $H_\delta(D|B)$ with regard to the size of the attribute subset B holds.

Proposition 9. *Suppose that a neighborhood decision system $NS =< U, C, D, \delta >$ with $B \subseteq C$ exists. Then, the monotonicity of $MI_\delta(D; B)$ is uncertain, that is, the mutual information does not satisfy monotonicity.*

Proof. Suppose $B_1 \subseteq B_2 \subseteq C$, and it follows immediately from Definition 4 that

$$\begin{aligned}\Delta &= MI_\delta(D;B_2) - MI_\delta(D;B_1)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n_{B_2}^\delta(x_i)||[x_i]_D|^2}{|U||n_{B_2}^\delta(x_i)\cap[x_i]_D|^2}\right) + \frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n_{B_1}^\delta(x_i)||[x_i]_D|^2}{|U||n_{B_1}^\delta(x_i)\cap[x_i]_D|^2}\right)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{|n_{B_2}^\delta(x_i)||[x_i]_D|^2}{|U||n_{B_2}^\delta(x_i)\cap[x_i]_D|^2}\cdot\frac{|U||n_{B_1}^\delta(x_i)\cap[x_i]_D|^2}{|n_{B_1}^\delta(x_i)||[x_i]_D|^2}\right)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\left(\frac{|n_{B_1}^\delta(x_i)\cap[x_i]_D|}{|n_{B_2}^\delta(x_i)\cap[x_i]_D|}\right)^2\cdot\frac{|n_{B_2}^\delta(x_i)|}{|n_{B_1}^\delta(x_i)|}\right)\\ &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\left(\log\left(\frac{|n_{B_2}^\delta(x_i)\cap[x_i]_D|}{|n_{B_1}^\delta(x_i)\cap[x_i]_D|}\right)^2 + \log\frac{|n_{B_1}^\delta(x_i)|}{|n_{B_2}^\delta(x_i)|}\right)\\ &= \frac{1}{|U|}\sum_{i=1}^{|U|}\left(\log\left(\frac{|n_{B_1}^\delta(x_i)\cap[x_i]_D|}{|n_{B_2}^\delta(x_i)\cap[x_i]_D|}\right)^2 - \log\frac{|n_{B_1}^\delta(x_i)|}{|n_{B_2}^\delta(x_i)|}\right).\end{aligned}$$

Similar to Proposition 8, let $\frac{|n_{B_1}^\delta(x_i)|}{|n_{B_2}^\delta(x_i)|} = f_1(x_i)$; then, $f_1(x_i) \geq 1$. Let $\frac{|n_{B1}^\delta(x_i)\cap[x_i]_d|}{|n_{B_2}^\delta(x_i)\cap[x_i]_d|} = f_2(x_i)$; then, $f_2(x_i) \geq 1$. It follows that $\Delta = \frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{(f_2(x_i))^2}{f_1(x_i)}$. Thus, the size of $\frac{(f_2(x_i))^2}{f_1(x_i)}$ is uncertain, and the plus or minus sign of Δ cannot be determined. That is, the monotonicity of $MI_\delta(D;\, B)$ is uncertain. □

Example 2. Consider the neighborhood decision system $NS =< U, C, D, \delta >$ in Table 1, where $U = \{x_1, x_2, x_3, x_4\}$, $C = \{a, b, c\}$, $D = \{d\}$, and $\delta = 0.3$.

Assume that $B_1 = \{b\}$, $B_2 = \{a, b\}$; then, $B_1 \subseteq B_2 \subseteq C$. Based on the calculation results of Example 1, it follows that $MI_\delta(D; \{b\}) - MI_\delta(D; \{a, b\}) = H_\delta(D; \{a, b\}) - H_\delta(D; \{b\}) < 0$. However, if it is assumed that $B_1 = \{b\}, B_3 = \{b, c\}$, then $MI_\delta(D; \{b\}) - MI_\delta(D; \{b, c\}) = H_\delta(D; \{b, c\}) - H_\delta(D; \{b\}) > 0$. Thus, the monotonicity of $MI_\delta(D;\, B)$ is uncertain.

Proposition 9 states that whether the mutual information decreases or increases as the distinguishing ability of an attribute subset increases cannot be concluded. On the basis of the relationship between the decision neighborhood entropy and the mutual information, it can be observed that the monotonicity of $H_\delta(D|B)$ and $MI_\delta(D;\, B)$ does not hold based on the size of the attribute subset B.

Definition 5. Suppose that a neighborhood decision system $NS =< U, C, D, \delta >$ with $B \subseteq C$ and any $a \in B$ exist. If $H_\delta(D|B) \leq H_\delta(D|B - \{a\})$, then a is called redundant in B with respect to D; otherwise, a is indispensable in B with respect to D; and B is called dependent if any attribute in B is indispensable in B with respect to D, and B is called a reduct of C with respect to D if it satisfies the following two conditions:

(1) $H_\delta(D|B) \geq H_\delta(D|C)$,
(2) $H_\delta(D|B - \{a\}) < H_\delta(D|B)$, where any $a \in B$.

In a neighborhood decision system $NS =< U, C, D, \delta >$ with any $a \in B \subseteq C$. It follows from Definition 5 and Proposition 8 that if $MI_\delta(D;\, B) \leq MI_\delta(D;\, C)$ and $MI_\delta(D; B - \{a\}) > MI_\delta(D; B)$, then B is also called a reduct of C with respect to D.

It is obvious that a reduct of C with respect to D is a minimal attribute subset to retain or improve the mutual information of B and D.

Definition 6. Suppose that a neighborhood decision system $NS =< U, C, D, \delta >$ with $B \subseteq C$ and any $a \in B$ exist. Then, the significance measure of a in B with respect to D is defined as

$$Sig_{inner}(a, B, D) = H_\delta(D|B) - H_\delta(D|B - \{a\}). \tag{19}$$

Definition 7. Suppose that a neighborhood decision system $NS =< U, C, D, \delta >$ with any $a \in C$; if $Sig_{inner}(a,\, B,\, D) > 0$, then the attribute a is a core of C relative to D.

Definition 8. Suppose that a neighborhood decision system $NS =< U, C, D, \delta >$ with $B \subseteq C$ and any $a \in C - B$ exist. Then, the significance measure of a for B with respect to D is defined as

$$Sig_{outer}(a, B, D) = H_\delta(D|B \cup \{a\}) - H_\delta(D|B). \tag{20}$$

When $B = \emptyset$, $Sig_{outer}(a, B, D) = H_\delta(D|\{a\})$. From Definition 8, it can be seen that the significance measure of a indicates the increment of the distinguishing information after adding a into B. Then, the larger the value of $Sig_{outer}(a,\, B,\, D)$, the more important gene a is for B with respect to D.

Note that in $NS =< U, C, D, \delta >$ with $B \subseteq C$ and any $a \in C - B$, when calculating $Sig_{outer}(a,\, B,\, D)$, every time any testing attribute a with the maximum of $Sig_{outer}(a,\, B,\, D)$ is calculated, the calculation is in fact for the maximum of $H_\delta(D|B \cup \{a\})$ because $H_\delta(D|B)$ is a constant. That is, we need to calculate only the neighborhood classes of neighborhood entropy-based uncertainty measures. Therefore, according to Definition 8, the calculation of the maximum of $Sig_{outer}(a,\, B,\, D)$ is equivalent to that of the maximum of $H_\delta(D|B \cup \{a\})$.

4.2. Feature selection algorithm with the Fisher score for gene expression data

The classical rough set will discretize continuous datasets, which could result in the loss of some important information, and the original properties of the gene expression datasets also change after discretization [16]. In the neighborhood rough set, the proposal of neighborhood relations can solve the above problem [17]. In the gene expression datasets, the measured gene expression levels are presented by continuous-valued datasets at different magnitudes [30]. Hence, to avoid the discretization of continuous datasets, neighborhood rough sets can be used. Notably, the neighborhood decision system can be used to describe the gene expression dataset, where an object corresponds to a sample, a conditional attribute shows a gene, and a decision attribute expresses a subclass of cancer.

Given a neighborhood decision system $NS =< U, C, D, \delta >$ of gene expression data, where $U = \{x_1, x_2, \cdots, x_n\}$ is a gene sample set, n denotes the number of samples, $C = \{a_1, a_2, \cdots, a_m\}$ is a gene set that describes samples, and m denotes the number of genes. The corresponding matrix of the original gene expression data is formalized as $X \in R^{m\times n}$. To address this time-consuming issue, a heuristic strategy is usually employed to calculate a score for each gene independently using certain criteria. Here, the Fisher score method [47] is introduced to calculate the Fisher score of the j-th gene by

$$f(j) = \frac{\sum_{i=1}^{Q} n_i(\mu_i^j - \mu^j)^2}{\sum_{i=1}^{Q} n_i(\sigma_i^j)^2}, \tag{21}$$

where Q is the number of the sample classes, and n_i denotes the sample number of the i-th class, μ_i^j and σ_i^j are the mean and standard deviation of the samples from the i-th class corresponding to the j-th gene, respectively, and μ^j denotes the mean of the samples corresponding to the j-th gene. Then, the Fisher score method, as a preprocessing method, can significantly reduce the dimensionality of the genes and effectively distinguish gene expression datasets. After obtaining the Fisher score of each gene, the genes with the top-l highest scores are selected to construct a candidate gene subset. The process of preliminary dimension reduction is illustrated by Algorithm 1.

