International Journal of Approximate Reasoning 187 (2025) 109539

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

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

Attribute reduction based on weighted neighborhood distribution entropy

Peng Xu [a], Ping Zhu [a,b,*]

[a] School of Mathematical Sciences, Beijing University of Posts and Telecommunications, Beijing 100876, China
[b] Key Laboratory of Mathematics and Information Networks (Beijing University of Posts and Telecommunications), Ministry of Education, China

ARTICLE INFO

Keywords:
Attribute reduction
Neighborhood rough sets
Label distribution
Uncertainty measure

ABSTRACT

Neighborhood rough sets theory is a crucial tool for processing numerical data, and many studies have combined it with information entropy to address attribute reduction tasks. However, most related algorithms do not fully consider the critical role of the label distribution information in evaluating the uncertainty of attribute subsets, which may lead to their poor adaptability to different datasets. In this paper, we propose an attribute reduction algorithm based on weighted neighborhood distribution entropy to fully capture the uncertainty information caused by label distribution. Firstly, the concepts of neighborhood granularity and granular chaos are proposed based on the prior and posterior evaluation of the label distribution within neighborhood granules. Then, we construct a monotonic neighborhood distribution entropy, which is composed of neighborhood granularity and granular chaos, and can effectively measure the uncertainty of attribute subsets. Finally, we introduce the Davies-Bouldin Index to measure the label distribution between decision classes. Using it as a weight for the neighborhood distribution entropy, we define the significance of candidate attributes and design the corresponding greedy forward attribute reduction algorithm. Comparative experiments on 16 public datasets demonstrate the superiority and effectiveness of the proposed algorithm.

1. Introduction

In the era of big data, massive amounts of data provide rich information in various fields. However, as the dimension continues to increase, the computational complexity continues to increase, eventually leading to the curse of dimensionality. Attribute reduction, as an effective technology to solve this problem, has attracted the attention of many scholars. It can maintain or approach the distribution of the original data while removing redundant attributes. Due to its excellent dimensionality reduction ability, it has been widely used in various fields such as image processing [1,2], bioinformatics [3,4] and machine learning [5–7].

Rough sets theory is a data analysis theory proposed by Pawlak [8], which is an important mathematical tool for studying imprecise or incomplete information. The core idea of rough sets is to construct the upper and lower approximation operators of the sample sets [9]. Due to its excellent interpretability and the fact that it does not require any prior information, currently, rough sets have been widely applied in fields such as data mining [10,11], medical diagnosis [12,13], rule extraction [14,15], and so on. Among them, the most widely used application is attribute reduction [16–18]. However, the classical rough sets model requires discretization

* Corresponding author at: School of Mathematical Sciences, Beijing University of Posts and Telecommunications, Beijing 100876, China.
E-mail addresses: xupeng23@bupt.edu.cn (P. Xu), pzhubupt@bupt.edu.cn (P. Zhu).

https://doi.org/10.1016/j.ijar.2025.109539
Received 10 January 2025; Received in revised form 14 July 2025; Accepted 30 July 2025

Available online 6 August 2025

before processing continuous data, which may lead to information loss. To address this issue, some scholars introduced neighborhood relationship into rough sets and formed the neighborhood rough sets theory [19–22].

Hu et al. [23] introduced the neighborhood rough sets model into the attribute reduction task in 2008. This model generalizes the equivalence relationship to a neighborhood relationship by introducing the distance metric. When the neighborhood radius is 0, the neighborhood rough sets model will degenerate into the classic rough sets model. Thereafter, many scholars conducted further research on attribute reduction based on neighborhood rough sets. The traditional model considers the positive region composed of the lower approximation. Fan et al. [24] noticed that the boundary region may also contain classification information, and they expanded the positive region by adding samples that are closest to the decision attribute. Subsequently, Hu et al. [25] constructed the weighted neighborhood rough sets model. They considered that different attributes have different importance for decision attribute, and assigned corresponding weights to the attributes by solving a optimization problem. Considering the neighborhood radius is uniformly given, Zhang et al. [26] designed a variable radius neighborhood rough sets model. They focused on the location of each sample and assigned different neighborhood radius to each sample according to the surrounding environment of them. However, this process increases the computational resource consumption of the algorithm. To enhance the efficiency of algorithms, Liu et al. [27] defined ordered buckets based on hashing methods and organized the samples into them. Compared with [23], only the samples in the adjacent buckets need to be considered when constructing the neighborhood granules. Wang et al. [28] combined local rough sets with neighborhood rough sets and proposed the concept of local neighborhood rough sets. The model only considers the samples in the decision class when calculating the lower approximation of the decision class, which makes the algorithm have a linear time complexity. In addition, Xia et al. [29] used rectangular-based neighborhood granules instead of the distance-based neighborhood granules. The new model can effectively reduce the search space of neighborhood radius. The above reduction algorithms are all designed based on the positive region, which may lead to ignoring the impact of other non-positive samples on the uncertainty of attribute subsets.

Currently, scholars have introduced uncertainty measures into neighborhood rough sets model for attribute reduction [30–33]. Information entropy, as an important tool for data processing, comprehensively considers the impact of all samples on the uncertainty of attribute subsets. In 2011, Hu et al. [34] combined the information entropy with neighborhood relations and proposed neighborhood mutual information. It is a natural extension of the Shannon entropy in neighborhood decision system. Wang et al. [35] proposed neighborhood discrimination index with Shannon entropy properties to measure the importance of attribute subsets. The neighborhood discrimination index is defined based on the cardinality of neighborhood relations rather than neighborhood similarity classes. Zhang et al. [36] proposed a heterogeneous attribute reduction algorithm based on the conditional neighborhood combination entropy, which reflects the probability that neighborhood granules can be distinguished from each other. However, these measures lack monotonicity when measuring the uncertainty of attribute subsets, which may lead to unreasonable reduction subsets. Wang et al. [37] considered the classification information in the upper approximation and proposed an attribute reduction algorithm based on neighborhood self-information. Furthermore, Ji et al. [38] proposed fusion information entropy, which can dynamically adjust the contribution of the upper and lower approximations to the decision. Although these uncertainty measures have better interpretability, the limitation of them is that the constructed entropies do not consider the label distribution information. Different datasets have different distribution characteristics, which restricts the generalization ability of these entropy measures across various datasets.

Recently, Dai et al. [39] defined neighborhood complementary entropy based on the mutual inclusion of samples between neighborhood granules. However, they ignore the label distribution information between decision classes. For imbalanced datasets, the reduction subsets selected by the algorithm may ignore the attributes that are discriminative for the minority classes. In addition, Xu et al. [40] focused on the decision distribution information of samples and proposed a composite entropy. However, for datasets with highly fuzzy decision boundaries, different attributes have tiny impact on the overall distribution. It is difficult for the algorithm to select optimal attribute subsets. As far as we know, the existing entropy measures do not simultaneously consider the distribution information at both the granule and decision class levels, which still limit the algorithm's generalization performance on specific types of datasets. Therefore, the motivation of this study is to develop a measure that deeply mines label distribution information within the neighborhood granules and between decision classes to further improve the generalization ability across different datasets.

Specifically, we define the evaluation measures neighborhood granularity and granular chaos based on the neighborhood rough sets model. They quantify the uncertainty of the corresponding attribute subsets by measuring the label distribution information within neighborhood granules from the perspective of prior and posterior respectively. Based on this, we introduce the concept of neighborhood distribution entropy (NDE) and prove its monotonicity with respect to attribute subsets. Similar to traditional conditional entropy, a lower NDE value indicates a stronger ability of the attribute subset to distinguish samples between different decision classes. Subsequently, we introduce the clustering evaluation index Davies-Bouldin Index (DBI) as a weighting factor for NDE to define the significance of an attribute, and from this foundation propose a greedy forward attribute reduction algorithm ARWNDE. Notably, DBI evaluates the label distribution of samples from a global perspective, which enables ARWNDE to select attribute subsets with clearer decision boundaries and higher classification accuracy. The experimental results show that the proposed method can more accurately quantify the uncertainty of attribute subsets by fully mining the label distribution information. The main contributions of the paper can be summarized as follows:

(1) We define the measures to measure the distribution information of all class labels within neighborhood granules, including neighborhood granularity and granular chaos, and propose a monotonic attribute subsets measure NDE by combining them;
(2) We introduce the clustering evaluation index DBI to measure the distribution between decision classes. We use it as a weighting factor for NDE and propose the significance of attributes.

(3) We propose the greedy forward attribute reduction algorithm ARWNDE. Experimental results on 16 public datasets and 9 comparison algorithms show that ARWNDE performs significantly better than its competitors.

The remainder of this paper is organized as follows. In Section 2, we briefly review the concepts related to this paper. Then we construct the neighborhood distribution entropy to measure the uncertainty of attribute subsets in Section 3. In Section 4, we design an attribute reduction algorithm ARWNDE based on the weighted neighborhood distribution entropy. In Section 5, we verify the effectiveness and robustness of our algorithm using public datasets. Finally, we summarize our findings and discuss future work in Section 6.

2. Preliminaries

2.1. Neighborhood rough sets model

The neighborhood rough sets can be regarded as an extension of the classical rough sets. This section mainly introduces the neighborhood rough sets model proposed by Hu et al. [23].

Formally, a dataset can be represented as an information system, denoted as $IS = \langle U, A \rangle$, where $U = \{x_1, x_2, \ldots, x_n\}$ represents a sample set called a universe, A is a set of attributes used to describe the samples in U. An IS can be regarded as a decision system (DS) if $A = C \cup D$, where C represents the set of condition attributes and D represents the decision attribute.

Given an attribute subset $B \subseteq C$, a mapping d_B from $U \times U$ to the set of nonnegative real numbers $\mathbf{R}^+$ is called a distance function if for $\forall x, y, z \in U$, it satisfies:

(1) Nonnegative: $d_B(x, y) \geq 0$, $d_B(x, x) = 0$;
(2) Symmetry: $d_B(x, y) = d_B(y, x)$;
(3) Triangle inequality: $d_B(x, z) \leq d_B(x, y) + d_B(y, z)$.

For $\forall x \in U$ and $\forall a \in A$, let $a(x)$ represent the value of sample x under attribute a. Given $\forall B \subseteq A$, the distance function d_B can be written in the form of the Minkowski distance, denoted as

$$d_B(x, y) = \left(\sum_{a \in B} |a(x) - a(y)|^p \right)^{1/p}. \tag{1}$$

Specifically, when $p = 1$, d_B is called the Manhattan distance; when $p = 2$, it is called the Euclidean distance, and when $p = \infty$, it is called the Chebyshev distance.

Given $\forall B \subseteq A$ and a parameter $\delta \geq 0$, then a neighborhood relation N_B^δ on U is represented by a relation matrix $M(N_B^\delta) = (r_{ij}^B)_{n \times n}$, where

$$r_{ij}^B = \begin{cases} 1, & d_B(x_i, x_j) \leq \delta \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

It is obvious that the matrix $M(N_B^\delta)$ satisfies:

(1) Reflexivity: $r_{ii}^B = 1$;
(2) Symmetry: $r_{ij}^B = r_{ji}^B$.

A DS becomes a neighborhood decision system (NDS) if a neighborhood relation is defined on it.

