Approximate Supplement-Based Neighborhood Rough Set Model in Incomplete Hybrid Information Systems

Xiong Meng[1,2], Jilin Yang[1,2(✉)], Die Wu[1,2], and Tang Liu[1,2]

[1] Department of Computer Science, Sichuan Normal University, Chengdu, China
jilinyang@sicnu.edu.cn
[2] Visual Computing and Virtual Reality Key Laboratory of Sichuan Province, Sichuan Normal University, Chengdu, China

Abstract. Incomplete hybrid information systems (IHISs) contain hybrid data (e.g., categorical data, numerical data) and incomplete data. With the development of big data, IHISs widely exist in various practical applications. Due to the heterogeneity of hybrid data and the complex semantics of incomplete data, effectively processing the IHIS has become a significant challenge. The established indiscernibility relations of the existing studies for dealing with IHIS over-amplify the uncertainty of missing values, which may achieve unsatisfactory results. In this paper, we propose an approximate supplement-based neighborhood rough set model (AS-NRSM) to deal with the data of IHISs. Specifically, we propose a method to approximate supplement missing values with known values or constructed interval values, and the original IHIS is becoming the constructed IHIS*. Then, we formulate a novel similarity function to construct the improved neighborhood tolerance relation and the corresponding neighborhood tolerance classes. Finally, we design two experiments on 5 UCI data sets by introducing three performance metrics. Experimental results illustrate that the proposed AS-NRSM has higher classification performance than the two representative models.

Keywords: Incomplete data · Hybrid data · Neighborhood rough sets · Incomplete hybrid information systems

1 Introduction

With the fast development of the information era, more and more data in the real world show heterogeneous and incomplete. For example, the medical records contain categorical data, e.g., blood type (O, A, B, AB), gender (Male, Female) and marital status (Married, Unmarried), and numerical data, e.g., blood lipid (mmol/L), body temperature ($°C$) and height (cm). In addition, due to some unpredictable factors like human omission or equipment failure, missing data

F. Liu et al. (Eds.): PRICAI 2023, LNAI 14327, pp. 281–293, 2024.
https://doi.org/10.1007/978-981-99-7025-4_25

will occur occasionally. At present, in the fields of data mining and knowledge discovery, methods of efficiently processing these heterogeneous and incomplete data have become a research hotspot.

Rough set theory is an effective tool for addressing vague and uncertain issues, and has been used successfully in many realms [6,8,12,19–21]. The classical rough set model can only deal with complete and categorical data in information systems (ISs) based on equivalence relation. ISs with incomplete and hybrid data are called incomplete hybrid information systems (IHISs), which exist widely in practical applications. Thus, many investigations have extended the rough set model to process the information systems with hybrid data [1,3,11,14,18] and incomplete data [9,10,13,15] separately.

Recently, there has been an increasing interest in simultaneously considering hybrid data and incomplete data [2,5,16,22,23]. Huang et al. [5] given two pseudo-distance functions according to two semantics of missing value (i.e., "lost value" and "do not care") in IHISs. For only one semantic of missing value, namely, "lost value", Ge et al. [2] proposed an improved neighborhood rough set model (NRSM) in an IHIS to process the missing data. Zhang et al. [22] proposed an attribute reduction algorithm in IHIS by introducing the Dempster-Shafer evidence theory in the distance function. Wang et al. [16] proposed the decision-theoretic rough set model in IHIS with image and employed it in a medical diagnosis example. However, in these models, missing values are generally considered to be equal to all known values in the corresponding domain. Intuitively, the lost value should be similar to part of values in the corresponding domain, rather than be equal to all known values. Therefore, the existing models amplify the uncertainty of lost values, which may acquire unreasonable classification results. The specific analysis and statement of the problem are presented in Sect. 2.

To better describe the uncertainty of lost values and improve the classification performance in IHIS, we propose an Approximate Supplement-based Neighborhood Rough Set Model (AS-NRSM). Considering the heterogeneity of categorical and numerical data, we firstly approximately supplement lost values with known categorical or interval values, to avoid the uncertainty amplification caused by the above-mentioned models. Then the IHIS is becoming a constructed IHIS* with only one semantic (i.e., "do not care"). In the constructed IHIS*, we define the corresponding similarity function to simultaneously deal with three data types: categorical, numerical, and the replaced interval values. Then, we construct the AS-NRSM based on the similarity function. Finally, two experiments are designed and implemented to verify the effectiveness of the proposed AS-NRSM.

The main contributions are presented as follows. (1) To describe the uncertainty of missing values as accurately as possible, we propose a method of approximate supplement for the lost values in the original IHIS. (2) To address three types of data in the constructed IHIS*, we define a novel similarity function to measure the similarity between objects. (3) We design an algorithm of AS-NRSM and introduce three performance metrics into the AS-NRSM to verify the

performance of the proposed model. This paper is organized as follows. In Sect. 2, some basic notions of IHIS are reviewed and a specific analysis of the existing models is given. In Sect. 3, we give a method of approximating supplements and discuss the similarity function in IHIS*. The AS-NRSM will be established eventually. In Sect. 4, several experiments are carried out on 5 UCI data sets to compare the performance of the proposed method and two existing methods. Finally, the conclusion of this paper and the possible future works are presented in Sect. 5.

