Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 33(11): 6672–6699.
DOI: 10.3934/era.2025295
Received: 15 September 2025
Revised: 22 October 2025
Accepted: 27 October 2025
Published: 13 November 2025

Research article

Feature selection in partially labeled multiset-valued decision information systems based on local conditional entropy

Dongliang Li[1,2] **and Yanlan Zhang**[1,2,*]

[1] School of Computer Science, Minnan Normal University, Zhangzhou 363000, China

[2] Key Laboratory of Data Science and Intelligence Application, Fujian Province University, Zhangzhou 363000, China

* **Correspondence:** Email: zhangppff@126.com.

Abstract: As the cost of data labeling continues to escalate, research on feature selection for a partially labeled multiset-valued decision information system (p-MSVDIS) has emerged as a core challenge in the field of data mining. Information entropy, as an uncertainty measurement tool, can be used for feature selection via global equivalence relations. However, it fails to capture features within local data regions. Furthermore, in partially labeled scenarios, its limitations in dealing with missing labels cause poor accuracy in feature selection. In contrast, local conditional entropy can accurately characterize the discriminative ability of features in local regions by quantifying the information of the dataset. To address the problem of feature selection in a p-MSVDIS, this paper proposed two algorithms for feature selection in a p-MSVDIS. First, we utilized the Hellinger distance in a p-MSVDIS to define the tolerance classes of different attributes. Second, we introduced local conditional entropy and designed two feature selection algorithms for a p-MSVDIS with predicted labels. Finally, comparative experimental results demonstrated that the proposed algorithms significantly improved classification performance and reduced redundant features in partially labeled data scenarios.

Keywords: p-MSVDIS; local conditional entropy; feature selection; predicted label

1. Introduction

1.1. Research background

In the interdisciplinary field of granular computing [1] and data mining [2], feature selection in an incomplete information system [3] is a core challenge underpinning accurate decision-making in complex scenarios. Redundant features not only reduce computational efficiency but also degrade classification performance due to noise interference; thus, efficient feature selection methods have

become increasingly important.

With the escalating cost of data annotation [4], missing-valued attributes and partially labeled characteristics have become the norm. The multiset-valued decision information system (MSVDIS) [5] regards the values under each attribute as multisets, thereby completely recording the occurrence frequencies of attribute values. This method effectively addresses the issue that traditional methods may lose frequency information. Feature selection of an MSVDIS is also one of the important research contents. As an information measurement tool, entropy [6] provides an intuitive and efficient basis for feature selection by quantifying uncertainty and variable correlations. Feature selection based on entropy can select the most valuable features for the dataset, simplify the model while improving its performance, and thus stand as one of the classic methods in feature selection within machine learning [7]. Therefore, introducing entropy measures into MSVDISs for feature selection is one of the key research focuses. In 2023, Huang et al. [8] employed fuzzy conditional information entropy metrics to perform feature selection on multiset-valued data in supervised scenarios [9], capturing multiset information via iterative models and matrix operations. Li et al. [10] proposed a method for feature selection of p-MSVDISs that constructs dependence and conditional entropy via predicted label strategies, and designed semi-supervised attribute reduction algorithms based on dependence and conditional entropy. In 2025, Guo et al. [11] incorporated an information granulation method based on the θ-tolerance relation into the multiset-valued information system model, and subsequently defined uncertainty measures including θ-information entropy and θ-information quantity. He et al. [12] divided p-MSVDISs into labeled multiset-valued decision information systems and unlabeled multiset-valued decision information systems, then introduced two adaptive semi-supervised attribute selection algorithms. MSVDISs preserve multi-valued frequency information and can handle missing labels. Feature selection of MSVDISs based on entropy can quantify the information content and discriminative power of features, and thus prioritizes the selection of features that can effectively reduce system uncertainty and improve model performance. However, these methods in MSVDISs all utilize global entropy for feature selection on datasets, and, thus, have limitations of relatively high computational complexity and the need to set parameters.