Algorithm 1

```
Input: The original high-dimensional gene expression data matrix $X \in R^{m\times n}$ and the number $l$ of expected genes
Output: The preliminary candidate gene subset $S$
1: for each gene in high-dimensional space do
2:     Evaluate the corresponding Fisher score by Eq. (3) in [47] and record the score in a score array.
3: end for
4: Sort the score in descending order with the radix sorting algorithm in [34].
5: Select the top-$l$ genes with a high score, and place their gene indicesinto the set $T$.
6: Obtain the dimension-reduction gene subset $S$ in terms of $T$.
7: return  the preliminary candidate gene subset $S$.
```

The time complexity of Algorithm 1 is determined by step 1 through step 3. Given m genes, the time complexity of step 1 through step 3 is $O(m)$. At step 4, the time complexity of the radix sorting algorithm in [34] is $O(m)$, which is linearly complex. Thus, the time complexity of Algorithm 1 is $O(m)$.

To achieve efficient dimensionality reduction, many heuristic search methods have been developed, and the forward greedy search strategy is usually employed [5,10,16,31]. A decision neighborhood entropy-based heuristic attribute reduction (DNEAR) algorithm for neighborhood decision systems is described in Algorithm 2.

Algorithm 2

```
Input: A neighborhood decision system $NS =< U, C, D, \delta >$
Output: A reduction attribute set $R$
1: Initialize $R = \emptyset$.
2: while $Sig_{outer}(C, R, D) \le 0$ do
3:     Let $Agent = R$, and $h = 0$.
4:     for any $a \in C - R$ do
5:         Compute $H_\delta(D|R \cup \{a\})$.
6:         if $H_\delta(D|R \cup \{a\}) > h$ then
7:             Let $Agent = R \cup a$ and $h = H_\delta(D|R \cup \{a\})$.
8:         end if
9:     end for
10:    Let $R = Agent$.
11: end while
12: return  A reduction attribute set $R$.
```

Table 2
Description of the ten experimental data sets.

No.	Data sets	Genes	Samples	Classes
1	Brain_Tumor2	10367	50	4 (14/7/14/15)
2	Colon	2000	62	2(40/22)
3	DLBCL	5469	77	2(58/19)
4	Leukemia	7129	72	2(47/25)
5	Leukemia1	5327	72	3(9/38/25)
6	Lung	12533	181	2(31/150)
7	Prostate	12600	136	2(59/77)
8	Prostate1	10509	102	2(52/50)
9	SRBCT	2308	63	4(23/8/12/20)
10	9_Tumors	5726	60	9(9/7/8/6/6/8/8/2/6)

In the DNEAR algorithm, the neighborhood classes need to be frequently calculated in a neighborhood decision system. The process of achieving neighborhood classes largely affects the time complexity of selecting attributes. The sorting algorithm of buckets in [34] seems feasible in practice; thus, the time complexity of calculating the neighborhood class is $O(mn)$. Then, the computational time complexity of the decision neighborhood entropy is $O(m)$. Obviously, $O(m) < O(mn)$; therefore, the worst complexity of calculating decision neighborhood entropy is $O(mn)$. In such a case, there are two loops at step 2 through step 11 of the DNEAR algorithm, and so the worst time complexity of DNEAR is $O(m^3n)$. Suppose that for the calculation of neighborhood classes, the number of selected attributes is m_R. We consider only the candidate attributes that do not involve the entire attribute subset. Thus, the time complexity of calculating all neighborhood classes is $O(m_Rn)$. Since the number of outer loop iterations is m_R and the number of inner loop iterations is $(m - m_R)$ in DNEAR, the total time complexity of the DNEAR algorithm is approximately $O(nm_R(m - m_R)m_R)$. It is well known that $m_R \ll m$ in most cases. Therefore, the time complexity of DNEAR is close to $O(mn)$. Furthermore, its space complexity is $O(mn)$.

In what follows, the above two preceding subalgorithms are used to construct a feature selection algorithm with the Fisher score based on decision neighborhood entropy (FSDNE), which is described in Algorithm 3.

Algorithm 3

```
Input: The original high-dimensional gene expression data matrix $X \in R^{m \times n}$, and the number $l$ of expected genes
Output: A selected gene subset $R$
1: Initialize $R = \emptyset$.
2: Compute the preliminary selected candidate gene subset $S$ with Algorithm 1.
3: Let $NS = < U, S \cup D, \delta >$.
4: Obtain $R$ in $NS$ with Algorithm 2.
5: return  A selected gene subset $R$.
```

By using the FSDNE algorithm, the time complexity of feature selection for the gene expression dataset is a polynomial. Suppose that l genes are selected at step 2 through step 3 of the FSDNE algorithm to form a candidate gene subset. Then, the time complexity of step 2 through step 3 is approximate to that of Algorithm 1. Since $l \ll m$ in most cases, the time complexity of step 4 is $O(ln)$. Thus, the total time complexity of FSDNE is $O(m + ln)$. Therefore, the time complexity of FSDNE is approximately $O(m)$. In Algorithm 3, through the dimensionality reduction with the Fisher score method, the time complexity is effectively decreased. Thus far, FSDNE appears to be more efficient than some of the existing algorithms for feature selection [5,16,31,42,46,50] in neighborhood decision systems.

5. Experimental results and analysis

5.1. Experiment preparation

In this section, the performance of our gene selection algorithm given in Section 3.2 is demonstrated. The gene expression datasets shown in Table 2 are described in detail as follows:

(1) A brain tumor [18] occurs when abnormal cells form within the brain. Brain tumors may produce symptoms that vary depending on the part of the brain involved. The Brain_Tumor2 gene expression dataset contains 10,367 genes and 50 samples with four subtypes.

(2) Colon cancer [27] is the development of cancer in the colon or rectum. Most colon cancers are due to old age and lifestyle factors, and only a small number of cases are due to underlying genetic disorders. The dataset contains 2000 genes and 62 samples, including 40 patient samples and 22 healthy samples.

(3) Diffuse large B-cell lymphoma (DLBCL) [34] is a cancer of B cells, a type of white blood cell responsible for producing antibodies. It is the most common type of non-Hodgkin lymphoma among adults, with an annual incidence of 7–8

cases per 100,000 people per year in the USA and the UK. The dataset contains 5469 genes and 77 samples, including 58 patient samples and 19 healthy samples.

(4) Leukemia [36] is a group of cancers that usually begin in the bone marrow and result in high numbers of abnormal white blood cells. These white blood cells are not fully developed and are called blasts or leukemia cells. The dataset contains 7129 genes and 72 samples, including 47 patient samples and 25 healthy samples.

(5) Leukemia1 [8] is derived from the characteristic high white blood cell count that presents in most afflicted people before treatment. The dataset contains 5327 genes and 72 samples. There are three subtypes of leukemia, including 9 samples for ALL-T (acute lymphoblastic leukemia, T-cell), 38 samples for ALL-B (acute lymphoblastic leukemia, B-cell), and 25 samples for acute myeloid leukemia.

(6) Lung cancer [39] is a disease in which certain cells in the lungs become abnormal and multiply uncontrollably to form a tumor, which is categorized by the size and appearance of the malignant cells. The dataset contains 12,533 genes and 181 samples, including 31 patient samples and 150 healthy samples.

(7) Prostate cancer [38] is cancer that occurs in the prostate, and although no single gene is responsible for prostate cancer, many different genes have been implicated. Mutations in BRCA1 and BRCA2 have been implicated in prostate cancer. The dataset contains 12,600 genes and 136 samples, including 59 patient samples and 77 healthy samples.

(8) Prostate1 cancer [48] is the part of the prostate cancer that occurs in the prostate. The dataset contains 10,509 genes and 102 samples, including 52 patient samples and 50 healthy samples.

(9) Small-round-blue-cell tumor (SRBCT) [48] is any one of a group of malignant neoplasms that have a characteristic appearance under the microscope. The dataset contains 2308 genes and 63 samples with four subtypes, including 23 Ewing Sarcoma, 8 Burkitt Lymphoma, 12 Neuroblastoma and 20 Rhabd omyosarcoma.

(10) 9_Tumors [48] contains 5726 genes and 60 samples with nine subtypes: nonsmall cell lung cancer, colon cancer, breast cancer, ovarian cancer, leukemia, kidney cancer, melanoma, prostate cancer and central nervous system cancer.

The experiments were performed on a personal computer running Windows 7 with an Intel(R) Core(TM) i5-3470 CPU operating at 3.20 GHz with 4 GB memory. All simulation experiments were performed in MATLAB R2016a, and the classifiers (KNN, C4.5 and SVM) were selected to verify the classification accuracy in WEKA, where the parameter $k = 5$ in KNN and the linear kernel function was selected in SVM. The following experimental comparisons for classification on the selected genes are implemented with 10-fold cross-validations on all test datasets.