Definition 2.1 *([23]).* Given $NDS = \langle U, C, D \rangle$, for any $B \subseteq C$, $x \in U$ and $\delta > 0$, the parameterized neighborhood granule of x induced by B is defined as

$$\delta_B(x) = \{ y \mid d_B(x, y) \leq \delta \}. \tag{3}$$

$\delta_B(x_i)$ represents the neighborhood granule generated by x_i, which consists of samples at a distance from x_i less than or equal to δ.

Proposition 2.1 *([35]). Given $NDS = \langle U, C, D \rangle$, $B_1 \subseteq B_2 \subseteq C$, and $\delta > 0$. For $\forall x \in U$, if the distance between samples in U calculated using Euclidean distance, then $\delta_{B_2}(x) \subseteq \delta_{B_1}(x)$.*

According to Proposition 2.1, as the number of attributes increases, the samples within the neighborhood granules will decrease.

Definition 2.2 *([23]).* Given $NDS = \langle U, C, D \rangle$, $B \subseteq C$, $\delta > 0$, and $U/D = \{D_1, D_2, \ldots, D_m\}$ represents m equivalence classes formed by the partitioning of U under D. Then, the lower and the upper approximations of D under B are defined as

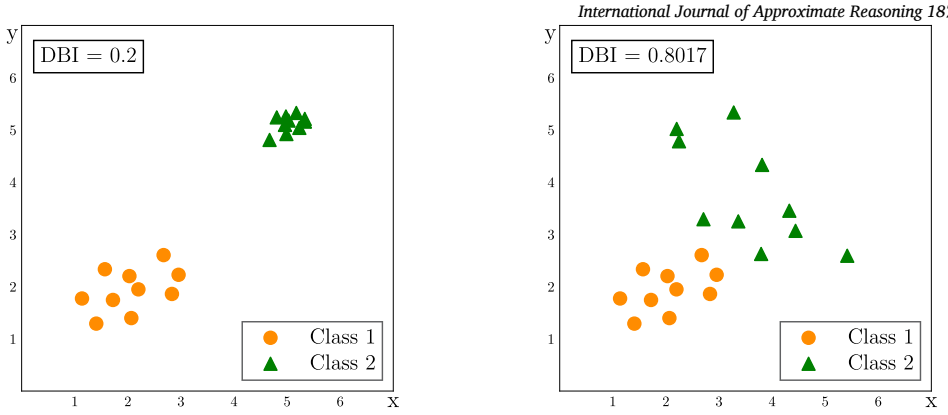

Fig. 1. Different distribution types of datasets with different DBI values. (a) High intra-class compactness and high inter-class separation. (b) Low intra-class compactness and low inter-class separation.

$$\underline{N_B^\delta} D = \bigcup_{i=1}^{m} \underline{N_B^\delta} D_i; \tag{4}$$

$$\overline{N_B^\delta} D = \bigcup_{i=1}^{m} \overline{N_B^\delta} D_i, \tag{5}$$

where $\underline{N_B^\delta} D_i = \{x \in U \mid \delta_B(x) \subseteq D_i\}$ and $\overline{N_B^\delta} D_i = \{x \in U \mid \delta_B(x) \cap D_i \neq \varnothing\}$. The positive region is defined as $POS_B(D) = \underline{N_B^\delta} D$. According to Definition 2.2, x_i will be classified into the positive region when the labels of samples in $\delta_B(x_i)$ are the same. The larger the positive region, the lower the uncertainty of the current attribute subset.

2.2. Davies-Bouldin Index

Most neighborhood rough sets models focus on the uncertainty at the granular level. As a clustering evaluation index, Davies-Bouldin Index (DBI) can reflect the distribution of data from an overall perspective. Therefore, this paper introduces DBI to better evaluate the uncertainty of attribute subsets.

Definition 2.3 *([41])*. Given $NDS = \langle U, C, D \rangle$, $B \subseteq C$, let c_j^B denote the center of the samples in D_j under B. Then, the Davies-Bouldin Index of U under B is defined as

$$I_B = \frac{1}{m} \sum_{i=1}^{m} \max_{j \neq i} \frac{s_i^B + s_j^B}{\|c_i^B - c_j^B\|_2}, \tag{6}$$

where $s_i^B = \frac{1}{|D_i|} \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - c_i^B\|_2$ represents the average Euclidean distance between the samples in D_i and c_j^B under B.

The value of DBI can reflect the degree of data separation [42]. As shown in Fig. 1, a small DBI value corresponds to data with high separation. In other words, the distribution of labels between decision classes is well separated.

3. Uncertainty measures based on label distribution

In this section, we propose the concepts of neighborhood granularity and granular chaos. Based on them, we define neighborhood distribution entropy, which quantifies the uncertainty of attribute subsets using all label information within granules.

3.1. Neighborhood granularity and granular chaos

To some extent, the discernibility of an IS can be reflected by the value of knowledge granularity [43]. Inspired by [43], the concept of knowledge granularity is extended to NDS.

Definition 3.1. Given $NDS = \langle U, C, D \rangle$, $B \subseteq C$ and $\delta > 0$, then the neighborhood granularity of B on U is defined as

$$NG_B^\delta = \sum_{i=1}^{n} \left(P_B^\delta(x_i) \right)^2, \tag{7}$$

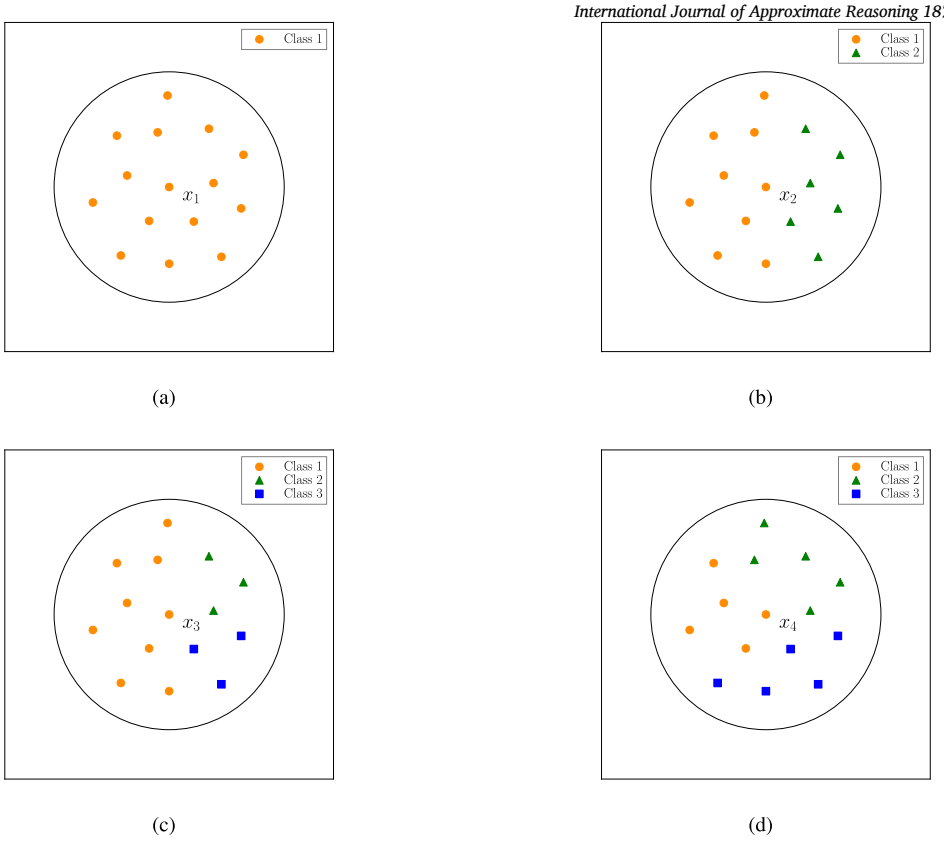

Fig. 2. The local scatter plot of four neighborhood decision systems.

where $P_B^{\delta}(x_i) = |\delta_B(x_i)|/|U|$ represents the proportion of the cardinality of $\delta_B(x_i)$ to the cardinality of U.

Obviously, when all neighborhood granules only contain one sample, the neighborhood granularity reaches its minimum of $1/|U|$; when all neighborhood granules cover U, the neighborhood granularity reaches its maximum of $|U|$. Beyond these cases, if a neighborhood granule contains fewer samples, they are more likely to have the same label from a prior perspective. According to Definition 2.2, the center of the granule is more likely to belong to the positive region. It can be considered that smaller neighborhood granularity may correspond to a larger positive region and a lower uncertainty of the attribute subset.

The neighborhood granularity performs preliminary prior estimation of the quality of attribute subsets without using label information, which may lead to inaccurate results. In fact, there may exist a granule containing fewer samples, whose center does not belong to positive region due to the presence of heterogeneous label samples. Therefore, from the perspective of posterior estimation considering the distribution of all labels within a single granule, the granular chaos is defined as follows.

Definition 3.2. Given $NDS = \langle U, C, D\rangle$, $B \subseteq C$, $\delta > 0$, and $U/D = \{D_1, D_2, \ldots, D_m\}$, then the granular chaos of $\delta_B(x_i)$ under D is defined as

$$GC_B^{\delta}(D|x_i) = -\sum_{j=1}^{m} \frac{|\delta_B(x_i) \cap D_j|}{|\delta_B(x_i)|} \log \frac{|\delta_B(x_i) \cap D_j|}{|\delta_B(x_i)|}. \tag{8}$$

The granular chaos uses all label information to provide a more reasonable posterior evaluation of the label distribution within the granule. Next, we will demonstrate the calculation process of granular chaos through an example and analyze its rationality.

Example 3.1. The local scatter plots of four neighborhood decision systems are shown in Fig. 2. For x_3 in Fig. 2(c), the circle represents its neighborhood range. From the figure we can get $|\delta_{B_3}(x_3)| = 15$, and the number of samples in decision class 1, class 2, and class 3 in the granule is 9, 3, and 3, respectively. Therefore, the granular chaos $GC_{B_3}^{\delta}(D|x_3) = -\frac{9}{15}\log\frac{9}{15} - \frac{3}{15}\log\frac{3}{15} - \frac{3}{15}\log\frac{3}{15} = 0.951$. Similarly, the granular chaos of $\delta_{B_1}(x_1)$, $\delta_{B_2}(x_2)$ and $\delta_{B_4}(x_4)$ is 0, 0.673, and 1.099, respectively. From Fig. 2, we can infer that x_1 is completely in the decision positive region and has the smallest uncertainty, while the label distribution in $\delta_{B_2}(x_2)$, $\delta_{B_3}(x_3)$ and $\delta_{B_4}(x_4)$ is becoming more and more chaotic, which indicates that x_2, x_3 and x_4 may progressively approach the decision boundary. The increasing uncertainty is consistent with their granular chaos values.

Proposition 3.1. *Given* $NDS = \langle U, C, D\rangle$, $B \subseteq C$, $\delta > 0$, *and* $U/D = \{D_1, D_2, \ldots, D_m\}$, *then* $0 \le GC_B^\delta(D|x_i) \le \log m$.

