Applied Intelligence
https://doi.org/10.1007/s10489-018-1320-1

Joint neighborhood entropy-based gene selection method with fisher score for tumor classification

Lin Sun[1,2] · Xiao-Yu Zhang[1] · Yu-Hua Qian[3] · Jiu-Cheng Xu[1] · Shi-Guang Zhang[1] · Yun Tian[4]

Abstract
Tumor classification is one of the most vital technologies for cancer diagnosis. Due to the high dimensionality, gene selection (finding a small, closely related gene set to accurately classify tumor) is an important step for improving gene expression data classification performance. Traditional rough set model as a classical attribute reduction method deals with discrete data only. As for the gene expression data containing real-value or noisy data, they are usually employed by a discrete preprocessing, which may result in poor classification accuracy. In this paper, a novel neighborhood rough sets and entropy measure-based gene selection with Fisher score for tumor classification is proposed, which has the ability of dealing with real-value data whilst maintaining the original gene classification information. First, the Fisher score method is employed to eliminate irrelevant genes to significantly reduce computation complexity. Next, some neighborhood entropy-based uncertainty measures are investigated for handling the uncertainty and noisy of gene expression data. Moreover, some of their properties are derived and the relationships among these measures are established. Finally, a joint neighborhood entropy-based gene selection algorithm with the Fisher score is presented to improve the classification performance of gene expression data. The experimental results under an instance and several public gene expression data sets prove that the proposed method is very effective for selecting the most relevant genes with high classification accuracy.

Keywords Rough sets · Neighborhood rough sets · Gene selection · Neighborhood entropy · Tumor classification

1 Introduction

Tumor is a chronic disease that is caused mainly by irregularities in genes, and it is important to identify such oncogenes that cause cancer [10, 17, 35]. Biological data like gene expressions, protein sequences, RNA-sequences, pathway analysis, Pan- cancer analysis and structural biomarkers could aid in cancer diagnosis, classification and prognosis [39]. The rapid development of DNA microarray technology helps the researchers to analyze thousands of gene expression data in an efficient manner. Gene expression profiling by microarray method has appeared as a capable technique for classification and certain diagnosis, therapy, and cancer prognosis [10, 29, 35]. Currently, the number of gene expression samples remains in the hundreds, compared to tens of thousands of genes involved [17]. Although gene expres-

✉ Jiu-Cheng Xu
jiuchxu@gmail.com

Lin Sun
linsunok@gmail.com

Xiao-Yu Zhang
zhangxy0903@126.com

Yu-Hua Qian
jinchengqyh@126.com

Shi-Guang Zhang
zhangshiguang@htu.edu.cn

Yun Tian
tianyun@bnu.edu.cn

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

[2] Post-doctoral Mobile Station of Biology, College of Life Science, Henan Normal University, Xinxiang 453007, China

[3] Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China

[4] College of Information Science and Technology, Beijing Normal University, Beijing 100875, China

Published online: 03 November 2018

sion data sets are high-dimensional, only a few of these dimensions are beneficial to tumor classification, and a large number of irrelevant and redundant genes would deteriorate the performance of the classifiers [10]. Gene selection as an important data preprocessing technique for cancer classification is one of the most challenging issues in the field of microarray data analysis [33]. It aims to select the most representative gene subset with a high resolution by eliminating redundant and unimportant genes [18]. Its application in the study of cancer has proved to be successful in revealing the pathological mechanism, with the potential of altering clinical practice through individualized cancer care and contributing to the battle against cancer [10]. This study has significant influence on bioinformatics, tumor or cancer classification, disease diagnosis, and so on [4].

Considering whether the evaluation criterion involves classification models, feature selection methods can be divided into three categories [13, 19]: the filter, the wrapper and the embedded methods. Based on the intrinsic properties of the dataset, the filter methods have been directed to discriminate or filter out features by estimating their relevance scores to state a cut-off schema [27]. Lyu et al. [24] studied a filter feature selection method based on the maximal information coefficient for biomedical data mining. Wang et al. [43] proposed a hybrid feature selection algorithm which combines minimum redundancy maximum relevance with imperialist competition algorithm for cancer classification from microarray gene expression data. The wrapper methods use a classifier to find the most discriminant feature subset by minimizing an error prediction function. However, the wrapper methods not only exhibit sensitivity to the classifier and unstable performance, but also tend to consume a lot of runtime [13]. So, they are not extensively used in microarray tasks, and few works in the field have employed these methods [3]. The embedded methods integrate gene selection in the training process to reduce the total time required for reclassifying different subsets [6, 26]. Our gene selection method is based on the filter approach, in which a heuristic search algorithm is used to find an optimal gene subset with neighborhood rough sets for the gene expression data.

Rough set theory is a useful tool to deal with vague, uncertain and incomplete information [4, 26, 47]. However, traditional rough set based on equivalence relation could only handle data with categorical attributes, and it could be useless for continuous numerical data [14, 42, 46]. To overcome this drawback, Hu et al. [16] established a neighborhood rough set model to process both numerical and categorical data sets via neighborhood relation. This model will not break the neighborhood structure and order structure of data sets in real spaces. Moreover, in order to evaluate the uncertainty of discrete sample spaces, information entropy as a significant uncertainty measure tool can characterize the distinguished information of feature subsets [40, 44]. Then, based on neighborhood systems, information entropy and its variants have been established and adapted for feature selections. For instance, Chen et al. [5] studied a gene selection method for tumor classification using neighborhood rough sets and entropy measures, which has the ability of dealing with real-value data while maintaining the original gene classification information. Lin et al. [23] researched multi-label feature selection based on neighborhood mutual information. Zhao et al. [48] combined adaptive neighborhood granularity with multi-level confidence to process cost-sensitive feature selection. Dong et al. [7] proposed a hybrid genetic algorithm with granular information for feature selection. Garcia-Torres et al. [9] utilized feature grouping to increase the effectiveness of search and proposed a variable neighborhood search approach. According to the idea of neighborhood, a gene expression data set is granulated by neighborhood parameters and some entropy measures based on neighborhood rough sets are developed in this paper. Then, based on neighborhood entropy measures, a gene selection method in the frame of neighborhood rough sets is presented to address the uncertainty and noisy of gene expression data.

To avoid the high computational complexity of attribute reduction algorithms and obtain genes which facilitate classification and prediction, the Fisher score method is used to carry out preliminary dimensionality reduction [12]. As a feature relevance criterion, the Fisher score is a kind of supervised learning with many advantages, such as few calculations, high accuracy, and strong operability, which can reduce computation complexity [20]. Hancer et al. [11] presented a filter criterion inspired by mutual information, ReliefF and Fisher score. However, the Fisher score often selects redundant features, which in turn affects the classification result [12]. In this paper, the Fisher score method is combined with entropy measure to reduce initial dimensions of gene expression data, and improve the classification performance of high-dimensional gene expression data sets. The appropriate genes are selected to form a candidate gene subset.

The remainder of this paper is organized as follows: Section 2 reviews some basic concepts. In Section 3, neighborhood entropy based-uncertainty measures of neighborhood rough sets are presented, and a gene selection method with Fisher score for gene expression data is developed. Section 4 analyzes the experiments conducted on several public gene expression data sets. Finally, the conclusions are summarized in Section 5.

2 Previous knowledge

In this section, we briefly review some basic concepts about rough sets, information entropy measures and neighborhood rough sets. The following notations have been given in [16, 25, 28, 30, 40].

2.1 Rough sets

Given a discrete-value data, which can be formalized as a decision system $DS = (U, C, D)$. $U = \{x_1, x_2, \cdots, x_n\}$ is a nonempty finite set named universe, $C = \{a_1, a_2, \cdots, a_m\}$ is a set of all conditional attributes, and D is the set of decision attributes. $V = \cup_{a\in\{C\cup D\}} V_a$, where V_a is a value set of attribute a. $f: U \times \{C \cup D\} \rightarrow V$ is a map function, and $f(a, x)$ represents the value of x on attribute $a \in C \cup D$.

Given a decision system $DS = (U, C, D)$, for any two samples $x, y \in U$ and attribute subset $B \subseteq C$, the equivalence relation is described as

$$IND(B) = \{(x, y)|\forall a \in B, f(a, x) = f(a, y)\}. \quad (1)$$

For any sample $x \in U$, $[x]_B = \{y|\ y \in U, (x, y) \in IND(B)\}$ is an equivalence class of x, and $U/IND(B)$ is a partition that is compose of the equivalence classes.