5.2. The effect of Fisher score method

The Fisher score method is used to achieve preliminary dimensionality reduction. For each gene expression dataset, the Fisher score of each gene is calculated and sorted, and then l genes are selected to constitute a candidate gene subset. The classification accuracy of Algorithm 1 with the different number of genes was verified in WEKA. Fig. 1 illustrates the changing trend of the classification accuracy versus the number of genes for the ten gene expression datasets with Algorithm 1. As shown in Fig. 1, the classification accuracies for each gene expression dataset with different values of l are very similar in most situations. Furthermore, the cardinality and classification accuracy of the candidate gene subset are two important indices for evaluating the classification performance of feature selection algorithms. Therefore, the appropriate values of l are selected from Fig. 1, and the value of l is set to 300 in the Brain_Tumor2 dataset, to 200 in the Colon, DLBCL, Leukemia, Leukemia1, Prostate, Prostate1 and 9_Tumors datasets, and to 50 in the Lung and SRBCT datasets.

5.3. Effect of different neighborhood parameter values

The following part of our experiments concerns the reduction rate and the classification accuracy with the different neighborhood parameter values. For gene expression data, feature selection aims to improve classification accuracy by eliminating redundant genes. The neighborhood parameters influence the size of granulated gene data, which affects the classification accuracy of selected genes. Therefore, the different neighborhood parameter values should be set in the process of feature selection of gene expression datasets. Moreover, the altered neighborhood parameters also affect the cardinality of the selected gene subset. To obtain a suitable neighborhood parameter and a good gene subset, the cardinality and classification accuracy of a gene subset for different neighborhood parameters should be discussed in detail. To explain the reduction in performance and classification accuracy with different neighborhood parameters, Chen et al. [5] designed a reduction rate to evaluate attribute redundancy. However, for our proposed Algorithm 3, the test curves of Chen's reduction rate were nearly horizontal; therefore, this reduction rate cannot efficiently illustrate the performance of our method. To identify better neighborhood parameters, we present a novel reduction rate to evaluate gene redundancy for our feature selection method and identify better neighborhood parameters.

Definition 8. A reduction rate for gene expression datasets is defined as

$$Rate_{\delta} = 1 - \frac{|R_{\delta}|}{\max(|R_{\delta}|)}, \tag{22}$$

where $|R_{\delta}|$ represents the number of selected genes generated by a given δ. A higher reduction rate indicates that the algorithm has a stronger reduction ability for gene expression datasets. Thus, a higher reduction rate indicates lower redundancy.

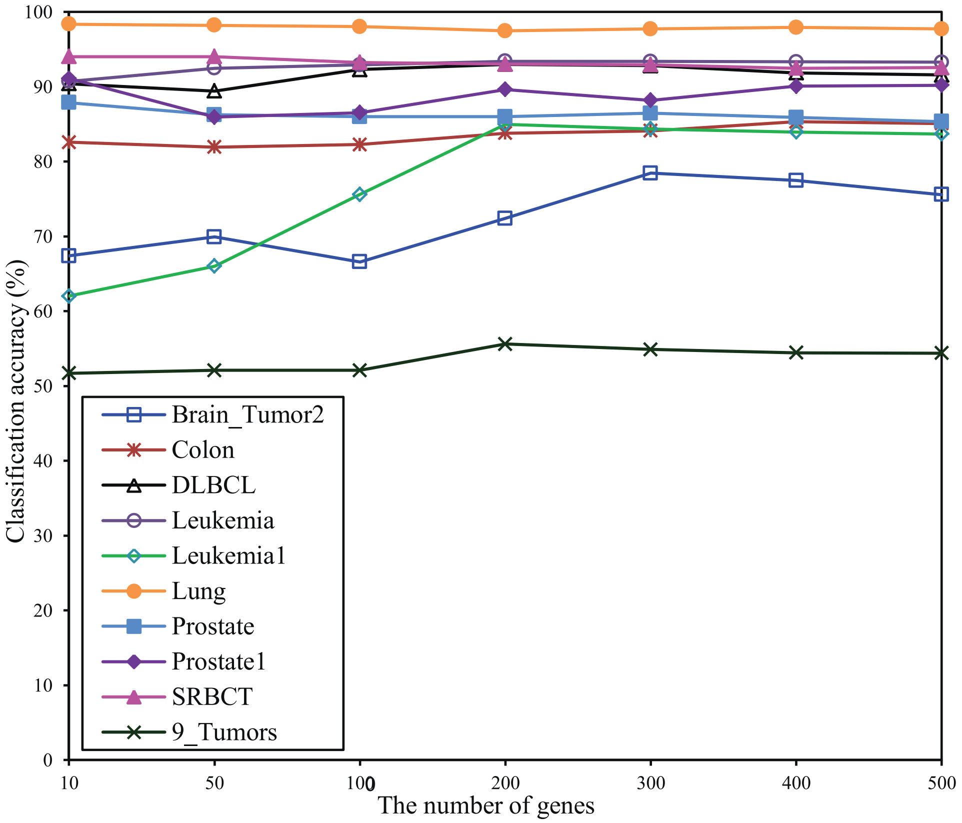


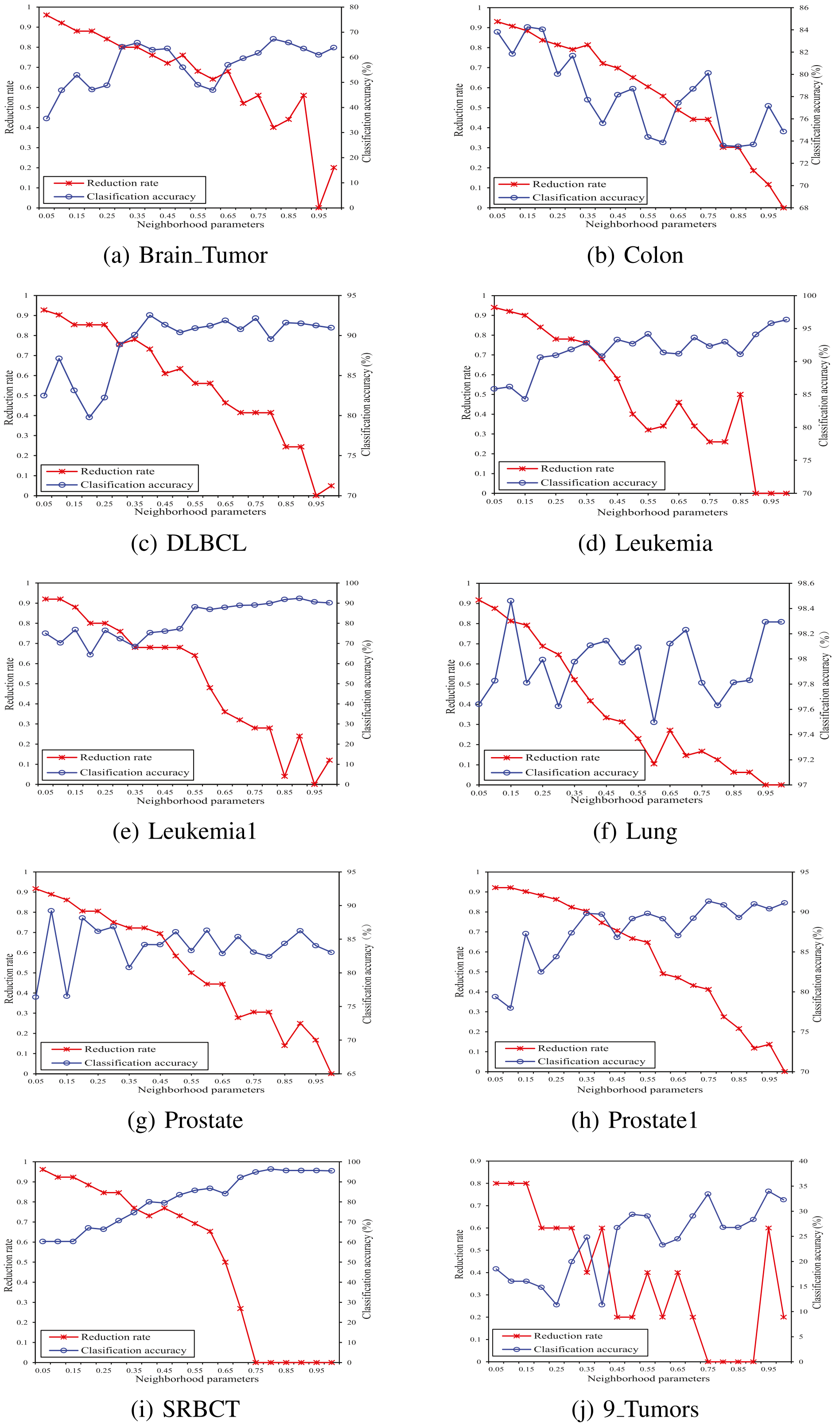
Fig. 1. Classification accuracy versus the number of genes in the five gene expression datasets with Algorithm 1.