Proof. When all samples in $\delta_B(x_i)$ have the same decision attribute assumed to be D_p, then $\delta_B(x_i) \cap D_p = \delta_B(x_i)$ and $\delta_B(x_i) \cap D_q = \varnothing$ $(q \neq p)$. So we have

$$GC_B^\delta(D|x_i) = -\frac{|\delta_B(x_i) \cap D_p|}{|\delta_B(x_i)|} \log \frac{|\delta_B(x_i) \cap D_p|}{|\delta_B(x_i)|} - \sum_{q \neq p} \frac{|\delta_B(x_i) \cap D_q|}{|\delta_B(x_i)|} \log \frac{|\delta_B(x_i) \cap D_q|}{|\delta_B(x_i)|} = 0.$$

Let $\mu_{ij} = |\delta_B(x_i) \cap D_j|/|\delta_B(x_i)|$, we have $\sum_{j=1}^m \mu_{ij} = 1$. Under this constraint, we seek the maximum value of $GC_B^\delta(D|x_i)$. Using the Lagrange multiplier method, the converted unconstrained objective function is as follows:

$$L(\mu_{i1}, \mu_{i2}, \ldots, \mu_{im}, \lambda_i) = -\sum_{j=1}^m \mu_{ij} \log \mu_{ij} + \lambda_i \left(\sum_{j=1}^m \mu_{ij} - 1 \right),$$

where λ_i is the Lagrange multiplier. To find the stationary point of the function, we calculate the partial derivatives of L with respect to μ_{ij} and λ_i:

$$\begin{cases} \partial L/\partial \mu_{ij} = -\log \mu_{ij} - 1 + \lambda_i; \\ \partial L/\partial \lambda_i = \sum_{j=1}^m \mu_{ij} - 1. \end{cases}$$

Setting $[\partial L/\partial \mu_{ij}] = 0$ and $[\partial L/\partial \lambda_i] = 0$, we can obtain $\mu_{ij} = \frac{1}{m}$. Bringing it into (8), the maximum value of $GC_B^\delta(D|x_i)$ is $\log m$.

Remark 3.1. In the granule $\delta_B(x_i)$, a highly homogeneous label distribution indicates that x_i is far from the decision boundary, carrying lower uncertainty. In particular, when the labels of samples in $\delta_B(x_i)$ are exactly the same, x_i carries the lowest uncertainty. According to Proposition 3.1, $GC_B^\delta(D|x_i)$ reaches its minimum value of 0. On the contrary, when the labels of samples in $\delta_B(x_i)$ are chaotically distributed, it means x_i is close to the decision boundary, carrying higher uncertainty. Specifically, when the labels of samples in $\delta_B(x_i)$ are completely chaotic in distribution, x_i carries the highest uncertainty. According to Proposition 3.1, $GC_B^\delta(D|x_i)$ reaches its maximum value of $\log m$. In summary, granular chaos uses the label information within granules to infer the uncertainty caused by the degree to which samples deviate from the decision boundary.

3.2. Neighborhood distribution entropy

By combining neighborhood granularity and granular chaos, the neighborhood distribution entropy can be defined as follows. It can comprehensively measure the uncertainty of attribute subsets.

Definition 3.3. Given $NDS = \langle U, C, D\rangle$, $B \subseteq C$, $\delta > 0$, and $U/D = \{D_1, D_2, \ldots, D_m\}$, then the neighborhood distribution entropy of D under B is defined as

$$H_B^\delta(D) = \sum_{i=1}^n \left(P_B^\delta(x_i) \right)^2 GC_B^\delta(D|x_i). \tag{9}$$

Note the formula of neighborhood granularity NG_B^δ in (7). The neighborhood distribution entropy $H_B^\delta(D)$ can be formally regarded as multiplying each component of NG_B^δ by the corresponding sample's granular chaos $GC_B^\delta(D|x_i)$. The cardinality of a granule estimates a prior whether its center belongs to decision positive region without using label information, and the granular chaos infers a posterior the extent to which its center deviates from the decision boundary using all label information in the granule. Therefore, neighborhood distribution entropy (NDE) measures the uncertainty of corresponding attribute subset by comprehensively measuring the label distribution of neighborhood granules.

Proposition 3.2. *Given* $NDS = \langle U, C, D\rangle$, $B \subseteq C$, $\delta > 0$, *and* $U/D = \{D_1, D_2, \ldots, D_m\}$, *then* $0 \le H_B^\delta(D) \le |U| \log |U|$.

Proof. When all neighborhood granules contain no heterogeneous samples, according to Proposition 3.1 and (9), we have $GC_B^\delta(D|x_i) = 0$ and thus $H_B^\delta(D) = 0$. In other case, we have

$$H_B^\delta(D) = \sum_{i=1}^n \left(P_B^\delta(x_i) \right)^2 GC_B^\delta(D|x_i) \le \sum_{i=1}^n \frac{|\delta_B(x_i)|^2}{|U|^2} \times \log m \le |U| \log |U|.$$

The equality in last inequality holds when the following conditions are met: (1) $m = |U|$, i.e., all samples in U have different labels. (2) $|\delta_B(x_i)| = |U|$, meaning the neighborhood granule of every sample can cover U.

To prove that $H_B^\delta(D)$ is monotonic with respect to attribute subsets, we first prove the following lemma.

Table 1
Neighborhood decision system.

U	a_1	a_2	a_3	d	U	a_1	a_2	a_3	d
x_1	0.41	0.60	0.50	1	x_{10}	0.76	0.76	0.46	2
x_2	0.46	0.54	0.52	1	x_{11}	0.67	0.67	0.35	2
x_3	0.25	0.25	0.55	1	x_{12}	0.60	0.60	0.34	2
x_4	0.24	0.24	0.46	1	x_{13}	0.50	0.50	0.68	3
x_5	0.29	0.29	0.39	1	x_{14}	0.43	0.43	0.78	3
x_6	0.33	0.33	0.35	1	x_{15}	0.62	0.62	0.73	3
x_7	0.54	0.46	0.52	2	x_{16}	0.56	0.56	0.80	3
x_8	0.60	0.40	0.50	2	x_{17}	0.40	0.40	0.68	3
x_9	0.76	0.76	0.55	2	x_{18}	0.58	0.58	0.68	3

Lemma 3.1. *The function* $f(x,y)=(x+y)x\log\frac{x}{x+y}$ *decreases monotonically with respect to* $x>0$ *and* $y>0$.

Proof. Notice that $f_y = x(\log\frac{x}{x+y}-1)$, and $f_y<0$ holds for $\forall x,y>0$. Next we prove that f_x has the same property as f_y.

The inequality $f_x=(2x+y)\log\frac{x}{x+y}+y<0$ is equivalent to $\log\frac{x}{x+y}<-\frac{y}{2x+y}$. Let $v=x$ and $u=x+y$, we need to prove $\log\frac{v}{u}<\frac{v-u}{v+u}=\frac{2v}{u+v}-1$. Note that when $u>v>0$, $\frac{v}{u}=\frac{2v}{u+u}<\frac{2v}{u+v}$, so we only need to prove $\log\frac{v}{u}<\frac{v}{u}-1$. Let $t=\frac{u}{v}$ and $g(t)=\log t-t+1$, and now we need to prove that $g(t)<0$ when $0<t<1$. Since $g'(t)=\frac{1}{t}-1>0$, we have $g(t)<g_{max}=g(1)=0$. So we complete the proof.

Theorem 3.1. *Given* $NDS=\langle U,C,D\rangle$, $B_1\subseteq B_2\subseteq C$, $\delta>0$, *and* $U/D=\{D_1,D_2,\dots,D_m\}$, *then* $H^{\delta}_{B_1}(D)\geq H^{\delta}_{B_2}(D)$.

Proof. Let $H^{\delta}_{B_1}(D)-H^{\delta}_{B_2}(D)$ be denoted as ΔH. According to (9), we have

$$\begin{aligned}\Delta H&=\sum_{i=1}^{n}\left(P^{\delta}_{B_1}(x_i)\right)^2 GC^{\delta}_{B_1}(D|x_i)-\sum_{i=1}^{n}\left(P^{\delta}_{B_2}(x_i)\right)^2 GC^{\delta}_{B_2}(D|x_i)\\&=\sum_{i=1}^{n}\sum_{j=1}^{m}\left(|\delta_{B_2}(x_i)||\delta_{B_2}(x_i)\cap D_j|\log\frac{|\delta_{B_2}(x_i)\cap D_j|}{|\delta_{B_2}(x_i)|}-|\delta_{B_1}(x_i)||\delta_{B_1}(x_i)\cap D_j|\log\frac{|\delta_{B_1}(x_i)\cap D_j|}{|\delta_{B_1}(x_i)|}\right):=\sum_{i=1}^{n}\sum_{j=1}^{m}\Delta_{ij}.\end{aligned}$$

Here, $:=$ denotes "is defined as". Now we prove $\Delta_{ij}\geq 0$ holds for $\forall i,j$. Notice that $|\delta_B(x_i)|=|\delta_B(x_i)\cap D_j|+|\delta_B(x_i)\cap D_j^c|$. Let $x_B=|\delta_B(x_i)\cap D_j|\geq 0$ and $y_B=|\delta_B(x_i)\cap D_j^c|\geq 0$, we have $\Delta_{ij}=(x_{B_2}+y_{B_2})x_{B_2}\log\frac{x_{B_2}}{x_{B_2}+y_{B_2}}-(x_{B_1}+y_{B_1})x_{B_1}\log\frac{x_{B_1}}{x_{B_1}+y_{B_1}}$. According to the Proposition 2.1, we have $\delta_{B_2}(x_i)\subseteq\delta_{B_1}(x_i)$, and thus $\delta_{B_2}(x_i)\cap D_j\subseteq\delta_{B_1}(x_i)\cap D_j$ and $\delta_{B_2}(x_i)\cap D_j^c\subseteq\delta_{B_1}(x_i)\cap D_j^c$. In other words, $x_{B_1}\geq x_{B_2}$ and $y_{B_1}\geq y_{B_2}$, according to the Lemma 3.1, we have $\Delta_{ij}\geq 0$ and thus $\Delta H\geq 0$. The equality holds if and only if $\delta_{B_1}(x_i)=\delta_{B_2}(x_i)$ for $\forall x_i\in U$.

Theorem 3.1 shows that the NDE decreases monotonically with the increase of attribute subsets. The larger the attribute subset, the smaller the value of NDE. This means that the better the label distribution within granules and the lower the uncertainty of attribute subset.

4. Attribute reduction method based on weighted neighborhood distribution entropy

Since the NDE is monotonic, the value of the entropy function decreases monotonically as new attributes are added during the attribute reduction process. Therefore, the decrease in the entropy function can be used to measure the significance of the candidate attributes.

Definition 4.1. Given $NDS=\langle U,C,D\rangle$, $B\subseteq C$, $\delta>0$, and $a\in C-B$, then the significance of attribute a under B is defined as

$$SIG^{(1)}_{\delta}(a,B,D)=H^{\delta}_{B}(D)-H^{\delta}_{B\cup\{a\}}(D). \tag{10}$$

The significance of attribute a is determined by the decrease of the NDE when a is added to B. A large value of $SIG^{(1)}_{\delta}(a,B,D)$ means a is more important for B under D.

As mentioned above, $SIG^{(1)}_{\delta}$ measures the significance of an attribute by comparing changes in the data distribution within neighborhood granules before and after adding it to the reduction subset. In the following, we use a specific example to demonstrate the calculation process and discuss the possible shortcomings of this significance measure.