The idea of local rough sets [13] is to eliminate redundant calculations involving objects irrelevant to the target concept during the reduction process, so as to improve computational efficiency. Inspired by this, researchers have combined localized analysis approach with the information measurement property of entropy, proposing to calculate entropy values in local regions of the dataset. Calculating entropy values in local regions of the dataset is used to measure the information contribution of features within local scopes to the target variable, thereby achieving more accurate feature selection. In 2019, Wang et al. [14] proposed the concept of local conditional entropy, investigated its properties, and defined attribute reduction based on local conditional entropy. In 2023, Xie et al. [15] proposed a fuzzy rough attribute reduction method based on local information entropy. This method addressed the problems that traditional information entropy had relatively low computational efficiency in attribute reduction and would lead to overfitting when dealing with large datasets. In 2025, Chen et al. [16] aimed at the problems of ”global computational redundancy” and ”sensitivity to class distribution” existing in the traditional dominance rough set model within imbalanced ordered decision systems, and proposed a new feature selection framework based on local fuzzy dominance neighborhood composite entropy. The advantage of local entropy is that it can measure information uncertainty from a local perspective, capture local data distribution characteristics that global entropy struggles to reflect, and thereby improve

performance and achieve higher efficiency in tasks such as feature selection and classification.

When processing complete datasets, feature selection methods based local conditional entropy exhibit significant efficiency advantages. However, when dealing with a p-MSVDIS, the feature selection performance of traditional conditional entropy methods degrades significantly. The core challenge is how to effectively handle missing labels under incomplete information systems to optimize the selection of feature subsets. Thus, how to introduce local conditional entropy definition in a p-MSVDIS, thereby effectively reducing misclassification costs and improving the efficiency of feature selection, has become a key challenge in current research.

1.2. Motivation and innovation

To address missing values in datasets. Feature selection of partial labeled information systems is one of the research hotspots. Dai et al. [17] constructed a method applicable to both supervised and unsupervised frameworks, and developed two semi-supervised algorithms for attribute selection in partially labeled classification data. Zhang et al. [18] defined consistency entropy, monotonic neighborhood entropy, and maximum neighborhood entropy based on neighborhood rough sets, and designed two algorithms for addressing the attribute selection problem in partially labeled heterogeneous data. Luo et al. [19] proposed a three-transaction model based on rough sets for partially labeled data, and constructed a semi-supervised discernibility matrix that incorporates both labeled and unlabeled data to realize attribute selection for such data. Zhou et al. [20] proposed a multiscale fuzzy information fusion mechanism, transformed decision with missing labels into fuzzy decision for label distribution, constructed multiscale fuzzy rough set models for label distribution from both global and local perspectives, and accordingly developed two semisupervised feature selection algorithms. All these studies focus on partially labeled information systems, but they all assume that there are no missing conditional attributes; however, in real-world scenarios, the absence of conditional attributes is quite common. Therefore, based on this practical need, this paper adopts the multiset-valued method to handle the missing conditional attributes.

Although previous studies have shown promising prospects in handling missing labels, most of them rely on global entropy or dependency measures. Such methods not only incur high computational costs but also may lead to a decline in computational efficiency and a decrease in classification performance when dealing with large-scale datasets. To address these issues, this paper proposes two feature selection algorithms based on local conditional entropy for p-MSVDISs, aiming to enhance computational efficiency and robustness in partially labeled scenarios. The main contributions are as follows: 1) introducing a local conditional entropy framework into p-MSVDIS; 2) designing two feature selection strategies that can target the local information of high-dimensional, partially labeled data.

1.3. Organization

Section 2 reviews some basic concepts and introduces the fundamental concepts of p-MSVDISs and the conversion process from a p-MSVDIS to an MSVDIS. Section 3 describes the local conditional entropy measures for p-MSVDISs with predicted labels. Section 4 presents two feature selection algorithms based on local conditional entropy and total local conditional entropy. Section 5 conducts classification experiments and parameter experiments on real-world datasets. Section 6 presents the conclusions of this paper. Figure 1 reveals the framework of this paper.