The equivalence class defines two classical sets, named upper and lower approximation sets, as the elementary units. In a decision system $DS = (U, C, D)$ with $B \subseteq C$ and $X \subseteq U$, the lower approximation set and the upper approximation set of X with respect to B can be described as, respectively

$$\underline{B}(X) = \{x|[x]_B \subseteq X, x \in U\}, \quad (2)$$

$$\bar{B}(X) = \{x|[x]_B \cap X \neq \emptyset, x \in U\}. \quad (3)$$

Thus, for a sample set X related to $B \subseteq C$, its lower approximation set is the set of all elements which can be classified with certainty as equivalence class of X related to B, and its upper approximation set is a set of all elements which are possibly classified to equivalence class of X related to B.

2.2 Information entropy measures

Given a decision system $DS = (U, C, D)$, for an attribute subset $B \subseteq C$, $U/B = \{X_1, X_2, \cdots, X_n\}$, then the information entropy of B is described as

$$H(B) = -\sum_{i=1}^{n} p(X_i) \log p(X_i), \quad (4)$$

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 X_i.

Given a decision system $DS = (U, C, D)$, for any two attribute subsets $B_1, B_2 \subseteq C$, $U/B_1 = \{X_1, X_2, \cdots, X_n\}$, and $U/B_2 = \{Y_1, Y_2, \cdots, Y_m\}$, then the joint entropy of B_1 and B_2 is expressed as

$$H(B_1 \cup B_2) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(X_i \cap Y_j) \log p(X_i \cap Y_j), \quad (5)$$

where $p\left(X_i \cap Y_j\right) = \frac{|X_i \cap Y_j|}{|U|}$.

2.3 Neighborhood rough sets

The continuous data set must be discretized when processing continuous data with the classical rough set, but the original property of the gene expression data will change after discretization, and some useful information will be lost. The neighborhood rough set is a method to solve the problem that classical rough sets cannot handle continuous numerical data [16, 41]. All data in gene selection are numerical. Moreover, in the gene expression data sets, the measured gene expression levels and pharmaceutical tests are presented by continuous-valued data at different magnitudes [21]. By utilizing neighborhood rough sets, the discretization of continuous data can be avoided. Note that the gene expression data set can be described by a neighborhood decision system, where an object is corresponding to a sample, a conditional attribute contains a gene, and a decision attribute corresponds to a subclass of cancer.

Given a real-value gene expression data set, which is formalized as a neighborhood decision system $NS = (U, C, D, V, f, \Delta, \delta)$. $U = \{x_1, x_2, \cdots, x_n\}$ is a sample set, $C = \{a_1, a_2, \cdots, a_m\}$ is a set of all genes, and D is a decision set. $V = \cup_{a\in\{C\cup D\}} V_a$, where V_a is a value set of gene a. $f: U \times \{C \cup D\} \rightarrow V$ is a map function. $\Delta \rightarrow [0, \infty)$ is a distance function, and δ is a neighborhood parameter, where $0 \leq \delta \leq 1$. In the following, $NS = (U, C, D, V, f, \Delta, \delta)$ is simply noted as $NS = (U, C, D, \delta)$.

For any three samples $x, y, z \in U$ on a gene subset B, the distance function $\Delta_B(x, y)$ satisfies the following three conditions:

(1) $\Delta_B(x, y) \geq 0$,
(2) $\Delta_B(x, y) = \Delta_B(y, x)$,
(3) $\Delta_B(x, y) + \Delta_B(y, z) \geq \Delta_B(x, z)$.

The first item states that the distance function of two samples is non-negative, and that the equality holds if and only if the two samples are the same. The second item indicates that the distance function is irrespective of the order of the samples, i.e., it satisfies symmetry. The last item demonstrates that the distance function satisfy triangular inequality.

It is well known that Manhattan, Euclidean and Chebychev distance functions are three classical metrics. Since the Euclidean distance function effectively reflects the basic information of the unknown data [5], it is introduced into this paper, and expressed as

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

Given a neighborhood decision system $NS = (U, C, D, \delta)$ and a distance function $\Delta \rightarrow [0, \infty)$, for any gene subset $B \subseteq C$ and neighborhood parameter $\delta \in [0, 1]$, the similarity relation resulting by B is described as

$$NR_\delta(B) = \{(x, y) \in U \times U | \Delta_B(x, y) \leq \delta\}. \tag{7}$$

For any $x \in U$, the neighborhood class of x with respect to B is described as

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

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

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

$$\bar{B}_\delta(X) = \{x_i | n_B^\delta(x_i) \cap X \neq \emptyset, x_i \in U\}. \tag{10}$$

Thus, for a set X with respect to any gene subset $B \subseteq C$, its neighborhood lower approximation set is the set of all elements which can be with certainty classified as neighborhood class of X with respect to B, and its neighborhood upper approximation set of a set X with respect to B is the set of all elements which are possibly classified to neighborhood class of X with respect to B.

3 Gene selection using neighborhood entropy-based uncertainty measures for gene expression data

3.1 Neighborhood entropy-based uncertainty measures

The information entropy is not suitable for measuring the neighborhood class in the numeric data sets. To solve this issue, the concept of neighborhood is combined with information theory measures. Based on the neighborhood relation, neighborhood entropy-based uncertainty measures are investigated for continuous numerical gene expression data.

Given a neighborhood decision system $NS = (U, C, D, \delta)$ with $B \subseteq C$, and $n_B^\delta(x_i)$ is a neighborhood class of $x_i \in U$, then Hu et al. [15] described the neighborhood entropy of x_i as follows

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

Given a neighborhood decision system $NS = (U, C, D, \delta)$ with $B \subseteq C$, then Hu et al. [15] and Chen et al. [5] computed the average neighborhood entropy of the sample set as follows

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

Proporty 1 Given a neighborhood decision system $NS = (U, C, D, \delta)$ and $n_C^\delta(x_i) \subseteq U$ for $\forall x_i \in U$, then it follows that $\frac{1}{|U|} \leq \frac{|n_C^\delta(x_i)|}{|U|} \leq 1$. Thus, $0 \leq H_\delta(C) \leq \log |U|$.

Proof Since $n_C^\delta(x_i) \subseteq U$ for $\forall x_i \in U$, then it follows that $\frac{1}{|U|} \leq \frac{|n_C^\delta(x_i)|}{|U|} \leq 1$, and one has $0 \leq H_\delta(C) \leq \log |U|$. □

Proposition 1 *Given a neighborhood decision system $NS = (U, C, D, \delta)$ with $B \subseteq C$, and two different neighborhood parameters δ and δ'. For $\forall x_i \in U$, and any gene subset $B \subseteq C$, if $\delta \leq \delta'$, then $n_B^\delta(x_i) \subseteq n_B^{\delta'}(x_i)$ and $H_\delta(B) \geq H_{\delta'}(B)$.*

Proof Let $\delta \leq \delta'$, then one has that $\Delta_B(x, y) \leq \delta \leq \delta'$. For $\forall x_i \in U$, it follows from (8) that $n_B^\delta(x_i) = \{y | x_i, y \in U, \Delta_B(x_i, y) \leq \delta\}$ and $n_B^{\delta'}(x_i) = \{y | x_i, y \in U, \Delta_B(x_i, y) \leq \delta'\}$. It can be easily obtained that $n_B^\delta(x_i) \subseteq n_B^{\delta'}(x_i)$ and $\left|n_B^\delta(x_i)\right| \leq \left|n_B^{\delta'}(x_i)\right|$. Hence, by (12), $H_\delta(B) \geq H_{\delta'}(B)$ holds. □

Proposition 2 *Given a neighborhood decision system $NS = (U, C, D, \delta)$, for $\forall x_i \in U$, if $B_1 \subseteq B_2 \subseteq C$, then $n_{B_1}^\delta(x_i) \supseteq n_{B_2}^\delta(x_i)$ and $H_\delta(B_2) \geq H_\delta(B_1)$.*

Proof Let $B_1 \subseteq B_2 \subseteq C$, and similar to the proof of Proposition 1 in [5], one has that$n_{B_1}^\delta(x_i) \supseteq n_{B_2}^\delta(x_i)$. Since $n_{B_1}^\delta(x_i) \supseteq n_{B_2}^\delta(x_i)$, then $\left|n_{B_1}^\delta(x_i)\right| \geq \left|n_{B_2}^\delta(x_i)\right|$. Hence, by (12), $H_\delta(B_2) \geq H_\delta(B_1)$. □