Example 4.1. Given a constructed neighborhood decision system $\langle U,C,D\rangle$ detailed in Table 1, the universe U is $\{x_1,x_2,\dots,x_{18}\}$ and condition attributes C is $\{a_1,a_2,a_3\}$. Let $\delta=0.15$, and $B=\{a_3\}$ be the reduction subset of the current iteration. Firstly, we compute

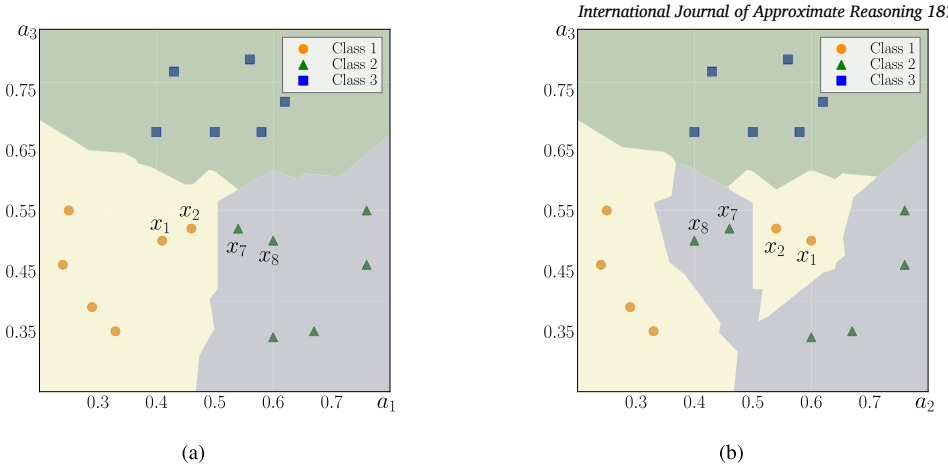

Fig. 3. Classification regions of KNN under different attribute subsets. (a) $\{a_1, a_3\}$. (b) $\{a_2, a_3\}$.

the NDE of attribute subset $B_1 = \{a_1, a_3\}$ and $B_2 = \{a_2, a_3\}$. Since the granular chaos of the neighborhood granules containing homogeneous samples is 0, we only consider the neighborhood granules under B_1 and B_2 that contain heterogeneous samples:

$$\delta_{B_1}(x_1) = \delta_{B_2}(x_1) = \{x_1, x_2, x_8\};$$
$$\delta_{B_1}(x_7) = \delta_{B_2}(x_7) = \{x_2, x_7, x_8\};$$
$$\delta_{B_1}(x_2) = \delta_{B_1}(x_8) = \delta_{B_2}(x_2) = \delta_{B_2}(x_8) = \{x_1, x_2, x_7, x_8\}.$$

The sizes of the granules and their corresponding granular chaos values are computed as follows:

$$P^{\delta}_{B_1}(x_1) = P^{\delta}_{B_1}(x_7) = 0.1667, GC^{\delta}_{B_1}(D|x_1) = GC^{\delta}_{B_1}(D|x_7) = 0.6365;$$
$$P^{\delta}_{B_1}(x_2) = P^{\delta}_{B_1}(x_8) = 0.2222, GC^{\delta}_{B_1}(D|x_2) = GC^{\delta}_{B_1}(D|x_8) = 0.6931;$$
$$P^{\delta}_{B_2}(x_1) = P^{\delta}_{B_2}(x_7) = 0.1667, GC^{\delta}_{B_2}(D|x_1) = GC^{\delta}_{B_2}(D|x_7) = 0.6365;$$
$$P^{\delta}_{B_2}(x_2) = P^{\delta}_{B_2}(x_8) = 0.2222, GC^{\delta}_{B_2}(D|x_2) = GC^{\delta}_{B_2}(D|x_8) = 0.6931.$$

The NDE under B_1 and B_2 can be computed as follows:

$$H^{\delta}_{B_1}(D) = \sum_{i=1}^{18} \left(P^{\delta}_{B_1}(x_i)\right)^2 GC^{\delta}_{B_1}(D|x_i) = 0.1038;$$
$$H^{\delta}_{B_2}(D) = \sum_{i=1}^{18} \left(P^{\delta}_{B_2}(x_i)\right)^2 GC^{\delta}_{B_2}(D|x_i) = 0.1038.$$

Similarly, we can compute that $H^{\delta}_{B}(D) = 2.9982$. Therefore, we have $SIG^{(1)}_{\delta}(a_1, \{a_3\}, D) = SIG^{(1)}_{\delta}(a_2, \{a_3\}, D) = 2.8944$.

One of the tasks of attribute reduction is to find the best attribute subset to achieve higher classification accuracy. Fig. 3 shows the KNN ($k = 3$) classification region of Table 1 under attribute subsets $\{a_1, a_3\}$ and $\{a_2, a_3\}$. Although the significance of attributes a_1 and a_2 is the same according to the calculation in Example 4.1, Fig. 3 demonstrates that the KNN classification performance of subset $\{a_1, a_3\}$ is obviously robust than $\{a_2, a_3\}$. One of the reasons for this is that the construction of the entropy function is based on the perspective of each neighborhood granule. In other words, it lacks an evaluation of the attribute subsets from a macroscopic perspective.

As mentioned in Section 2.2, DBI can measure the label distribution between decision classes. The smaller DBI values generally correspond to attribute subsets with clearer decision boundaries. Therefore, we set DBI as the weight of the NDE, and the new significance of attribute is introduced as follows:

Definition 4.2. Given $NDS = \langle U, C, D\rangle$, $B \subseteq C$, $\delta > 0$, $a \in C - B$, and I_B denotes the DBI of U under B. Then, the weighted significance of attribute a corresponding to B is defined as

$$SIG^{(2)}_{\delta}(a, B, D) = I_B \times H^{\delta}_{B}(D) - I_{B\cup\{a\}} \times H^{\delta}_{B\cup\{a\}}(D). \tag{11}$$

The term $I_B \times H^{\delta}_{B}(D)$ is called the weighted neighborhood distribution entropy (WNDE), which uses the label distribution information from both the granule and decision classes perspectives to measure the uncertainty of the attribute subsets.

Algorithm 1 Attribute reduction algorithm based on weighted neighborhood distribution entropy (ARWNDE).

Input: $NDS = \langle U, C, D\rangle$, δ, and ϵ.
Output: Attribute subset red.
1: Initialize: $red = \emptyset$, start $= 1$.
2: **while** start **do**
3: **for** each $a \in C - red$ **do**
4: Compute $H^{\delta}_{red\cup\{a\}}(D)$ by (9).
5: Compute $SIG^{(2)}_{\delta}(a, red, D)$ by (11).
6: **end for**
7: Select $a_0 = \arg\max\{SIG^{(2)}_{\delta}(a, red, D), a \in (C - red)\}$.
8: **if** $SIG^{(2)}_{\delta}(a_0, red, D) > \epsilon$ **then**
9: $red \leftarrow red \cup \{a_0\}$.
10: **else**
11: start $= 0$.
12: **end if**
13: **end while**
14: **return** red.

Example 4.2 *(Continuation of Example 4.1).* For the NDS, B, B_1 and B_2 given in Example 4.1, the DBI of U under B, B_1, and B_2 can be calculated as 10.8549, 0.6734, and 1.1747 according to (6). Therefore, the weighted significance of a_1 and a_2 corresponding to B can be computed as follows:

$$SIG^{(2)}_{\delta}(a_1, B, D) = I_B \times H^{\delta}_B(D) - I_{B_1} \times H^{\delta}_{B_1}(D) = 10.8549 \times 2.9982 - 0.6734 \times 0.1038 = 32.4753;$$
$$SIG^{(2)}_{\delta}(a_2, B, D) = I_B \times H^{\delta}_B(D) - I_{B_2} \times H^{\delta}_{B_2}(D) = 10.8549 \times 2.9982 - 1.1747 \times 0.1038 = 32.4232.$$

By adding the weighting factor DBI to NDE, the significance of attribute that lead to poor overall distribution is reduced. According to the above calculations, we obtain that $SIG^{(2)}_{\delta}(a_1, B, D) > SIG^{(2)}_{\delta}(a_2, B, D)$, which is consistent with their classification performance. In summary, WNDE provides a more comprehensive measure of uncertainty for attribute subsets than NDE.

Based on Definition 4.2, we propose the greedy forward attribute reduction algorithm ARWNDE to choose the optimal attribute subset, as detailed in Algorithm 1. The parameter δ is used to control the size of the neighborhood radius of samples, and the parameter ϵ is used to control the stopping criterion for attribute reduction. Assume that the sample space size is n, the number of decision classes is m, and the reduction subset is red in the current iteration. Steps 3-7 traverse all attributes in $C - red$ and calculate their significance. The time complexity of computing distance $d_{red}(x, y)$ and the value of $GC^{\delta}_{red}(D|x)$ defined by (8) are $O(n^2|red|)$ and $O(n^2 m)$, respectively. In addition, the time complexity of computing the value of s^{red} and $\|c^{red}_i - c^{red}_j\|_2$ defined by (6) are $O(n|red|)$ and $O(m^2|red|)$, respectively. Compared with the sample space and the number of the reduction subsets, the size of m can be ignored. Therefore, the time complexity of adding an optimal attribute a_0 to red can be approximated as $O(n^2|red| + n|red|)$. In the worst case, we need to add $|C - red|$ attributes to the reduction subset red, which means the process of step 2 to step 7 can be executed at most $|C - red|$ times. In summary, the time complexity of Algorithm 1 is $O(|C - red|(n^2|red| + n|red|))$.

5. Experimental analysis

In this section, we conduct several experiments to evaluate the performance of ARWNDE. All algorithms are implemented in Python 3.10 on the PyCharm platform. All experiments are run on a 64-bit Windows 11 OS with an AMD Ryzen 7-5800H processor @ 3.20 GHz, and 16 GB RAM.

5.1. Experimental setup

We select 16 public datasets from https://archive.ics.uci.edu/datasets and https://networkrepository.com/. The detailed information is presented in Table 2. All numerical data are normalized to the interval [0, 1] using the max-min method. To evaluate the effectiveness of ARWNDE, experiments are conducted on the above datasets using several existing representative attribute reduction algorithms, including NRS [23], NMI [34], HANDI [35], SFSS [44], SemiFREE [31], FScNCE [36], WKNRS [22], SNCMI [45]. To show the effect of the weight we introduced to the NDE, we set the comparison algorithm ARNDE. The specific information of these comparison algorithms is as follows:

(1) *Raw Data*: Select all attributes.
(2) NRS: An attribute dependency function based on neighborhood rough sets is proposed, which reflects the percentage of the decision positive region.
(3) NMI: Classic mutual information is extended to NDS. The proposed entropy function is used to calculate the correlation between attributes.
(4) $HANDI$: A neighborhood discriminant index is proposed based on neighborhood relations, which reflects the discriminant information of attributes.
(5) $SFSS$: A forward search attribute subset strategy is proposed, which maximizes the separability of the information system.

Table 2
Description of datasets.