2 Preliminaries

In this section, we briefly review some concepts about IHIS. Then, we shortly analyze the irrationality of two existing NRSMs when processing some cases in IHIS.

2.1 Incomplete Hybrid Information Systems (IHISs)

An IHIS can be denoted as $\Omega = (U, A, V, f, ?, *)$, where $A = A^C \cup A^N$, $A^C \cap A^N = \emptyset$, and A is a non-empty finite set of attribute, A^C and A^N are the categorical and numerical attribute sets, respectively. "?" and "*" denote two semantics of missing values, namely, "lost value" and "do not care".

Table 1. $\Omega = (U, A, V, f, ?, *)$

U	a_1	a_2	a_3	a_4	a_5
x_1	S	A	0.6	0.8	0.9
x_2	M	*	*	0.2	0.5
x_3	*	C	0.3	0.2	0.6
x_4	?	B	?	0.2	?
x_5	S	?	0.1	?	0.1
x_6	M	?	?	?	?
x_7	M	?	?	?	?
x_8	M	*	*	0.9	0.1

Example 1. We use Table 1 to illustrate an IHIS, where $U = \{x_1, \cdots, x_8\}$, $A = \{a_1, \cdots, a_5\}$. For categorical attribute set $A^C = \{a_1, a_2\}$, $V_{a_1} = \{S, M\}$ and $V_{a_2} = \{A, B, C\}$, we have some objects with missing values, e.g., $f(x_3, a_1) = *$, $f(x_4, a_1) =?$. For numerical attribute set $A^N = \{a_3, a_4, a_5\}$, $V_{a_3}, V_{a_4}, V_{a_5} \in [0, 1]$, there are some missing values, e.g., $f(x_2, a_3) = *$, $f(x_4, a_3) =?$.

To cope with the numerical data, the theory of neighborhood rough sets (NRSs) is introduced [1,4,17]. For numerical attribute A^N, through a neighborhood radius $\delta \in [0, 1]$. The neighborhood relation $N^{\delta}_{A^N}$ is defined by:

$$N^{\delta}_{A^N} = \{(x, y) \in U^2 \mid \forall a \in A^N, d_a(x, y) \leq \delta\}, \tag{1}$$

where $d_a(x, y)$ is the distance of x and y on the attribute $a \in A^N$. For any $x \in U$, the neighborhood class $N^{\delta}_{A^N}(x)$ is defined by:

$$N^{\delta}_{A^N}(x) = \{(x, y) \in U^2 \mid \forall a \in A^N, (x, y) \in N^{\delta}_{A^N}\}. \tag{2}$$

In IHIS, Huang et al. [5] proposed Three-Way Neighborhood Decision Model (TWNDM). Ge et al. [2] constructed an Improved Neighborhood Rough Set Model (INRSM). The lost value can be equivalent to any one of the known values in the corresponding domain in TWNDM and INRSM. The specific analysis of the two models are given in Example 2.

Example 2. Consider the IHIS in Example 1, we assume the neighborhood radius $\delta = 0.2$, the neighborhood classes $N^{\delta}_{A}(x)$ are respectively constructed based on TWNDM [5] and INRSM [2] as Table 2.

Table 2. Neighborhood classes

Method	TWNDM	INRSM
$N^{\delta}_{A}(x_1)$	$\{x_1\}$	$\{x_1, x_6, x_7\}$
$N^{\delta}_{A}(x_2)$	$\{x_2, x_3, x_5, x_6, x_7\}$	$\{x_2, x_3, x_5, x_6, x_7\}$
$N^{\delta}_{A}(x_3)$	$\{x_2, x_3, x_6, x_7\}$	$\{x_2, x_3, x_4, x_5, x_6, x_7\}$
$N^{\delta}_{A}(x_4)$	$\{x_2, \boldsymbol{x_4}, \boldsymbol{x_5}, x_6, x_7\}$	$\{x_2, x_3, \boldsymbol{x_4}, \boldsymbol{x_5}, x_6, x_7, x_8\}$
$N^{\delta}_{A}(x_5)$	$\{\boldsymbol{x_4}, \boldsymbol{x_5}\}$	$\{x_3, \boldsymbol{x_4}, \boldsymbol{x_5}, x_6, x_7, x_8\}$
$N^{\delta}_{A}(x_6)$	$\{x_2, x_3, x_5, \boldsymbol{x_6}, \boldsymbol{x_7}, x_8\}$	$\{x_1, x_2, x_3, x_4, x_5, \boldsymbol{x_6}, \boldsymbol{x_7}, x_8\}$
$N^{\delta}_{A}(x_7)$	$\{x_2, x_3, x_5, \boldsymbol{x_6}, \boldsymbol{x_7}, x_8\}$	$\{x_1, x_2, x_3, x_4, x_5, \boldsymbol{x_6}, \boldsymbol{x_7}, x_8\}$
$N^{\delta}_{A}(x_8)$	$\{x_6, x_7, x_8\}$	$\{x_3, x_4, x_6, x_7, x_8\}$

We note that TWNDM and INRSM are not very reasonable when dealing with the following two special cases.

