entropy

Article

An Attribute Reduction Method Using Neighborhood Entropy Measures in Neighborhood Rough Sets

Lin Sun [1,2,*], **Xiaoyu Zhang** [1], **Jiucheng Xu** [1,2] **and Shiguang Zhang** [1,2]

[1] College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China; zhangxy0903@126.com (X.Z.); jiuchxu@gmail.com (J.X.); sg201106@163.com (S.Z.)
[2] Engineering Technology Research Center for Computing Intelligence and Data Mining, Henan 453007, China
[*] Correspondence: linsunok@gmail.com or sunlin@htu.edu.cn

Received: 8 December 2018; Accepted: 1 February 2019; Published: 7 February 2019

Abstract: Attribute reduction as an important preprocessing step for data mining, and has become a hot research topic in rough set theory. Neighborhood rough set theory can overcome the shortcoming that classical rough set theory may lose some useful information in the process of discretization for continuous-valued data sets. In this paper, to improve the classification performance of complex data, a novel attribute reduction method using neighborhood entropy measures, combining algebra view with information view, in neighborhood rough sets is proposed, which has the ability of dealing with continuous data whilst maintaining the classification information of original attributes. First, to efficiently analyze the uncertainty of knowledge in neighborhood rough sets, by combining neighborhood approximate precision with neighborhood entropy, a new average neighborhood entropy, based on the strong complementarity between the algebra definition of attribute significance and the definition of information view, is presented. Then, a concept of decision neighborhood entropy is investigated for handling the uncertainty and noisiness of neighborhood decision systems, which integrates the credibility degree with the coverage degree of neighborhood decision systems to fully reflect the decision ability of attributes. Moreover, some of their properties are derived and the relationships among these measures are established, which helps to understand the essence of knowledge content and the uncertainty of neighborhood decision systems. Finally, a heuristic attribute reduction algorithm is proposed to improve the classification performance of complex data sets. The experimental results under an instance and several public data sets demonstrate that the proposed method is very effective for selecting the most relevant attributes with great classification performance.

Keywords: rough sets; neighborhood rough sets; attribute reduction; neighborhood entropy; classification

1. Introduction

Attribute reduction in rough set theory has been recognized as an important feature selection method, aimed to select the most representative attribute subset with a high resolution by eliminating redundant and unimportant attributes [1]. The attribute reduction methods can be widely implemented in the fields of data classification, data mining, machine learning, and pattern recognition [2–6]. Due to the development of the internet, the scale of data becomes bigger and bigger. Even thousands of attributes may be acquired in some real-world databases. In order to shorten the processing time and obtain better generalization, the attribute reduction problem attracts more and more attention in recent years [5,7,8].

In the classical rough set theory, there are two forms of definition for attribute reduction. The one is the algebra definition based on approximate precision, which determines whether certain conditional attributes can be removed according to the variation of approximate precision and considers the effect of attributes on the deterministic subsets in the field. The other is the definition of information view based on information entropy, which determines whether certain conditional attributes can be removed according to the changes of conditional entropy and considers the effect of attributes on the indeterminate subsets in the field [9]. Many attribute reduction algorithms are based on the algebra definition so far. Mi et al. [10] introduced the concepts of a lower distribution reduct and an upper distribution reduct based on the variable precision rough sets, and obtained an approach for knowledge reduction in variable precision rough sets. Syau et al. [11] provided the characterizations of lower and upper approximations for the connection between the concepts of variable precision generalized rough set model and neighborhood systems by introducing minimal neighborhood systems. What is more, as a measure to evaluate the uncertainty of discrete sample spaces, information entropy is a significant tool for characterizing the distinguishment information of attributes subsets [12]. Information entropy based on neighborhood systems has been established, and the extension of information entropy and its variants are adapted for attribute reduction. Gao et al. [13] developed a heuristic attribute reduction algorithm based on the maximum decision entropy in the decision-theoretic rough set model. Dai et al. [14] proposed a framework for attribute reduction in interval-valued data from the information view. It is known that there is a strong complementarity between the algebra view and the information view of attribute importance, and the two views can be combined to produce a more comprehensive measurement mechanism [15]. Wang [15] summarized the reduct in rough sets from algebra view and information view. This inspires the authors to investigate new attribute reduction methods from algebra view and information view in this paper.

The classical rough set theory is established on the equivalence approximate space and only compatible for discrete data set, and it could be useless for continuous numerical data [13,15,16]. In general, it needs to discretize when processing continuous numerical data, which will lead to the loss of information (including the neighborhood structure information and order structure information in real spaces) [17,18]. To overcome this drawback, many extensions of classical rough set theory have been presented [19–26], such as fuzzy rough set [21,22], tolerance approximate models [23], similarity rough approximate model [24], covering approximation model [25], and neighborhood granular model [26]. Among all the extensions, Hu et al. [18] developed a neighborhood rough set model to process both numerical and categorical data sets via neighborhood relation. Then, the neighborhood rough set model can process both numerical and discrete data sets via neighborhood parameters. Wang et al. [27] presented a local neighborhood rough set combining the neighborhood rough set and local rough set, to be applied to rough data analysis in big data. Fan et al. [28] designed an attribute reduction algorithm based on the max-decision neighborhood rough set model. Chen et al. [29] investigated approaches to attribute reduction in parallel using dominance-based neighborhood rough sets. Therefore, this paper studies an attribute reduction algorithm based on neighborhood rough sets by making full use of the advantages for rough sets.

It is noted that the reduction calculation of decision neighborhood systems is a key problem in neighborhood rough set theory. In addition, the reducts of an information system need to be achieved to further extract rule-like knowledge from information systems [29]. In practical application of decision-making, both the certainty factor and the object coverage factor of rule are two important standards of evaluating the decision ability of decision systems [30,31]. However, some of these existing knowledge reduction methods cannot reflect the change of decision ability for classification objectively. It is known that the credibility degree and the coverage degree can efficiently reflect the classification ability of conditional attributes with respect to the decision attribute [30]. The conditional attributes with higher credibility and coverage degrees are more important with respect to the decision attribute. Therefore, it is necessary to investigate a new uncertainty measure and an effective heuristic search algorithm. Sun et al. [32] proposed a concept of decision degree based on the notions of the certainty factor and the coverage factor of rule in rough sets, which includes the degree of sufficiency of a proposition and the degree of its necessity. Until now, the works are not considered in neighborhood rough sets. This inspires the authors to investigate a new measure to effectively illustrate the classification ability and the decision ability of neighborhood decision systems. Based on this idea, the concepts of coverage and credibility degrees are introduced into neighborhood decision systems as measures to reflect the classification ability of conditional attributes, with respect to decision attributes in this paper. In order to fully reflect the decision ability of attributes, the credibility degree and the coverage degree based on neighborhood rough sets are integrated into neighborhood entropy measures. Then, a heuristic attribute reduction method based on decision neighborhood entropy is presented to address the uncertainty and noisiness of complex data sets in neighborhood rough sets.

The remainder of this paper is organized as follows: Section 2 reviews some basic concepts of rough sets, information entropy measures, and neighborhood rough sets. In Section 3, some uncertainty measures based on neighborhood entropy in neighborhood decision systems are investigated, their properties are derived, and the relationships among these measures are established. An attribute reduction algorithm based on decision neighborhood entropy with complexity analysis is designed. Section 4 analyzes the classification experiments conducted on several public data sets. Finally, Section 5 summarizes the study.

2. Previous Knowledge

In this section, we briefly review several basic concepts of rough sets, information entropy measures and neighborhood rough sets in [12,13,18,33–35].

2.1. Rough Sets

Given a decision system $DS = (U, C, D, V, f)$, 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 set of classification attributes, f: $U \times \{C \cup D\} \to V$ is an information function which associates a unique value of each attribute with every object belonging to U, and $f(a, x)$ represents the value of $x \in U$ on attribute $a \in C \cup D$. For any $B \subseteq C$, two samples $x, y \in U$, the equivalence relation is described as

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

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

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 expressed, respectively, as

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

$$\overline{B}(X) = \{x|[x]_B \cap X \neq \varnothing, x \in U\}. \tag{3}$$

2.2. Information Entropy Measures

Given a decision system $DS = (U, C, D)$ with $B \subseteq C$, and $U/B = \{X_1, X_2, \cdots, X_N\}$, then the information entropy of B is described as

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

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

Given a decision system $DS = (U, C, D)$ with 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 denoted as

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

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 $B_1, B_2 \subseteq C$, $U/B_1 = \{X_1, X_2, \cdots, X_N\}$, and $U/B_2 = \{Y_1, Y_2, \cdots, Y_M\}$, then the conditional information entropy of B_2 with respect to B_1 is defined as

$$H(B_2|B_1) = -\sum_{i=1}^{N} p(X_i) \sum_{j=1}^{M} p(Y_j|X_i) \log p(Y_j|X_i), \tag{6}$$

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 $B_1, B_2 \subseteq C$, the mutual information between B_1 and B_2 is defined as

$$I(B_1; B_2) = H(B_2) - H(B_2 \mid B_1). \tag{7}$$

Given a decision system $DS = (U, C, D)$ with $B_1, B_2 \subseteq C$, the mutual information has the following properties:

(1) $I(B_1; B_2) \geq 0$,
(2) $I(B_1; B_2) = I(B_2; B_1)$,
(3) $I(B_1; B_2) = H(B_1) + H(B_2) - H(B_1 \cup B_2)$,
(4) $I(B_1; B_2) = H(B_1) - H(B_1 \mid B_2) = H(B_2) - H(B_2 \mid B_1)$.

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

2.3. Neighborhood Rough Sets

Given a neighborhood decision system $NDS = (U, C, D, V, f, \Delta, \delta)$, usually written more simply as $NDS = (U, C, D, \delta)$, 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 decision attribute, $V = \cup_{D\}} V_a$, V_a is a value set of attribute 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 with $0 \leq \delta \leq 1$.