Proposition 3 *Given a neighborhood decision system $NS = (U, C, D, \delta)$ with $B \subseteq C$, if $\delta = 0$, $H_\delta(B) = H(B)$.*

Proof For $B \subseteq C$, if $\delta = 0$, then $n_B^\delta(x)$ can be viewed as an equivalent class $[x]_B$ in rough sets, i.e., $n_B^\delta(x) = [x]_B$. For $\forall x_i \in U$, $\frac{|n_B^\delta(x_i)|}{|U|}$ is the probability distribution of all the possible combinations of genes. Thus, it follows from (12) that

$$-\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|n_B^\delta(x_i)|}{|U|} = -\frac{|[x_i]_B|}{|U|}\sum_i \log\frac{|[x_i]_B|}{|U|} = H(B),$$

i.e., $H_\delta(B) = H(B)$. □

Definition 1 Given a neighborhood decision system $NS = (U, C, D, \delta)$ with $B \subseteq C$, $n_B^\delta(x_i)$ 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 $IND(D)$, then a joint neighborhood entropy of B and D is defined as

$$H_\delta(D, B) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{\left|n_B^\delta(x_i)\cap[x_i]_D\right|^2}{|U|\,|[x_i]_D|}\right). \tag{13}$$

Proposition 4 *Given a neighborhood decision system NS* $= (U, C, D, \delta)$ *with* $B_1 \subseteq B_2 \subseteq C$, *then* $H_\delta(D, B_1) \le H_\delta(D, B_2)$, *where* $H_\delta(D, B_1) = H_\delta(D, B_2)$ *holds if and only if* $n_{B_1}^\delta(x_i) = n_{B_2}^\delta(x_i)$ *for* $\forall x_i \in U$.

Proof Let $B_1 \subseteq B_2 \subseteq C$, according to Proposition 2, it follows that $n_{B_1}^\delta(x_i) \supseteq n_{B_2}^\delta(x_i)$. Then, $U \supseteq n_{B_1}^\delta(x_i)\cap[x_i]_D \supseteq n_{B_2}^\delta(x_i)\cap[x_i]_D \supseteq \{x_i\}$. It is easily obtained that $|U| \ge \left|n_{B_1}^\delta(x_i)\cap[x_i]_D\right| \ge \left|n_{B_2}^\delta(x_i)\cap[x_i]_D\right| \ge |\{x_i\}|$.

So, one has that $\frac{|U|}{|[x_i]_D|} \ge \frac{\left|n_{B_1}^\delta(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|} \ge \frac{\left|n_{B_2}^\delta(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|} \ge \frac{1}{|U||[x_i]_D|}$ and $\log\left(\frac{|U|}{|[x_i]_D|}\right) \ge \log\left(\frac{|n_{B_i}^\delta(x_i)\cap[x_i]_D|^2}{|U||[x_i]_D|}\right) \ge \log\left(\frac{\left|n_{B_2}^\delta(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|}\right) \ge \log\left(\frac{1}{|U||[x_i]_D|}\right)$. Then, it is clear that $-\sum_{i=1}^{m}\log\left(\frac{|U|}{|[x_i]_D|}\right) \le -1 \le -\sum_{i=1}^{m}\log\left(\frac{\left|n_{B_2}^\delta(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|}\right) \le -\sum_{i=1}^{m}\log\left(\frac{1}{|U||[x_i]_D|}\right)$.

It can be concluded from Definition 1 that $H_\delta(D, B_1) \le H_\delta(D, B_2)$. When $n_{B_1}^\delta(x_i) = n_{B_2}^\delta(x_i)$, then $\frac{\left|n_{B_1}^\delta(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|} = \frac{\left|n_{B_2}^\delta(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|}$ holds. Hence, $H_\delta(D, B_1) \le H_\delta(D, B_2)$. □

Proposition 5 *Given a neighborhood decision system NS* $= (U, C, D, \delta)$ *with* $B \subseteq C$, *then* $H_\delta(D, B) \ge H_\delta(B)$.

Proof It follows immediately from Definition 1 and the average neighborhood entropy in [5, 15] that

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

Since there exists $n_B^\delta(x_i)\cap[x_i]_D \subseteq n_B^\delta(x_i)$ and $n_B^\delta(x_i)\cap[x_i]_D \subseteq [x_i]_D$, it is easily obtained that $\left|n_B^\delta(x_i)\cap[x_i]_D\right| \le \left|n_B^\delta(x_i)\right|$ and $\left|n_B^\delta(x_i)\cap[x_i]_D\right| \le |[x_i]_D|$. Then, $\frac{|n_B^\delta(x_i)\cap[x_i]_D|}{|[x_i]_D|} \le 1$ and $\frac{|n_B^\delta(x_i)\cap[x_i]_D|}{|n_B^\delta(x_i)|} \le 1$. Therefore, $H_\delta(D, B) - H_\delta(B) \ge 0$, i.e., $H_\delta(D, B) \ge H_\delta(B)$. □

Definition 2 Given a neighborhood decision system $NS = (U, C, D, \delta)$ with $B \subseteq C$ and $\forall a \in B$, if $H_\delta(D, B) \le H_\delta(D, B - \{a\})$, then a is redundant in B with respect to D, otherwise, the a is indispensable in B with respect to D. B is dependent if any gene in B is indispensable in B with respect to D. B is called a reduction of C with respect to D if it satisfies the following two conditions:

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

Obviously, a reduction of C with respect to D is the minimal gene subset to retain the joint neighborhood entropy of C and D.

Definition 3 Given a neighborhood decision system $NS = (U, C, D, \delta)$ with $B \subseteq C$ and $\forall a \in C - B$, then the significance measure of a in B with respect to D is defined as

$$Sig(a, B, D) = H_\delta(D, B \cup \{a\}) - H_\delta(D, B). \tag{14}$$

When $B = \emptyset$, $Sig(a, B, D) = H_\delta(D, \{a\})$. From Definition 3, the significance of gene a is the increment of the distinguishing information after adding a into B. The larger value of $Sig(a, B, D)$ is, the more importance of gene a for B with respect to D is.

3.2 Gene selection algorithm with Fisher score for gene expression data

In this subsection, Fisher score method as a key pretreatment method can significantly reduce the dimension of the gene, and it is briefly described as follows.

Given a neighborhood decision system $NG = (U, C, D, \delta)$ about gene expression data. Its corresponding matrix is $X \in R^{m\times n}$, where m denotes the number of genes, and n denotes the number of samples. Then, the Fisher score is computed by

$$f(Z) = \frac{tr(A_b)}{tr(A_w)}, \tag{15}$$

where $tr()$ denotes the trace of a matrix, A_b is the between-class scatter matrix, and A_w is the within-class scatter matrix.

To address the time-consuming issue of traditional combination optimization methods, a heuristic strategy is usually employed to calculate a score for each gene independently by using some criteria. Then, the Fisher score of the j-th gene is calculated by

$$f(i) = \frac{\sum_{i=1}^{C} n_i(\mu_i^j - \mu^j)^2}{\sum_{i=1}^{C} n_i(\sigma_i^j)^2}, \tag{16}$$

where 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.

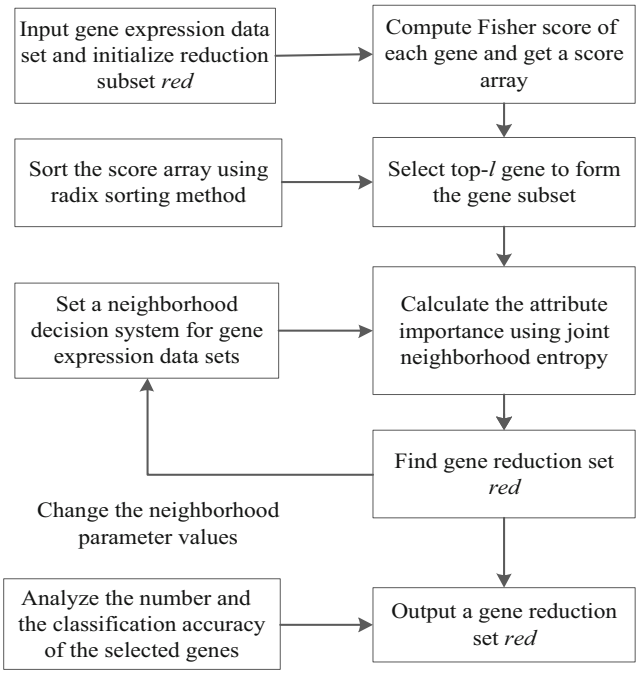

Fig. 1 Flow char of gene selection with the Fisher score

To facilitate the understanding of the gene selection algorithm with Fisher score for gene expression data, the process of gene selection with the Fisher score for gene expression data is shown in Fig. 1.