The corresponding experiments are performed to graphically illustrate the reduction performance and classification accuracy of Algorithm 3 under different neighborhood parameter values. The results are shown in Fig. 2, where the horizontal axis denotes the neighborhood parameters with the neighborhood parameter $\delta \in [0.05, 1]$ at intervals of 0.05, and the left and right vertical axes represent the reduction rate and classification accuracy, respectively.

From Fig. 2, it can be observed that the neighborhood parameter value greatly influences the classification performance of FSDNE. Furthermore, Fig. 2 illustrates that the thinner the granule, the smaller the roughness value of the granule when the values of different neighborhood parameters are smaller. It follows that the reduction rate increases as the roughness of the granule decreases. The neighborhood parameter is usually set to make both the reduction rate and the classification accuracy high. Thus, the appropriate parameter values can be selected for each dataset from Fig. 2. In Fig. 2(a), for the Brain_Tumor2 dataset, as the neighborhood parameters increase, the reduction rate decreases and the classification accuracy tends to increase. When the neighborhood parameter δ is 0.35, the reduction rate and classification accuracy are at a high level. In Fig. 2(b), the classification accuracy of the Colon dataset is higher when $\delta \in [0.05, 0.2]$ and the reduction rate reaches a maximum when the neighborhood parameter δ is 0.05. Similar to the Brain_Tumor2 dataset, in Figs. 2(c)-(e), the δ of the DLBCL, Leukemia and Leukemia 1 datasets can be set to 0.4, 0.35 and 0.55, respectively. In Fig. 2(f), for the Lung dataset, the classification accuracy reaches a maximum when the neighborhood parameter δ is 0.15 and the reduction rate is high. Hence, the δ of Lung can be set to 0.15. Similar to the Lung dataset, in Figs. 2(g) and (j), the δ of the Prostate and 9_Tumors datasets can be set to 0.1 and 0.95, respectively. In Figs. 2(h) and (i), for the Prostate1 and SRBCT datasets, the reduction rate and classification accuracy are at a high level when the values of the parameters are 0.35 and 0.6, respectively; thus, the δ of the Prostate1 and SRBCT datasets can be set to 0.35 and 0.6, respectively.

By analyzing Fig. 2, the appropriate neighborhood parameters of different datasets are determined. In addition, the classification results of the original data and the reduced data using Algorithm 3 on the ten gene expression datasets obtained from the three classifiers (KNN, C4.5 and SVM) with 10-fold cross-validation are shown in Table 3, where the neighborhood parameter values are listed in the last column. Here, it is noted that the bold font means the best value in the following subsections.

Table 3 shows that our proposed algorithm can greatly reduce the dimensionality of all gene expression datasets without a loss of classification accuracy, and most of the redundant genes are reduced. For the ten gene expression datasets, the average classification accuracies on the KNN and C4.5 classifiers are higher than those on the original datasets by 5.6%



(a) Brain_Tumor

(b) Colon

(c) DLBCL

(d) Leukemia

(e) Leukemia1

(f) Lung

(g) Prostate

(h) Prostate1

(i) SRBCT

(j) 9_Tumors

Fig. 2. Reduction rate and classification accuracy for the ten gene expression datasets with different neighborhood parameter values.

Table 3
Classification results of the original data and the reduced data under the three classifiers.

Datasets	Original data				Reduced data using Algorithm 3				δ
	Genes	KNN	C4.5	SVM	Genes	KNN	C4.5	SVM	
Brain_Tumor2	10367	0.61	0.56	**0.736**	5	**0.634**	**0.732**	0.606	0.35
Colon	2000	0.776	**0.82**	0.811	3	**0.84**	0.796	**0.838**	0.05
DLBCL	5469	0.896	0.809	0.925	11	**0.946**	**0.903**	**0.927**	0.4
Leukemia	7129	0.842	0.814	0.913	9	**0.952**	**0.905**	**0.929**	0.35
Leukemia1	5327	0.821	0.846	0.831	9	**0.902**	**0.882**	**0.861**	0.55
Lung	12533	0.935	0.939	**1**	8	**0.987**	**0.979**	0.988	0.15
Prostate	12600	0.796	0.791	**0.916**	4	**0.895**	**0.898**	0.884	0.1
Prostate1	10509	0.843	0.846	**0.907**	10	**0.876**	**0.912**	**0.907**	0.35
SRBCT	2308	0.808	0.78	0.924	9	**0.846**	**0.821**	**0.936**	0.6
9_Tumors	5726	0.357	0.268	0.327	2	**0.36**	**0.31**	**0.35**	0.95
Average	7397	0.768	0.747	**0.829**	7	**0.824**	**0.814**	0.823	

Table 4
Selected gene subsets on the ten gene expression datasets using FSDNE.

Datasets	Gene subset after reduction
Brain_Tumor2	{ 9413, 7844, 642, 9794, 7169 }
Colon	{765, 627, 1668 }
DLBCL	{ 3127, 3942, 874, 1600, 3264, 4588, 4094, 2949, 2971, 3304, 889 }
leukemia	{ 4196, 1144, 758, 5552, 1630, 2659, 3897, 6584, 6471 }
Leukemia1	{ 4688, 3256, 1610, 568, 848, 5032, 861, 3358, 2197 }
Lung	{ 2255, 11957, 12298, 4815, 1673, 8709, 4772, 2421 }
Prostate	{ 6185, 8330, 4483, 5155 }
Prostate1	{ 10349, 7652, 2718, 2596, 2792, 10130, 7515, 785, 7266, 6745 }
SRBCT	{ 758, 545, 836, 1884, 1954, 74, 1327, 1974, 1319 }
9_Tumors	{ 2590, 1677 }

Table 5
Classification accuracy of the selected Brain_Tumor2 genes with the six algorithms.

Gene selection methods	Genes	KNN	C4.5	SVM	Average
ODP	10367	0.61	0.56	**0.736**	0.635
MEAR	–	–	–	–	–
EGGS	9	0.492	0.492	0.538	0.507
DNEAR	3	0.478	0.464	0.514	0.485
EGGS-FS	5	0.492	0.392	0.514	0.466
FSDNE	5	**0.634**	**0.732**	0.606	**0.657**

and 6.7%, respectively. However, differences in the classification accuracy are observed for the SVM classifier. On the KNN and C4.5 classifiers, the classification accuracies of almost all datasets are higher than those of the original datasets. On the SVM classifier, the classification accuracies of the Colon, DLBCL, Leukemia and Leukemia1 datasets are 2.7%, 0.2%, 1.6% and 3% higher than those of the original datasets, respectively. However, for the Brain_Tumor2, Lung and Prostate datasets, the classification accuracies are slightly lower than those of the original datasets. This situation may be due to the loss of some genes that have important information during reduction. Therefore, the FSDNE algorithm is efficient in reducing the dimensionality of low-dimensional and high-dimensional datasets.

By using the FSDNE algorithm and the neighborhood parameters set in the previous part, the results of selected gene subsets from the ten gene expression datasets are shown in Table 4.

5.4. Comparisons of the classification accuracy of entropy-based feature selection algorithms

This portion of our experiment evaluates the classification performance of our proposed algorithm in terms of classification accuracy for the selected genes. The state-of-the-art entropy-based feature selection algorithms used in the comparison include the following: (1) the mutual entropy-based attribute reduction algorithm (MEAR) [46], (2) the entropy gain-based gene selection algorithm (EGGS) [5], (3) our proposed DNEAR algorithm, and (4) the EGGS algorithm [5] combined with the Fisher score [47] (EGGS-FS). Following the experimental techniques designed in [5,46,47], the classification performance of the FSDNE algorithm is compared with four reduction algorithms using the ten gene expression datasets in Table 2. Tables 5,6,7,8,9,10,11,12,13–14 show the experimental results of the six different reduction methods, where ODP describes the original data processing method, the symbol - denotes that no results were obtained using the corresponding algorithm, and the neighborhood parameter values were determined as discussed in Subsection 4.3.

Table 6
Classification accuracy of the selected Colon genes with the six algorithms.

Gene selection methods	Genes	KNN	C4.5	SVM	Average
ODP	2000	0.776	0.82	0.811	0.802
MEAR	5	0.77	**0.822**	**0.849**	0.814
EGGS	11	0.649	0.646	0.556	0.617
DNEAR	15	0.579	0.566	0.628	0.591
EGGS-FS	2	0.702	0.672	0.621	0.665
FSDNE	3	**0.84**	0.796	0.838	**0.825**

Table 7
Classification accuracy of the selected DLBCL genes with the six algorithms.

Gene selection methods	Genes	KNN	C4.5	SVM	Average
ODP	5469	0.896	0.809	0.925	0.877
MEAR	2	0.765	0.778	0.777	0.773
EGGS	20	0.854	0.826	0.781	0.82
DNEAR	10	0.698	0.718	0.692	0.703
EGGS-FS	3	0.87	0.801	0.841	0.837
FSDNE	11	**0.946**	**0.903**	**0.927**	**0.925**

Table 8
Classification accuracy of the selected Leukemia genes with the six algorithms.