For any samples $x, y, z \in U$ on a subset B, the distance function $\Delta_B(x, y)$ satisfies the following three conditions: $\Delta_B(x, y) \geq 0$, $\Delta_B(x, y) = \Delta_B(y, x)$, and $\Delta_B(x, y) + \Delta_B(y, z) \geq \Delta_B(x, z)$.

It is well known that there are three classical metrics including Manhattan, Euclidean, and Chebychev distance functions, where the Euclidean distance function effectively reflects the basic information of the unknown data [33]. Given a neighborhood decision system $NDS = (U, C, D, \delta)$ with $B \subseteq C$, for any $x, y \in U$, the Euclidean distance function between x and y is expressed as

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

Given a neighborhood decision system $NDS = (U, C, D, \delta)$ with $B \subseteq C$, the similarity relation resulting by B is defined as

$$NR_\delta(B) = \{(x, y) \in U \times U | \Delta_B(x, y) \leq \delta\}. \tag{9}$$

Given a neighborhood decision system $NDS = (U, C, D, \delta)$ with $B \subseteq C$, for any $x \in U$, the neighborhood class of x with respect to B is described as

$$n_B^\delta(x) = \{y | x, y \in U, \Delta_B(x, y) \leq \delta\}. \tag{10}$$

Given a neighborhood decision system *NDS* = (*U*, *C*, *D*, *δ*) with *B*⊆*C* and *X*⊆*U*, the neighborhood lower approximation set and the neighborhood upper approximation set of *X* with respect to *B* are described, respectively, as

$$\underline{B}_\delta(X) = \left\{ x_i \middle| n_B^\delta(x_i) \subseteq X,\ x_i \in U,\ i = 1,\ 2,\ \cdots,\ n \right\}, \tag{11}$$

$$\overline{B}_\delta(X) = \left\{ x_i \middle| n_B^\delta(x_i) \cap X \neq \varnothing, x_i \in U,\ i = 1,\ 2,\ \cdots,\ n \right\}. \tag{12}$$

Given a neighborhood decision system *NDS* = (*U*, *C*, *D*, *δ*) with *B*⊆*C* and *X*⊆*U*, $\underline{B}_\delta(X)$ is the neighborhood lower approximation set of *X* with respect to *B*, and $\overline{B}_\delta(X)$ is the neighborhood upper approximation set of *X* with respect to *B*, then the approximate precision of *X* with respect to *B* is described as

$$p_B(X) = \frac{|\underline{B}_\delta(X)|}{\left|\overline{B}_\delta(X)\right|}. \tag{13}$$

Given a neighborhood decision system *NDS* = (*U*, *C*, *D*, *δ*) with *B*⊆*C*, *U*/*D* = {X_1, X_2, · · · , X_N}, then the neighborhood lower approximation set and the neighborhood upper approximation set of *D* with respect to *B* are described respectively as

$$\underline{B}_\delta(D) = \bigcup_{i=1}^{N} \underline{B}_\delta(X_i), \tag{14}$$

$$\overline{B}_\delta(D) = \bigcup_{i=1}^{N} \overline{B}_\delta(X_i), \tag{15}$$

where $\underline{B}_\delta(X_i)$ respects the neighborhood lower approximation set of X_i with respect to *B*, $\overline{B}_\delta(X_i)$ respects the neighborhood upper approximation set of X_i with respect to *B*, and *i* = 1, 2, · · · , *N*.

Given a neighborhood decision system *NDS* = (*U*, *C*, *D*, *δ*) with *B*⊆*C*, then the neighborhood approximate precision of *D* with respect to *B* is described as

$$p_B(D) = \frac{|\underline{B}_\delta(D)|}{\left|\overline{B}_\delta(D)\right|}. \tag{16}$$

The neighborhood approximate precision can be used to reflect the complete degree of the knowledge of a set, but this precision measure does not take into account the size of the particles that are included in the lower approximation set completely. Therefore, it is not sufficient to only consider attribute reduction from the algebra view.

3. Attribute Reduction Method Using Neighborhood Entropy Measures in Neighborhood Decision Systems

Attribute reduction is a core part of the rough set theory [13]. In the classical rough set theory, there are two forms of definition for attribute reduction: One is the algebra definition based on set theory; the other is the definition of information view based on information entropy. There is a strong complementarity between the algebra definition of attribute significance and the definition of information view. The former considers the influence of attributes on the defined subset in the domain of theory, while the latter considers the influence of attributes on the uncertain subset in the domain of theory. Therefore, the two views can be combined to produce a more comprehensive measurement mechanism. In rough sets, the equivalence classes-based information entropy does not work for numerical data. Then, for continuous data sets, a discretization should be performed before further processing. However, the discretization may result in information loss, and it is difficult to employ mutual information in attribute evaluation due to the difficulty in estimating the probability density of attributes [36]. To address this issue, the concept of neighborhood can be combined with information theory to extend Shannon entropy, and then some correlative concepts of neighborhood entropy are defined to measure the uncertainty of numerical data. Then, some concepts of neighborhood entropy-based uncertainty measures are presented to measure the uncertainty of knowledge in neighborhood decision systems, some important properties and relationships of these measures are deduced respectively as well, and a heuristic attribute reduction algorithm is investigated to improve the classification performance of complex data sets.

3.1. Neighborhood Entropy-Based Uncertainty Measures

Given a neighborhood decision system $NDS = (U, C, D, \delta)$ with $B \subseteq C$, $n_B^\delta(x_i)$ is a neighborhood class of $x_i \in U$, then Hu et al. [37] described the neighborhood entropy of x_i as

$$H_\delta^{x_i}(B) = -\log \frac{|n_B^\delta(x_i)|}{|U|}. \tag{17}$$

Given a neighborhood decision system $NDS = (U, C, D, \delta)$ with $B \subseteq C$, Hu et al. [37] and Chen et al. [33] computed the average neighborhood entropy of the sample set as

$$H_\delta(B) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log \frac{|n_B^\delta(x_i)|}{|U|}. \tag{18}$$

The concept of neighborhood entropy is defined based on the information entropy theory, which granulates the space of the domain by neighborhood relation and is used to measure the uncertainty and classification ability of the numerical knowledge classification system [38]. In this paper, the neighborhood approximate precision is combined with the neighborhood entropy to reflect the uncertainty of knowledge, and then a new average neighborhood entropy is defined as follows.

Definition 1. *Given a neighborhood decision system $NDS = (U, C, D, \delta)$ with $B \subseteq C$, a new average neighborhood entropy of the sample set is defined as*

$$H_\delta(B) = -\frac{p_B(D)}{|U|}\sum_{i=1}^{|U|} \log \frac{|n_B^\delta(x_i)|}{|U|}. \tag{19}$$

From Definition 1, the average neighborhood entropy combines the neighborhood precision with the average neighborhood entropy, and it can make full use of the advantages of algebra and information view, and overcomes the drawbacks of traditional precision measurement.

Property 1. *Given a neighborhood decision system $NDS = (U, C, D, \delta)$ with $x_i \in U$, then $0 \le H_\delta\ (C) \le \log | U |$.*

Proof. It follows from Equation (10) that $n_C^\delta(x_i) \subseteq U$ for any $x_i \in U$, then one has that $\frac{1}{|U|} \le \frac{|n_C^\delta(x_i)|}{|U|} \le 1$. From Equation (16), it is obtained that $0 \le p_B(D) \le 1$, and then $0 \le \frac{p_B(D)}{|U|} \le \frac{1}{|U|}$. Thus, it is obvious from Definition 1 that $0 \le H_\delta\ (C) \le \log | U |$. □

Proposition 1. *Given a neighborhood decision system $NDS = (U, C, D, \delta)$ with $x_i \in U$, if $B_1 \subseteq B_2 \subseteq C$, then $H_\delta(B_2) \ge H_\delta(B_1)$.*

Proof. Suppose that $B_1 \subseteq B_2 \subseteq C$, and similar to the proof of Proposition 1 in [33], one has that $n_{B_1}^\delta(x_i) \supseteq n_{B_2}^\delta(x_i)$. Then, $\left|n_{B_1}^\delta(x_i)\right| \ge \left|n_{B_2}^\delta(x_i)\right|$ holds. It follows from Equations (11) and (12) that $\underline{B_1}_\delta(X) \subseteq \underline{B_2}_\delta(X)$ and $\overline{B_1}_\delta(X) \supseteq \overline{B_2}_\delta(X)$. By Equation (16), one has that $p_{B_1}(D) \le p_{B_2}(D)$. Hence, it can be obtained from Equation (19) that $H_\delta(B_2) \ge H_\delta(B_1)$. □

Definition 2. *Given a neighborhood decision system NDS = (U, C, D, δ) with B⊆C, $n_B^{\delta}(x_i)$ is a neighborhood class of x_i∈U generated by NR_{δ}(B), and $[x_i]_D$ is an equivalence class of x_i∈U generated by IND(D), then a decision neighborhood entropy of B and D is defined as*

$$H_{\delta}(D,B) = -\frac{p_B(D)}{|U|}\sum_{i=1}^{|U|}\log(\frac{\left|n_B^{\delta}(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|}). \tag{20}$$

In a decision system $DS = (U, C, D)$ with any $x_i \in U$, Pawlak et al. [34] and Wang et al. [30] express a decision rule as $d_{x_i} : des([x_i]_C) \Rightarrow des([x_i]_D)(\beta_{x_i}, \lambda_{x_i})$, where $des([x_i]_C)$ and $des([x_i]_D)$ are the descriptions of x_i under the equivalence relations $IND(C)$ and $IND(D)$, respectively. $\beta_{x_i} = \frac{|[x_i]_C \cap [x_i]_D|}{|[x_i]_C|}$ is called the credibility degree of decision rule d_{x_i}, and $\lambda_{x_i} = \frac{|[x_i]_C \cap [x_i]_D|}{|U|}$ is called the coverage degree of decision rule d_{x_i}. Wang et al. [30] declared that the credibility degree and the coverage degree can reflect the classification ability of conditional attributes with respect to the decision attribute, and the conditional attributes with higher credibility and coverage degrees are more important with respect to the decision attribute. Furthermore, Tsumoto [31] emphasized that the credibility degree indicates the adequacy of the proposition, and the coverage degree describes the necessity of the proposition. Then, in order to fully reflect the decision ability and the classification ability of neighborhood decision systems, this paper investigates some neighborhood entropy-based uncertainty measures by combining the credibility degree with the coverage degree in neighborhood rough sets.