To support efficient gene selection, a gene selection algorithm with the Fisher score based on joint neighborhood entropy (GSFSJNE) is constructed and described as Algorithm 1.

Algorithm 1

Input: The original gene expression data matrix $X \in R^{m\times n}$, and the number l of the expected genes;
Output: A gene reduction set *red*;

1. Initialize $red = \emptyset$;
2. **for** each gene in the gene space **do**
3. Evaluate the corresponding Fisher Score by (16), and record the score in a score array;
4. **end for**
5. Sort the score by descending order with the radix sorting algorithm in [33];
6. Select the top-l genes with a high score, and place their gene indexes into the set T;
7. Obtain the dimension-reduction gene subset S in terms of T;
8. Let $NS = (U, S \cup D, \delta)$;
9. **while** $Sig(S, red, D) = 0$ **do**
10. Let $Temp = red$, and $g = 0$;
11. **for** $\forall a \in (S - red)$ **do**
12. Compute $Sig(a, red, D)$;
13. **if** $Sig(a, red, D) > g$ **then**
14. Let $Temp = red \cup \{a\}$, and $g = Sig(a, red, D)$;
15. **end if**
16. **end for**
17. Let $red = Temp$;
18. **end while**
19. **Return** A gene reduction set *red*.

3.3 Complexity analysis

For the GSFSJNE algorithm, the time complexity of feature selection from a neighborhood decision system is polynomial. When giving m genes and n samples, the time complexity of step 2 through step 4 is $O(m)$. At step 5, the time complexity of the sorting method by using the radix sorting algorithm in [31] is $O(m)$, which is a linear complexity. Thus, the time complexity of step 2 through step 7 is $O(m)$. Joint neighborhood entropy-based gene selection algorithm is given from step 8 to step 19, where the neighborhood classes induced by the conditional attributes and the joint neighborhood entropies need to be frequently calculated in neighborhood decision systems.

The above computational process largely affects time complexity of selecting genes. Suppose that the number of selected genes at step 7 is l, then, based on the type of bucket employed [22], the time complexity of achieving neighborhood classes is $O(ln)$, and the complexity of calculating joint neighborhood entropy is $O(l)$. Since $O(l) < O(ln)$, the computational complexity of calculation of joint neighborhood entropy is $O(ln)$, and then there are two loops in step 8 through step 19, so the worst time complexity of GSFSJNE is $O(l^3n)$. Suppose that the number of selected genes at step 19 is l_R, for the calculation of the neighborhood classes, we only consider the candidate genes without involving the entire gene set. Then, the time complexity of achieving neighborhood classes is $O(l_Rn)$. The times of the outer loop are l_R, and the times of the inner loop are $l - l_R$. Then, the time complexity from step 8 to step 19 is $O(nl_R^2l - nl_R^3)$. It is well known that $l_R \ll l$ in most cases, and the time complexity is close to $O(ln)$. Then, the total time complexity of GSFSJNE is $O(m+ ln)$. Thus, the time complexity of GSFSJNE is $O(m)$ approximately. Therefore, in our proposed algorithm, the time complexity is effectively decreased through dimensionality reduction with the Fisher score method. Furthermore, its space complexity is $O(mn)$.

3.4 An illustrative example

In order to demonstrate the performance of GSFSJNE algorithm, a neighborhood decision system $NS = (U, C, D, \delta)$ is as an example [42], where $U = \{x_1, x_2, x_3, x_4\}$, $C = \{a, b, c\}$, $D = \{d\}$, and $\delta = 0.3$, as shown in Table 1.

For Table 1, an example of gene selection using proposed algorithm without Fisher score is given. Then, the neighborhood class of each gene in Table 1 is calculated using the Euclidean distance function.

For a gene subset $\{a\}$, one has that $\Delta_{\{a\}}(x_1, x_2) = 0.09$, $\Delta_{\{a\}}(x_1, x_3) = 0.19$, $\Delta_{\{a\}}(x_1, x_4) = 0.49$, $\Delta_{\{a\}}(x_2, x_3) = 0.1$, $\Delta_{\{a\}}(x_2, x_4) = 0.4$, and $\Delta_{\{a\}}(x_3, x_4) = 0.3$.

Then the neighborhood classes of $\forall x_i \in U$ can be computed as follows:

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

Table 1 A gene expression data set

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

According to the decision $D = \{d\}$ in Table 1, a partition can be obtained, i.e., $U/\{d\} = \{D_1, D_2\} = \{\{x_1, x_2\}, \{x_3, x_4\}\}$. Then, the joint neighborhood entropy of D with respect to $\{a\}$ can be calculated by

$$\begin{aligned} H_{\delta}(D, \{a\}) &= -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\left(\frac{\left|n_a^{\delta}(x_i) \cap [x_i]_D\right|^2}{|U|\,|[x_i]_D|}\right) \\ &= -\frac{1}{4}\left(\log\left(\frac{2^2}{4\times 2}\right) + \log\left(\frac{2^2}{4\times 2}\right)\right. \\ &\quad \left. + \log\left(\frac{2^2}{4\times 2}\right) + \log\left(\frac{2^2}{4\times 2}\right)\right) \\ &= 0.301. \end{aligned}$$

Similarly, the other joint neighborhood entropies can be calculated respectively as follows:

$H_{\delta}(D, \{b\}) = 0.301$, $H_{\delta}(D, \{c\}) = 0.602$, $H_{\delta}(D, \{a, b\}) = 0.903$, $H_{\delta}(D, \{a, c\}) = 0.903$, $H_{\delta}(D, \{b, c\}) = 0.753$, and $H_{\delta}(D, \{a, b, c\}) = 0.903$.

The significance measures are calculated as follows:

$Sig(a, \emptyset, D) = H_{\delta}(D, \{a\}) = 0.301$, $Sig(b, \emptyset, D) = H_{\delta}(D, \{b\}) = 0.301$, and $Sig(c, \emptyset, D) = H_{\delta}(D, \{c\}) = 0.602$.

From the above calculated results, it can be observed that $Sig(c, \emptyset, D) > Sig(a, \emptyset, D) = Sig(b, \emptyset, D)$. Since the gene c has the maximum with significance measure. Then, the gene c should be added to the candidate, i.e., $red = \{c\}$. By computing, we have that $Sig(C, red, D) = H_{\delta}(D, C) - H_{\delta}(D, \{c\}) = 0.903-0.602 = 0.301 \neq 0$, so the searching need to be continued. It is easily computed that $H_{\delta}(D, \{a, c\}) > H_{\delta}(D, \{b, c\})$. Hence, the gene a is added to the candidate reduction set, i.e., $red = \{a, c\}$, because the joint neighborhood entropy of D with respect to $\{a, c\}$ is greater.

In the next step, the candidate reduction set red is checked if it satisfies the termination criterion. By computing, we know that $H_{\delta}(D, red) = H_{\delta}(D, C)$, i.e., $Sig(C, red, D) = 0$, which satisfies the termination criterion. Thus, a selected gene subset $\{a, c\}$ is achieved.

4 Experimental results and analysis

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

(1) Colon cancer [36] is the development of cancer from the Colon or rectum. Most Colon cancers are due to old age and lifestyle factors, with only a small number of

Table 2 Description of five gene expression data sets

No.	Data set	Genes	Samples	Classies
1	Colon	2000	62	2 (40/22)
2	SRBCT	2308	63	4 (23/8/12/20)
3	DLBCL	5469	77	2 (58/19)
4	Brain_Tumor1	5920	90	5 (60/10/10/4/6)
5	Leukemia	7129	72	2 (47/25)

cases due to underlying genetic disorders. The Colon gene expression data set contains 2000 genes and 62 samples, including 40 patient samples and 22 healthy samples.

(2) Small-round-blue-cell tumor (SRBCT) [37] is any one of a group of malignant neoplasms that have a characteristic appearance under the microscope, i.e. consisting of small round cells that stain blue on routine H&E stained sections. These tumors are seen more often in children than in adults. They typically represent undifferentiated cells, which contain 2308 genes and 63 samples with four subtypes, including 23 Ewing Sarcoma, 8 Burkitt Lymphoma, 12 Neuroblastoma and 20 Rhabd omyosarcoma.