(1) The lost values of two objects always alternate. As shown in Table 1, for $\forall a \in A$, objects x_4 and x_5 always satisfy two conditions: (1) if $f(x_4, a) \neq ?$ then $f(x_5, a) = ?$, (2) if $f(x_5, a) \neq ?$ then $f(x_4, a) = ?$. Intuitively, objects x_4 and x_5 are unlikely to be similar. Hence, it is unreasonable to classify x_4 and x_5 into the same neighborhood class. However, by TWNDM and INRSM, we have $x_4 \in N^{\delta}_{A}(x_5)$ and $x_5 \in N^{\delta}_{A}(x_4)$.

(2) If two objects have the same or similar known values under very few attributes, and the rest values are all lost. As shown in Table 1, objects x_6 and x_7 have only one known value in a_1, and the rest of their values are all lost values. The possibility that x_6 and x_7 to be similar is very low. Therefore, it's unreasonable to classify x_6 and x_7 into the same neighborhood class. Still, we have $x_6 \in N^{\delta}_{A}(x_7)$ and $x_7 \in N^{\delta}_{A}(x_6)$ by TWNDM and INRSM.

Herein, the lost value is assumed to be equivalent to all known values in the corresponding domain, which increase the uncertainty of the lost value, and further leads to unreasonable classification results. This problem has been explicated and shown in Example 1 and 2. In the following, we propose an approximate supplement method for solving this problem.

3 Approximate Supplement-Based NRSM

3.1 Approximate Supplement in IHIS

In this subsection, we propose the method of approximate supplement in IHIS. Let x be an object with the lost value "?" in IHIS. Then the lost value will be approximately supplemented by the known categorical value or a constructed interval value.

In an IHIS, $\forall a \in A^N$, let $V_a^\dagger = \{v_a^1, v_a^2, v_a^3, \cdots, v_a^n\}$ be a known value set of attribute a, and $n = |V_a^\dagger|$ be the number of known values in attribute a. For any $a \in A^N$, we have the standard deviation in attribute a:

$$Std_a = \sqrt{\frac{\sum_{i=1}^{n}(v_a^i - \mathrm{AVG_a})^2}{n}}, \tag{3}$$

where $\mathrm{AVG_a}$ is the average value of $V_a^\dagger$, i.e., $\mathrm{AVG_a} = \sum_{i=1}^{n} v_a^i/n$.

$$\delta_a = \frac{Std_a}{\lambda}, \tag{4}$$

where λ is a parameter for the neighborhood radius.

Definition 1. *Suppose $\Omega = (U, A, V, f, *, ?)$ is an IHIS, $A = A^C \cup A^N$. For any $x, y \in U$, the distance function in the categorical attribute $a \in A^C$ is defined by*

$$d_a(x,y) = \begin{cases} 0, & x = y \vee f(x,a) = f(y,a); \\ 0, & f(x,a) = * \vee f(y,a) = *; \\ 0, & f(x,a) =? \vee f(y,a) =?; \\ 1, & otherwise. \end{cases} \tag{5}$$

And the distance function under the numerical attribute $a \in A^N$ is defined by

$$d_a(x,y) = \begin{cases} 0, & x = y \vee f(x,a) = * \vee f(y,a) = *; \\ \delta_a, & (f(x,a) =? \wedge f(y,a) \in V_a^\dagger) \\ & \vee (f(x,a) \in V_a^\dagger \wedge f(y,a) =?) \\ & \vee (f(x,a) =? \wedge f(y,a) =?); \\ \frac{|f(x,a) - f(y,a)|}{|max(a) - min(a)|}, & otherwise, \end{cases} \tag{6}$$

where $max(a)$ and $min(a)$ are the maximum and minimum values in attribute a.

In Eq. (6), when two objects have a lost value "?" or a do not care value "*", the distance between them is no longer considered to be 0, but is considered to be δ_a. Intuitively, two objects have a slim probability to be equal under the numerical attribute in these cases. Therefore, the neighborhood radius δ_a is introduced.

Definition 2. *Let $\Omega = (U, A, V, f, *, ?)$ be an IHIS. The distance function under the attribute set A is defined as follows:*

$$d_A(x, y) = \frac{\sum_{k=1}^{|A|} d_{a_k}(x, y)}{|A|}, \tag{7}$$

where $|\cdot|$ denotes the cardinality of a set.

Definition 3. *Let $\Omega = (U, A, V, f, *, ?)$ be an IHIS, for any $a \in A$, we have*

$$X_c(a) = \{x | x \in U, f(x, a) \neq * \wedge f(x, a) \neq ?\}, \tag{8}$$

where the complete class $X_c(a)$ denotes the set of all objects with known values in attribute a.