Property 2. *Given a neighborhood decision system NDS = (U, C, D, δ), if* $\beta_{x_i} = \frac{\left|n_B^{\delta}(x_i)\cap[x_i]_D\right|}{|[x_i]_D|}$ *and* $\lambda_{x_i} = \frac{\left|n_B^{\delta}(x_i)\cap[x_i]_D\right|}{|U|}$ *for any x_i∈U, then one has that* $H_{\delta}(D,B) = -\frac{p_B(D)}{|U|}\sum_{i=1}^{|U|}\log(\beta_{x_i}\lambda_{x_i})$.

Proof. It follows immediately from Definition 2 that

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

Property 2 shows that the decision neighborhood entropy of B and D combines the credibility degree and the coverage degree in the neighborhood decision system, which can fully reflect the decision ability of the neighborhood decision system.

Proposition 2. *Given a neighborhood decision system NDS = (U, C, D, δ) with B_1⊆B_2⊆C, then $H_{\delta}(D,B_1) \le H_{\delta}(D,B_2)$, where the equal sign holds if and only if $n_{B_1}^{\delta}(x_i) = n_{B_2}^{\delta}(x_i)$ for any x_i∈*U.

Proof. Suppose that $B_1 \subseteq B_2 \subseteq C$, it follows from Proposition 1 that $n_{B_1}^{\delta}(x_i) \supseteq n_{B_2}^{\delta}(x_i)$. Then, it is obvious that $U \supseteq n_{B_1}^{\delta}(x_i)\cap[x_i]_D \supseteq n_{B_2}^{\delta}(x_i)\cap[x_i]_D \supseteq \{x_i\}$. It is easily obtained that $|U| \ge \left|n_{B_1}^{\delta}(x_i)\cap[x_i]_D\right| \ge \left|n_{B_2}^{\delta}(x_i)\cap[x_i]_D\right| \ge |\{x_i\}| = 1$. Thus, one has that $\frac{|U|^2}{|U||[x_i]_D|} \ge \frac{\left|n_{B_1}^{\delta}(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|} \ge \frac{\left|n_{B_2}^{\delta}(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|} \ge \frac{1}{|U||[x_i]_D|}$. So, $\log(\frac{|U|}{|[x_i]_D|}) \ge \log(\frac{\left|n_{B_1}^{\delta}(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|}) \ge \log(\frac{\left|n_{B_2}^{\delta}(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|}) \ge \log(\frac{1}{|U||[x_i]_D|})$ obviously holds. In addition, from Equations (11) and (12), it follows that $\underline{B_1}_{\delta}(X) \subseteq \underline{B_2}_{\delta}(X)$ and $\overline{B_1}_{\delta}(X) \supseteq \overline{B_2}_{\delta}(X)$. According to Equation (16), one has that $p_{B_1}(D) \le p_{B_2}(D)$. Hence, it can be obtained from Definition 2 that $H_\delta(D, B_1) \le H_\delta(D, B_2)$. When $n^\delta_{B_1}(x_i) = n^\delta_{B_2}(x_i)$ for any $x_i \in U$, it is obvious that $\frac{\left|n^\delta_{B_1}(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|} = \frac{\left|n^\delta_{B_2}(x_i)\cap[x_i]_D\right|^2}{|U||[x_i]_D|}$, and one has $p_{B_1}(D) = p_{B_2}(D)$. From Definition 2, it follows that $H_\delta(D, B_1) = H_\delta(D, B_2)$. Therefore, $H_\delta(D, B_1) \le H_\delta(D, B_2)$ holds. □

The monotonicity is one of the most important properties for an effective uncertainty measure of attribute reduction. According to Proposition 2, it is quite obvious that the decision neighborhood entropy is monotonic, decreasing when adding the condition attributes, which validates the monotonicity of the proposed uncertainty measure.

Proposition 3. *Given a neighborhood decision system NDS = (U, C, D, δ) with B⊆C, then H_δ (D, B) ≥ H_δ (B).*

Proof. It follows immediately from Definitions 1 and 2 that:

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

Since there exists $0 \le \frac{|\beta_B(D)|}{|U|} \le \frac{1}{|U|}$, it follows that $n^\delta_B(x_i) \cap [x_i]_D \subseteq n^\delta_B(x_i)$ and $n^\delta_B(x_i) \cap [x_i]_D \subseteq [x_i]_D$. Then, it is easily obtained that $\left|n^\delta_B(x_i) \cap [x_i]_D\right| \le \left|n^\delta_B(x_i)\right|$ and $\left|n^\delta_B(x_i) \cap [x_i]_D\right| \le |[x_i]_D|$. Thus, obviously, $\frac{\left|n^\delta_B(x_i)\cap[x_i]_D\right|}{\left|n^\delta_B(x_i)\right|} \le 1$ and $\frac{\left|n^\delta_B(x_i)\cap[x_i]_D\right|}{|[x_i]_D|} \le 1$ hold. Therefore, H_δ(*D*, *B*) − H_δ(*B*) ≥ 0 can be obtained, in essence, H_δ(*D*, *B*) ≥ H_δ(*B*). □

Definition 3. *Given a neighborhood decision system NS = (U, C, D, δ) with B⊆C, and any a∈B, if H_δ(D, B) ≤ H_δ(D, B−{a}), then a is redundant in B with respect to D; otherwise, a is indispensable in B with respect to D. B is dependent if any attribute in B with respect to D is indispensable. B is called a reduct of C with respect to D if it satisfies the following two conditions:*

(1) H_δ(D, B) = H_δ(D, C);

(2) H_δ(D, B − {a}) < H_δ(D, B), where any a∈B.

Obviously, a reduct of *C* with respect to *D* is the minimal attribute subset to retain the decision neighborhood entropy of *C* with respect to *D*.

Definition 4. *Given a neighborhood decision system NDS = (U, C, D, δ) with B⊆C and any attribute a∈B, then the significance measure of a in B with respect to D is defined as*

$$Sig_{in}(a, B, D) = H_\delta(D, B) - H_\delta(D, B - \{a\}). \tag{21}$$

Definition 5. *Given a neighborhood decision system NDS = (U, C, D, δ) with B⊆C and any attribute a∈C − B, then the significance measure of a with respect to D is defined as*

$$Sig_{out}(a, B, D) = H_\delta(D, B \cup \{a\}) - H_\delta(D, B). \tag{22}$$

When $B = \varnothing$, $Sig_{out}(a, B, D) = H_\delta(D, \{a\})$. From Definition 5, the significance of attribute a is the increment of the distinguishing information after adding a into B. The larger the value of $Sig(a, B, D)$ is, the greater the importance of attribute a for B with respect to D is.

3.2. Attribute Reduction Algorithm Based on Decision Neighborhood Entropy

The process of the attribute reduction method for classification is shown in Figure 1.

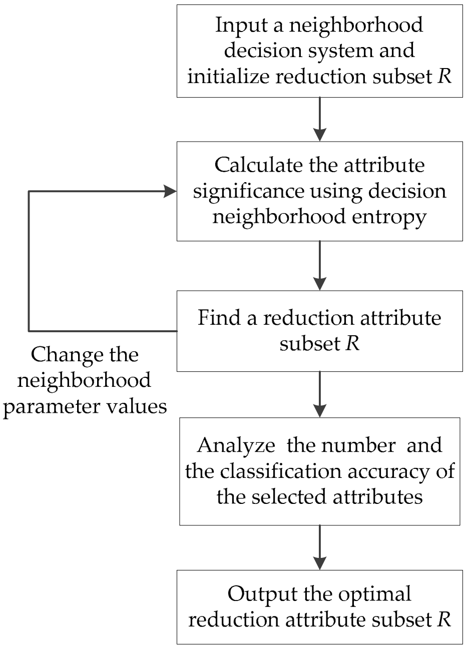

Input a neighborhood decision system and initialize reduction subset R
Calculate the attribute significance using decision neighborhood entropy
Find a reduction attribute subset R
Change the neighborhood parameter values
Analyze the number and the classification accuracy of the selected attributes
Output the optimal reduction attribute subset R

Figure 1. The process of the attribute reduction method based on decision neighborhood entropy.

To support efficient knowledge reduction, an attribute reduction algorithm based on decision neighborhood entropy (ARDNE) is constructed and described as Algorithm 1.