No.	Datasets	Sample	Attribute	Class
1	Wine	178	13	3
2	Pop	540	18	2
3	Statlog	846	18	4
4	Parkinsons	195	22	2
5	Wdbc	569	30	2
6	Wpbc	198	33	2
7	Iono	351	34	2
8	Sonar	208	60	2
9	Wine_red	1599	11	6
10	Wine_white	4898	12	7
11	Segment	2310	19	7
12	COLON	62	2000	2
13	DLBCL	77	5469	2
14	Leukemia	72	11225	3
15	MLL	72	12582	3
16	Ovarian	253	15154	2

(6) $SemiFREE$: A semi-supervised attribute reduction algorithm is proposed, which uses an extended fuzzy entropy to measure the correlation and redundancy between attributes.
(7) $FScNCE$: A reduction algorithm is proposed based on neighborhood combination entropy, which can reflect the probability of distinguishing granules from each other.
(8) $WKNRS$: By weighting and evaluating each sample based on its k-nearest neighbors, the constructed lower approximation helps to solve the noise sensitivity problem in neighborhood rough sets.
(9) $SNCMI$: A soft neighborhood rough sets model with dynamically adjusted neighborhood radius is proposed. In addition, an attribute measure that comprehensively considers the correlation, redundancy, complementarity, and synergy between attributes is proposed.
(10) $ARNDE$: Use $SIG_{\delta}^{(1)}(a, red, D)$ instead of $SIG_{\delta}^{(2)}(a, red, D)$ to calculate steps 5, 7, and 8 in Algorithm 1. The other calculation processes are the same as Algorithm 1.

We use three classifiers to evaluate the classification performance of different attribute reduction algorithms, i.e., KNN ($k=3$), SVM (rbf-SVM) and CART. To ensure the reliability of the experimental results, all experiments use the ten-fold cross validation method. It divides the dataset into ten independent subdatasets, and takes nine of them as training sets in each round, the remaining one as the validation set. The average accuracy and standard deviation after ten rounds are taken as the final classification results.

In order to ensure the fairness and effectiveness of the experimental results, we apply a unified grid search strategy for the neighborhood radius. The δ parameter for NRS, NMI, HANDI, FScNCE, ARNDE and ARWNDE is set to [0, 0.7] with step size of 0.05. According to the search strategy of the original paper, the number of neighbors parameter in WKNRS is searched in the range of 1 to 15 with a step size of 1, while that in SNCMI is searched in the range of 1 to 21 with a step size of 2. SFSS and SemiFREE are both forward sequential search filtering methods, and the number of reduction subset is their parameter. In order to better compare with ARWNDE, we set the interval with a radius of 6, centered on the number of attribute subsets obtained by ARWNDE, as the parameter search interval of SFSS and SemiFREE.

In addition, ARWNDE includes the stop criterion ϵ. The algorithm terminates when the significance of the optimal attribute selected in the current iteration falls below it, which means that the attributes selected subsequently have little effect on reducing the uncertainty of NDS. The parameter ϵ is set to 0.01 and 0.001 in high-dimensional and low-dimensional datasets, respectively. The larger stopping criterion for high-dimensional datasets is to prevent the algorithm from selecting redundant attributes beyond the expected number for it. In this experiment, datasets with more than 1000 attributes are called high-dimensional datasets.

5.2. Experimental results

Table 3 presents the size of reduced attributes by different algorithms, and the results with the shortest reduction size are marked in bold. In ARWNDE, the neighborhood radius δ affects the size of neighborhood granules, finally influences the classification accuracy. The last column of Table 3 shows the best δ selected by ARWNDE, which means the attribute subset selected by it can maximize the average classification accuracy of the three classifiers. Compared to the original datasets, all algorithms can effectively reduce the number of attributes. Notably, although ARNDE and ARWNDE only select the minimum number of attributes on few datasets, their average reduction size still rank first and second among all algorithms, respectively. This shows the stopping criterion ϵ effectively prevents the selection of less important attributes.

Table 4 shows a comparison of the runtime of different algorithms on 16 datasets. We use the optimal parameters of each algorithm to perform attribute reduction on the complete datasets, and the results with the shortest running time are marked in bold. Since ARWNDE considers all label information within the granules and between decision classes, its running time is relatively long. In contrast, ARNDE has a shorter runtime than ARWNDE because it avoids calculating the DBI weighting factor. This difference is particularly evident on high-dimensional datasets. The complex method for constructing significance of attributes leads to relatively

Table 3
The number of reduced attributes by different algorithms.

Dataset	Raw Data	NRS	NMI	HANDI	SFSS	SemiFREE	FScNCE	WKNRS	SNCMI	ARNDE	ARWNDE	δ
Wine	13	7	7.9	6.7	7	7	7.6	7.2	7	**6**	**6**	0.2
Pop	18	6	9.2	6	5	**4**	5.8	6.3	6	6	6	0.3
Statlog	18	12.2	13.8	14	12	11	9.7	11.8	13	**9.3**	9.4	0.15
Parkinsons	22	**5.7**	7.7	**5.7**	7	8	7	6.4	6	5.8	5.8	0.15
Wdbc	30	9.3	**7.1**	7.4	9	8	13.7	8.9	8	7.4	7.3	0.15
Wpbc	33	9.6	7.3	9.5	9	**7**	15	9.2	9	7.5	8.8	0.2
Iono	34	12.5	11.6	9.3	13	12	12.8	10.2	12	**9.2**	9.8	0.2
Sonar	60	**6.6**	23.7	17.9	21	20	22.3	13.8	12	10.8	11.1	0.3
Wine_red	11	9.9	10	9	9	**8**	9.8	8.7	**8**	10	10	0.1
Wine_white	12	8	7.9	7.9	8	7	**5**	7.6	7	11	11.2	0.1
Segment	19	11.2	12.6	9	10	10	8.8	10.2	10	**7.5**	8	0.1
COLON	2000	12	12.8	10.7	10	8	11.8	9.3	9	**7.6**	8	0.45
DLBCL	5469	7.3	6.8	**4.7**	10	7	7.3	6.2	7	6.5	5.1	0.35
Leukemia	11225	10.3	5.9	6.9	**5**	10	5.8	6	6	6.7	7.1	0.55
MLL	12582	5.6	**5.1**	6	7	6	6.2	6.8	6	5.5	5.7	0.5
Ovarian	15154	4.2	4	**3**	**3**	4	3.7	3.9	4	4	4	0.3
Average	2918.75	8.59	9.59	8.36	9.06	8.56	9.52	8.28	8.13	**7.55**	7.71	

Table 4
The runtime(s) of different algorithms.

Dataset	NRS	NMI	HANDI	SFSS	SemiFREE	FScNCE	WKNRS	SNCMI	ARNDE	ARWNDE
Wine	0.4412	0.4817	**0.3159**	1.7174	3.7622	1.1678	0.9622	1.0855	0.507	0.739
Pop	3.0888	5.0878	5.5473	**2.9766**	38.8946	6.8509	3.2343	12.3416	4.004	4.244
Statlog	13.7721	**12.5992**	14.4772	17.4839	220.362	19.9047	15.6033	33.6238	14.0518	16.5102
Parkinsons	**0.8023**	1.0499	0.8566	3.0502	17.2155	1.847	1.2277	5.4077	0.8564	1.3902
Wdbc	13.1934	7.481	**2.6565**	9.5767	187.2881	23.2016	9.0509	42.5295	9.4973	11.421
Wpbc	2.1743	**1.5860**	2.2999	5.0043	22.9226	4.6347	2.7261	9.8112	2.2858	3.2528
Iono	6.4574	6.1276	**6.0340**	10.3545	155.4959	9.6093	6.4497	27.252	6.5213	6.3973
Sonar	**3.0337**	10.4272	10.7709	21.9197	284.6912	10.2038	9.8459	38.8146	6.148	8.0999
Wine_red	23.3275	**13.4184**	17.5743	19.0545	290.583	39.4177	17.6038	58.326	22.1557	25.7832
Wine_white	152.9489	101.8047	180.6095	**62.8484**	2348.7235	383.4107	190.1104	403.4686	239.448	234.239
Segment	122.7684	75.5038	**60.6706**	70.0617	1409.7897	134.3183	90.4467	185.3262	72.1486	99.7206
COLON	65.1711	78.0708	**39.6904**	224.0068	258.4092	147.9602	52.1631	176.9612	47.7898	79.7875
DLBCL	125.782	110.2158	**64.339**	949.1146	840.7781	262.7363	122.7335	294.158	102.5286	184.4848
Leukemia	324.461	239.8571	**219.0195**	1550.2139	2305.1666	689.2387	252.0978	877.0596	274.7825	534.2308
MLL	202.4187	210.9277	**195.7817**	3439.4724	1978.2475	821.9164	245.5509	940.4954	313.3893	588.523
Ovarian	701.6874	600.8316	**489.6838**	3795.5772	4231.8716	2533.9732	975.8537	2513.6031	925.0226	1185.4166
Average	110.0955	92.2169	**81.8954**	636.402	912.1376	318.1495	124.7288	351.2665	127.571	186.515

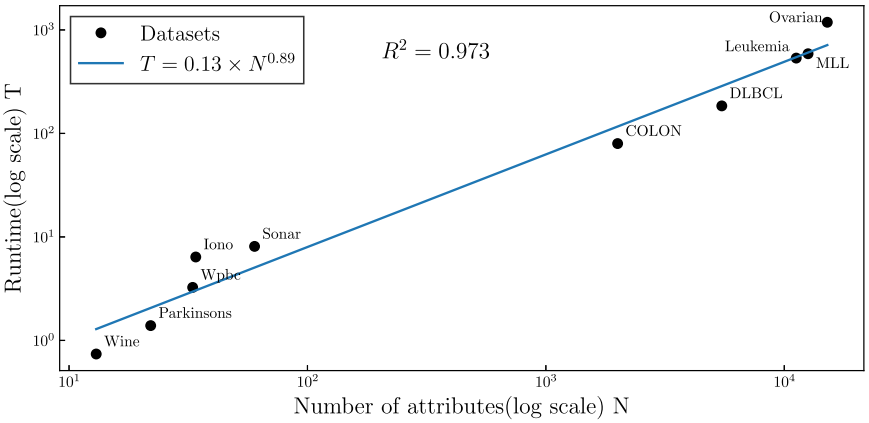

Fig. 4. Results of power law fitting between runtime and number of attributes across datasets.

longer runtimes for SemiFREE, FScNCE, and SNCMI. Notably, HANDI is defined based on the cardinality of neighborhood relations, which allows it to achieve the shortest runtime.