For any $x \in U$, we can easily find an object from $X_c(a)$ which is the most similar to x, i.e.,

$$sim(x) = \{y \in X_c(a) \mid min(d_A(x, y))\}, \tag{9}$$

where $min(d_A(\cdot))$ means the minimal distance under the attribute set A. When $f(x, a) =?$ and $a \in A^C$, the known categorical value of the object which is the most similar to x can be used to supplement the "lost value", i.e.,

$$f^*(x, a) = f(sim(x), a). \tag{10}$$

When $f(x, a) =?$ and $a \in A^N$, a constructed interval value is used to replace the "lost value", i.e.,

$$f^*(x, a) = [max(0, f(sim(x), a) - \delta_a), min(f(sim(x), a) + \delta_a), 1], \tag{11}$$

where $f^*(x, a) \in [0, 1]$ when $a \in A^N$. By approximately supplementing the lost value "?", we can maintain the uncertainty of the lost value "?" to some extent.

Example 3. We continue with the IHIS in Example 1 and assume $\lambda = 2$. For all of the "lost value" in IHIS, such as $f(x_4, a_1) =?$ and $f(x_6, a_3) =?$, we have $sim(x_4) = x_2$ and $sim(x_6) = x_3$, according to Eqs. (7), (8) and (9). Therefore, we can obtain $f^*(x_4, a_1) = f(x_2, a_1) =$ **[*M*]** and $f^*(x_6, a_3) = [f(x_3, a_3) - \delta_{a_3}, f(x_3, a_3) + \delta_{a_3}] =$ **[0.197, 0.403]** to instead of "?", respectively, according to Eqs. (10) and (11). Similarly, we can approximately supplement all of the "lost value" in IHIS to get a constructed IHIS* as shown in Table 3.

3.2 Construction of AS-NRSM in IHIS*

The constructed IHIS* contains four types of data: (1) the categorical attribute values; (2) the crisp value under the numerical attribute; (3) the "do not care" value, i.e., "*"; (4) the supplemented interval value under the numerical attribute. In this subsection, a novel similarity function is given for dealing

Table 3. A constructed IHIS*

U	a_1	a_2	a_3	a_4	a_5
x_1	S	A	0.6	0.8	0.9
x_2	M	*	*	0.2	0.5
x_3	*	C	0.3	0.2	0.6
x_4	[M]	B	**[0, 0.203]**	0.2	**[0.346, 0.654]**
x_5	S	[B]	0.1	**[0.04, 0.36]**	0.1
x_6	M	[B]	**[0.197, 0.403]**	**[0.04, 0.36]**	**[0.346, 0.654]**
x_7	M	[C]	**[0.197, 0.403]**	**[0.74, 1]**	**[0, 0.254]**
x_8	M	*	*	0.9	0.1

with three types of data in the constructed IHIS*, i.e., (1), (2) and (4). Moreover, considering (3) in IHIS*, we would discuss the neighborhood tolerance relation and neighborhood tolerance classes based on the defined similarity function. Consequently, we propose AS-NRSM in the constructed IHIS*.

Considering the supplemented values in categorical attributes are known categorical values, and in numerical attributes, they are interval values within a width of 2δ, we proposed the following similarity function.

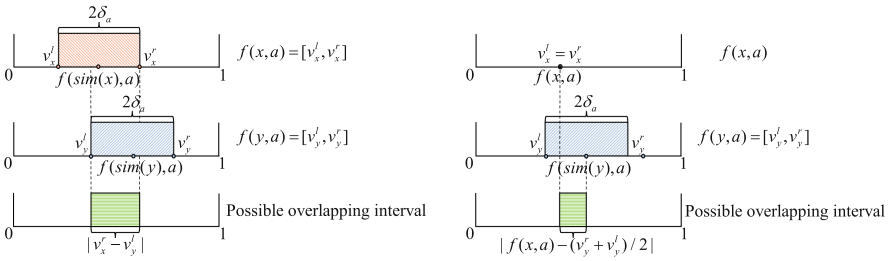
Definition 4. *Let $\Omega = (U, A, V, f, *)$ be a constructed IHIS*, $x, y \in U$, the similarity function under attribute $a \in A^C \cup A^N$ is defined by*

$$S_a^*(x,y) = \begin{cases} 1, & a \in A^C \wedge f(x,a) = f(y,a); \\ 0, & a \in A^C \wedge f(x,a) \neq f(y,a); \\ 1, & a \in A^N \wedge |f(x,a) - f(y,a)| \leq \delta_a; \\ \frac{2 \star min(|v_x^l - v_y^r|, |v_y^l - v_x^r|)}{(v_x^r - v_x^l) + (v_y^r - v_y^l)}, & a \in A^N \wedge (v_x^l \neq v_x^r \wedge v_y^l \neq v_y^r) \wedge \\ & |f(sim(x),a) - f(sim(y),a)| \leq 2\delta_a; \\ 1 - \frac{|f(x,a) - (v_y^l + v_y^r)/2|}{(v_y^r - v_y^l)}, & a \in A^N \wedge (v_x^l = v_x^r \wedge v_y^l \neq v_y^r) \wedge \\ & (v_y^l \leq f(x,a) \leq v_y^r); \\ 0, & otherwise, \end{cases} \tag{12}$$

where $|\cdot|$ denotes the absolute value.