Gene selection methods	Genes	KNN	C4.5	SVM	Average
ODP	7129	0.842	0.814	0.913	0.856
MEAR	3	0.928	**0.934**	0.920	0.927
EGGS	8	0.629	0.733	0.802	0.721
DNEAR	8	0.533	0.671	0.691	0.632
EGGS-FS	5	0.801	0.813	0.680	0.765
FSDNE	9	**0.952**	0.905	**0.929**	**0.929**

Table 9
Classification accuracy of the selected Leukemia1 genes with the six algorithms.

Gene selection methods	Genes	KNN	C4.5	SVM	Average
ODP	5327	0.821	0.846	0.831	0.833
MEAR	5	0.83	0.889	0.881	0.867
EGGS	3	0.513	0.558	0.546	0.539
DNEAR	11	0.5	0.512	0.542	0.518
EGGS-FS	2	0.88	**0.902**	0.886	**0.889**
FSDNE	9	**0.902**	0.882	**0.861**	0.882

Table 10
Classification accuracy of the selected Lung genes with the six algorithms.

Gene selection methods	Genes	KNN	C4.5	SVM	Average
ODP	12533	0.935	0.939	**1**	0.958
MEAR	6	0.958	0.964	0.929	0.950
EGGS	12	0.859	0.966	0.960	0.928
DNEAR	6	0.822	0.819	0.833	0.825
EGGS-FS	6	0.979	0.955	0.990	0.975
FSDNE	8	**0.987**	**0.979**	0.988	**0.985**

Table 11
Classification accuracy of the selected Prostate genes with the six algorithms.

Gene selection methods	Genes	KNN	C4.5	SVM	Average
DP	12600	0.796	0.791	**0.916**	0.834
MEAR	4	0.512	0.566	0.564	0.547
EGGS	8	0.639	0.591	0.532	0.587
DNEAR	5	0.611	0.570	0.657	0.613
EGGS-FS	14	0.849	0.863	0.878	0.863
FSDNE	4	**0.895**	**0.897**	0.884	**0.892**

Table 12
Classification accuracy of the selected Prostate1 genes with the six algorithms.

Gene selection methods	Genes	KNN	C4.5	SVM	Average
ODP	10509	0.843	0.846	**0.907**	0.865
MEAR	–	–	–	–	–
EGGS	20	0.632	0.703	0.637	0.657
DNEAR	9	0.722	0.606	0.698	0.675
EGGS-FS	5	0.849	**0.931**	0.900	0.893
FSDNE	10	**0.876**	0.912	**0.907**	**0.898**

Table 13
Classification accuracy of the selected SRBCT genes with the six algorithms.

Gene selection methods	Genes	KNN	C4.5	SVM	Average
ODP	2308	0.808	0.78	0.924	0.837
MEAR	1	0.389	0.365	0.364	0.373
EGGS	12	0.575	0.513	0.703	0.597
DNEAR	12	0.383	0.418	0.428	0.41
EGGS-FS	1	0.637	0.626	0.651	0.638
FSDNE	9	**0.846**	**0.821**	**0.936**	**0.868**

Table 14
Classification accuracy of the selected 9_Tumors genes with the six algorithms.

Gene selection methods	Genes	KNN	C4.5	SVM	Average
ODP	5726	0.357	0.268	0.327	0.317
MEAR	–	–	–	–	–
EGGS	1	0.105	0.12	**0.667**	0.297
DNEAR	10	0.183	0.183	0.175	0.18
EGGS-FS	1	0.21	0.202	0.292	0.235
FSDNE	2	**0.36**	**0.31**	0.35	**0.34**

As shown in Table 5, the FSDNE algorithm selects 5 important genes from the original Brain_Tumor2 genes and obtains 63.4% and 73.2% better classification accuracy on the KNN and C4.5 classifiers, respectively. The classification accuracies of genes selected by the EGGS, DNEAR and EGGS-FS algorithms are lower than those of our FSDNE algorithm on the KNN, C4.5 and SVM classifiers. Since the process of discretization generally causes a loss of genes with significant information, MEAR cannot acquire a reduction subset from Brain_Tumor2, and its results are denoted by the symbol -. For the FSDNE algorithm, although its classification accuracy is close to that of the original data on SVM, the number of selected genes is substantially lower than that of ODP. Thus, our algorithm can effectively remove noise from the original Brain_Tumor2 dataset.

Table 6 shows that the FSDNE algorithm achieves the highest average classification accuracy for the selected Colon genes. On the KNN classifier, the classification accuracy of genes selected by our algorithm is 84%, which is higher than those of the other five methods. On the C4.5 and SVM classifiers, the classification accuracies of genes selected by the FSDNE algorithm are slightly lower than those of the MEAR algorithm. However, the FSDNE algorithm selects fewer genes. Therefore, our algorithm can effectively remove noise from the original data and improve the classification accuracy of the selected Colon genes.

As shown in Table 7, the FSDNE algorithm achieves the highest average classification accuracy for the selected DLBCL genes. On the three classifiers (KNN, C4.5 and SVM), the classification accuracies of genes selected by our algorithm are higher than those of the other five reduction algorithms. Although the number of genes selected by the FSDNE algorithm is higher than that selected by the MEAR and EGGS-FS algorithms, the classification accuracies of genes selected by our algorithm are 18.1%, 12.5% and 15% higher than those of MEAR and 7.6%, 10.2% and 8.6% higher than those of EGGS-FS on the KNN, C4.5 and SVM classifiers, respectively, because the MEAR and EGGS-FS algorithms lose some important genes during reduction, thereby resulting in a reduction of classification accuracy.

Table 8 shows that the FSDNE algorithm obtains the highest average classification accuracy for the selected Leukemia genes. The average classification accuracy of the FSDNE algorithm is higher than that of the EGGS, DNEAR and EGGS-FS algorithms by 20.8%, 29.67% and 16.4%, respectively. Since the MEAR algorithm may result in the loss of some useful genes, the classification accuracies of MEAR are lower than those of FSDNE on the KNN and SVM classifiers. Therefore, our proposed algorithm exhibits better classification performance than the other five methods.

As shown in Table 9, the FSDNE algorithm achieves the highest classification accuracy for the selected Leukemia1 genes on the KNN and SVM classifiers, and the classification accuracies of selected genes are 90.2% and 86.1%, respectively. On the C4.5 classifier, the classification accuracy of genes selected by our algorithm is slightly lower than that of the MEAR and EGGS-FS algorithms. However, for the average classification accuracy, FSDNE is similar to EGGS-FS. Thus, both FSDNE and EGGS-FS can effectively remove noise from the original Leukemia1 dataset.

Table 10 shows that the FSDNE algorithm achieves the highest average classification accuracy for the selected Lung genes. On the KNN and C4.5 classifiers, the classification accuracies of genes selected by our algorithm are 98.7% and 97.9%, respectively, which are higher than those of the other five methods. However, at the expense of more genes for the original Lung dataset, the classification accuracy of the FSDNE algorithm on the SVM classifier is slightly lower than that of ODP. Hence, our algorithm can efficiently eliminate noise and improve the classification accuracy of the original Lung dataset.

Similar to the classification results in Tables 8 and 10, Table 11 illustrates that the highest average classification accuracy is achieved with the FSDNE algorithm for the selected Prostate genes. On the KNN and C4.5 classifiers, the classification accuracies of genes selected by the FSDNE algorithm are 89.5% and 89.7%, respectively, which are higher than those of the other five methods. Compared with ODP, our FSDNE algorithm selects the lowest number of genes, although its classification accuracy on the SVM classifier is slightly lower than that of ODP. Furthermore, both the FSDNE and EGGS algorithms select 4 genes from the original dataset. However, the average classification accuracy of FSDNE is 89.2%, which is nearly 30.5% higher than that of EGGS. Therefore, our algorithm is superior to the other five methods in terms of number and classification accuracy for the selected Prostate genes, and it has better classification performance.

Table 12 shows that the FSDNE algorithm achieves the highest average classification accuracy for the selected Prostate1 genes. On the KNN and SVM classifiers, the classification accuracies of genes selected by our algorithm are 87.6% and 90.7%, respectively, which are higher than those of the EGGS, DNEAR and EGGS-FS algorithms. On the C4.5 classifier, the classification accuracy of genes selected by FSDNE is slightly lower than that of EGGS-FS and higher than those of EGGS and DNEAR. Although the number of genes selected by FSDNE is higher than that of EGGS-FS, both algorithms can effectively reduce the dimensionality of the original Prostate1 dataset while improving the classification accuracy of the original dataset. Similar to Table 5, MEAR cannot acquire a reduction subset from the Prostate1 dataset.

Table 13 clearly shows that the FSDNE algorithm achieves the highest classification accuracy for the selected SRBCT genes on the KNN, C4.5 and SVM classifiers. Compared with the MEAR and EGGS-FS algorithms, the FSDNE algorithm retains genes with classification information and achieves higher classification accuracy. Compared with the classification results of ODP, this algorithm eliminates a large number of redundant and noise genes to effectively improve the classification accuracy. Therefore, our algorithm achieves the best classification performance on the SRBCT dataset.