(3) Diffuse large B-cell lymphoma (DLBCL) [37] 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 100000 people per year in the USA and the UK. It contains 5469 genes and 77 samples, including 58 patient samples and 19 healthy samples.

(4) Brain tumor [37] occurs when abnormal cells form within the brain. All types of brain tumors may produce symptoms that vary depending on the part of the brain involved. The Brain_Tumor1 gene expression data set contains 5920 genes and 90 samples with five subtypes, including 60 Medulloblastoma, 10 Malignant glioma, 10 AT/RT, 4 Normal cerebellum and 6 PNET.

(5) Leukemia [37] 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. Diagnosis is typically made by blood tests or bone marrow biopsy. The Leukemia gene expression data set contains 7129 genes and 72 samples, including 47 patient samples and 25 healthy samples.

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, and 4 GB memory. All the simulation experiments were implemented in Matlab R2014a, and the three classifiers (KNN, C4.5 and LibSVM) were selected to verify the classification accuracy in WEKA software, whose parameter k in KNN was set to 5, and the linear kernel functions were selected in LibSVM.

In the first part of our experiments, the Fisher score method is used to achieve preliminary dimensionality reduction. For each gene expression data set, the Fisher score value of each gene is calculated and sorted, and then l genes are selected to constitute a candidate gene subset. The classification accuracy with different numbers of genes was verified in WEKA. Figure 2 illustrates the changing trend of the classification accuracy versus the number of genes on five gene expression data sets.

It can be observed from Fig. 2 that the classification accuracies of each gene expression data set with different values of l are very similar in most situations. Furthermore, it is well known that the cardinality and the classification accuracy of the candidate gene subset are two important indices for evaluating the performance of gene selection algorithms. Then, the appropriate values of l need to be selected from Fig. 2. Hence, the l value is set to 200.

In what follows, the second part of our experiments concerns the effect of different neighborhood parameter values. The value of the neighborhood parameters decides the granularity of data manipulation, which affects both the cardinality of the dataset and the classification accuracy of selected gene subset. Then, reduction rate is introduced to evaluate the gene redundancy gene selection algorithms. In this paper, a new reduction rate for gene expression data sets in neighborhood decision systems is defined as

$$Rate_\delta = \frac{\max(|R_\delta|) - |R_\delta|}{\max(|R_\delta|)}, \tag{17}$$

where $|R_\delta|$ represents the number of selected genes generated by a given δ.

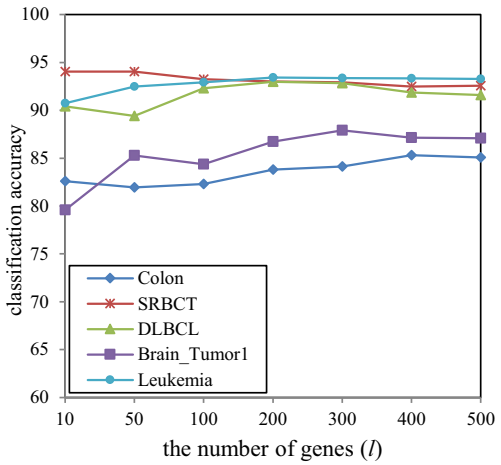

Fig. 2 The classification accuracy versus the number of genes on five gene expression data sets with GSFSJNE algorithm

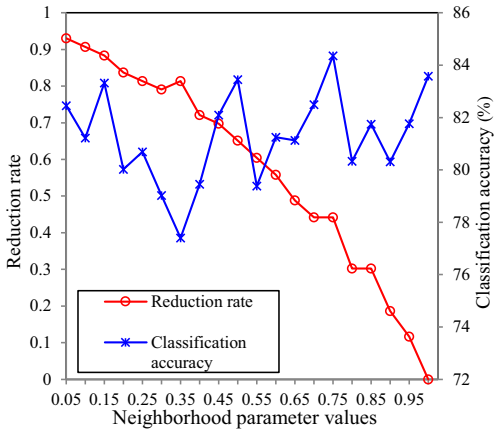

Fig. 3 The reduction rate and the classification accuracy of Colon data sets with different neighborhood parameter values

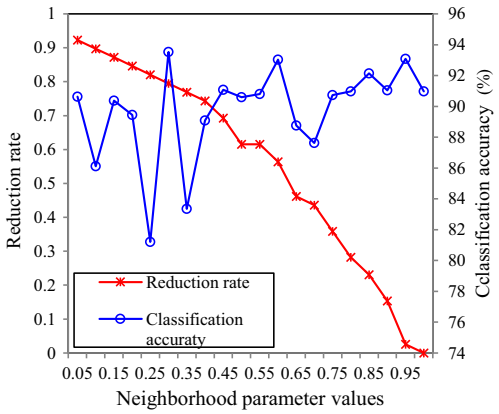

Fig. 5 The reduction rate and the classification accuracy of DLBCL data sets with different neighborhood parameter values

The reduction rate represents the degree of redundancy. A higher reduction rate indicates that the algorithm has stronger reduction ability for gene expression data sets. That is, the higher reduction rate is, the lower redundancy is.

The reduction rate and classification accuracy of a gene subset for different neighborhood parameter values is discussed to obtain a suitable neighborhood parameter value and a better genes subset. The classification accuracy of the gene expression datasets given in Table 2 was obtained by using the GSFSJNE algorithm with different neighborhood parameters. The results are shown in Figs. 3, 4, 5, 6 and 7, where the horizontal coordinate denotes the neighborhood parameter values, the neighborhood parameter $\delta \in [0.05,1]$ at intervals of 0.05, and the left and right vertical coordinates describe the reduction rate and the classification accuracy, respectively.

Figures 3–7 show that the classification accuracy of the selected gene subsets with GSFSJNE algorithm is increasing, and the reduction rate is decreasing with

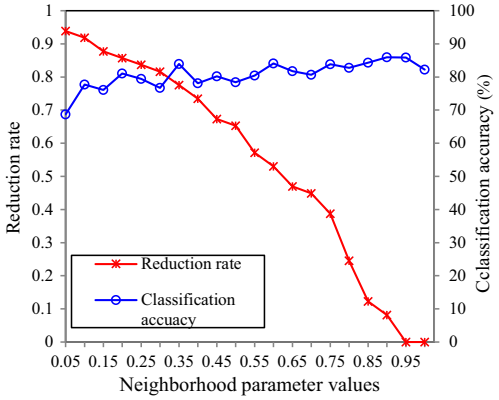

Fig. 6 The reduction rate and the classification accuracy of Brain_Tumor1 data sets with different neighborhood parameter values

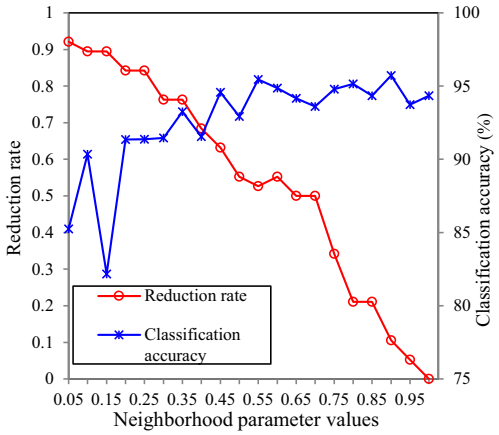

Fig. 4 The reduction rate and the classification accuracy of SRBCT data sets with different neighborhood parameter values

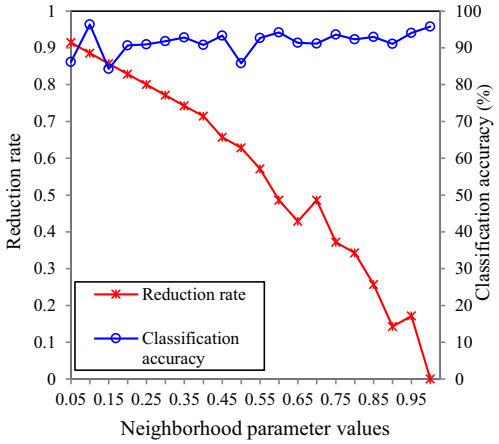

Fig. 7 The reduction rate and the classification accuracy of Leukemia data sets with different neighborhood parameter values