Herein, v_x^l and v_x^r are the left and right endpoints of the interval $f(x,a)$, respectively. When $v_x^l = v_x^r$, $f(x,a)$ is a crisp value in the original IHIS. When $v_x^l < v_x^r$, $f(x,a)$ is an interval value in the constructed IHIS*. To more intuitively understand the proposed functions for the similarity of interval values in Eq. (12), the possible overlapping interval between the two objects with two interval values is shown in Fig. 1-(a), one interval value and one crisp value is shown in Fig. 1-(b).

In the light of tolerance relation [7], we give the neighborhood tolerance relation based on the defined similarity function as follows.



(a) Two objects with two interval values (b) Two objects with one interval value and one crisp value

Fig. 1. The possible overlapping interval of two different cases.

Definition 5. *Let $\Omega = (U, A, V, f, *)$ be a constructed IHIS*, δ_a is a neighborhood radius. For $\forall a \in A$, neighborhood tolerance relation NT_a^δ is defined by*

$$NT_a^\delta = \{(x, y) \in U^2 \mid S_a^*(x, y) \geq \zeta \vee (f(x, a) = * \vee f(y, a) = *)\}, \tag{13}$$

where $\zeta \in [0, 1]$ is a threshold for similarity and NT_a^δ satisfies symmetry and reflexivity.

Definition 6. *Let $\Omega = (U, A, V, f, *)$ be a constructed IHIS*, for $\forall a \in A$, the neighborhood tolerance class of any object $x \in U$ is defined by*

$$NT_a^\delta(x) = \{y \in U \mid (x, y) \in NT_a^\delta\}. \tag{14}$$

The neighborhood tolerance class under the attribute set A is defined by

$$NT_A^\delta(x) = \{y \in U \mid \forall a \in A, (x, y) \in NT_a^\delta\} = \bigcap_{a \in A} NT_a^\delta(x). \tag{15}$$

Example 4. We continue with Example 3 and suppose $\zeta = 0.2$. According to Eqs. (13), (14) and (15), we can build neighborhood classes of every object as follows:

As shown in Table 4, the neighborhood classes of TWNDM, INRSM are far looser than AS-NRSM:

(1) Objects x_4 and x_5 are in the same neighborhood class, i.e., $x_4 \in N_A^\delta(x_5), x_5 \in N_A^\delta(x_4)$ in TWNDM and INRSM, but by AS-NRSM they are irrelevant.
(2) Objects x_6 and x_7 are in the same neighborhood class, i.e., $x_6 \in N_A^\delta(x_7), x_7 \in N_A^\delta(x_6)$ in TWNDM and INRSM, but by AS-NRSM, they are irrelevant.

According to analysis in Example 2, the possibility that objects x_4 and x_5, x_6 and x_7 belong to the same neighborhood class is very low. By the method of AS-NRSM, these objects are classified into the appropriate neighborhood tolerance classes, namely, the classification based on AS-NRSM we proposed is more reasonable. The algorithm of AS-NRSM can be described as follows:

Table 4. Neighborhood classes of three methods

Method	TWNDM	INRSM	AS-NRSM
$N_A^\delta(x_1)$	$\{x_1\}$	$\{x_1, x_6, x_7\}$	$\{x_1\}$
$N_A^\delta(x_2)$	$\{x_2, x_3, x_5, x_6, x_7\}$	$\{x_2, x_3, x_5, x_6, x_7\}$	$\{x_2, x_3, x_4, x_6\}$
$N_A^\delta(x_3)$	$\{x_2, x_3, x_6, x_7\}$	$\{x_2, x_3, x_4, x_5, x_6, x_7\}$	$\{x_2, x_3\}$
$N_A^\delta(x_4)$	$\{x_2, \boldsymbol{x_4}, \boldsymbol{x_5}, x_6, x_7\}$	$\{x_2, x_3, \boldsymbol{x_4}, \boldsymbol{x_5}, x_6, x_7, x_8\}$	$\{x_2, \boldsymbol{x_4}\}$
$N_A^\delta(x_5)$	$\{\boldsymbol{x_4}, \boldsymbol{x_5}\}$	$\{x_3, \boldsymbol{x_4}, \boldsymbol{x_5}, x_6, x_7, x_8\}$	$\{\boldsymbol{x_5}\}$
$N_A^\delta(x_6)$	$\{x_2, x_3, x_5, \boldsymbol{x_6}, \boldsymbol{x_7}, x_8\}$	$\{x_1, x_2, x_3, x_4, x_5, \boldsymbol{x_6}, \boldsymbol{x_7}, x_8\}$	$\{x_2, \boldsymbol{x_6}\}$
$N_A^\delta(x_7)$	$\{x_2, x_3, x_5, \boldsymbol{x_6}, \boldsymbol{x_7}, x_8\}$	$\{x_1, x_2, x_3, x_4, x_5, \boldsymbol{x_6}, \boldsymbol{x_7}, x_8\}$	$\{\boldsymbol{x_7}, x_8\}$
$N_A^\delta(x_8)$	$\{x_6, x_7, x_8\}$	$\{x_3, x_4, x_6, x_7, x_8\}$	$\{x_7, x_8\}$

Algorithm 1. The Algorithm of AS-NRSM

Input: (1) An IHIS $\Omega = (U, A, V, f, *, ?)$, where $U = \{x_i, x_j \mid 1 \leq i, j \leq n\}$ and $A = \{a_k \mid 1 \leq k \leq m\}$; (2) The neighborhood radius parameter λ.