To test the scalability of the proposed algorithm on high-dimensional datasets, we perform a power law fit of the runtime of ARWNDE in Table 4 and the number of attributes for the datasets with less than 500 samples. As shown in Fig. 4, the coefficient of determination $R^2 = 0.973$ indicates that the model fits well, while the exponent $0.89 < 1$ indicates that the running time grows

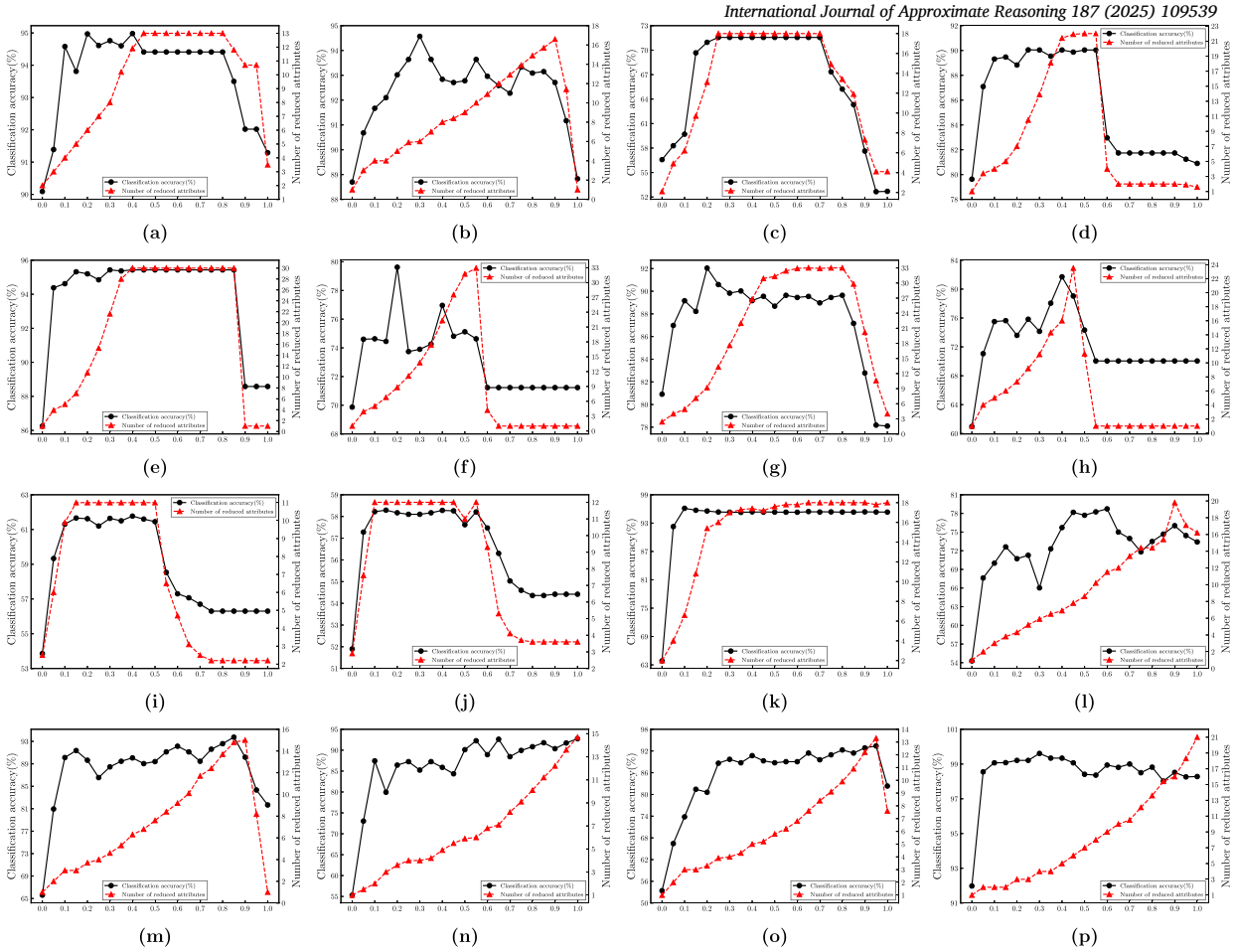

Fig. 5. Average accuracy of three classifiers and number of reduction subsets varying with neighborhood radius δ on 16 datasets. (a) Wine. (b)Pop. (c) Statlog. (d) Parkinsons. (e) Wdbc. (f) Wpbc. (g) Iono. (h) Sonar. (i) Wine_red. (j)Wine_white. (k) Segment. (l) COLON. (m) DLBCL. (n) Leukemia. (o) MLL. (p) Ovarian.

sublinearly with the number of attributes. The fitting results verify that ARWNDE has good scalability on ten-thousand dimensional datasets.

To verify the superiority of ARWNDE in classification accuracy, Tables 5-7 show the classification performance of 10 algorithms and raw datasets under KNN, SVM and CART classifiers, where the best results are marked in bold. Based on the reported results, the following can be clearly observed:

(1) According to Tables 5-7, ARWNDE achieve the highest accuracy on 75%, 75%, and 68% of the datasets, respectively, and has the highest average accuracy. This indicates that compared to competing algorithms, ARWNDE exhibits stronger generalization performance on real-world datasets with imbalanced distributions or ambiguous decision boundaries. As mentioned above, ARWNDE considers the label distribution information both within neighborhood granules and between decision classes more comprehensively and provides a more complete and accurate measurement of uncertainty.

(2) ARWNDE shows better performance on 5 high-dimensional dataset related to gene expression. A small number of strongly expressed genes can effectively distinguish classes, which allows ARWNDE to maintain high classification accuracy while selecting a small number of attributes. For datasets with severely overlapping decision boundaries such as Statlog and Sonar, the highly close class centers lead to abnormally high DBI values. According to (11), this will cause ARWNDE to lose the ability to measure uncertainty from a local perspective and lead to poor performance.

(3) ARWNDE achieves higher classification accuracy on almost all datasets under the three classifiers, while ARNDE has an advantage in runtime. Compared to ARNDE, ARWNDE can select attribute subsets that lead to clearer decision boundaries and more robust classification regions. Combined with the runtime results analysis in Table 4, ARWNDE successfully balances computational efficiency and classification performance.

Table 5
Classification accuracies of comparison algorithms under classifier KNN.

Dataset	Raw Data	NRS	NMI	HANDI	SFSS	SemiFREE	FScNCE	WKNRS	SNCMI	ARNDE	ARWNDE
Wine	96.08±3.57	97.16±3.78	97.78±3.69	97.75±3.72	97.75±3.72	95.42±5.1	95.44±4.29	97.78±2.72	97.84±4.14	**98.3±2.6**	**98.3±2.6**
Pop	91.67±2.78	93.52±2.23	91.67±4.24	93.33±3.33	**94.63±2.68**	93.15±3.42	92.26±3.18	93.74±1.66	92.85±4.77	94.26±1.93	94.44±2.75
Statlog	69.27±1.89	67.74±3.25	69.38±5.03	69.62±3.79	66.67±3.78	**70.3±4.27**	69.79±4.02	68.7±4.22	69.08±4.57	68.92±2.87	**70.3±4.44**
Parkinsons	92.82±5.24	95.39±3.59	92.32±5.24	92.82±4.74	89.32±5.64	88.68±7.49	89.24±5.34	88.68±6.58	91.26±5.19	94.34±4.33	**95.84±3.78**
Wdbc	96.66±2.14	94.72±2.61	95.44±1.95	95.08±2.19	94.02±2.86	95.26±2.48	95.44±2.42	94.85±2.34	95.73±2.61	96.13±1.91	**96.2±3.12**
Wpbc	74.32±9.1	78.26±7.22	70.63±5.38	71.68±9.21	77.66±6.62	73.82±7.82	77.26±7.25	72.77±7.92	76.23±7.55	77.37±12.56	**78.29±6.73**
Iono	86.88±3.9	88.33±3.46	89.2±6.15	89.74±4.47	88.63±5.49	87.47±5.73	88.33±5.06	89.3±5.75	87.62±6.27	89.73±3.45	**90.05±3.8**
Sonar	82.74±6.38	82.17±7.58	81.74±5.53	**83.77±10.08**	81.69±8.02	74.05±10.89	77.95±8.42	78.37±8.45	79.83±8.12	79.81±7.02	83.6±7.05
Wine_red	58.85±3.23	58.1±4.31	58.1±3.35	57.16±3.02	56.47±1.19	57.97±2.65	58.1±2.91	57.6±3.13	55.53±2.53	57.26±3.52	**58.16±3.5**
Wine_white	57.1±2.34	55.33±2.08	55.96±2.21	**56.19±1.79**	54.12±1.56	53.53±1.21	54.35±1.77	53.33±1.18	51.78±1.47	54.23±2.46	54.39±1.79
Segment	95.93±1.21	96.45±0.67	96.32±1.26	96.28±1.05	96.19±1.21	94.68±1.42	95.06±1.12	96.02±1.17	95.54±1.2	96.62±1.57	**96.75±1.15**
COLON	72.62±12.2	80.48±10.2	80.71±12.11	85.48±11.64	84.05±15.23	83.33±17.82	80.57±13.44	85.71±4.28	82.51±15.16	87.14±6.49	**87.16±12.83**
DLBCL	89.46±13.93	93.75±6.25	89.64±7.65	91.07±11.32	82.14±13.31	81.32±13.92	92.16±10.49	89.72±10.81	90.72±9.54	90.06±12.98	**94.64±6.59**
Leukemia	87.68±9.46	84.46±10.04	90.18±9.08	91.79±10.84	93.04±9.46	89.24±12.92	90.89±10.47	**93.61±10.86**	91.39±11.56	88.93±5.58	93.18±5.23
MLL	84.46±10.04	84.64±9.59	88.93±10.16	90.36±12.76	90.54±8.38	81.96±8.91	83.62±9.96	86.72±11.96	84.15±9.59	89.28±12.64	**94.3±8.28**
Ovarian	93.71±5.27	99.22±1.57	**100±0**	**100±0**	97.25±2.49	96.83±3.45	98.28±6.03	99.41±4.88	99.36±2.01	**100±0**	**100±0**
Average	83.14±12.54	84.36±13.26	84.25±13.33	85.13±13.47	84.01±13.46	82.31±12.85	83.67±12.9	84.14±13.69	85.21±13.24	85.15±13.65	**86.6±13.76**

Table 6
Classification accuracies of comparison algorithms under classifier SVM.