Algorithm 1

Input: A neighborhood decision system $NDS = (U, C, D, \delta)$
Output: A reduction attribute subset R

```
Initialize $R = \varnothing$.
while $Sig(S, R, D) = 0$ do
    Let $Agent = R$, and $h = 0$.
    for any $a \in (S - R)$ do
        Compute $H_\delta(D, R \cup \{a\})$.
        if $H_\delta(D, R \cup \{a\}) > h$ then
            Let $Agent = R \cup \{a\}$, and $h = H_\delta(D, B \cup \{a\})$.
        end if
    end for
    Let $R = Agent$.
end while
Let $r = |R|$.
for $i = 1$ to $r$ do
    Select $a_i \in \mathrm{R}$.
    Calculate $H_\delta(D, R - \{a_i\})$.
    if $H_\delta(D, R - \{a_i\}) \geq H_\delta(D, C)$ then
        Let $R = R - \{a_i\}$.
end
Return a reduction attributes subset $R$.
```

3.3. Complexity Analysis of ARDNE Algorithm

From Algorithm 1, the decision neighborhood entropy and the neighborhood classes induced by the conditional attributes need to be frequently calculated in the computation of the attribute significance measure. The above computational process largely affects the time complexity of selecting attributes. Suppose that the number of attributes is m, and the number of samples is n. The complexity of calculating neighborhood classes is $O(mn)$, and the computational complexity of calculating decision neighborhood entropy is $O(n)$. Since $O(n) < O(mn)$, the computational complexity of calculation of significance measure is $O(mn)$. There are two loops in step 2 through step 11, then the worst time complexity of ARDNE is $O(m^3n)$. Suppose that the number of selected attributes is m_R, for the calculation of the neighborhood classes, only the candidate attributes are considered instead of the entire attribute set. Then, the time complexity of achieving neighborhood classes is $O(m_Rn)$. The outer loop times are m_R, and the inner loop times are $m - m_R$. Thus, the time complexity of this part is $O(m_R(m - m_R)m_Rn)$. Similar to the last steps, the time complexity of step 12 through step 18 is $O(m_Rn)$. It is well known that $m_R \ll m$ in most cases. Therefore, the time complexity of ARDNE is close to $O(mn)$. So far, ARDNE appears to be more efficient than some of the existing algorithms for attribute reduction in [33,39–41] in neighborhood decision systems. Furthermore, its space complexity is $O(mn)$.

3.4. An Illustrative Example

In the following, the performance of the ARDNE algorithm is shown through an illustrative example in [42]. A neighborhood decision system $NDS = (U, C, D, \delta)$ is employed, 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.

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

For Table 1, an example for attribute reduction using Algorithm 1 is given. Then, the neighborhood class of each attribute in Table 1 is calculated by using the Euclidean distance function as follows.

For an attribute 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 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 $n^{\delta}_{\{a\}}(x_4) = \{x_3, x_4\}$.

Due to $D = \{d\}$ in Table 1, it follows that $U/\{d\} = \{X_1, X_2\} = \{\{x_1, x_2\}, \{x_3, x_4\}\}$. Then, from Equation (20), one has that $H_\delta(D, \{a\}) = -\frac{1}{4\times 4}(\log(\frac{2^2}{4\times 2}) + \log(\frac{2^2}{4\times 2}) + \log(\frac{2^2}{4\times 2}) + \log(\frac{2^2}{4\times 2})) = 0.0753$.

Similarly, $H_\delta(D, \{b\}) = 0.0753$, $H_\delta(D, \{c\}) = 0.1505$, $H_\delta(D, \{a, b\}) = 0.0753$, $H_\delta(D, \{a, c\}) = 0.1505$, $H_\delta(D, \{b, c\}) = 0.1505$, and $H_\delta(D, \{a, b, c\}) = 0.1505$.

From the above calculated results, it can be observed that $H_\delta(D, \{c\}) > H_\delta(D, \{a\}) = H_\delta(D, \{b\})$. Since the decision neighborhood entropy of $\{c\}$ and D is maximum, the attribute c should be added to the candidate attribute set (i.e., $R = \{c\}$). By computing, one has that $Sig(C, R, \mathrm{D}) = H_\delta(D, C) - H_\delta(D, \{c\}) = 0$, which satisfies the termination criterion. Thus, a selected attribute subset $\{c\}$ is achieved.

4. Experimental Results and Analysis

4.1. Experiment Preparation

It is known that the objective of an attribute reduction algorithm usually has two aspects: One is to select a small attribute subset and the other is to maintain high classification accuracy. To demonstrate the classification performances of our proposed attribute reduction algorithm described in Section 3.2

on several public data sets, the more comprehensive results of all contrasted algorithms should be achieved and analyzed. The selected four UCI (University of California at Irvine) Machine Learning Repository data sets with low-dimensional attributes include Ionosphere, Wisconsin Diagnostic Breast Cancer (Wdbc), Wisconsin Prognostic Breast Cancer (Wpbc), and Wine, which were downloaded from https://archive.ics.uci.edu/ml/datasets.html. The selected seven microarray gene expression data sets with high-dimensional attributes included Brain_Tumor1, Diffuse Large B Cell Lymphoma (DLBCL), Leukemia, Small Round Blue Cell Tumor (SRBCT), Colon, Lung, and Prostate, where the four gene expression data sets (Brain_Tumor1, DLBCL, Leukemia and SRBCT) can be downloaded at http://www.gems-system.org, the Colon gene expression data set can be downloaded at http://eps.upo.es/bigs/datasets.html, the Lung data set can be downloaded at http://bioinformatics.rutgers.ed/Static/Supple-mens/CompCancer/datasets, and the Prostate gene expression data set can be downloaded at http://www.gems-system.org. All of the data sets above are summarized in Table 2.

Table 2. Description of the eleven public data sets.

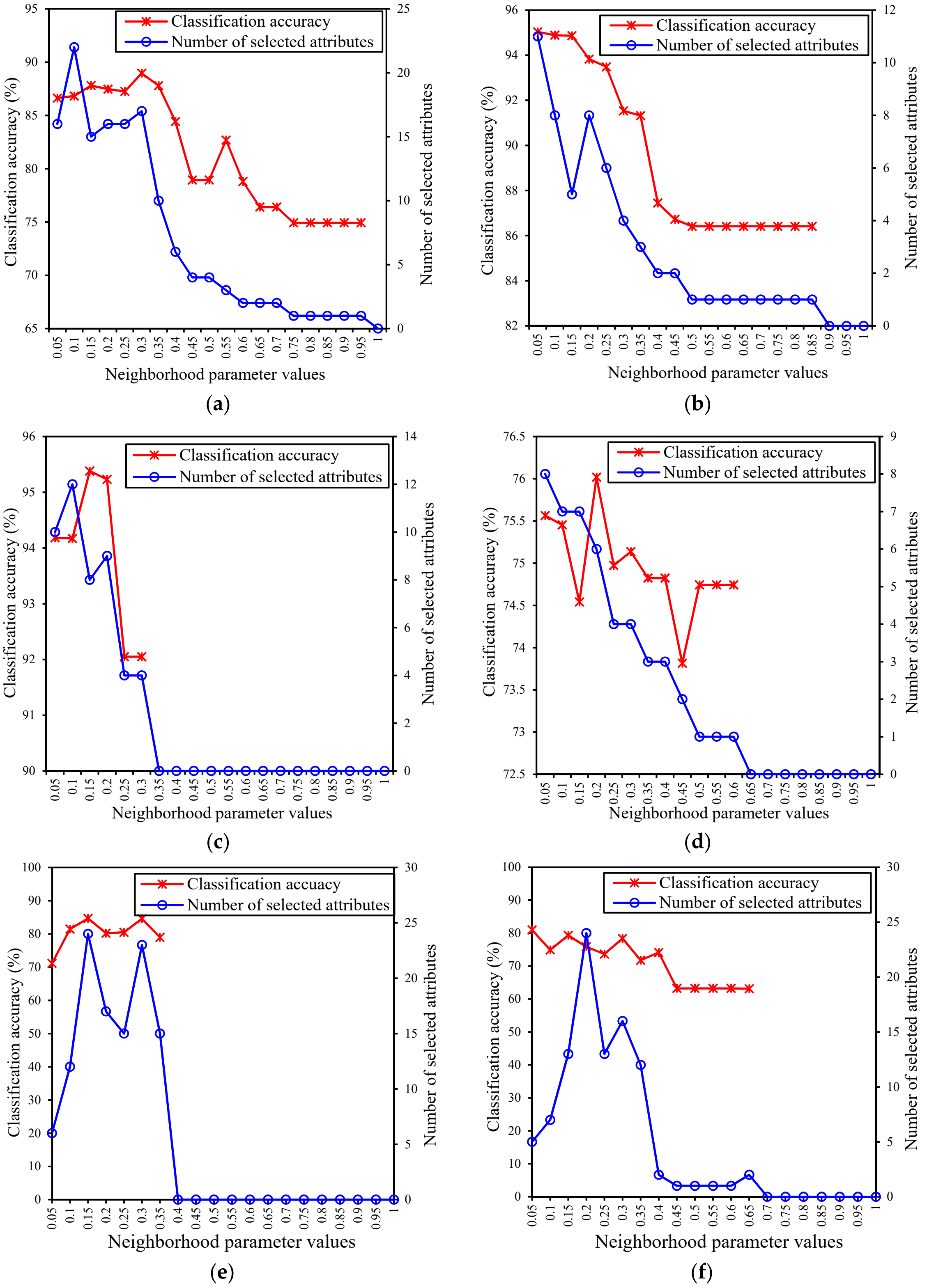
No.	Data Sets	Samples	Attributes	Classes	Reference
1	Ionosphere	351	33	2	Fen et al. [28]
2	Wdbc	569	31	2	
3	Wine	178	13	3	
4	Wpbc	198	33	2	
5	Brain_Tumor1	90	5920	5	Huang et al. [43]
6	Colon	62	2000	2	Mu et al. [44]
7	DLBCL	77	5469	2	Wang et al. [45]
8	Leukemia	72	7129	2	Dong et al. [46]
9	Lung	181	12,533	2	Sun et al. [47]
10	Prostate	136	12,600	2	
11	SRBCT	63	2308	4	Tibshirani et al. [48]

Wdbc—Wisconsin Diagnostic Breast Cancer; Wpbc—Wisconsin Prognostic Breast Cancer; DLBCL—Diffuse Large B Cell Lymphoma; and SRBCT—Small Round Blue Cell Tumor.

The experiments were performed on a personal computer running Windows 7 with an Intel(R) Core(TM) i5-3470 CPU operating at 3.20 GH, and 4 GB memory. All the simulation experiments were implemented in MATLAB R2014a, and the k-nearest neighbors (KNN) classifier and the support vector machine (SVM) classifier were selected to verify the classification accuracy in WEKA software, where the parameter k in KNN was set to 3 and the linear kernel functions were selected in SVM. All of the following experimental comparisons for classification on the selected attributes are implemented using a 10-fold cross-validation with all the test data sets, where every data set is first randomly divided into ten portions which are the same size subset of data each other, one data subset is used as the testing data set, the rest nine data subsets are used as the training data set, and each of the ten data subsets only is employed exactly once as the testing data set; secondly, the operation of the cross-validation is repeated ten times; finally, the average of ten test results is as the obtained classification accuracy.