Output: Neighborhood tolerance classes $NT_A^\delta(x)$.

for $1 \leq k \leq m$ **do**
  for $1 \leq i \leq n$ **do**
    if $f(x_i, a_k) \neq$ "?" $\wedge f(x_i, a_k) \neq$ "*" **then**
      $f(x_i, a_k) \in V_{a_k}^\dagger$; // $V_{a_k}^\dagger$ is the known values set in attribute a_k;
Compute the neighborhood radius δ_{a_k};
for $1 \leq i \leq n$ **do** // Approximately replacing the lost values by Definition 3;
  for $1 \leq j \leq n$ **do**
    if $\forall a_k \in A, f(x_i, a_k) =?$ **then**
      According to Equations (7), (8) and (9), compute $d_A(x_i, x_j)$, $X_c(a_k)$ and $sim(x_i)$;
      According to Equations (10) and (11), replace the lost value by $f^*(x_i, a_k)$;
for $1 \leq i \leq n$ **do** // calculating the distances by Definition 4;
  for $1 \leq j \leq n$ **do**
    for $1 \leq k \leq m$ **do**
      According to Equation (12), compute $d_{a_k}^*(x_i, x_j)$ in IHIS*;
According to Definition 5, compute NT_a^δ;
According to Definition 6, compute $NT_a^\delta(x)$ and $NT_A^\delta(x)$.

The time complexity of Algorithm 1 is $O(mn^2)$.

4 Experiments and Analysis

4.1 Performance Comparisons of Different Algorithms

To better reflect the performance of the proposed algorithm, we conducted simulation experiments. The 5 data sets are downloaded from the University of California at Irvine (UCI) data sets (http://archive.ics.uci.edu/ml/) and displayed in Table 5. The applicability and performance of TWNDM, INRSM and AS-NRSM were evaluated for different types of data.

Table 5. The description of data sets

No.	Datasets	Objects	Conditional attributes (A)			Decision attribute (d)
			Categorical(A^C)	Numerical(A^N)	Total($\|A\|$)	
1	Segment	2310	0	19	19	1
2	Heart	270	0	12	12	1
3	Annealing	798	9	6	15	1
4	MPG	398	4	5	9	1
5	Abalone	4177	1	8	9	1

We constructed three levels of missing values in the complete data set, that is, (1) Replacing 0%, 5%, and 10% known values as the low-missing level. (2) Replacing 15%, 20%, and 25% known values as the medium-missing level. (3) Replacing 30%, 35%, and 40% known values as the high-missing level. Especially, the missing values are composed of "lost value" and "do not care" values with a ratio of 4:1 when carrying out AS-NRSM and TWNDM algorithms. Besides, all of the missing values are "lost value" when performing INRSM algorithms because the semantic of "do not care" is not considered in INRSM.

Herein, we introduce three metrics for measuring the neighborhood class quality, namely, Precision (P), Recall (R) and F1-score ($F1$). When the classification is too loose, it will get low P and $F1$ and high R. When classification is too strict, R and $F1$ will be low, and conversely, P will be high. We apply the three metrics to the 5 data sets of Table 5, the experimental results are shown in Table 6. We can get P of AS-NRSM is 0.95 ± 0.05 in the low-missing level of the Segment data set, where 0.95 is the average performance in 0%, 5%, and 10% missing values and 0.05 is the standard deviation. Similarly, we can obtain the average performance metrics of algorithms at the medium-missing level and high-missing level. Herein, the optimal metrics are highlighted in bold. It is observed that three metrics: P, R and $F1$ of three algorithms decrease with the level of missing values increase. However, as the level of missing values increases, the $F1$ of TWNDM and INRSM is lower than AS-NRSM in most data sets.

Table 6. Performance comparison by three algorithms in different data sets.