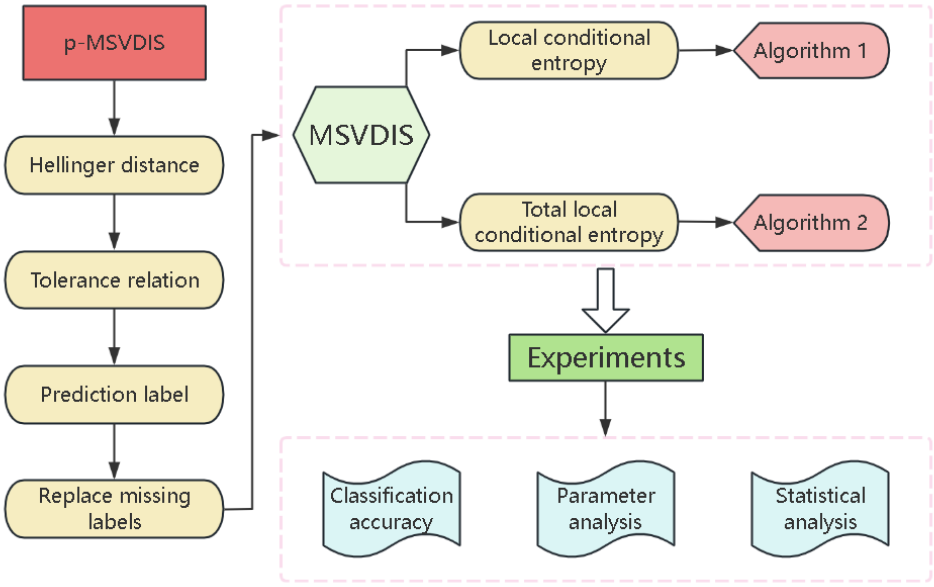


Figure 1. The framework of this paper.

2. Preliminaries

In this paper, U denotes a finite universe set, 2^U represents the power set of U, $|W|$ indicates the cardinality of W, N stands for the set of natural numbers, and Q denotes the set of rational numbers. Let $W_1, W_2 \in 2^U$, $p(W_1) = \frac{|W_1|}{|U|}$, and $p(W_1 \mid W_2) = \frac{|W_1 \cap W_2|}{|W_2|}$.

2.1. Multisets and sets of probability distributions

Unlike ordinary sets, a multiset can represent the number of occurrences of elements within the set. A set of probability distributions is a set containing multiple probability distributions, where each distribution describes the probability of events associated with a random variable.

Definition 1. [21] Given a finite set $Y = \{y_1, y_2, \cdots, y_l\}$, the mapping of the multiset M on Y is denoted as $M : Y \to N \cup \{0\}$, where $M(y) = m$ means that y occurs m times in M, recorded as $m/y \in M$. Denote $M = \{m_1/y_1, m_2/y_2, \cdots, m_l/y_l\}$.

Definition 2. [22] Let $Y = \{y_1, y_2, \cdots, y_l\}$, $S = \left\{\frac{y_1, y_2, \cdots, y_l}{s_1, s_2, \cdots, s_l}\right\}$. If $0 \le s_i \le 1$ for all $i \in \{1, 2, \cdots, l\}$ and $\sum_{i=1}^{l} s_i = 1$, then S is called the set of probability distributions on Y. If $s_i \in Q$ for all $i \in \{1, 2, \cdots, l\}$, then S is called the set of rational probability distributions on Y.

Definition 3. [22] Let $S = \left\{\frac{y_1, y_2, \cdots, y_l}{s_1, s_2, \cdots, s_l}\right\}$, $T = \left\{\frac{y_1, y_2, \cdots, y_l}{t_1, t_2, \cdots, t_l}\right\}$. Define the Hellinger distance between S and T as

$$HD(S, T) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{l} \left(\sqrt{s_i} - \sqrt{t_i}\right)^2}. \tag{2.1}$$

Definition 4. [23] Let $Y = \{y_1, y_2, \cdots, y_l\}$, $M = \{m_1/y_1, m_2/y_2, \cdots, m_l/y_l\}$. If $s_i = \frac{m_i}{m_1+m_2+\cdots+m_l}$ for all $i \in \{1, 2, \cdots, l\}$, then $S_M = \left\{\frac{y_1, y_2, \cdots, y_l}{s_1, s_2, \cdots, s_l}\right\}$ is the set of rational probability distributions on Y, and S_M is said to be the set of probability distributions drawn from M.

2.2. The definition of MSVDIS

An MSVDIS is an extended model of a decision information system (DIS), which is mainly used to deal with complex decision problems. The information in an MSVDIS is represented not only by a traditional single value or a categorical label, but also by a collection of multiple values. The MSVDIS is particularly well suited for representing problems involving multiple choices, ambiguity, or uncertainty.