Table 3 The selected gene subsets of five gene expression data sets with the GSFSJNE algorithm

Data sets	Selected gene subsets
Colon	{765, 590, 384, 266, 1541}
SRBCT	{1606, 758, 246, 1954, 165, 823, 1066, 1980, 1327}
DLBCL	{3127, 584, 2639, 1570, 4588, 4094, 2949, 3304}
Brain_Tumor1	{467, 2295, 4151, 5175, 5413, 1879, 2095, 3401, 5713, 820, 5633}
Leukemia	{461, 1962, 5552, 2131}

the neighborhood parameters changing from 0.05 to 1 in the majority of cases. It indicates that the granule is thinner, and the roughness of granule is smaller as the neighborhood parameter is smaller, and the reduction rate increases as the granule roughness decreases. Then, the optimal neighborhood parameters can be selected for each gene expression data sets. Figure 3 illustrates the classification accuracy of the Colon data set for different neighborhood parameters. The reduction rate decreases as the neighborhood parameters increase, and the classification accuracy of the selected genes first reaches a maximal value when the neighborhood parameter is 0.15. Thus, the neighborhood parameter δ of the Colon data set can be set to 0.15. Similarly, for SRBCT and Brain_Tumor1 data sets with different neighborhood parameters in Figs. 4 and 6, the neighborhood parameter δ can be set to 0.35. Figure 5 shows the classification accuracy of the DLBCL data set with different neighborhood parameters. The reduction rate decreases as the neighborhood parameters increase, and the classification accuracy of the selected genes first reaches a maximum when the neighborhood parameter is 0.3. Thus, the neighborhood parameter δ of the DLBCL data set can be set to 0.3. Likewise, from Fig. 7, the neighborhood parameter δ of the Leukemia data set can be set to 0.1.

In the third part of our experiments, by using the GSFSJNE algorithm and the above set neighborhood parameters, the results of gene selection on the five gene expression data sets was obtained, which are shown in Table 3.

The fourth part of our experiments is to test the performance of our proposed algorithm in terms of the classification accuracy for the selected genes. In the current experiments, the classification performance of GSFSJNE is compared with those of the other four related gene selection methods on the five gene expression data sets. The methods used in the comparison include: (1) the original data processing method (ODP), (1) mutual entropy-based gene selection algorithm (MEGS) [45], (2) entropy gain-based gene selection algorithm (EGGS) [5], and (3) joint neighborhood entropy-based gene selection algorithm (JNEGS). The KNN, C4.5 and SVM classifiers are employed in WEKA. The objective of these further experiments is to show the classification performance of the proposed approach to gene selection. Tables 4, 5, 6, 7 and 8 show the experimental results of five different methods, in which the minimal cardinality (in terms of gene subset length) of each algorithm is given, and the bold numbers indicate the optimal values, namely the highest accuracy.

From Tables 4–8, the classification accuracies of the five gene expression data sets by five methods are verified respectively in the KNN, C4.5 and SVM classifiers. The classification effectiveness of the classifiers is affected by the data sets. For example, the SVM classifier exhibits good classification performance for the original data set, and the classification accuracies of the GSFSJNE algorithm in the KNN classifier are higher. In addition, for the uniform treatment of all the algorithms, the average classification accuracies of the three classifiers are given. By comparing the average classification accuracy, it can be observed that the GSFSJNE algorithm significantly improves the classification accuracy in most cases. On the five different data sets, the cardinalities of gene subset with MEGS algorithm are minimal, however, the classification

Table 4 Classification accuracy for the selected Colon genes (δ = 0.15)

Methods	Genes	KNN	C4.5	SVM	Average
ODP	2000	77.64	81.95	81.14	80.24
MEGS	5	77.00	82.24	**84.86**	81.37
EGGS	5	54.00	59.40	64.26	59.22
JNEGS	5	55.45	64.12	60.55	60.04
GSFSJNE	5	**84.10**	81.57	84.26	**83.31**

Table 5 Classification accuracy for the selected SRBCT genes (δ = 0.35)

Methods	Genes	KNN	C4.5	SVM	Average
ODP	2308	80.79	77.98	**98.40**	85.72
MEGS	4	53.69	42.33	53.88	49.97
EGGS	8	50.29	42.36	53.48	48.71
JNEGS	8	50.29	42.36	53.48	48.71
GSFSJNE	9	**97.43**	**87.83**	94.50	**93.25**

Table 6 Classification accuracy for the selected DLBCL genes (δ =0.3)

Methods	Genes	KNN	C4.5	SVM	Average
ODP	5469	89.59	80.89	**96.54**	89.01
MEGS	2	76.52	77.77	77.73	77.34
EGGS	20	75.21	83.09	86.23	81.51
JNEGS	7	75.68	71.89	66.32	71.30
GSFSJNE	8	**95.59**	**90.98**	94.04	**93.54**

Table 8 Classification accuracy for the selected Leukemia genes (δ =0.1)

Methods	Genes	KNN	C4.5	SVM	Average
ODP	7129	84.16	81.43	**97.27**	87.62
MEGS	3	**92.77**	**93.39**	92.04	**92.73**
EGGS	3	58.71	64.70	53.59	59.00
JNEGS	3	58.71	64.70	53.59	59.00
GSFSJNE	4	91.98	77.64	88.89	86.17

accuracies of SRBCT, DLBCL and Brain_Tumor1 data sets are lower than that of the original data set, which indicates that the useful information will lost after discretizing, and the part of the effective genes will be deleted. The GSFSJNE algorithm not only overcomes this defect, but also selects genes that are related to the decision and more conducive to classification performance. A comparison of the results of the average classification accuracy of the GSFSJNE algorithm with the EGGS algorithm illustrates that the GSFSJNE algorithm has higher average classification accuracy for the five gene expression data sets, which indicates that the proposed GSFSJNE algorithm can select genes that contribute the most to the classification. Thus, the GSFSJNE algorithm is more suitable for processing the large-scale gene expression data sets.

In order to further verify the classification performance of our proposed method, the objective of the last part of our experiments is to evaluate the ten methods in terms of the number and the classification accuracy of selected genes. Our GSFSJNE algorithm is compared with the nine related recent dimensionality reduction methods, which include: (1) ODP, (1) the Fisher score algorithm [12], (2) the Lasso [49], (3) the neighborhood rough sets (NRS) [8], (4) the gene selection algorithm based on the fisher linear discriminant and the neighborhood rough set (FLD-NRS) [34], (5) the gene selection algorithm based on the locally linear embedding and neighborhood rough set (LLE-NRS) [32], (6) the RelieF [38] combined with the NRS [8] (RelieF+NRS), (7) the fuzzy backward feature elimination (FBFE) [2], and (8) the binary differential evolution (BDE) [1]. Lib-SVM classifier in WEKA tool is used to simulation experiment. The number and classification accuracies of selected genes are shown in Tables 9 and 10, respectively.

Table 7 Classification accuracy for the selected Brain_Tumor1 genes (δ =0.35)

Methods	Genes	KNN	C4.5	SVM	Average
ODP	5920	78.33	75.11	**86.00**	79.81
MEGS	2	68.33	68.89	69.11	68.78
EGGS	8	66.67	59.44	66.56	64.22
JNEGS	9	71.11	65.11	63.22	66.48
GSFSJNE	11	**88.33**	**82.22**	81.22	**83.92**

According to the results of the number of selected genes and classification accuracy in Tables 9 and 10, the difference among ten methods can be clearly found. For Colon data, though FLD-NRS acquires higher classification accuracy 88%, the number of genes of GSFSJNE is less than that of FLD-NRS, LLE-NRS and RelieF+NRS. For Leukemia dataset, FBFE has higher classification accuracy than GSFSJNE but failure in number of selected genes. For Lung data set with FBFE method, the number selected genes is as high as 80, but for some methods such as NRS, FLD-NRS, BDE, and GSFSJNE, the number selected genes is less than 10. As for the classification accuracy, it can be observed that the GSFSJNE algorithm obtains higher classification accuracy than NRS. The classification accuracy produced by our method is 99.8% for Lung data, and the result is higher than that of other algorithms. The NRS algorithm selects less number of genes, where the number is only 5 at most for Leukemia data and the number of rest of the data is lower. However, some genes with classification information also are deleted, which leads to low classification accuracy of the NRS The three extended NRS methods (FLD-NRS, LLE-NRS and RelieF+NRS) overcome this drawback to increase the number of selected genes and improve the classification accuracy Compared with these methods, the GSFSJNE algorithm holds higher classification accuracy for Leukemia, Lung and Prostate data sets Compared with FBFE and BDE, the GSFSJNE algorithm has slightly improved classification accuracy for Colon and Lung datasets Therefore, our proposed approach obviously reduces the dimension of gene expression data sets, outperforms the other related methods of gene selection, and can provide an efficient dimensionality reduction technique for high-dimensional large-scale data sets.