According to the classification results in Table 14, the FSDNE algorithm selects 2 important genes from the original genes and obtains 36% and 31% classification accuracy on the KNN and C4.5 classifiers, respectively. The classification accuracies of genes selected by the DNEAR and EGGS-FS algorithms are lower than those of our FSDNE algorithm on the KNN, C4.5 and SVM classifiers. Similar to Tables 5 and 12, the MEAR algorithm still cannot acquire a reduction subset from the 9_Tumors dataset, which shows that this algorithm is unsteady. For the EGGS algorithm, the classification accuracies have large differences on three classifiers, and the average classification accuracy is lower than that of the FSDNE algorithm.

In summary, compared with state-of-the-art entropy-based feature selection methods, our DNEAR algorithm can fully reflect the decision-making ability of attributes, avoid the loss of useful genes caused by discretization, tackle the uncertainty and noise in gene expression data, and efficiently improve the classification performance.

5.5. Comparison of classification performance of dimensionality reduction algorithms

To further verify the classification performance of our proposed FSDNE algorithm, the goal of this section of our experiments is to test fifteen dimensionality reduction methods in terms of the number of selected genes and the classification accuracy on selected genes. These fourteen state-of-the-art reduction methods include the following: (1) the Fisher score algorithm [47], (2) the Lasso algorithm [50], (3) the neighborhood rough set-based reduction algorithm (NRS) [10], (4) the gene selection algorithm based on Fisher linear discriminant and neighborhood rough set (FLD-NRS) [31], (5) the gene selection algorithm based on locally linear embedding and neighborhood rough set (LLE-NRS) [32], (6) the Relief algorithm [49] combined with the NRS algorithm [10] (Relief + NRS), (7) the fuzzy backward feature elimination algorithm (FBFE) [2], (8) the binary differential evolution algorithm (BDE) [1], (9) the sequential forward selection algorithm (SFS) [21], (10) the sparse group Lasso algorithm (SGL) [29], (11) the adaptive sparse group Lasso algorithm based on conditional mutual information (ASGL-CMI) [20], (12) the Spearman's rank correlation coefficient algorithm (SC^2) [45], (13) the mutual information maximization algorithm (MIM) [21], and (14) the distributed ranking filter approach removing the features with information gain zero from the ranking and correlation-based feature selection algorithm (DRF0-CFS) [20]. The SVM classifier in WEKA is employed for simulation experiments. Following the experimental techniques in [1–3,5,20,21,29,45], the four representative gene expression datasets (Colon, Leukemia, Lung, and Prostate) are selected from Table 2, and then the number and classification accuracy of selected genes are shown in Tables 15 and 16, respectively, where the symbol - denotes that no results were obtained for Leukemia using the SGL and ASGL-CMI algorithms.

According to the results of Tables 15 and 16, the differences among the fifteen dimensionality reduction methods can be clearly identified. As shown in Table 15, the performances of the NRS, FLD-NRS, BDE, SFS, SC^2, MIM and FSDNE algorithms are very similar for the four gene expression datasets, and the seven reduction algorithms perform markedly better than the other eight algorithms, with the Fisher score exhibiting the worst performance. As shown in Table 16, although the classification accuracy of the FSDNE algorithm for the Colon dataset is similar to that of the Fisher score, Lasso, FLD-NRS, LLE-NRS, ASGL-CMI and DRF0-CFS algorithms, our algorithm selects the lowest number of genes and achieves higher classification accuracy than the remaining eight algorithms. For the Leukemia dataset, the classification accuracy of the FSDNE algorithm is similar to that of the Fisher score, Lasso and SFS algorithms and higher than that of the remaining nine algo-

Table 15
Number of selected genes by fifteen dimensionality reduction algorithms.

Algorithms	Colon	Leukemia	Lung	Prostate	Average
Fisher score	200	200	200	200	200
Lasso	5	23	8	63	24.75
NRS	4	5	3	4	4
FLD-NRS	6	6	3	4	4.75
LLE-NRS	16	22	16	19	18.25
Relief + NRS	9	17	23	16	16.25
FBFE	35	30	80	50	48.75
BDE	3	7	3	3	4
SFS	19	7	3	3	8
SGL	55	–	43	34	44
ASGL-CMI	33	–	32	29	31.33
SC^2	4	5	3	5	4.25
MIM	19	7	3	3	8
DRF0-CFS	10	13	17	113	38.25
FSDNE	3	9	8	4	6

Table 16
Classification accuracy of selected genes by fifteen dimensionality reduction algorithms.

Algorithms	Colon	Leukemia	Lung	Prostate	Average
Fisher score	0.838	0.934	0.975	0.86	0.902
Lasso	0.887	0.986	0.995	0.961	0.957
NRS	0.611	0.645	0.641	0.647	0.636
FLD-NRS	0.88	0.828	0.889	0.8	0.849
LLE-NRS	0.84	0.868	0.907	0.711	0.832
Relief + NRS	0.564	0.563	0.919	0.642	0.672
FBFE	0.833	0.912	0.852	0.832	0.857
BDE	0.75	0.824	0.98	0.941	0.874
SFS	0.521	0.969	0.833	0.84	0.791
SGL	0.826	–	0.827	0.834	0.829
ASGL-CMI	0.851	–	0.841	0.858	0.85
SC^2	0.805	0.852	0.806	0.795	0.815
MIM	0.653	0.727	0.795	0.865	0.76
DRF0-CFS	0.9	0.912	0.987	0.853	0.913
FSDNE	0.838	0.928	0.988	0.883	0.909

rithms; however, this algorithm selects substantially fewer genes than the Fisher score and Lasso algorithms. For the Lung dataset with the FBFE algorithm in Table 15, the number of selected genes is as high as 80, although for some algorithms, such as NRS, FLD-NRS, BDE, SFS, SC^2, MIM, and FSDNE, the number of selected genes is fewer than 10. For the classification accuracy in Table 16, the FSDNE algorithm has the highest classification accuracy on the Lung dataset when compared the other thirteen algorithms, excluding the Lasso algorithm. For the Prostate dataset, although the Lasso and BDE algorithms exhibit slightly better classification accuracy than the FSDNE algorithm, indicating that the performance of the FSDNE algorithm is somewhat different for different datasets, the performances of the FSDNE and BDE algorithms are almost the same in terms of the number of selected genes. Compared with the FBFE and BDE algorithms, the FSDNE algorithm shows slightly improved classification accuracy for the Colon, Leukemia and Lung datasets. Table 15 shows that the NRS and BDE algorithms select the smallest number of genes. Additionally, since some genes with classification information are deleted, the NRS algorithm achieves low classification accuracy. Then, the three extended NRS algorithms (FLD-NRS, LLE-NRS and Relief + NRS) overcome this drawback, increase the number of selected genes and improve the classification accuracy. Compared with the four related NRS algorithms, the FSDNE algorithm achieves higher classification accuracy for the Leukemia, Lung and Prostate datasets. Regarding the number of selected genes in Table 15 and the average classification accuracy in Table 16, the FSDNE algorithm performs the best. Hence, our proposed approach can reduce the dimensionality of gene expression datasets and outperforms other related high-dimensional feature selection algorithms.

In the following subsection, to demonstrate the classification performance with more algorithms on more datasets, the five representative gene expression datasets (Colon, DLBCL, Leukemia, Lung, and SRBCT) are selected from Table 2, and six state-of-the-art dimensionality reduction methods are selected for comparison with our FSDNE algorithm, including the following: (1) the conditional mutual information maximization algorithm (CMIM) [11], (2) CMIM and the adaptive genetic algorithm (CMIMAGA) [28], (3) the double input symmetrical relevance algorithm (DISR) [44], (4) the joint mutual information algorithm (JMI) [6], (5) the maximum relevance of minimal redundancy algorithm (mRMR) [23], and (6) the Relief-F algorithm [28]. Following the experimental techniques used by Shukla et al. [28], the SVM classifier in WEKA is employed

Table 17
Number of selected genes of the seven dimensionality reduction algorithms.

Datasets	CMIM	CMIMAGA	DISR	JMI	mRMR	Relief-F	FSDNE
Colon	5	10	5	10	5	20	3
DLBCL	30	20	20	10	10	20	11
Leukemia	10	10	40	20	10	50	9
Lung	10	20	30	20	10	50	8
SRBCT	30	10	30	30	10	50	9
Average	17	14	25	18	9	38	8

Table 18
Classification accuracy of the selected genes of the seven dimensionality reduction algorithms.