Data sets \ Missing Level		AS-NRSM			TWNDM			INRSM		
		P	R	F1	P	R	F1	P	R	F1
Segment	Low	**0.95 ± 0.05**	0.95 ± 0.05	**0.95 ± 0.05**	0.88 ± 0.12	**0.96 ± 0.04**	0.91 ± 0.08	0.80 ± 0.06	0.95 ± 0.01	0.87 ± 0.03
	Medium	**0.82 ± 0.05**	0.82 ± 0.05	**0.82 ± 0.05**	0.60 ± 0.06	0.90 ± 0.01	0.73 ± 0.04	0.62 ± 0.05	**0.93 ± 0.00**	0.75 ± 0.02
	High	**0.70 ± 0.02**	0.68 ± 0.06	**0.69 ± 0.04**	0.35 ± 0.07	0.85 ± 0.01	0.50 ± 0.07	0.49 ± 0.06	**0.90 ± 0.01**	0.68 ± 0.03
Heart	Low	**1.00 ± 0.00**	0.99 ± 0.01	**0.99 ± 0.01**	0.93 ± 0.08	**1.00 ± 0.00**	0.96 ± 0.04	0.98 ± 0.02	0.91 ± 0.02	0.94 ± 0.02
	Medium	**0.98 ± 0.01**	0.97 ± 0.00	**0.97 ± 0.01**	0.49 ± 0.19	**0.99 ± 0.01**	0.65 ± 0.17	0.83 ± 0.08	0.88 ± 0.01	0.85 ± 0.05
	High	**0.90 ± 0.02**	0.96 ± 0.01	**0.93 ± 0.01**	0.15 ± 0.08	**0.98 ± 0.00**	0.25 ± 0.12	0.63 ± 0.04	0.86 ± 0.01	0.73 ± 0.02
Annealing	Low	**1.00 ± 0.00**	**0.98 ± 0.02**	**0.99 ± 0.01**	0.99 ± 0.01	**0.98 ± 0.02**	**0.99 ± 0.01**	0.96 ± 0.03	0.89 ± 0.02	0.93 ± 0.02
	Medium	**0.99 ± 0.00**	**0.93 ± 0.01**	**0.96 ± 0.01**	**0.99 ± 0.00**	0.92 ± 0.01	**0.96 ± 0.00**	0.89 ± 0.02	0.86 ± 0.01	0.87 ± 0.02
	High	0.98 ± 0.00	**0.93 ± 0.01**	**0.95 ± 0.00**	**0.99 ± 0.01**	0.92 ± 0.01	**0.95 ± 0.00**	0.82 ± 0.02	0.82 ± 0.01	0.82 ± 0.01
MPG	Low	**0.96 ± 0.03**	**0.99 ± 0.01**	**0.97 ± 0.02**	0.94 ± 0.06	0.96 ± 0.05	0.95 ± 0.05	0.87 ± 0.02	0.81 ± 0.02	0.83 ± 0.02
	Medium	**0.85 ± 0.06**	0.88 ± 0.02	**0.86 ± 0.04**	0.78 ± 0.04	**0.89 ± 0.01**	0.83 ± 0.02	0.83 ± 0.01	0.77 ± 0.01	0.80 ± 0.01
	High	0.74 ± 0.03	0.82 ± 0.03	**0.78 ± 0.03**	0.67 ± 0.03	**0.86 ± 0.02**	0.76 ± 0.03	**0.77 ± 0.02**	0.73 ± 0.01	0.75 ± 0.01
Abalone	Low	**0.89 ± 0.10**	0.94 ± 0.05	0.92 ± 0.08	**0.89 ± 0.10**	**0.96 ± 0.06**	**0.95 ± 0.05**	0.76 ± 0.04	**0.96 ± 0.00**	0.86 ± 0.02
	Medium	**0.82 ± 0.03**	0.86 ± 0.01	**0.84 ± 0.03**	0.75 ± 0.03	0.87 ± 0.03	0.80 ± 0.03	0.63 ± 0.03	**0.95 ± 0.00**	0.79 ± 0.02
	High	**0.72 ± 0.03**	0.85 ± 0.03	**0.78 ± 0.00**	0.64 ± 0.00	0.86 ± 0.02	0.74 ± 0.01	0.55 ± 0.02	**0.94 ± 0.01**	0.75 ± 0.01

Due to $F1$ being a more comprehensive metric than P and R to evaluate the performance of the algorithms, we choose $F1$ as the judging metric. To make a more intuitive comparison of different algorithms, we take 0% - 40% of missing values as x-axis to observe the metric $F1$ changes of three algorithms, as shown in Fig. 2. The performance $F1$ of the three algorithms decreases with the increase of missing ratio, while the proposed algorithm AS-NRSM can achieve optimal performance in most datasets.

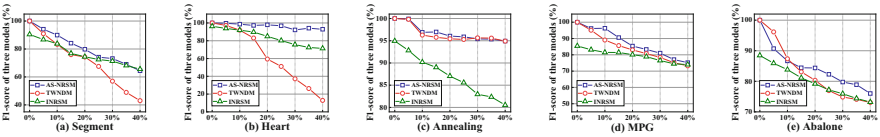


Fig. 2. F1-score of algorithms in different ratios of missing values

In order to explore the influence of different neighborhood parameters λ on the classification performance, we conduct a comparative experiment in a low missing level. Specifically, we move λ from 1 to 5 with a step of 0.5 to compare the $F1$ performance of three algorithms under different δ neighborhoods constructed by λ.

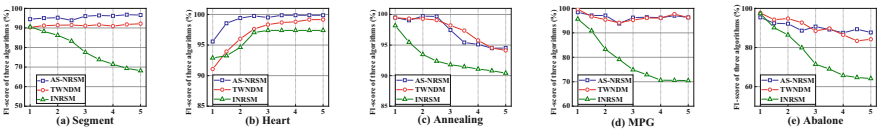


Fig. 3. $F1$ of algorithms in different neighborhood parameter λ

As shown in Fig. 3, the performance $F1$ of INRSM is sensitive to parameter λ, which decreases as the λ becomes larger in (a) Segment, (c) Annealing, (d) MPG and (e) Abalone, and increases as the parameters become larger in (b) Heart. In addition, the $F1$ of AS-NRSM and TWNDM in the 5 data sets are basically synchronous floating. Still, it's observed that AS-NRSM can achieve the optimal performance $F1$ in most data sets, and is more stable than the other two algorithms.