Denote (U, At, d) as a DIS [24], where

1) U is the universe of objects, At is the set of conditional attribute sets, and d is the decision attribute;
2) for each $a \in At,\ a\colon U \to V_a$, $V_a = \{a(x) \mid x \in U\}$;
3) $d\colon U \to V_d$, $V_d = \{d(x) \mid x \in U\}$.

The system (U, At, d) is called a complete decision information system if $a(x) \neq *$ for all $a \in At$, $x \in U$ ($*$ denotes a missing information value). Conversely, if $\exists\ a \in At$ and $x \in U$ such that $a(x) = *$, then (U, At, d) is an incomplete decision information system (IDIS). We define $V_a^* = \{a(x) \mid x \in U,\ a(x) \neq *\}$.

Definition 5. [10] Given a DIS (U, At, d), let $U = \{x_1, x_2, \cdots, x_n\}$. For each $a \in At$, $a(x_1), a(x_2), \cdots, a(x_n)$ are multisets drawn from the same set, then (U, At, d) is called an MSVDIS.

Definition 6. [10] Given an IDIS (U, At, d), let $U = \{x_1, x_2, \cdots, x_n\}$. Given an attribute $a \in At$, denote $V_a^* = \{y_1, y_2, \cdots, y_l\}$. For each $i \in \{1, 2, \cdots, n\}$, let m_i denote the number of occurrences of y_i in $\{a(x_1), a(x_2), \cdots, a(x_n)\} \setminus \{*\}$. The substitution rules are as follows:

- if $a(x) = *$, replace $*$ with $\left\{\frac{m_1}{y_1}, \frac{m_2}{y_2}, \cdots, \frac{m_l}{y_l}\right\}$,
- if $a(x) = y_j$, replace y_j with $\left\{\frac{0}{y_1}, \cdots, \frac{1}{y_j}, \cdots, \frac{0}{y_l}\right\}$.

The resulting system is called the MSVDIS derived from the IDIS (U, At, d).

Example 1. An IDIS (U, At, d) is given in Table 1, where $U = \{x_1, x_2, \cdots, x_8\}$ represents the set of 8 customers, $At = \{\text{age}(a_1), \text{shopping frequency}(a_2), \text{discount}(a_3), \text{amount}(a_4)\}$ is the conditional attribute set with $V_{a_1} = \{Youth(Y), Adult(A), Elderly(E)\}$, $V_{a_2} = \{High(H), Medium(M), Low(L)\}$, $V_{a_3} = \{Strong(S), Moderate(M), Weak(W)\}$, and $V_{a_4} = \{High(H), Medium(M), Low(L)\}$. The decision attribute d indicates purchase intention and $V_d = \{Low(L), Medium(M), High(H)\}$.

In Table 1, $V_{a_1}^* = \{A, Y, E\}$, $V_{a_2}^* = \{H, M, L\}$, $V_{a_3}^* = \{S, M, W\}$, $V_{a_4}^* = \{H, M, L\}$, $V_d^* = V_d = \{L, M, H\}$.

Table 2 shows the MSVDIS corresponding to the IDIS in Table 1.

2.3. The p-MSVDIS and its predicted labels

This subsection introduces the p-MSVDIS and proposes to obtain an MSVDIS after predicting the missing decision values based on the information of the existing conditional attribute values.

2.3.1. The definition of p-MSVDIS

Definition 7. [10] Let (U, At, d) be an MSVDIS, $U = \{x_1, x_2, \cdots, x_n\}$, $U^* = \{x \in U \mid d(x) \neq *\}$, $U_* = \{x \in U \mid d(x) = *\}$, and $V_d^* = \{d(x) \mid x \in U^*\}$. If $U^* \neq \emptyset$ and $U_* \neq \emptyset$, then (U, At, d) is called a p-MSVDIS. If $C \neq \emptyset$ and $C \subseteq At$, then (U, C, d) is a subsystem of (U, At, d).

Table 1. An IDIS (U, At, d).