Dataset	Raw Data	NRS	NMI	HANDI	SFSS	SemiFREE	FScNCE	WKNRS	SNCMI	ARNDE	ARWNDE
Wine	98.33±3.56	97.71±2.8	97.75±3.72	98.3±2.6	97.75±2.76	97.12±3.9	97.19±5.14	97.19±3.81	97.75±2.43	98.28±2.19	**98.33±2.19**
Pop	92.96±1.61	94.07±1.39	93.89±1.19	95.19±2.06	94.81±1.81	93.33±2.22	92.23±2.85	93.48±1.91	93.15±2.82	94.63±2.68	**95.56±1.89**
Statlog	75.54±3.59	72.94±3.53	72.7±4.01	73.53±3.13	72.22±4.25	70.22±5.05	74±3.53	72.17±4.13	73.45±3.32	**75.72±3.82**	74.47±2.8
Parkinsons	87.08±5.5	87.63±4.29	86.74±6.41	87.68±4.69	86.18±5.55	86.18±4.44	87.18±5.24	86.66±4.69	87.67±5.76	87.29±6.77	**88.92±5.77**
Wdbc	97.54±1.95	96.3±2.55	96.14±2.81	96.13±1.32	95.25±2.94	94.56±3.08	**97.37±2.39**	96.43±2.78	94.1±1.83	95.43±1.95	96.48±1.75
Wpbc	79.34±5.11	82.84±7.1	78.34±7.3	80.79±4.96	78.76±3.17	76.29±1.99	80.82±2.95	82.32±5.08	83.13±2.27	82.89±7.75	**84.32±5.42**
Iono	93.44±3.15	92.87±3.45	94.33±4.35	93.44±4.44	93.46±3.35	94.02±4.14	94.03±4.09	93.88±4.23	93.76±3.87	94.59±3.24	**94.89±4.34**
Sonar	80.74±8.65	78.43±8.73	**82.19±2.33**	80.33±9.25	79.26±7.68	69.74±9.21	77.98±10.63	79.88±8.82	75.99±9.84	79.86±4.98	82.14±7.43
Wine_red	61.35±3.9	61.29±3.76	60.85±1.58	**61.42±2.23**	60.91±3.75	60.35±3.85	60.98±4.36	60.16±3.04	60.41±4.02	61.1±3.22	60.6±2.85
Wine_white	55.06±1.28	54.55±1.94	54.53±1.54	54.72±0.83	54.21±1.33	53.94±1.8	54±2.34	54.13±2.42	52.31±1.42	54.28±2.32	**54.96±2.15**
Segment	93.9±0.98	94.5±1.33	94.33±1.43	94.89±1.52	93.55±1.35	91.56±1.7	92.29±1.46	94.47±1.37	94.58±1.93	94.89±1.24	**94.92±1.2**
COLON	79.29±12.28	82.14±9.11	80.71±14.22	83.81±10.58	82.38±14.16	80.95±13.88	83.33±15.59	81.95±2.38	82.52±12.82	84.29±11.13	**85.12±13.84**
DLBCL	84.29±8.52	93.75±6.25	90.89±8.2	88.57±8.79	89.64±8.07	77.88±5.9	81.65±11.17	90.44±9.26	93.72±8.48	92.32±8.87	**97.32±5.37**
Leukemia	94.46±9.33	91.61±9.38	90.18±9.08	91.79±10.84	93.04±9.46	87.12±14.36	90.32±10.83	91.89±11.56	91.89±13.17	91.61±6.87	**94.46±6.8**
MLL	91.43±9.48	88.93±10.16	90.18±9.08	91.79±12.96	91.79±6.74	77.86±12.86	78.16±6.24	92.11±9.69	93.33±11.89	92.13±9.37	**96.62±4.72**
Ovarian	98.45±1.9	99.62±1.15	**100±0**	**100±0**	96.45±3.28	96.43±3.74	98.26±2.97	99.41±1.96	**100±0**	**100±0**	**100±0**
Average	85.2±12.48	85.57±12.66	85.23±12.72	85.77±12.62	84.97±12.6	81.72±12.78	83.74±12.4	85.41±12.83	85.49±13.23	86.21±12.64	**87.44±13.12**

Table 7
Classification accuracies of comparison algorithms under classifier CART.

Dataset	Raw Data	NRS	NMI	HANDI	SFSS	SemiFREE	FScNCE	WKNRS	SNCMI	ARNDE	ARWNDE
Wine	90.52±6.58	90.92±6.96	92.16±5.06	89.38±8.04	89.84±5.59	92.03±6.99	88.14±10.33	91.58±7.53	93.57±8.17	92.26±4.46	**94.93±4.62**
Pop	92.22±2.72	91.3±2.49	92.04±3.42	90.56±2.68	92.41±1.54	91.67±2.23	90.05±3.28	91.63±4.15	90±2.54	92.78±2.1	**93.7±3.12**
Statlog	70.1±5.3	**69.64±4.88**	68.81±5.07	69.02±3.78	67.5±4.51	68.32±3.77	69.03±5.98	67.27±4.21	65.84±4.32	66.66±2.88	67.64±5.32
Parkinsons	88.66±6.51	87.05±8.28	88.79±7.17	90.29±5.25	77.03±7.59	87.74±6.37	85.71±7.36	89.16±8.19	89.15±11.31	87.93±8.42	**91.34±6.81**
Wdbc	92.62±2.33	93.39±3.15	92.79±2.01	93.15±3.26	92.08±3.19	91.74±3.24	91.22±3.23	92.02±3.62	92.10±2.84	92.96±4.09	**93.49±3.64**
Wpbc	69.26±12.99	68.18±8.96	65.16±5.71	68.16±6.44	66.24±9.31	64.03±9.88	70.26±6.27	67.15±8.16	72.13±7.68	72.21±6.46	**76.24±9.56**
Iono	87.73±5.6	91.14±6.05	89.74±4.65	91.44±4.96	89.19±5.63	**91.73±3.16**	88.61±4.58	90.62±5.15	90.76±3.37	91.44±4.05	91.18±6.78
Sonar	72.62±7.6	71.17±7.91	76.48±7.83	75.45±8.45	76.86±11.07	66.4±13.84	74.6±9.87	76.15±8.97	77.99±12.84	72.9±12.87	**80.24±5.66**
Wine_red	65.17±2.39	63.54±2.1	63.85±2.06	63.6±4.21	63.41±2.22	62.1±3.62	63.92±2.05	61.54±2.77	63.41±4.02	65.36±5.12	**66.36±2.35**
Wine_white	62.74±2.46	62.19±1.57	62.31±1.63	62.66±2.23	61.82±2.41	59.84±1.06	61.98±1.51	61.28±1.85	62.31±3.42	61.56±2.97	**63.16±2.53**
Segment	96.23±1.36	95.8±1.34	96.36±1.41	96.71±0.59	96.15±1.18	96.54±1.12	96.23±1.53	96.71±1.44	96.58±0.93	96.54±0.93	**96.92±1.02**
COLON	75.71±13.38	74.05±14.3	70.95±18.98	74.29±19.63	67.38±16.07	71.43±24.74	72.14±11.88	76.19±14.76	72.52±28.29	73.81±21.24	**78.57±16.84**
DLBCL	83.21±9.59	89.82±10.95	86.07±14.19	85.71±15.15	80.36±9.37	74.04±11	88.31±6.18	89.18±10.34	91.72±18.48	**93.39±9.15**	92.32±5.89
Leukemia	87.32±9.99	**95.71±6.55**	91.61±6.87	94.64±8.64	89.11±11.96	84.04±8.91	91.07±5.13	87.17±8.23	89.79±7.51	87.92±8.18	92.27±6.38
MLL	87.14±10	90±9.15	81.61±18.18	91.43±13.09	86.43±7.95	72.14±14.36	81.20±2.21	86.11±12.22	89.33±8.90	**91.61±13.73**	90.52±8.92
Ovarian	96.85±2.94	97.25±2.49	98.05±2.6	98±3.69	95.66±2.77	96.05±3.54	94.55±2.54	96.81±3.81	98.64±2.94	98.05±2.6	**98.82±11.26**
Average	82.38±10.98	83.2±12.18	82.3±12.06	83.41±11.96	80.72±11.81	79.37±12.82	81.69±10.93	82.54±11.96	83.49±11.96	83.59±12.06	**85.48±11.34**

Table 8
Statistical test of seven methods under three classifiers.

Classifiers	Mean rankings										χ_F^2	F_F
	NRS	NMI	HANDI	SFSS	SemiFREE	FScNCE	WKNRS	SNCMI	ARNDE	ARWNDE		
KNN	5.969	5.594	4.688	6.531	8.188	6.75	6.313	5	4.313	1.656	48.625	7.6471
SVM	5.156	5.875	4.125	6.5	9.281	6.656	6.375	5.594	3.688	1.75	64.4111	12.14
CART	5.25	5.563	4.313	7.469	7.656	7.188	6.344	5	4.406	1.813	49.964	7.9699

5.3. Sensitivity analysis of neighborhood radius

To test the impact of neighborhood radius δ on the performance of ARWNDE, we conduct neighborhood radius sensitivity experiments on 16 datasets. Since all values are normalized to the interval [0,1] during the data preprocessing stage, we set δ from 0 to 1 with steps of 0.05. As shown in Fig. 5, the horizontal axis represents the δ value, and the vertical axes represent the average reduction length and accuracy of the KNN, SVM, and CART classifiers corresponding to δ. By analyzing Fig. 5, we can draw the following conclusions:

(1) On most datasets, the trend of classification accuracy first increases with the increase of δ, and then begins to decrease after reaching the maximum value. This indicates that when δ is too small or too large, the values of NDE will also be too small or too large, which makes the value of NDE and DBI fall on different scales. Therefore, the value of WNDE is dominated by one of the components, which will eventually lead to the poor classification performance.
(2) On the 16 datasets, the trend of reduction length is correlated with the trend of classification accuracy. Especially on Statlog, Wine_red, Wine_white, and Segment, the trend of the curves of reduction length and classification accuracy is consistent.
(3) When the value of δ is within a specific range, the average classification accuracy can reach or approach the maximum value on most datasets. Especially on most low-dimensional datasets, the average classification accuracy and reduction length can remain stable within a certain range of δ. This indicates that ARWNDE has a certain robustness in the selection of neighborhood radius.

The relationship between classification accuracy and δ cannot be generalized due to the diversity of datasets. Notably, the location of the accuracy peak may indicate where the scales of NDE and DBI become relatively balanced. Therefore, we can still provide the recommended value of δ based on the analysis of Fig. 5. For low-dimensional datasets, we recommend setting δ to values in the interval [0, 0.3], and for high-dimensional datasets, we recommend setting δ to values in the interval [0.3, 0.6]. In other words, ARWNDE may achieve better performance in these intervals.

5.4. Statistical analysis

In order to evaluate whether there are statistically significant differences in the classification performance of the 10 algorithms, we first conduct the Friedman test [46]. The Friedman test is used to verify whether the performance of algorithms is similar. Before using it, we rank all algorithms on the datasets in descending order of accuracy. If multiple algorithms have the same classification accuracy, they are assigned an average rank. Generally, assuming we are comparing k algorithms across N datasets, let r_i represent the average rank of the i-th algorithm. The formula for the Friedman test can be expressed as follows:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4}\right); \tag{12}$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}, \tag{13}$$

where χ_F^2 follows a χ^2 distribution with degree of freedom $k-1$, while F_F follows an F-distribution with degrees of freedom $k-1$ and $(k-1)(N-1)$.

In this experiment we have $k = 10$ and $N = 16$. Under the null hypothesis that all algorithms perform the same, we calculate the average ranking of each algorithm and further obtain χ_F^2 and F_F. As shown in Table 8, when the significance level α is 0.05, the values of F_F under the three classifiers are all greater than the critical value 1.95 for $F(9, 135)$. Therefore, we accept the alternative hypothesis, that is, there are significant differences in the performance of the algorithms under the three classifiers.

Next we employ the Nemenyi post hoc test [46] to identify the specific differences between the algorithms. This test can calculate the critical range of average rank differences, represented by the following formula:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \tag{14}$$

where q_α is the critical value at the significance level α.