4.2. Effect of Different Neighborhood Parameter Values

Since the value of neighborhood parameter decides the granularity of data manipulation, which affects both the cardinality of the data set and the classification accuracy of the attribute subset, in this subsection, our experiments concern the number of selected attributes and the classification accuracy with the different neighborhood parameter values. Following the experimental techniques designed by Chen et al. [33], the number of selected attributes and the classification accuracy of selected attribute subset for the different neighborhood parameter values is discussed to obtain a suitable neighborhood parameter value and a better attribute subset. The classification results of the data sets given in Table 2 were obtained by using the ARDNE algorithm with the different neighborhood parameters, shown in Figure 2, where the horizontal coordinates denotes the neighborhood parameters with $\delta \in [0.05, 1]$ at intervals of 0.05, and the left and right vertical axes represent the classification accuracy and the number of selected attributes, respectively.



(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

Figure 2. *Cont.*

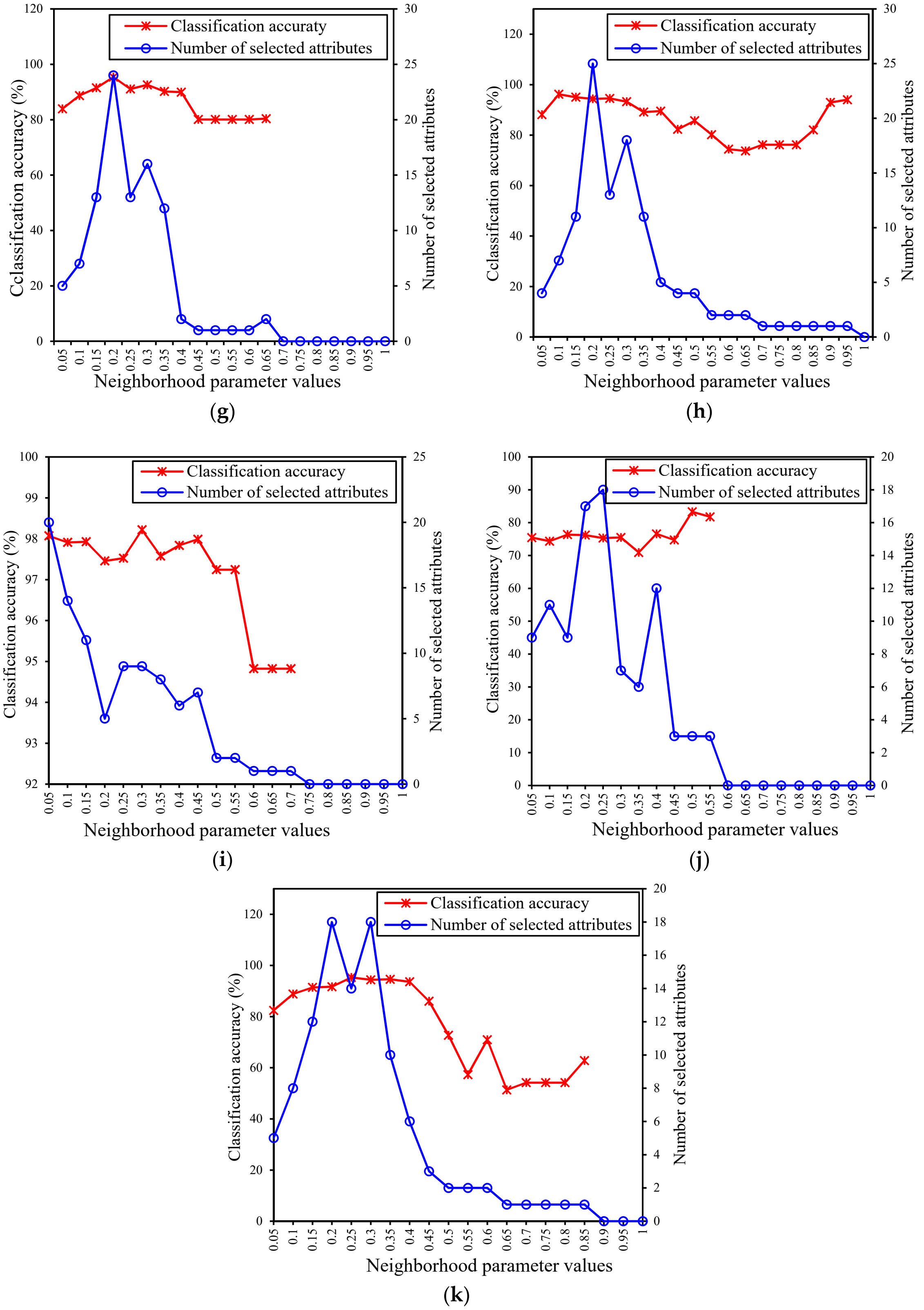


Figure 2. The number of selected attributes and the classification accuracy of the eleven data sets with the different neighborhood parameter values. (**a**) Ionosphere data set; (**b**) Wdbc data set; (**c**) Wine data set; (**d**) Wpbc data set; (**e**) Brain_Tumor1 data set; (**f**) Colon data set; (**g**) DLBCL data set; (**h**) Leukemia data set; (**i**) Lung data set; (**j**) Prostate data set; (**k**) SRBCT data set.

Figure 2a–k show the number of selected attributes and the classification accuracy of eleven data sets with the different neighborhood parameter values. For the Ionosphere data set in Figure 2a, the classification accuracy reached its maximum when the parameter was 0.3. As the parameter value continued to increase, the number of selected attributes decreased, resulting in a rapid decrease in classification accuracy. For the Wdbc data set in Figure 2b, when the parameter took values in the interval [0.05, 0.15], there was little change in the classification accuracy, and the number of selected attributes was less when the parameter was 0.15. For the Wine data set in Figure 2c, the classification accuracy reached its maximum when the parameter was 0.15. Similarly, for the Wpbc data set in Figure 2d, the classification accuracy reached its maximum when the parameter was 0.2. For the Brain_Tumor1 data set in Figure 2e, the classification accuracy achieved the maximum when the parameter was set as 0.15. For the Colon data set in Figure 2f, as the parameter value continued to increase, the number of selected attributes increased first and in turn decreased, and then the classification accuracy reached its maximum when the parameter was 0.05. For the DLBCL data set in Figure 2g, when the parameter took the values in the interval [0.15, 0.3], there was a slight difference in the classification accuracy, and then the number of selected attributes was less when the parameter was 0.15. For the Leukemia, Lung, Prostate, and SRBCT data sets in Figure 2h–k, the classification accuracy reached their maximum when the parameters were 0.1, 0.3, 0.5, and 0.25, respectively. In addition, when the neighborhood parameter was about 0.5, the number of selected attributes would be close to zero. Therefore, the appropriate neighborhood parameters of eleven data sets should take values in the interval [0.05, 0.5].

4.3. Classification Results of ARDNE Algorithm under Different Neighborhood Parameter Values

In this part of our experiments, by using the above selected neighborhood parameters in Section 4.2, the classification results of the raw data and the reduced data using Algorithm 1 on the eleven gene expression data sets in Table 2 could be obtained. Then, the number of the attributes selected by the ARDNE algorithm and the corresponding classification accuracy with SVM and KNN based on 10-fold cross validation are shown in Table 3, respectively. The corresponding neighborhood parameter values are listed in the last column.

Table 3. The number of selected attributes and the classification accuracy under the SVM and KNN classifiers on the raw data and the reduced data with Algorithm 1.

Data Sets	**Raw Data**			**Reduced Data using Algorithm 1**			δ
	Attributes	**SVM**	**KNN**	**Attributes**	**SVM**	**KNN**	
Ionosphere	33	0.874	0.857	11	0.909	0.893	0.3
Wdbc	31	0.538	0.896	10	0.959	0.959	0.15
Wine	13	0.401	0.69	7	0.959	0.96	0.15
Wpbc	33	0.667	0.752	6	0.772	0.753	0.2
Brain_Tumor1	5920	0.86	0.783	13	0.83	0.897	0.15
DLBCL	2000	0.965	0.896	10	0.993	0.998	0.05
Colon	5469	0.811	0.776	5	0.808	0.818	0.15
Lung	7129	0.979	0.975	6	0.99	0.99	0.1
Leukemia	12,533	0.973	0.842	6	0.967	0.981	0.3
Prostate	12,600	0.916	0.796	3	0.829	0.858	0.5
SRBCT	2308	0.984	0.808	6	1	1	0.25
Average	4369.9	0.815	0.825	7.5	0.911	0.919	

SVM—support vector machine and KNN—*k*-nearest neighbors.

From Table 3, it can be found that our proposed algorithm can greatly reduce the attributes for all the data sets without loss of classification accuracy, and most of the redundant attributes are reduced. In the four low-dimensional data sets, the classification accuracy of the SVM and KNN classifiers were higher. In the seven high-dimensional gene expression data sets, the classification accuracy of the KNN classifier was higher than that of the raw data, while there were some differences in the classification accuracy of SVM classifier. On the KNN classifier, the classification accuracy of all the data sets was higher than that of the raw data. On the SVM classifier, the classification accuracy of the Brain_Tumor1 data set was 3% less than that of the raw data, and the classification accuracy of the Colon and Leukemia data sets were slightly lower than that of the raw data. It shows that the reduced attribute set can maintain the classification accuracy of the raw data. However, on the SVM classifier, for the Prostate data set, the classification accuracy was 8.7% less than that of the raw data. The reason is that some attributes with important information during reduction are lost. What is more, for the average classification accuracy, our ARDNE algorithm obtained 91.04% and 91.86% on the SVM and KNN classifiers, respectively, which was higher by 10% than that of raw data sets. Therefore, the proposed ARDNE algorithm was efficient in dimension reduction of low-dimensional and high-dimensional data sets.