U	a_1	a_2	a_3	a_4	d
x_1	$Adult(A)$	$*$	$Strong(S)$	$High(H)$	$Low(L)$
x_2	$Youth(Y)$	$Medium(M)$	$Moderate(M)$	$Medium(M)$	$Medium(M)$
x_3	$Elderly(E)$	$Low(L)$	$*$	$Low(L)$	$Low(L)$
x_4	$*$	$High(H)$	$Strong(S)$	$High(H)$	$Medium(M)$
x_5	$Adult(A)$	$*$	$Moderate(M)$	$Medium(M)$	$Low(L)$
x_6	$Youth(Y)$	$Low(L)$	$Weak(W)$	$*$	$High(H)$
x_7	$Elderly(E)$	$Medium(M)$	$Strong(S)$	$Medium(M)$	$Low(L)$
x_8	$Youth(Y)$	$High(H)$	$*$	$Low(L)$	$Medium(M)$

Table 2. The MSVDIS corresponding to the (U, At, d) in Table 1.

U	a_1	a_2	a_3	a_4	d
x_1	$\{1/A, 0/Y, 0/E\}$	$\{2/H, 2/M, 2/L\}$	$\{1/S, 0/M, 0/W\}$	$\{1/H, 0/M, 0/L\}$	L
x_2	$\{0/A, 1/Y, 0/E\}$	$\{0/H, 1/M, 0/L\}$	$\{0/S, 1/M, 0/W\}$	$\{0/H, 1/M, 0/L\}$	M
x_3	$\{0/A, 0/Y, 1/E\}$	$\{0/H, 0/M, 1/L\}$	$\{3/S, 2/M, 1/W\}$	$\{0/H, 0/M, 1/L\}$	L
x_4	$\{2/A, 3/Y, 2/E\}$	$\{1/H, 0/M, 0/L\}$	$\{1/S, 0/M, 0/W\}$	$\{1/H, 0/M, 0/L\}$	M
x_5	$\{1/A, 0/Y, 0/E\}$	$\{2/H, 2/M, 2/L\}$	$\{0/S, 1/M, 0/W\}$	$\{0/H, 1/M, 0/L\}$	L
x_6	$\{0/A, 1/Y, 0/E\}$	$\{0/H, 0/M, 1/L\}$	$\{0/S, 0/M, 1/W\}$	$\{2/H, 3/M, 2/L\}$	H
x_7	$\{0/A, 0/Y, 1/E\}$	$\{0/H, 1/M, 0/L\}$	$\{1/S, 0/M, 0/W\}$	$\{0/H, 1/M, 0/L\}$	L
x_8	$\{0/A, 1/Y, 0/E\}$	$\{1/H, 0/M, 0/L\}$	$\{3/S, 2/M, 1/W\}$	$\{0/H, 0/M, 1/L\}$	M

All subsequent discussions are conducted under the condition that $C \neq \emptyset$.

Clearly, $U_* = U \setminus U^*$, $|V_d^*| + |U_*| \leq n$, $V_d^* \neq \emptyset \Leftrightarrow U^* \neq \emptyset$. Denote $[x]_d^* = \{x' \in U^* \mid d(x) = d(x')\}$ $(x \in U^*)$, $U^*/d = \{[x]_d^* \mid x \in U^*\}$.

A tolerance relation can be defined from a p-MSVDIS by introducing a threshold $\lambda \in [0, 1]$ based on the Hellinger distance.

Definition 8. [10] Let (U, At, d) be a p-MSVDIS and $a \in At$. The Hellinger distance matrix on the attribute a is defined as

$$HD(a) = \left(HD(S_{a(x_i)}, S_{a(x_j)})\right)_{n\times n}, \tag{2.2}$$

where $S_{a(x_i)}$ and $S_{a(x_j)}$ are the probability distribution sets derived from $a(x_i)$ and $a(x_j)$, respectively.

Definition 9. [10] Let (U, At, d) be a p-MSVDIS, $C \subseteq At$, and $\lambda \in [0, 1]$. Define a tolerance relation under C as

$$S_\lambda^C = \left\{(x, x') \in U \times U \,\middle|\, HD(S_{a(x)}, S_{a(x')}) \leq \lambda \text{ for all } a \in C\right\}. \tag{2.3}$$

The tolerance class of x under C is described as

$$S_\lambda^C(x) = \{x' \in U : (x, x') \in S_\lambda^C\}. \tag{2.4}$$

For any $x, y \in U$, $y \in S_\lambda^C(x) \Leftrightarrow x \in S_\lambda^C(y)$. Clearly, if $C_1 \subseteq C_2 \subseteq At$, then $S_\lambda^{C_2}(x) \subseteq S_\lambda^{C_1}(x)$ for each $x \in U$.