During the above experiments, the rough ordering of these nine methods with respect to time complexity is as follows: $O(\text{GSFSJNE}) = O(\text{Fisher score}) < O(\text{FBFE}) <$

Table 9 The number of selected genes with ten related recent dimensionality reduction methods

Method	ODP	Fisher score	Lasso	NRS	FLD-NRS	LLE-NRS	RelieF+NRS	FBFE	BDE	GSFSJNE
Colon	2000	200	5	4	6	16	9	35	3	5
Leukemia	7129	200	23	5	6	22	17	30	7	4
Lung [37]	12533	200	8	3	3	16	23	80	3	6
Prostate [37]	12600	200	63	4	4	19	16	50	3	4

O(BDE) $<$ O(FLD-NRS) $<$ O(RelieF+NRS) $<$ O(NRS) $<$ O(LLE-NRS) $<$ O(Lasso), where O(A) denotes the time complexity of A algorithm. The time complexity of Lasso algorithm is $O(nm^3)$ [49]. For high-dimensional gene expression data, the Lasso algorithm has the highest time complexity. For the NRS algorithm and its extension forms, the time complexity of is $O(m^2n + m^2n\log n)$ for the LLE-NRS algorithm [32], and $O(m^2n\log n)$ for NRS algorithm [8]. Therefore, the time complexity of LLE-NRS is higher than that of NRS. Moreover, Relief+NRS has a time complexity of $O(mn+mn\log n)$ [8, 38], and FLD-NRS has the time complexity of $O(mn\log n)$ [34]. The time complexities of three extended NRS methods (FLD-NRS, LLE-NRS and RelieF+NRS) are lower than that of NRS. Since population initialization is main process of the BDE algorithm, the time complexity of BDE is close to$O(nm)$ [1]. For FBFE, the time is mainly spent on evaluating the relevance of the features using entropy, and its time complexity is not more than $O(nm)$ [2]. The time complexity of GSFSJNE and Fisher score [12] are approximately equal to $O(m)$ and lower than that of other seven algorithms. Through the above analyses of time complexity, Lasso algorithm costs significantly more time. Since in most cases $n \ll m$, the time complexity and the number of selected genes of Lasso are much larger than those of GSFSJNE, though Lasso has better classification accuracy. It is immediately apparent from these results that our proposed algorithm can effectively reduce dimension of gene expression data, increase the classification accuracy, and speed up the classification process with less time complexity.

5 Conclusion

Identifying tumor-related genes is helpful for earlier tumor diagnosis and drug design. Gene selection is one of the important steps in tumor classification. In this paper, a gene selection method using neighborhood entropy-based uncertainty measures is proposed to improve the classification accuracy of gene expression data. The Fisher score method first preliminarily reduces the dimension of gene expression data sets, which eliminates irrelevant genes, and significantly decreases the complexity of subsequent computations. Then, the neighborhood entropy-based uncertainty measures are investigated to measure the uncertainties of real-value gene expression data sets and exclude the redundancy genes. Furthermore, the properties and the relationships among these measures are derived. Thus, a heuristic algorithm is constructed to improve computational efficiency of selecting genes in neighborhood decision systems. The experimental results show that our proposed algorithm can obtain a small, effective gene subset with higher classification accuracy.

Acknowledgments This work was partially supported by the National Natural Science Foundation of China (Grants 61772176, 61402153, 61672332, 61370169, and 61472042), the China Postdoctoral Science Foundation (Grant 2016M602247), the Plan for Scientific Innovation Talent of Henan Province (Grant 184100510003), the Key Project of Science and Technology Department of Henan Province (Grants 182102210362), the Young Scholar Program of Henan Province (Grant 2017GGJS041), the Key Scientific and Technological Project of Xinxiang City (Grant CXGG17002), and the Ph.D. Research Foundation of Henan Normal University (Grants qd15132, qd15129).

Table 10 The classification accuracy of selected genes with ten related recent dimensionality reduction methods

Method	ODP	Fisher score	Lasso	NRS	FLD-NRS	LLE-NRS	RelieF+NRS	FBFE	BDE	GSFSJNE
Colon	0.645	0.838	0.887	0.611	0.88	0.84	0.564	0.833	0.75	0.843
Leukemia	0.653	0.934	0.986	0.645	0.828	0.868	0.563	0.912	0.824	0.889
Lung	0.916	0.975	0.995	0.641	0.889	0.907	0.919	0.852	0.98	0.998
Prostate	0.566	0.86	0.961	0.647	0.8	0.711	0.642	0.832	0.941	0.85
Avearage	0.695	0.902	0.957	0.636	0.849	0.832	0.672	0.857	0.874	0.892