4.4. Classification Results of UCI Data Sets with Low-Dimensions

This portion of our experiments was to evaluate the performance of our proposed algorithm in terms of classification accuracy, and the classification performance of the ARDNE algorithm was compared with those of the other four related state-of-the-art attribute reduction algorithms on the four UCI data sets, selected from Table 2. The algorithms used in the comparison included: (1) The classical rough set algorithm (RS) [34], (2) the neighborhood rough set algorithm (NRS) [49], (3) the covering decision algorithm (CDA) [50], and (4) the max-decision neighborhood rough set algorithm (MDNRS) [28]. Table 4 gives the numbers of selected attributes in the reduced data with the four different algorithms. Tables 5 and 6 show the comparison results of classification accuracy using the four different methods.

Table 4. The number of selected attributes of the five reduction algorithms on the four UCI data sets.

Data Sets	RS	NRS	CDA	MDNRS	ARDNE
Ionosphere	17	8	9	8	11
Wdbc	8	2	2	2	10
Wine	5	3	2	4	7
Wpbc	7	2	2	4	6
Average	9.25	3.75	3.75	4.5	8.5

RS—classical rough set algorithm; NRS—neighborhood rough set algorithm; CDA—covering decision algorithm; MDNRS—max-decision neighborhood rough set algorithm; and ARDNE— attribute reduction algorithm based on decision neighborhood entropy.

Table 5. Classification accuracy of the five reduction algorithms on the four UCI data sets with KNN.

Data Sets	RS	NRS	CDA	MDNRS	ARDNE
Ionosphere	0.866	0.859	0.848	0.891	0.893
Wdbc	0.911	0.923	0.923	0.930	0.959
Wine	0.863	0.752	0.727	0.911	0.960
Wpbc	0.743	0.738	0.738	0.761	0.753
Average	0.846	0.818	0.809	0.873	0.891

Table 6. Classification accuracy of the five reduction algorithms on the four UCI data sets with SVM.

Data Sets	RS	NRS	CDA	MDNRS	ARDNE
Ionosphere	0.881	0.872	0.878	0.870	0.909
Wdbc	0.589	0.595	0.595	0.861	0.959
Wine	0.640	0.402	0.643	0.910	0.959
Wpbc	0.778	0.757	0.757	0.692	0.772
Average	0.722	0.657	0.718	0.833	0.900

From Table 4, comparing the numbers of selected attributes, the NRS, CDA, MDNRS, and ARDNE algorithms were all superior to the RS algorithm, but the ARDNE algorithm was slightly inferior to NRS, CDA, and MDNRS algorithms. From Tables 5 and 6, it is obvious that the classification accuracy of the proposed ARDNE algorithm outperformed that of the other algorithms on most of UCI data sets, except for the Wpbc data set. Furthermore, the average classification accuracy of the ARDNE was the highest and greatly improved on the SVM and KNN classifiers. For the Wpbc data set, the number of attributes selected by the ARDNE algorithm was six, which was not far from the MDNRS algorithm, and its classification accuracy was 0.8% lower than the MDNRS algorithm on KNN, while the accuracy was 8% higher than the MDNRS algorithm on SVM. Meanwhile, for the Wpbc data set, the number of attributes selected by the RS was seven, and its classification accuracy was 1% lower than the ARDNE algorithm on KNN, while the accuracy was 0.6% higher than the ARDNE algorithm on SVM. For the RS, NRS, and CDA algorithms, the classification accuracy of the Wdbc and Wine data sets were unstable. The classification accuracy of the Wine data sets only was 40.23% on the SVM classifier, and the average classification accuracy of the NRS model was the lowest. It can be obtained that the classification accuracy of the ARDNE algorithm on the SVM and KNN classifiers were relatively steady. Based on the results in Table 4, it can be seen that some important information attributes were lost in the process of reduction for the RS, NRS, and CDA algorithms, resulting in the decrease of classification accuracy of the reduced data sets with fewer attributes. The experimental results show that our attribute reduction algorithm could greatly remove the redundant attributes, and improve the classification accuracy for most of the data sets.

4.5. Classification Results of Microarray Data Sets with High-Dimensions

This subsection of our experiments continued testing the classification performance of the ARDNE algorithm, compared with those of the other three state-of-the-art entropy-based attribute reduction algorithms on the five microarray gene expression data sets with high-dimensional attributes, selected from Table 2. The algorithms used in the comparison included: (1) The mutual entropy-based attribute reduction algorithm (MEAR) [50], (2) the entropy gain-based attribute reduction algorithm (EGAR) [33], and (3) the average decision neighborhood entropy-based attribute reduction algorithm (ADNEAR) [42]. The objective of these further experiments was to show the classification power of the proposed approach to gene selection. Tables 7 and 8 show the number of selected genes and the classification accuracy of the five high-dimensional gene expression data sets with the KNN and SVM classifiers, respectively.

From Tables 7 and 8 the ARDNE algorithm obtained 93.8% and 91.9% average classification accuracy on the KNN and SVM classifiers, respectively. The classification accuracy of genes selected by the MEAR, EGAR, and ADNEAR algorithms were far lower than that with the ARDNE algorithm. For the MEAR algorithm, since the process of discretization generally results in loss of extensive useful gene information, the MEAR algorithm acquired the lower classification accuracy. For the number of selected genes, there was no significant difference among the EGAR, ADNEAR, and ARDNE algorithms. However, the classification accuracy of the ARDNE algorithm was superior to the EGAR and ADNEAR algorithms. It shows that the proposed ARDNE algorithm was able to find the most informative genes for classification. For the Colon data set, the classification accuracy of ARDNE was 80.8%, which was slightly less than that of the MEAR algorithm. So, it indicates that MDNRS algorithm was greatly affected by the data set, and the classification results were not as stable as the ARDNE algorithm. For the SRBCT data set, the classification accuracy of the ARDNE algorithm was obviously higher than those of other algorithms, and its number of selected genes only was six. The result of the further experiments shows that the proposed method had significant classification ability on the five microarray gene expression data sets.

Table 7. Classification results of the four entropy-based reduction algorithms with KNN.

Data Sets	MEAR		EGAR		ADNEAR		ARDNE	
	Genes	Accuracy	Genes	Accuracy	Genes	Accuracy	Genes	Accuracy
Brain_Tumor1	2	0.683	8	0.667	9	0.711	13	0.897
Colon	5	0.77	5	0.540	5	0.555	5	0.817
DLBCL	2	0.765	20	0.752	7	0.757	10	0.998
Leukemia	3	0.928	3	0.587	3	0.587	6	0.981
SRBCT	4	0.537	8	0.503	8	0.503	6	1
Average	3.2	0.737	8.2	0.610	6.4	0.622	8	0.938

MEAR—mutual entropy-based attribute reduction algorithm; EGAR—entropy gain-based attribute reduction algorithm; and ADNEAR—average decision neighborhood entropy-based attribute reduction algorithm.

Table 8. Classification results of the four entropy-based reduction algorithms with SVM

Data Sets	MEAR		EGAR		ADNEAR		ARDNE	
	Genes	Accuracy	Genes	Accuracy	Genes	Accuracy	Genes	Accuracy
Brain_Tumor1	2	0.691	8	0.666	9	0.666	13	0.830
Colon	5	0.849	5	0.643	5	0.643	5	0.808
DLBCL	2	0.777	20	0.862	7	0.862	10	0.993
Leukemia	3	0.920	3	0.536	3	0.536	6	0.967
SRBCT	4	0.539	8	0.535	8	0.535	6	1
Average	3.2	0.755	8.2	0.648	6.4	0.648	8	0.919

4.6. Classification Results of Dimensionality Reduction Methods on Gene Expression Data Sets

To further verify the classification performance of our proposed method, the eight methods were employed to evaluate the number of selected genes and the classification accuracy on the four gene expression data sets selected from Table 2. The ARDNE algorithm was compared with the seven related state-of-the-art dimensionality reduction methods, which included: (1) The sequential forward selection algorithm (SFS) [51], (2) the sparse group lasso algorithm (SGL) [52], (3) the adaptive sparse group lasso based on conditional mutual information algorithm (ASGL-CMI) [53], (4) the Spearman's rank correlation coefficient algorithm (SC^2) [44], (5) the gene selection algorithm based on fisher linear discriminant and neighborhood rough set (FLD-NRS) [39], (6) the gene selection algorithm based on locally linear embedding and neighborhood rough set algorithm (LLE-NRS) [40], and (7) the RelieF algorithm [41] combined with the NRS algorithm [49] (RelieF+NRS). The SVM classifier in the WEKA tool was used to do some simulation experiments. The number of selected genes and the classification accuracy are shown in Tables 9 and 10 respectively, where the symbol (–) denotes no results obtained for Leukemia using the SGL and ASGL-CMI algorithms.

Table 9. The number of selected genes of the eight reduction algorithms on the four gene expression data sets.

Data Sets	SFS	SGL	ASGL-CMI	SC^2	FLD-NRS	LLE-NRS	RelieF+NRS	ARDNE
Colon	19	55	33	4	6	16	9	5
Leukemia	7	-	-	5	6	22	17	6
Lung	3	43	32	3	3	16	23	6
Prostate	3	34	29	5	4	19	16	3
Average	8	44	31.3	4.25	4.75	18.25	16.25	5

SFS—sequential forward selection algorithm; SGL—sparse group lasso algorithm; ASGL-CMI— adaptive sparse group lasso based on conditional mutual information algorithm; SC^2—Spearman's rank correlation coefficient algorithm; FLD-NRS—gene selection algorithm based on fisher linear discriminant and neighborhood rough set; LLE-NRS—gene selection algorithm based on locally linear embedding and neighborhood rough set algorithm, and RelieF+NRS—RelieF algorithm combined with NRS algorithm.