Table 3. A p-MSVDIS (U, At, d).

U	a_1	a_2	a_3	a_4	d
x_1	$\{1/A, 0/Y, 0/E\}$	$\{2/H, 2/M, 2/L\}$	$\{1/S, 0/M, 0/W\}$	$\{1/H, 0/M, 0/L\}$	L
x_2	$\{0/A, 1/Y, 0/E\}$	$\{0/H, 1/M, 0/L\}$	$\{0/S, 1/M, 0/W\}$	$\{0/H, 1/M, 0/L\}$	M
x_3	$\{0/A, 0/Y, 1/E\}$	$\{0/H, 0/M, 1/L\}$	$\{3/S, 2/M, 1/W\}$	$\{0/H, 0/M, 1/L\}$	$*$
x_4	$\{2/A, 3/Y, 2/E\}$	$\{1/H, 0/M, 0/L\}$	$\{1/S, 0/M, 0/W\}$	$\{1/H, 0/M, 0/L\}$	M
x_5	$\{1/A, 0/Y, 0/E\}$	$\{2/H, 2/M, 2/L\}$	$\{0/S, 1/M, 0/W\}$	$\{0/H, 1/M, 0/L\}$	L
x_6	$\{0/A, 1/Y, 0/E\}$	$\{0/H, 0/M, 1/L\}$	$\{0/S, 0/M, 1/W\}$	$\{2/H, 3/M, 2/L\}$	H
x_7	$\{0/A, 0/Y, 1/E\}$	$\{0/H, 1/M, 0/L\}$	$\{1/S, 0/M, 0/W\}$	$\{0/H, 1/M, 0/L\}$	$*$
x_8	$\{0/A, 1/Y, 0/E\}$	$\{1/H, 0/M, 0/L\}$	$\{3/S, 2/M, 1/W\}$	$\{0/H, 0/M, 1/L\}$	M

Example 2. A p-MSVDIS (U, At, d) is given in Table 3, where $*$ in d denotes a missing value (missing diagnostic information for this patient).

$U^* = \{x_1, x_2, x_4, x_5, x_6, x_8\}$, $U_* = \{x_3, x_7\}$, $U^*/d = \{D_1, D_2, D_3\}$, $D_1 = \{x_1, x_5\}$, $D_2 = \{x_2, x_4, x_8\}$, $D_3 = \{x_6\}$. Let $C = \{a_3, a_4\}$ and $\lambda = 0.6$. According to Definitions 3 and 4, the Hellinger distance matrices for a_1, a_2, a_3, and a_4 are obtained as follows:

$$HD(a_1) = \begin{pmatrix} 0 & 1 & 1 & 0.68 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0.59 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0.68 & 1 & 1 & 0 & 1 \\ 0.68 & 0.59 & 0.68 & 0 & 0.68 & 0.59 & 0.68 & 0.59 \\ 0 & 1 & 1 & 0.68 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0.59 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0.68 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0.59 & 1 & 0 & 1 & 0 \end{pmatrix},$$

$$HD(a_2) = \begin{pmatrix} 0 & 0.65 & 0.65 & 0.65 & 0 & 0.65 & 0.65 & 0.65 \\ 0.65 & 0 & 1 & 1 & 0.65 & 1 & 0 & 1 \\ 0.65 & 1 & 0 & 1 & 0.65 & 0 & 1 & 1 \\ 0.65 & 1 & 1 & 0 & 0.65 & 1 & 1 & 0 \\ 0 & 0.65 & 0.65 & 0.65 & 0 & 0.65 & 0.65 & 0.65 \\ 0.65 & 1 & 0 & 1 & 0.65 & 0 & 1 & 1 \\ 0.65 & 0 & 1 & 1 & 0.65 & 1 & 0 & 1 \\ 0.65 & 1 & 1 & 0 & 0.65 & 1 & 1 & 0 \end{pmatrix},$$