Datasets	CMIM	CMIMAGA	DISR	JMI	mRMR	Relief-F	FSDNE
Colon	0.574	0.832	0.601	0.631	0.684	0.658	**0.838**
DLBCL	0.882	**0.948**	0.702	0.87	0.914	0.92	0.927
Leukemia	0.774	0.88	0.786	0.71	0.782	0.864	**0.929**
Lung	0.774	0.856	0.769	0.788	0.81	0.817	**0.988**
SRBCT	0.619	0.924	0.776	0.649	0.92	0.709	**0.936**
Average	0.725	0.888	0.727	0.73	0.822	0.794	**0.924**

for simulation experiments. The number and classification accuracy of the selected genes are shown in Tables 17 and 18, respectively.

As shown in Table 17, the performances of the CMIM, CMIMAGA and JMI algorithms are very close on the five gene expression datasets and the number of genes selected is approximately 10. The number of genes selected by the DISR algorithm is approximately 20 and obviously less than that of the Relief-F algorithm. Compared with the other five algorithms in Table 17, the mRMR algorithm and the FSDNE algorithm proposed in this paper select the fewest genes. Specifically, the number of genes selected by mRMR is slightly greater than that selected by FSDNE on the five datasets. Table 18 shows that the FSDNE algorithm achieves the highest average classification accuracy. For the Colon dataset, the classification accuracy of the FSDNE algorithm is 83.8%, which is similar to that of the CMIMAGA algorithm, 20.7%-26.4% higher than the CMIM, DISR and JMI algorithms, and 15.4%-18% higher than the mRMR and Relief-F algorithms. For the DLBCL dataset, the classification accuracy of the FSDNE algorithm is similar to that of the Relief-F algorithm and higher than those of the CMIM, DISR, JMI and mRMR algorithms. Although the classification accuracy of FSDNE is slightly lower than that of CMIMAGA, the number of selected genes is substantially lower than that of CMIMAGA. For the Leukemia, Lung and SRBCT datasets, the FSDNE algorithm obtains the highest classification accuracies, which are far higher than those of the CMIM, DISR and JMI algorithms. From Tables 17 and 18, the FSDNE algorithm selects the smallest number of genes, with only 11 genes at most for the DLBCL dataset and even fewer genes for the other datasets; however, its classification accuracy is the highest in most cases. Therefore, our proposed approach can screen out genes with strong classification ability to significantly improve the classification accuracy and also select a very small number of genes, which verifies the effectiveness of our proposed algorithm.

In general, according to these experimental results compared with the above twenty state-of-the-art dimensionality reduction algorithms, our FSDNE algorithm with nonmonotonicity can achieve higher classification accuracy and select fewer genes for all gene expression datasets. Therefore, our proposed method is an efficient dimensionality reduction technique for high-dimensional, large-scale datasets.

5.6. Statistical analysis

The following part of our experiments further demonstrates the statistical significance of the classification results. To express the statistical importance of the above experimental results, the Friedman test [12] and the Bonferroni-Dunn test [9] are employed in this paper. The Friedman statistic is described as

$$\chi_F^2 = \frac{12N}{t(t+1)}\left(\sum_{i=1}^{t} R_i^2 - \frac{t(t+1)^2}{4}\right), \tag{23}$$

$$F_F = \frac{(N-1)\chi_F^2}{N(t-1)-\chi_F^2}, \tag{24}$$

where t is the number of algorithms, N is the number of datasets, R_i is the average ranking of algorithm i over all datasets, and F_F follows a Fisher distribution with $t-1$ and $(t-1)(N-1)$ degrees of freedom.

If the null hypothesis is rejected under the Friedman test statistic, a post hoc test such as the Bonferroni-Dunn test can be used to further explore which algorithms are different in statistical terms [36]. Based on the test results, if the average

Table 19
Ranking of the five feature selection algorithms under the KNN classifier.

Datasets	MEAR	EGGS	DNEAR	EGGS-FS	FSDNE
Brain_Tumor2	5	2.5	4	2.5	1
Colon	2	4	5	3	1
DLBCL	4	3	5	2	1
Leukemia	2	4	5	3	1
Leukemia1	3	4	5	2	1
Lung	3	4	5	2	1
Prostate	5	3	4	2	1
Prostate1	5	4	3	2	1
SRBCT	4	3	5	2	1
9_Tumors	5	4	3	2	1
Average	3.8	3.55	4.4	2.25	1

Table 20
Ranking of the five feature selection algorithms under the C4.5 classifier.

Datasets	MEAR	EGGS	DNEAR	EGGS-FS	FSDNE
Brain_Tumor2	5	2	3	4	1
Colon	1	4	5	3	2
DLBCL	4	2	5	3	1
Leukemia	1	4	5	3	2
Leukemia1	2	4	5	1	3
Lung	3	2	5	4	1
Prostate	4	5	3	2	1
Prostate1	5	3	4	1	2
SRBCT	5	3	4	2	1
9_Tumors	5	4	3	2	1
Average	3.5	3.3	4.2	2.5	1.5

Table 21
Ranking of the five feature selection algorithms under the SVM classifier.

Datasets	MEAR	EGGS	DNEAR	EGGS-FS	FSDNE
Brain_Tumor2	5	2	3.5	3.5	1
Colon	1	5	3	4	2
DLBCL	4	2	5	3	1
Leukemia	2	3	4	5	1
Leukemia1	2	4	5	1	3
Lung	4	3	5	1	2
Prostate	4	5	3	2	1
Prostate1	5	4	3	2	1
SRBCT	5	2	4	3	1
9_Tumors	5	1	4	3	2
Average	3.7	3.1	3.95	2.75	1.5

level of distance exceeds the critical distance, the performance of the two algorithms will be significantly different. The critical distance [35] is denoted as

$$CD_\alpha = q_\alpha \sqrt{\frac{t(t+1)}{6N}}, \tag{25}$$

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

In the first part of this experiment, we conducted two Friedman tests to investigate whether the classification performances of the five reduction algorithms in Subsection 5.4 are significantly different. In general, the classification accuracies of these five reduction algorithms are not very different in most cases. On the basis of the classification results in Tables 5–14, Tables 19,20–21 show the rankings of the five feature selection algorithms under the KNN, C4.5 and SVM classifiers.

From Tables 19,20–21, the Bonferroni-Dunn tests for the KNN, C4.5 and SVM classifiers indicate that the proposed FSDNE algorithm is statistically superior to the other four algorithms overall. Then, the values of the two evaluation measures (Friedman statistics χ_F^2 and Iman-Davenport test F_F) under the KNN, C4.5 and SVM classifiers can be obtained as follows: $\chi_F^2 = 26.5$ and $F_F = 29.86$ under KNN, $\chi_F^2 = 6.73$ and $F_F = 17.12$ under C4.5, and $\chi_F^2 = 5.32$ and $F_F = 14.86$ under SVM. When the significance level $\alpha = 0.1$, the critical value of $F(4, 36)$ is 2.11. The Friedman test rejects the null hypothesis under the

Table 22
Ranking of the fifteen dimensionality reduction algorithms with SVM.

Algorithms	Colon	Leukemia	Lung	Prostate	Average
Fisher score	6.5	3	5	5	4.9
Lasso	2	1	1	1	1.3
NRS	13	12	15	14	13.5
FLD-NRS	3	9	8	11	7.8
LLE-NRS	5	7	7	13	8
Relief + NRS	14	13	6	15	12
FBFE	8	5.5	9	10	8.1
BDE	11	10	4	2	6.8
SFS	15	2	11	8	9
SGL	9	14.5	12	9	11.1
ASGL-CMI	4	14.5	10	6	8.6
SC^2	10	8	13	12	10.8
MIM	12	11	14	4	10.3
DRF0-CFS	1	5.5	3	7	4.1
FSDNE	6.5	4	2	3	3.9

Table 23
Ranking of the seven dimensionality reduction algorithms with SVM.

Datasets	CMIM	CMIMAGA	DISR	JMI	mRMR	Relief-F	FSDNE
Colon	7	2	6	5	3	4	1
DLBCL	5	1	7	6	4	3	2
Leukemia	6	2	4	7	5	3	1
Lung	6	2	7	5	4	3	1
SRBCT	7	2	4	6	3	5	1
Average	6.2	1.8	5.6	5.8	3.8	3.6	1.2

three classifiers when all algorithms have the same performance. Therefore, two Bonferroni-Dunn tests need to be performed. The critical value $q_{0.1} = 2.241$ can be found in [36], and $CD = 1.58$ can be easily calculated.

In what follows, we perform a statistical analysis of the other selected dimensionality reduction algorithms. Table 22 shows the rankings of the fifteen dimensionality reduction algorithms in Table 16 under the SVM classifier in Subsection 5.5 because the SVM classifier was the only classifier used in [1–3,5,20,21,29,45]. Under the SVM classifier, $\chi_F^2 = 32.31$ and $F_F = 4.09$. Similarly, Table 23 shows the rankings of the seven dimensionality reduction algorithms in Table 18 under the SVM classifier, which was the only classifier used in [6,11,28,44]. The two indexes $\chi_F^2 = 25.2$ and $F_F = 21$ can be obtained under the SVM classifier.