References

1. Apolloni J, Leguizamon G, Alba E (2016) Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Appl Soft Comput 38:922–932
2. Aziz R, Verma CK, Srivastava N (2016) A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data. Genom Data 8:4–15
3. Bhola A, Singh S (2018) Gene selection using high dimensional gene expression data: An appraisal. Curr Bioinform 13(2):225–233
4. Chen HM, Li TR, Cai Y, Luo C, Fujitac H (2016) Parallel attribute reduction in dominance-based neighborhood rough set. Inf Sci 373:351–368
5. Chen YM, Zhang ZJ, Zheng JZ, Ma Y, Xue Y (2017) Gene selection for tumor classification using neighborhood rough sets and entropy measures. J Biomed Inform 67:59–68
6. Das AK, Sengupta S, Bhattacharyya S (2018) A group incremental feature selection for classification using rough set theory based genetic algorithm. Appl Soft Comput 65:400–411
7. Dong H, Li T, Ding R, Sun J (2018) A novel hybrid genetic algorithm with granular information for feature selection and optimization. Appl Soft Comput 65:33–46
8. Fan XD, Zhao WD, Wang CZ, Huang Y (2018) Attribute reduction based on max-decision neighborhood rough set model. Knowl-Based Syst 151:16–23
9. Garcia-Torres M, Gomez-Vela F, Melian-Batista B, Moreno-Vega JM (2016) High-dimensional feature selection via feature grouping: A variable neighborhood search approach. Inf Sci 326:102–118
10. Greenman CD (2012) Haploinsufficient gene selection in cancer. Science 337(6090):47–48
11. Hancer E, Xue B, Zhang MJ (2018) Differential evolution for filter feature selection based on information theory and feature ranking. Knowl-Based Syst 140:103–119
12. Hasanloei MAV, Sheikhpour R, Sarram MA, Sheikhpour E, Sharifi H (2018) A combined Fisher and Laplacian score for feature selection in QSAR based drug design using compounds with known and unknown activities. J Comput-Aided Mater 32(1):375–384
13. Hu L, Gao WF, Zhao K, Zhang P, Wang F (2018) Feature selection considering two types of feature relevancy and feature interdependency. Expert Syst Appl 93:423–434
14. Hu J, Pedrycz W, Wang GY, Wang K (2016) Rough sets in distributed decision information systems. Knowl-Based Syst 94:13–22
15. Hu QH, Pan W, An S, Ma PJ, Wei JM (2010) An efficient gene selection technique for cancer recognition based on neighborhood mutual information. Int J Mach Learn Cyb 1(1-4):63–74
16. Hu QH, Yu DR, Liu JF, Wu CX (2008) Neighborhood rough set based heterogeneous feature subset selection. Inf Sci 178(18):3577–3594
17. Huang XJ, Zhang L, Wang BJ, Li FZ, Zhang Z (2018) Feature clustering based support vector machine recursive feature elimination for gene selection. Appl Intell 48(2):594–607
18. Islam AKMT, Jeong BS, Bari ATMG, Lim CG, Jeon SH (2015) MapReduce based parallel gene selection method. Appl Intell 42(1):147–156
19. Ivica S, Jana K, Dragi K, Saso D (2018) HMC-ReliefF: Feature ranking for hierarchical multi-label classification. Comput Sci Inf Syst 15(1):187–209
20. Li JG, Su L, Pang ZN (2015) A filter feature selection method based on MFA score and redundancy excluding and it's application to tumor gene expression data analysis. Interdiscip Sci 7(3):391–396
21. Lin HY (2018) Reduced gene subset selection based on discrimination power boosting for molecular classification. Knowl-Based Syst 142:181–191
22. Liu Y, Huang WL, Jiang YL, Zeng ZY (2014) Quick attribute reduct algorithm for neighborhood rough set model. Inf Sci 271:65–81
23. Lin YJ, Hu QH, Liu JH, Chen JK, Duan J (2016) Multi-label feature selection based on neighborhood mutual information. Appl Soft Comput 38:244–256
24. Lyu HQ, Wan MX, Han JQ, Liu RL, Wang C (2017) A filter feature selection method based on the maximal information coefficient and gram-schmidt orthogonalization for biomedical data mining. Comput Biol Med 89:264–274
25. Pawlak Z (1982) Rough sets. Int J Comput Inform Sci 11(4):341–356
26. Qian YH, Liang XY, Wang Q, Liang JY, Liu B, Skowronef A, Yao YY, Ma JM, Dang CY (2018) Local rough set: a solution to rough data analysis in big data. Int J Approx Reason 97:38–63
27. Ramos J, Castellanos-Garzon JA, de Paz JF, Corchado JM (2018) A data mining framework based on boundary-points for gene selection from DNA-microarrays: Pancreatic Ductal Adenocarcinoma as a case study. Eng Appl Artif Intel 70:92–108
28. Sun L, Xu JC (2014) Information entropy and mutual information-based uncertainty measures in rough set theory. Appl Math Inform Sci 8(3):1973–1985
29. Sun L, Xu JC (2014) Feature selection using mutual information based uncertainty measures for tumor classification. Bio-Med Mater Eng 24:763–770
30. Sun L, Xu JC, Xu TH (2014) Information entropy and information granulation-based uncertainty measures in incomplete information systems. Appl Math Inform Sci 8(3):2073–2083
31. Sun L, Xu JC, Tian Y (2012) Feature selection using rough entropy-based uncertainty measures in incomplete decision systems. Knowl-Based Syst 36:206–216
32. Sun L, Xu JC, Wang W, Yin Y (2016) Locally linear embedding and neighborhood rough set-based gene selection for gene expression data classification. Genet Mol Res 15(2):15038990. gmr
33. Sun L, Xu JC, Yin Y (2015) Principal component-based feature selection for tumor classification. Bio-Med Mater Eng 26:S2011–S2017
34. Sun L, Zhang XY, Xu JC, Wang W, Liu RN (2018) A Gene selection approach based on the fisher linear discriminant and the neighborhood rough set. Bioengineered 9(1):144–151
35. Sun SQ, Peng QK, Zhang XK (2016) Global feature selection from microarray data using Lagrange multipliers. Knowl-Based Syst 110:267–274
36. The dataset is download from kent ridge bio-medical dataset. http://datam.i2r.a-star.edu.sg/datasets/krbd/
37. The dataset is download from gene expression model selector. http://www.gems-system.org
38. Urbanowicz RJ, Meeker M, La Cava W, Olsona RS, Moore JH (2018) Relief-based feature selection: introduction and review. J Biomed Inform 85:189–203
39. Venkataramana L, Jacob SG, Ramadoss R (2018) Parallelized classification of cancer sub-types from gene expression profiles using recursive gene selection. Stud Inform Control 27(1):215–224
40. Wang CZ, He Q, Shao MW, Xu YY, Hu QH (2017) A unified information measure for general binary relations. Knowl-Based Syst 135:18–28
41. Wang CZ, Hu QH, Wang XZ, Chen DG, Qian YH, Dong Z (2017) Feature selection based on neighborhood discrimination index. IEEE T Neur Net Lear 29(6):2986–2999

42. Wang CZ, Qi YL, Shao MW, Hu QH, Chen DG, Qian YH, Lin YJ (2017) A fitting model for feature selection with fuzzy rough sets. IEEE T Fuzzy Syst 25(3):741–753
43. Wang SQ, Kong W, Deng J, Gao S, Zeng WM (2018) Hybrid feature selection algorithm mRMR-ICA for cancer classification from microarray gene expression data. Comb Chem High T Scr 21(5):420–430
44. Wen LY, Min F, Wang SY (2017) A two-stage discretization algorithm based on information entropy. Appl Intell 47(3):1169–1185
45. Xu FF, Miao DQ, Wei L (2009) Fuzzy-rough attribute reduction via mutual information with an application to cancer classification. Comput Math Appl 57(5):1010–1017
46. Zhang BW, Min F, Ciucci D (2015) Representative- based classification through covering-based neighborhood rough sets. Appl Intell 43(3):840–854
47. Zhang XH, Miao DQ, Liu CH, Le ML (2016) Constructive methods of rough approximation operators and multigranulation rough sets. Knowl-Based Syst 91:114–125
48. Zhao H, Wang P, Hu QH (2016) Cost-sensitive feature selection based on adaptive neighborhood granularity with multi-level confidence. Inform Sciences 366:134–149
49. Zheng SF, Liu WX (2011) An experimental comparison of gene selection by lasso and dantzig selector for cancer classification. Comput Biol Med 41(10):1033–1040

Lin Sun is currently an Associate Professor at the College of Computer and Information Engineering, Henan Normal University, China. He received an M.S. degree in Computer Science and Technology from Henan Normal University in 2007 and a Ph.D. degree in Pattern Recognition and Intelligent Systems from Beijing University of Technology in 2015. He became a Postdoctoral Researcher with the Medical and Biological Engineering Research Group, Henan Normal University, China, in 2016. He has received funding from ten grants from the National Science Foundation of China, the China Postdoctoral Science Foundation, the Science and Technology Innovation Talents Project of Henan Province, and the Key Project of Science and Technology of Henan Province. He has authored or co-authored for over 70 articles. His main research interests include granular computing, rough sets, and big data mining. He has received the title of Henan's Distinguished Young Scholars for Science and Technology Innovation Talents, and has served as a reviewer in several prestigious peer-reviewed international journals.

Xiao-Yu Zhang is currently a postgraduate in Computer Science and Technology at the College of Computer and Information Engineering, Henan Normal University. She received a B.Sc. degree in Computer Science and Technology from Henan Normal University in 2016. Her main research interests include granular computing and data mining.

Yu-Hua Qian is a professor and Ph.D. supervisor of Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, China. He received a M.S. degree and an PhD degree in Computers with applications at Shanxi University in 2005 and 2011, respectively. He has published more than 70 articles in international journals. On professional services, he has served as program chairs or special issue chairs of RSKT, JRS, and ICIC, and PC members of many machine learning, data mining, and granular computing. He also served on the Editorial Board of International Journal of Knowledge-Based Organizations and the Editorial Board of Artificial Intelligence Research.

Jiu-Cheng Xu is currently a Professor at the College of Computer and Information Engineering, Henan Normal University. He received a B.S. degree in Mathematics from Henan Normal University in 1986, and an M.S. degree and a Ph.D. degree in Computer Science and Technology from Xi'an Jiaotong University in 1995 and 2004, respectively. He has received funding from grants from the National Science Foundation of China, the Key Scientific Research Project of Higher Education of Henan Province, and the Key Henan Provincial Science and Development Program. He has published over 100 articles. His research interests include granular computing, rough sets, intelligent information processing, and data mining. He has received the title of Henan's Distinguished High Profile Professional, and has served as a reviewer in several prestigious peer-reviewed international journals.

Shi-Guang Zhang is currently working at the College of Computer and Information Engineering, Henan Normal University, China, and is also completing his Postdoctoral studies at the School of Computer Science and Technology, Tianjin University, Tianjin, China. He received an M.S. degree in Mathematics from Guangxi University for Nationalities in 2007 and a Ph.D. degree in Applied Mathematics from Hebei Normal University in 2014. He has authored more than 10 papers and has served as a reviewer in several prestigious peer-reviewed international journals. His research interests include granular computing, knowledge discovery and machine learning.

Yun Tian is currently an associate professor at the College of Information Science and Technology, Beijing Normal University. He received a BSc degree in Computer Science and Technology from Henan Normal University in 2003 and a Ph.D. degree in Signal and Information Processing from Northwestern Polytechnic University in 2007. He has authored and co-authored more than 20 papers, and has served as a reviewer in several prestigious peer-reviewed international journals. His research interests include pattern recognition and image processing.