5 Conclusion and Future Work

Recently, a few studies have emerged to focus on dealing with incomplete hybrid data in IHISs. However, the indiscernibility relations of the existing studies are too loose, and may lead to unreasonable classification results. To describe the uncertainty of missing values as much as possible and enhance the performance of classification, we proposed the Approximate supplement-based Neighborhood

Rough Set Model (AS-NRSM). First, the lost values in IHIS have been approximately replaced by the known values or interval values. Then we obtained a constructed IHIS* with only one semantic from the original IHIS. Next, we defined a novel similarity function for the constructed IHIS* which contains three types of data. Then, the AS-NRSM is constructed for an IHIS*. Finally, comparative experiments are carried out to prove the performance of AS-NRSM. In the future, we will extend our work to a more comprehensive and realistic data environment to validate the effectiveness of the model.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (62072320) and the Natural Science Foundation of Sichuan Province (No. 2022NSFSC0569, No. 2022NSFSC0929).

References

1. Chen, H.M., Li, T.R., Fan, X., Luo, C.: Feature selection for imbalanced data based on neighborhood rough sets. Inf. Sci. **483**, 1–20 (2019)
2. Ge, H., Yang, C., Xu, Y.: Incremental updating three-way regions with variations of objects and attributes in incomplete neighborhood systems. Inf. Sci. **584**, 479–502 (2021)
3. Hu, Q., Xie, Z., Yu, D.: Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. Pattern Recogn. **40**(12), 3509–3521 (2007)
4. Hu, Q., Yu, D., Liu, J., Wu, C.: Neighborhood rough set based heterogeneous feature subset selection. Inf. Sci. **178**, 3577–3594 (2008)
5. Huang, Q.Q., Li, T.R., Huang, Y.Y., Yang, X.: Incremental three-way neighborhood approach for dynamic incomplete hybrid data. Inf. Sci. **541**, 98–122 (2020)
6. Kryszkiewicz, M.: Properties of incomplete information systems in the framework of rough sets. Rough Sets Knowl. Disc. **1**, 422–450 (1998)
7. Kryszkiewicz, M.: Rough set approach to incomplete information systems. Inf. Sci. **112**, 39–49 (1998)
8. Kryszkiewicz, M.: Rules in incomplete information systems. Inf. Sci. **113**, 271–292 (1999)
9. Luo, C., Li, T.R., Yao, Y.Y.: Dynamic probabilistic rough sets with incomplete data. Inf. Sci. **417**, 39–54 (2017)
10. Luo, J.F., Fujita, H., Yao, Y.Y., Qin, K.Y.: On modeling similarity and three-way decision under incomplete information in rough set theory. Knowl. Based Syst. **191**, 105251 (2020)
11. Pan, X.H., He, S.F., Wang, Y.M.: Multi-granular hybrid information-based decision-making framework and its application to waste to energy technology selection. Inf. Sci. **587**, 450–472 (2022)
12. Pawlak, Z.: Rough sets. IJICS **11**, 341–356 (1982)
13. Stefanowski, J., Tsoukiás, A.: Incomplete information tables and rough classification. Comput. Intell. **17**, 545–566 (2001)
14. Sun, B., Ma, W., Chen, D.: Rough approximation of a fuzzy concept on a hybrid attribute information system and its uncertainty measure. Inf. Sci. **284**, 60–80 (2014)
15. Sun, L., Wang, L.Y., Ding, W.P., Qian, Y.H., Xu, J.C.: Neighborhood multi-granulation rough sets-based attribute reduction using Lebasque and entropy measures in incomplete neighborhood decision systems. Knowl. Based Syst. **192**, 105373 (2019)

16. Wang, P., Zhang, P., Li, Z.: A three-way decision method based on gaussian kernel in a hybrid information system with images: An application in medical diagnosis. Appl. Soft Comput. **77**, 734–749 (2019)
17. Wang, Q., Qian, Y., Liang, X., Guo, Q., Liang, J.: Local neighborhood rough set. Knowl. Based Syst. **153**, 53–64 (2018)
18. Xu, J., Qu, K., Sun, Y., Yang, J.: Feature selection using self-information uncertainty measures in neighborhood information system. Appl. Intell. **53**, 4524–4540 (2022)
19. Yang, J.L., Yao, Y.Y.: Semantics of soft sets and three-way decision with soft sets. Knowl. Based Syst. **194**, 105538 (2020)
20. Yang, J.L., Yao, Y.Y.: A three-way decision-based construction of shadowed sets from Atanassov intuitionistic fuzzy sets. Inf. Sci. **577**, 1–21 (2021)
21. Yao, Y.Y.: Neighborhood systems and approximate retrieval. Inf. Sci. **176**, 3431–3452 (2006)
22. Zhang, Q., Qu, L., Li, Z.: Attribute reduction based on d-s evidence theory in a hybrid information system. Int. J. Approx. Reason. **148**, 202–234 (2022)
23. Zhang, X., Chen, X., Xu, W., Ding, W.: Dynamic information fusion in multi-source incomplete interval-valued information system with variation of information sources and attributes. Inf. Sci. **608**, 1–27 (2022)