Table 10. The classification accuracy of the eight reduction algorithms on the four gene expression data sets.

Data Sets	SFS	SGL	ASGL-CMI	SC^2	FLD-NRS	LLE-NRS	RelieF+NRS	ARDNE
Colon	0.521	0.826	0.851	0.805	0.88	0.84	0.564	0.81
Leukemia	0.969	-	-	0.852	0.828	0.868	0.563	0.967
Lung	0.833	0.827	0.841	0.806	0.889	0.907	0.919	0.987
Prostate	0.840	0.834	0.858	0.795	0.8	0.711	0.642	0.858
Average	0.791	0.829	0.85	0.815	0.849	0.832	0.672	0.898

According to the experimental results in terms of the number of selected genes and the classification accuracy in Tables 9 and 10, the differences among the eight methods could be clearly identified. For the SGL and ASGL-CMI methods, the number of selected genes was obviously higher than that the other six algorithms, and then the classification accuracy of the SGL and ASGL-CMI methods was not ideal. For some methods, such as the SFS, SC^2, FLD-NRS, and ARDNE algorithms, the average number of selected genes was less than 10. It follows that our proposed ARDNE algorithm selected fewer genes than the SFS, LLE-NRS, and RelieF+NRS algorithms, and it was roughly the same as SC^2 and FLD-NRS. For the Colon data set, the classification accuracy of the ARDNE algorithm was 81% which was slightly lower than the SGL, ASGL-CMI, FLD-NRS, and LLE-NRS methods, but for the Leukemia, Lung, and Prostate data sets, the classification accuracy of the ARDNE algorithm were 96.7%, 98.7%, and 85.8%, respectively, which were higher than the other methods. For the SFS, SC^2, LLE-NRS, and RelieF+NRS algorithms, their classification results were not as stable as the ARDNE algorithm. Thus, the classification effect of the algorithm for the four gene expression data sets would be slightly different, but the average classification ability of the ARDNE algorithm would not be affected. As for the average classification accuracy, the ARDNE algorithm obtained the highest accuracy. Therefore, our method was an efficient dimensionality reduction technique for high-dimensional, large-scale microarray data sets.

4.7. Statistical Analysis

The final part of our experiments was to further demonstrate the statistical significance of the results, and the Friedman test [53] and the Bonferroni–Dunn test [54] are employed in this paper.

The Friedman statistic is described as follows

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

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

where k is the number of algorithms, N is the number of data sets, and R_i is the average ranking of algorithm i over all the data sets. And the critical distance [55] is denoted as

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

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

In the following, based on the classification performance of the five attribute reduction algorithms in Tables 4, 11 and 12 show the rankings of the five algorithms under the KNN and SVM classifiers. The values of the different evaluation measures under the KNN and SVM classifiers are shown in Table 13. Tables 14 and 15 show the rankings of the four algorithms in Table 7 under the KNN and SVM classifiers. The values of the different evaluation measures under the KNN and SVM classifiers are shown in Table 16. Similarly, Table 17 shows the rankings of the eight attribute reduction algorithms in Table 10 under the SVM classifier. The values of the different evaluation measures under the SVM classifiers are shown in Table 18.

Table 11. Ranking of the five attribute reduction algorithms with KNN.

Data Sets	RS	NRS	CDA	MDNRS	ARDNE
Ionosphere	3	4	5	2	1
Wdbc	5	3.5	3.5	2	1
Wine	3	4	5	2	1
Wpbc	3	4.5	4.5	1	2
Average	3.5	4	4.5	1.75	1.25

Table 12. Ranking of the five attribute reduction algorithms with SVM.

Data Sets	RS	NRS	CDA	MDNRS	ARDNE
Ionosphere	2	4	3	5	1
Wdbc	5	3.5	3.5	2	1
Wine	4	5	3	2	1
Wpbc	1	3.5	3.5	5	2
Average	3	4	3.25	3.5	1.25

Table 13. F_F Values for the two classifiers.

	KNN	SVM
χ_F^2	13	7
F_F	13	2.33

F_F—Iman-Davenport test and χ_F^2—Friedman statistics.

Tables 11–18 show that the proposed ARDNE algorithm was statistically superior to the other algorithms in summary. It can be easily seen from Table 13, Table 16, and Table 18 that the values of F_F were 13 and 23.78 under the KNN classifier, respectively, and those of F_F were 2.33, 6.31, and 0.9 under the SVM classifier, respectively. When the significant level α = 0.1, the critical value of $F(4,12)$ was 2.48, $F(3,12)$ was 2.61, and $F(7,7)$ was 2.78. The critical value $q_{0.1}$= 2.241 can be found in [55], and it could be easily calculated from Equation (25) that the values of CD were 2.506, 1.83, and 1.093, respectively.

Table 14. Ranking of the four attribute reduction algorithms with KNN.

Data Sets	MEAR	EGAR	ADNEAR	ARDNE
Brain_Tumor1	3	4	2	1
Colon	2	4	3	1
DLBCL	2	4	3	1
Leukemia	2	3.5	3.5	1
SRBCT	2	3.5	3.5	1
Average	2.2	3.8	3	1

Table 15. Ranking of the four attribute reduction algorithms with SVM.

Data Sets	MEAR	EGAR	ADNEAR	ARDNE
Brain_Tumor1	2	3.5	3.5	1
Colon	1	3.5	3.5	2
DLBCL	4	2.5	2.5	1
Leukemia	2	3.5	3.5	1
SRBCT	2	3.5	3.5	1
Average	2.2	3.3	3.3	1.2

Table 16. F_F Values for the two classifiers.

	KNN	SVM
χ_F^2	12.84	9.18
F_F	23.78	6.31

Table 17. Ranking of the eight attribute reduction algorithms with SVM.

Data Sets	SFS	SGL	ASGL-CMI	SC^2	FLD-NRS	LLE-NRS	RelieF+NRS	ARDNE
Lung	6	7	5	8	4	3	2	1
Prostate	3	4	1.5	6	5	7	8	1.5
Average	4.5	5.5	3.25	7	4.5	5	5	1.25

Table 18. F_F Values for the two classifiers.

	SVM
χ_F^2	6.63
F_F	0.9

5. Conclusions

Attribute reduction is one of the important steps in data mining and classification learning. A number of measures for calculating the distinguishment ability of attribute subsets have been developed in recent years. Considering its effectiveness, neighborhood entropy is widely employed and discussed to evaluate attributes in neighborhood rough sets. In this paper, an attribute reduction method using neighborhood entropy measures in neighborhood rough sets is proposed. With the strong complementarity between the algebra definition of attribute importance and the definition of information view, some neighborhood entropy-based uncertainty measures in neighborhood decision systems are studied. Then, the significance measure is presented by combining the credibility degree with the coverage degree to analyze the classification ability of the selected attribute subset. On the basis of these theories, a heuristic attribute reduction algorithm is developed for the dimensionality reduction task to solve the practical problem. On the four UCI data sets with low-dimensional attributes and the seven microarray gene expression data sets with high-dimensional attributes, a series of experiments are carried out for verifying the effectiveness of the proposed method. The experimental results indicate that our algorithm is effective to remove the most redundant attributes without loss of classification accuracy. Comparing with the other related reduction algorithms, the reduction ability and the classification accuracy are more superior for knowledge reduction.

Author Contributions: L.S. and J.X. conceived the algorithm and designed the experiments; X.Z. implemented the experiments; S.Z. analyzed the results; X.Z. drafted the manuscript. All authors read and revised the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grants 61772176, 61402153, 61370169, and U1604154), the China Postdoctoral Science Foundation (Grant 2016M602247), the Plan for Scientific Innovation Talent of Henan Province (Grant 184100510003), the Key Scientific and Technological Project of Henan Province (Grants 182102210362 and 182102210078), the Natural Science Foundation of Henan Province (Grants 182300410130, 182300410368 and 182300410306), and the Ph.D. Research Foundation of Henan Normal University (Grant qd15132).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Islam, A.K.M.T.; Jeong, B.S.; Bari, A.T.M.G.; Lim, C.G.; Jeon, S.H. MapReduce based parallel gene selection method. *Appl. Intell.* **2015**, *42*, 147–156. [CrossRef]
2. Sun, L.; Xu, J.C.; Tian, Y. Feature selection using rough entropy-based uncertainty measures in incomplete decision systems. *Knowl. Based Syst.* **2012**, *36*, 206–216. [CrossRef]