Tables 22 and 23 show that the proposed FSDNE algorithm is statistically superior to the other algorithms overall. Tables 22 and 23 show that under SVM classifier, the values of F_F are 4.09 and 21, respectively. When the significance level $\alpha = 0.1$, the critical value of $F(14, 42)$ is 1.66 and $F(6, 24)$ is 2.04. The critical value $q_{0.1} = 2.241$, and it can be easily calculated from Eq. (25) that the values of CD are 28.35 and 15.31.

6. Conclusion

Reducing the redundant or irrelevant genes of gene expression datasets can effectively decrease the cost of cancer classification. In this paper, a feature selection algorithm using neighborhood entropy-based uncertainty measures is proposed to improve the classification performance of gene expression data. The neighborhood entropy-based uncertainty measures are first investigated to measure the uncertainty and exclude the noise in gene expression datasets. Then, the neighborhood credibility degree and the neighborhood coverage degree are introduced into decision neighborhood entropy and mutual information, which are proven to be nonmonotonic, to fully describe the decision-making ability of attributes in neighborhood decision systems. By using the Fisher score algorithm before reducing the dimensionality of gene expression datasets, a heuristic reduction algorithm for cancer classification is designed to efficiently decrease the computational complexity and improve the classification performance of gene expression data. The experimental results show that our proposed algorithm can find a small, effective subset of genes and obtain high classification accuracy in gene expression datasets. However, our proposed method cannot optimally balance on the size of the selected gene subset and classification accuracy in all high-dimensional gene expression datasets. Hence, further research into this problem will be helpful for the development of gene expression data classification. In future work, to make our algorithms more suitable for application areas, such as big data mining, pattern recognition and bioinformatics for biomarker discovery, and to further improve the classification performance and computational efficiency of the proposed algorithms for cancer classification, new search strategies and more efficient uncertainty measures based on neighborhood rough sets should be explored.

Declaration of the availability of data and materials

The datasets used during the study are available at the Kent Ridge Biomedical Dataset Repository and WEKA Collections of Datasets. (Last accessed: December 28, 2018) https://leo.ugr.es/elvira/DBCRepository/
https://www.cs.waikato.ac.nz/ml/weka/datasets.html

Declaration of Competing Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Nos. 61772176, 61402153, 61370169, 61672332), the Project Funded by China Postdoctoral Science Foundation (No. 2016M602247), the Plan of Scientific Innovation Talent of Henan Province (No. 184100510003), the Key Scientific and Technological Project of Henan Province (No. 182102210362), the Young Scholar Program of Henan Province (No. 2017GGJS041), and the Natural Science Foundation of Henan Province (No. 182300410130).

References

[1] J. Apolloni, G. Leguizamon, E. Alba, Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments, Appl. Soft Comput. 38 (2016) 922–932.

[2] R. Aziz, C.K. Verma, N. Srivastava, A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data, Genomics Data. 8 (2016) 4–15.

[3] V. Bolon-Canedo, N. Sanchez-Marono, A. Alonso-Betanzos, Distributed feature selection: an application to microarray data classification, Appl Soft Comput. 30 (2015) 136–150.

[4] H.M. Chen, T.R. Li, X. Fan, C. Luo, Feature selection for imbalanced data based on neighborhood rough sets, Inform. Sci. 483 (2019) 1–20.

[5] 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.

[6] A. Das, S. Das, Feature weighting and selection with a pareto-optimal trade-off between relevancy and redundancy, Pattern Recogn. Lett. 88 (2017) 12–19.

[7] 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.

[8] 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.

[9] O.J. Dunn, Multiple comparisons among means, J. Am. Stat. Assoc. 56 (293) (1961) 52–64.

[10] 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.

[11] F. Fleuret, Binary feature selection with conditional mutual information, J. Mach. Learn. Res. 5 (2004) 1531–1555.

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

[13] C.D. Greenman, Haploinsufficient gene selection in cancer, Science 337 (6090) (2012) 47–48.

[14] M.A.H. Valizade, R. Sheikhpour, M.A. Sarram, E. Sheikhpour, H. Sharifi, A combined fisher and laplacian score for feature selection in QSAR based drug design using compounds with known and unknown activities, J. Comput. Aid. Mol. Des. 32 (2) (2018) 375–384.

[15] L. Hu, W.F. Gao, K. Zhao, P. Zhang, F. Wang, Feature selection considering two types of feature relevancy and feature interdependency, Expert Syst. Appl. 93 (2018) 423–434.

[16] Q.H. Hu, W. Pan, S. An, P.J. Ma, J.M. Wei, An efficient gene selection technique for cancer recognition based on neighborhood mutual information, Int. J. Mach. Learn. Cyb. 1 (1–4) (2010) 63–74.

[17] Q.H. Hu, D.R. Yu, J.F. Liu, C.X. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inform. Sci. 178 (18) (2008) 3577–3594.

[18] X.J. Huang, L. Zhang, B.J. Wang, F.Z. Li, Z. Zhang, Feature clustering based support vector machine recursive feature elimination for gene selection, Appl. Intell. 48 (3) (2018) 594–607.

[19] H.X. Li, X.Z. Zhou, J.B. Zhao, D. Liu, Non-monotonic attribute reduction in decision-theoretic rough sets, Fund Inform. 126 (4) (2013) 415–432.

[20] J.T. Li, W.P. Dong, D.Y. Meng, Grouped gene selection of cancer via adaptive sparse group lasso based on conditional mutual information, IEEE/ACM T Comput. Bi. 15 (6) (2018) 2028–2038.

[21] H.J. Lu, J.Y. Chen, K. Yan, Q. Jin, Y. Xue, Z.G. Gao, A hybrid feature selection algorithm for gene expression data classification, Neurocomputing 256 (2017) 56–62.

[22] Z. Pawlak, Rough sets, Int. J. Comput. Inform. Sci. 11 (5) (1982) 341–356.

[23] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE T. Pattern Anal. 27 (8) (2005) 1226–1238.

[24] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science. 290 (5500) (2000) 2323–2326.

[25] C.E. Shannon, The mathematical theory of communication, AT&T Tech. J. 27 (3) (1948) 379–423.

[26] M.W. Shao, K.W. Li, Attribute reduction in generalized one-sided formal contexts, Inform. Sci. 378 (1) (2017) 317–327.

[27] A.K. Shukla, P. Singh, M. Vardhan, A hybrid gene selection method for microarray recognition, Biocybern. Biomed. Eng. 38 (4) (2018) 975–991.

[28] A.K. Shukla, P. Singh, M. Vardhan, A two-stage gene selection method for biomarker discovery from microarray data for cancer classification, Chemometr. Intell. Lab. 183 (2018) 47–58.

[29] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A sparsegroup lasso, J. Comput. Graph Stat. 22 (2) (2013) 231–245.

[30] 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.

[31] 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.

[32] L. Sun, J.C. Xu, W. Wang, Y. Yin, Locally linear embedding and neighborhood rough set-based gene selection for gene expression data classification, Genet. Mol. Res. 15 (3) (2016). Gmr.15038990

[33] 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.

[34] 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.

[35] L. Sun, X.Y. Zhang, J.C. Xu, S.G. Zhang, Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification, Entropy. 21 (2) (2019). Article ID 155

[36] 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). Article ID 138

[37] 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). Article ID 4182148

[38] 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.

[39] S.Q. Sun, Q.K. Peng, X.K. Zhang, Global feature selection from microarray data using lagrange multipliers, Knowl.-Based Syst. 110 (2016) 267–274.

[40] S. Tsumoto, Accuracy and coverage in rough set rule induction, International Conference on Rough Sets and Current Trends in Computing (2002) 373–380.

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

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

[43] G.Y. Wang, Rough set theory and knowledge acquisition, Xi'an Jiaotong University Press, 2001.

[44] J. Wang, J.M. Wei, Z.L. Yang, S.Q. Wang, Feature selection by maximizing independent classification information, IEEE Trans. Knowl. Data Eng. 29 (4) (2017) 828–841.

[45] J.C. Xu, H.Y. Mu, Y. Wang, F.Z. Huang, Feature genes selection using supervised locally linear embedding and correlation coefficient for microarray classification, Comput. Math. Method M. 2018 (2018). Article ID 5490513

[46] F.F. Xu, D.Q. Miao, L. Wei, Fuzzy-rough attribute reduction via mutual information with an application to cancer classification, Comput. Math. Appl. 57 (6) (2009) 1010–1017.

[47] J. Yang, Y.L. Liu, C.S. Feng, G.Q. Zhu, Applying the fisher score to identify alzheimer's disease-related genes, Genet. Mol. Res. 15 (2) (2016). Gmr.15028798

[48] C.C. Ye, J.L. Pan, Q. Jin, An improved SSO algorithm for cyber-enabled tumor risk analysis based on gene selection, Future Gener. Comp. Sy. 92 (2019) 407–418.

[49] W. Zhang, J.J. Chen, Relief feature selection and parameter optimization for support vector machine based on mixed kernel function, Int. J. Perform Eng. 14 (2) (2018) 280–289.

[50] S.F. Zheng, W.X. Liu, An experimental comparison of gene selection by lasso and dantzig selector for cancer classification, Comput. Biol. Med. 41 (11) (2011) 1033–1040.