Through [46] and calculations, we can get $q_\alpha = 3.164$ and $CD = 3.387$. If the difference between the average ranking of two algorithms exceeds the CD value, the hypothesis that "the performances of the two algorithms are the same" can be rejected with 95% confidence. The comparison between algorithms can be visualized using the Nemenyi's test figure, if there is no significant difference between the algorithms, they are connected by horizontal line segments. As shown in Fig. 6, ARWNDE, ARNDE, and HANDI rank high

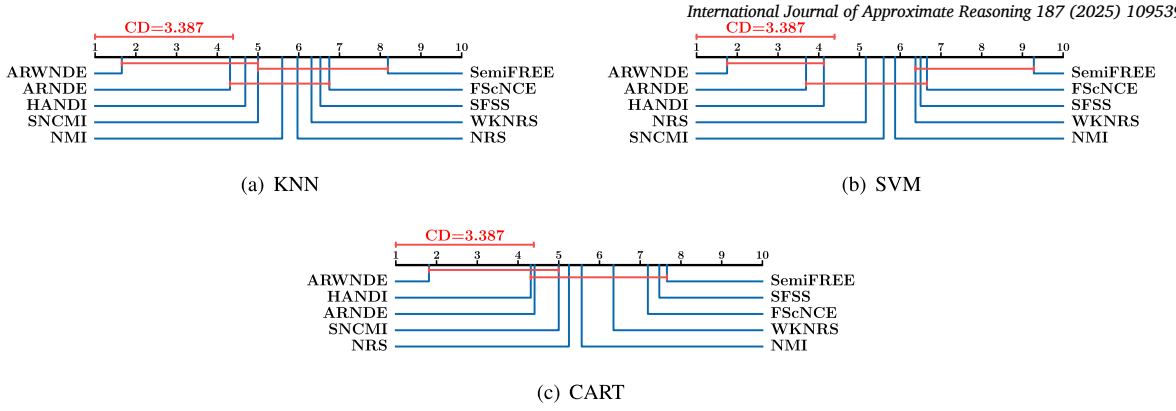

Fig. 6. Visualization of Nemenyi's test.

and have comparable classification performance. In addition, ARWNDE achieves the highest average ranking among all classifiers, which further demonstrates that the proposed algorithm exhibits strong generalization ability on datasets with different distributions.

6. Conclusion

This paper proposed an attribute reduction algorithm based on neighborhood rough sets. The core step of the algorithm was to use WNDE to measure the significance of attributes. Firstly, we defined the attribute evaluation function NDE and derived the relevant properties. Subsequently, we defined the significance function using DBI as the weighting factor of NDE and designed a greedy forward attribute reduction algorithm ARWNDE. The algorithm fully considered the label distribution information from multiple perspectives when selecting candidate attributes, and the theoretical analyses demonstrated that ARWNDE effectively selected attribute subsets with clearer decision boundaries. Finally, ARWNDE is compared with 9 comparison algorithms on 16 datasets. The experimental results showed that ARWNDE could select attribute subsets with short length and high classification accuracy on different types of datasets. Additionally, the parameter sensitivity analysis showed that the effect of neighborhood radius on classification accuracy had a certain degree of robustness.

It should be pointed out that the running time has a certain limit on the application of this algorithm. A feasible research direction is to adopt a reasonable granular ball computing method in the preprocessing stage to generate a small number of neighborhood granules covering the universe. This approach can also avoid the grid search process for the neighborhood radius in the experimental stage. In future work, we will design an effective method to granulate datasets to improve computational efficiency. In addition, we will further investigate attribute measurement functions based on neighborhood rough sets and look forward to apply this theory to multi label decision systems and incremental attribute reduction tasks.

CRediT authorship contribution statement

Peng Xu: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Ping Zhu:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ping Zhu reports financial support was provided by National Natural Science Foundation of China (No. 62172048). Ping Zhu reports financial support was provided by Fundamental Research Funds for the Central Universities (No. 2023ZCJH02). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the anonymous reviewers and Editor for their insightful comments and suggestions which greatly improve the quality of this paper. This work was supported by the National Natural Science Foundation of China under Grant (No. 62172048) and the Fundamental Research Funds for the Central Universities (No. 2023ZCJH02).

Data availability

The authors do not have permission to share data.

References

[1] V. Bolón-Canedo, B. Remeseiro, Feature selection in image analysis: a survey, Artif. Intell. Rev. 53 (4) (2020) 2905–2931.
[2] E.-S.M. El-Kenawy, A. Ibrahim, S. Mirjalili, M.M. Eid, S.E. Hussein, Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images, IEEE Access 8 (2020) 179317–179335.
[3] R.K. Singh, M. Sivabalakrishnan, Feature selection of gene expression data for cancer classification: a review, Proc. Comput. Sci. 50 (2015) 52–57.
[4] W. Ali, F. Saeed, Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data, Processes 11 (2) (2023) 562.
[5] P. Dhal, C. Azad, A comprehensive survey on feature selection in the various fields of machine learning, Appl. Intell. 52 (4) (2022) 4543–4581.
[6] Z. Wu, D. Rincon, P.D. Christofides, Real-time adaptive machine-learning-based predictive control of nonlinear processes, Ind. Eng. Chem. Res. 59 (6) (2020) 2275–2290.
[7] Y. Zhou, G. Cheng, S. Jiang, M. Dai, Building an efficient intrusion detection system based on feature selection and ensemble classifier, Comput. Netw. 174 (2020) 107247.
[8] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (5) (1982) 341–356.
[9] Z. Pawlak, S. Wong, W. Ziarko, Rough sets: probabilistic versus deterministic approach, Int. J. Man-Mach. Stud. 29 (1) (1988) 81–95.
[10] S. Greco, B. Matarazzo, R. Słowiński, Granular computing and data mining for ordered data: the dominance-based rough set approach, in: T. Lin, C. Liau, J. Kacprzyk (Eds.), Granular, Fuzzy, Soft Computing, Springer US, New York, 2023, pp. 117–145.
[11] J. Zhang, T. Li, H. Chen, Composite rough sets for dynamic data mining, Inf. Sci. 257 (2014) 81–100.
[12] H.H. Inbarani, A.T. Azar, G. Jothi, Supervised hybrid feature selection based on pso and rough sets for medical diagnosis, Comput. Methods Programs Biomed. 113 (1) (2014) 175–185.
[13] J. Sanabria, K. Rojo, F. Abad, A new approach of soft rough sets and a medical application for the diagnosis of coronavirus disease, AIMS Math. 8 (2) (2023) 2686–2707.
[14] X. Zhang, C. Mei, D. Chen, J. Li, Multi-confidence rule acquisition oriented attribute reduction of covering decision systems via combinatorial optimization, Knowl.-Based Syst. 50 (2013) 187–197.
[15] W. Zhu, Relationship among basic concepts in covering-based rough sets, Inf. Sci. 179 (14) (2009) 2478–2486.
[16] J.R. Anaraki, M. Eftekhari, Rough set based feature selection: a review, in: Information and Knowledge Technology (IKT), 2013 5th Conference on, IEEE, 2013, pp. 301–306.
[17] R. Cekik, A.K. Uysal, A novel filter feature selection method using rough set for short text data, Expert Syst. Appl. 160 (2020) 113691.
[18] J. Dai, Q. Hu, H. Hu, D. Huang, Neighbor inconsistent pair selection for attribute reduction by rough set approach, IEEE Trans. Fuzzy Syst. 26 (2) (2018) 937–950.
[19] T.Y. Lin, Neighborhood systems and relational databases, in: International Conference on Scientific Computing, 1988.
[20] T.Y. Lin, Neighborhood systems: mathematical models of information granulations, in: Proceedings of IEEE International Conference on Systems, Man Cybernetics, 2003, pp. 5–8.
[21] Y. Yao, Relational interpretations of neighborhood operators and rough set approximation operators, Inf. Sci. 111 (1) (1998) 239–259.
[22] N. Wang, E. Zhao, A new method for feature selection based on weighted k-nearest neighborhood rough set, Expert Syst. Appl. 238 (2024) 122324.
[23] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inf. Sci. 178 (18) (2008) 3577–3594.
[24] X. Fan, W. Zhao, C. Wang, Y. Huang, Attribute reduction based on max-decision neighborhood rough set model, Knowl.-Based Syst. 151 (2018) 16–23.
[25] M. Hu, E.C. Tsang, Y. Guo, D. Chen, W. Xu, A novel approach to attribute reduction based on weighted neighborhood rough sets, Knowl.-Based Syst. 220 (2021) 106908.
[26] D. Zhang, P. Zhu, Variable radius neighborhood rough sets and attribute reduction, Int. J. Approx. Reason. 150 (2022) 98–121.
[27] L. Yong, H. Wenliang, J. Yunliang, Z. Zhiyong, Quick attribute reduct algorithm for neighborhood rough set model, Inf. Sci. 271 (2014) 65–81.
[28] Q. Wang, Y. Qian, X. Liang, Q. Guo, J. Liang, Local neighborhood rough set, Knowl.-Based Syst. 153 (2018) 53–64.
[29] S. Xia, S. Wu, X. Chen, G. Wang, X. Gao, Q. Zhang, E. Giem, Z. Chen, GRRS: accurate and efficient neighborhood rough set for feature selection, IEEE Trans. Knowl. Data Eng. 35 (9) (2023) 9281–9294.
[30] Q. Hu, D. Yu, Z. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognit. Lett. 27 (5) (2006) 414–423.
[31] K. Liu, T. Li, X. Yang, H. Chen, J. Wang, Z. Deng, Semifree: semisupervised feature selection with fuzzy relevance and redundancy, IEEE Trans. Fuzzy Syst. 31 (10) (2023) 3384–3396.
[32] H. Zhao, K. Qin, Mixed feature selection in incomplete decision table, Knowl.-Based Syst. 57 (2014) 181–190.
[33] J. Dai, Z. Zhu, M. Li, X. Zou, C. Zhang, Attribute reduction for heterogeneous data based on monotonic relative neighborhood granularity, Int. J. Approx. Reason. 170 (2024) 109210.
[34] Q. Hu, L. Zhang, D. Zhang, W. Pan, S. An, W. Pedrycz, Measuring relevance between discrete and continuous features based on neighborhood mutual information, Expert Syst. Appl. 38 (9) (2011) 10737–10750.
[35] C. Wang, Q. Hu, X. Wang, D. Chen, Y. Qian, Z. Dong, Feature selection based on neighborhood discrimination index, IEEE Trans. Neural Netw. Learn. Syst. 29 (7) (2018) 2986–2999.
[36] P. Zhang, T. Li, Z. Yuan, C. Luo, K. Liu, X. Yang, Heterogeneous feature selection based on neighborhood combination entropy, IEEE Trans. Neural Netw. Learn. Syst. 35 (3) (2024) 3514–3527.
[37] C. Wang, Y. Huang, M. Shao, Q. Hu, D. Chen, Feature selection based on neighborhood self-information, IEEE Trans. Cybern. 50 (9) (2020) 4031–4042.
[38] X. Ji, J. Li, S. Yao, P. Zhao, Attribute reduction based on fusion information entropy, Int. J. Approx. Reason. 160 (2023) 108949.
[39] J. Dai, W. Chen, L. Xia, Feature selection based on neighborhood complementary entropy for heterogeneous data, Inf. Sci. 682 (2024) 121261.
[40] W. Xu, K. Yuan, W. Li, W. Ding, An emerging fuzzy feature selection method using composite entropy-based uncertainty measure and data distribution, IEEE Trans. Emerg. Top. Comput. Intell. 7 (2023) 76–88.
[41] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1 (2) (1979) 224–227.
[42] S. Bandyopadhyay, U. Maulik, Nonparametric genetic clustering: comparison of validity indices, IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. 31 (1) (2001) 120–125.
[43] J. Liang, Z. Shi, The information entropy, rough entropy and knowledge granulation in rough set theory, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 12 (1) (2004) 37–46.
[44] M. Hu, E.C.C. Tsang, Y. Guo, W. Xu, Fast and robust attribute reduction based on the separability in fuzzy decision systems, IEEE Trans. Cybern. 52 (6) (2022) 5559–5572.
[45] L. Chen, J. Chen, Y. Lin, Feature selection considering synergy between features based on soft neighborhood rough sets, Eng. Appl. Artif. Intell. 150 (2025) 110553.
[46] J. Demar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.