3. Sun, L.; Liu, R.N.; Xu, J.C.; Zhang, S.G.; Tian, Y. An affinity propagation clustering method using hybrid kernel function with LLE. *IEEE Access* **2018**, *6*, 68892–68909. [CrossRef]
4. Sun, L.; Meng, X.C.; Xu, J.C.; Zhang, S.G. An image segmentation method based on improved regularized level set model. *Appl. Sci.* **2018**, *8*, 2393. [CrossRef]
5. Jing, Y.G.; Li, T.R.; Fujita, H.; Wang, B.L.; Cheng, N. An incremental attribute reduction method for dynamic data mining. *Inf. Sci.* **2018**, *465*, 202–218. [CrossRef]
6. Sun, L.; Meng, X.C.; Xu, J.C.; Tian, Y. An image segmentation method using an active contour model based on improved SPF and LIF. *Appl. Sci.* **2018**, *8*, 2576. [CrossRef]
7. Sun, L.; Wang, L.Y.; Xu, J.C.; Zhang, S.G. A neighborhood rough sets-based attribute reduction method using Lebesgue and entropy measures. *Entropy* **2019**, *21*, 138. [CrossRef]
8. Fang, Y.; Min, F. Cost-sensitive approximate attribute reduction with three-way decisions. *Int. J. Approx. Reason.* **2019**, *104*, 148–165. [CrossRef]
9. Sang, B.B.; Guo, Y.T.; Shi, D.R.; Xu, W.H. Decision-theoretic rough set model of multi-source decision systems. *Int. J. Mach. Learn. Cybern.* **2018**, *9*, 1941–1954. [CrossRef]
10. Mi, J.S.; Wu, W.Z.; Zhang, W.X. Approaches to knowledge reduction based on variable precision rough set model. *Inf. Sci.* **2004**, *159*, 255–272. [CrossRef]
11. Syau, Y.R.; Lin, E.B.; Liau, C.J. Neighborhood systems and variable precision generalized rough sets. *Fundam. Inform.* **2017**, *153*, 271–290. [CrossRef]
12. Shannon, C.E. The mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 373–423. [CrossRef]
13. Gao, C.; Lai, Z.H.; Zhou, J.; Zhao, C.R.; Miao, D.Q. Maximum decision entropy-based attribute reduction in decision-theoretic rough set model. *Knowl. Based Syst.* **2018**, *143*, 179–191. [CrossRef]
14. Dai, J.H.; Hu, H.; Zheng, G.J.; Hu, Q.H.; Han, H.F.; Shi, H. Attribute reduction in interval-valued information systems based on information entropies. *Front. Inf. Technol. Electron. Eng.* **2016**, *17*, 919–928. [CrossRef]
15. Wang, G.Y. Rough reduction in algebra view and information view. *Int. J. Intell. Syst.* **2003**, *18*, 679–688. [CrossRef]
16. Wang, C.Z.; Qi, Y.L.; Shao, M.W.; Hu, Q.H.; Chen, D.G.; Qian, Y.H.; Lin, Y.J. A fitting model for feature selection with fuzzy rough sets. *IEEE Trans. Fuzzy Syst.* **2017**, *25*, 741–753. [CrossRef]
17. Sun, L.; Xu, J.C. Information entropy and mutual information-based uncertainty measures in rough set theory. *Appl. Math. Inf. Sci.* **2014**, *8*, 1973–1985. [CrossRef]
18. Hu, Q.H.; Yu, D.R.; Liu, J.F.; Wu, C.X. Neighborhood rough set based heterogeneous feature subset selection. *Inf. Sci.* **2008**, *178*, 3577–3594. [CrossRef]
19. Dai, J.H.; Hu, H.; Wu, W.Z.; Qian, Y.H.; Huang, D.B. Maximal discernibility pair based approach to attribute reduction in fuzzy rough sets. *IEEE Trans. Fuzzy Syst.* **2018**, *26*, 2174–2187. [CrossRef]
20. Luo, C.; Li, T.R.; Chen, H.M.; Fujita, H.; Yi, Z. Incremental rough set approach for hierarchical multicriteria classification. *Inf. Sci.* **2018**, *429*, 72–87. [CrossRef]
21. Lin, Y.J.; Li, Y.W.; Wang, C.X.; Chen, J.K. Attribute reduction for multi-label learning with fuzzy rough set. *Knowl. Based Syst.* **2018**, *152*, 51–61. [CrossRef]
22. Zhang, X.; Mei, C.L.; Chen, D.G.; Yang, Y.Y. A fuzzy rough set-based feature selection method using representative instances. *Knowl. Based Syst.* **2018**, *151*, 216–229. [CrossRef]
23. Peters, J.F.; Wasilewski, P. Tolerance spaces: Origins, theoretical aspects and applications. *Inf. Sci.* **2012**, *195*, 211–225. [CrossRef]
24. Slowinski, R.; Vanderpooten, D. A generalized definition of rough approximations based on similarity. *IEEE Trans. Knowl. Data Eng.* **2000**, *12*, 331–336. [CrossRef]
25. Yue, X.D.; Chen, Y.F.; Miao, D.Q.; Fujita, H. Fuzzy neighborhood covering for three-way classification. *Inf. Sci.* **2018**. [CrossRef]
26. El Barbary, O.G.; Salama, A.S.; Atlam, E.S. Granular information retrieval using neighborhood systems. *Math. Methods Appl. Sci.* **2018**, *41*, 5737–5753. [CrossRef]
27. Wang, Q.; Qian, Y.H.; Liang, X.Y.; Guo, Q.; Liang, J.Y. Local neighborhood rough set. *Knowl. Based Syst.* **2018**, *153*, 53–64. [CrossRef]
28. Fan, X.D.; Zhao, W.D.; Wang, C.Z.; Huang, Y. Attribute reduction based on max-decision neighborhood rough set model. *Knowl. Based Syst.* **2018**, *151*, 16–23. [CrossRef]

29. Chen, H.M.; Li, T.R.; Cai, Y.; Luo, C.; Fujita, H. Parallel attribute reduction in dominance-based neighborhood rough set. *Inf. Sci.* **2016**, *373*, 351–368. [CrossRef]
30. Wang, G.Y. *Rough Set Theory and Knowledge Acquisition*; Xi'an Jiaotong University Press: Xi'an, China, 2001.
31. Tsumoto, S. Accuracy and coverage in rough set rule induction. In Proceedings of the International Conference on Rough Sets and Current Trends in Computing, Malvern, PA, USA, 14–16 October 2002; Springer: Berlin/Heidelberg, Germnay, 2002; pp. 373–380.
32. Sun, L.; Xu, J.C.; Xue, Z.A.; Ren, J.Y. Decision degree-based decision tree technology for rule extraction. *J. Comput.* **2012**, *7*, 1769–1779. [CrossRef]
33. Chen, Y.M.; Zhang, Z.J.; Zheng, J.Z.; Ma, Y.; Xue, Y. Gene selection for tumor classification using neighborhood rough sets and entropy measures. *J. Biomed. Inform.* **2017**, *67*, 59–68. [CrossRef]
34. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. [CrossRef]
35. Sun, L.; Xu, J.C.; Xu, T.H. Information entropy and information granulation-based uncertainty measures in incomplete information systems. *Appl. Math. Inf. Sci.* **2014**, *8*, 2073–2083. [CrossRef]
36. Hancer, E.; Xue, B.; Zhang, M.J. Differential evolution for filter feature selection based on information theory and feature ranking. *Knowl. Based Syst.* **2018**, *140*, 103–119. [CrossRef]
37. Hu, Q.H.; Pan, W.; An, S.; Ma, P.J.; Wei, J.M. An efficient gene selection technique for cancer recognition based on neighborhood mutual information. *Int. J. Mach. Learn. Cybern.* **2010**, *1*, 63–74. [CrossRef]
38. Friedman, M. A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Stat.* **1940**, *11*, 86–92. [CrossRef]
39. Sun, L.; Zhang, X.Y.; Xu, J.C.; Wang, W.; Liu, R.N. A gene selection approach based on the fisher linear discriminant and the neighborhood rough set. *Bioengineered* **2018**, *9*, 144–151. [CrossRef]
40. Sun, L.; Xu, J.C.; Wang, W.; Yin, Y. Locally linear embedding and neighborhood rough set-based gene selection for gene expression data classification. *Genet. Mol. Res.* **2016**, *15*, 15038990. [CrossRef]
41. Urbanowicz, R.J.; Meeker, M.; La Cava, W.; Olsona, R.S.; Moore, J.H. Relief-based feature selection: Introduction and review. *J. Biomed. Inform.* **2018**, *85*, 189–203. [CrossRef]
42. Sun, L.; Zhang, X.Y.; Qian, Y.H.; Xu, J.C.; Zhang, S.G.; Tian, Y. Joint neighborhood entropy-based gene selection method with fisher score for tumor classification. *Appl. Intell.* **2018**. [CrossRef]
43. Huang, X.J.; Zhang, L.; Wang, B.J.; Li, F.Z.; Zhang, Z. Feature clustering based support vector machine recursive feature elimination for gene selection. *Appl. Intell.* **2018**, *48*, 594–607. [CrossRef]
44. Xu, J.C.; Mu, H.Y.; Wang, Y.; Huang, F.Z. Feature genes selection using supervised locally linear embedding and correlation coefficient for microarray classification. *Comput. Math. Methods Med.* **2018**, *2018*, 5490513. [CrossRef] [PubMed]
45. Wang, C.Z.; Hu, Q.H.; Wang, X.Z.; Chen, D.G.; Qian, Y.H.; Dong, Z. Feature selection based on neighborhood discrimination index. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 2986–2999. [CrossRef] [PubMed]
46. Dong, H.B.; Li, T.; Ding, R.; Sun, J. A novel hybrid genetic algorithm with granular information for feature selection and optimization. *Appl. Soft Comput.* **2018**, *65*, 33–46. [CrossRef]
47. Sun, S.Q.; Peng, Q.K.; Zhang, X.K. Global feature selection from microarray data using Lagrange multipliers. *Knowl. Based Syst.* **2016**, *110*, 267–274. [CrossRef]
48. Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 6567–6572. [CrossRef] [PubMed]
49. Yang, X.B.; Zhang, M.; Dou, H.L.; Yang, J.Y. Neighborhood systems-based rough sets in incomplete information system. *Knowl. Based Syst.* **2011**, *24*, 858–867. [CrossRef]
50. Xu, F.F.; Miao, D.Q.; Wei, L. Fuzzy-rough attribute reduction via mutual information with an application to cancer classification. *Comput. Math. Appl.* **2009**, *57*, 1010–1017. [CrossRef]
51. Reunanen, J. Overfitting in making comparisons between variable selection methods. *J. Mach. Learn. Res.* **2003**, *3*, 1371–1382.
52. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. A sparse-group lasso. *J. Comput. Graph. Stat.* **2013**, *22*, 231–245. [CrossRef]
53. Li, J.T.; Dong, W.P.; Meng, D.Y. Grouped gene selection of cancer via adaptive sparse group lasso based on conditional mutual information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**. [CrossRef] [PubMed]

54. Dunn, O.J. Multiple comparisons among means. *J. Am. Stat. Assoc.* **1961**, *56*, 52–64. [CrossRef]
55. Demsar, J.; Schuurmans, D. Statistical comparison of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.