$$HD(a_3) = \begin{pmatrix} 0 & 1 & 0.54 & 0 & 1 & 1 & 0 & 0.54 \\ 1 & 0 & 0.65 & 1 & 0 & 1 & 1 & 0.65 \\ 0.54 & 0.65 & 0 & 0.54 & 0.65 & 0.77 & 0.54 & 0 \\ 0 & 1 & 0.54 & 0 & 1 & 1 & 0 & 0.54 \\ 1 & 0 & 0.65 & 1 & 0 & 1 & 1 & 0.65 \\ 1 & 1 & 0.77 & 1 & 1 & 0 & 1 & 0.77 \\ 0 & 1 & 0.54 & 0 & 1 & 1 & 0 & 0.54 \\ 0.54 & 0.65 & 0 & 0.54 & 0.65 & 0.77 & 0.54 & 0 \end{pmatrix},$$

$$HD(a_4) = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 0.68 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0.59 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0.68 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0.68 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0.59 & 0 & 1 \\ 0.68 & 0.59 & 0.68 & 0.68 & 0.59 & 0 & 0.59 & 0.68 \\ 1 & 0 & 1 & 1 & 0 & 0.59 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0.68 & 1 & 0 \end{pmatrix}.$$

According to Definition 9, the $S_\lambda^{a_j}(x_i)$ ($i = 1, 2, \cdots, 8$, $j = 1, 2, 3, 4$) can be obtained and is listed in Table 4. Therefore,

$$S_\lambda^C(x_1) = S_\lambda^C(x_4) = \{x_1, x_4\},\ S_\lambda^C(x_2) = S_\lambda^C(x_5) = \{x_2, x_5\},$$
$$S_\lambda^C(x_3) = S_\lambda^C(x_8) = \{x_3, x_8\},\ S_\lambda^C(x_6) = \{x_6\},\ S_\lambda^C(x_7) = \{x_7\}.$$

Table 4. The tolerance class $S_\lambda^{a_j}(x_i)$ ($i = 1, 2, \cdots, 8$, $j = 1, 2, 3, 4$).

i	$S_\lambda^{a_1}(x_i)$	$S_\lambda^{a_2}(x_i)$	$S_\lambda^{a_3}(x_i)$	$S_\lambda^{a_4}(x_i)$	$S_\lambda^C(x_i)$
1	$\{x_1, x_5\}$	$\{x_1, x_5\}$	$\{x_1, x_3, x_4, x_7, x_8\}$	$\{x_1, x_4\}$	$\{x_1, x_4\}$
2	$\{x_2, x_4, x_6, x_8\}$	$\{x_2, x_7\}$	$\{x_2, x_5\}$	$\{x_2, x_5, x_6, x_7\}$	$\{x_2, x_5\}$
3	$\{x_3, x_7\}$	$\{x_3, x_6\}$	$\{x_1, x_3, x_4, x_7, x_8\}$	$\{x_3, x_8\}$	$\{x_3, x_8\}$
4	$\{x_2, x_4, x_6, x_8\}$	$\{x_4, x_8\}$	$\{x_1, x_3, x_4, x_7, x_8\}$	$\{x_1, x_4\}$	$\{x_1, x_4\}$
5	$\{x_1, x_5\}$	$\{x_1, x_5\}$	$\{x_2, x_5\}$	$\{x_2, x_5, x_6, x_7\}$	$\{x_2, x_5\}$
6	$\{x_2, x_4, x_6, x_8\}$	$\{x_3, x_6\}$	$\{x_6\}$	$\{x_2, x_5, x_6, x_7\}$	$\{x_6\}$
7	$\{x_3, x_7\}$	$\{x_2, x_7\}$	$\{x_1, x_3, x_4, x_7, x_8\}$	$\{x_2, x_5, x_6, x_7\}$	$\{x_7\}$
8	$\{x_2, x_4, x_6, x_8\}$	$\{x_4, x_8\}$	$\{x_1, x_3, x_4, x_7, x_8\}$	$\{x_3, x_8\}$	$\{x_3, x_8\}$

2.3.2. The predicted labels for a p-MSVDIS

Since there are missing labels in the p-MSVDIS, this part handles the missing labels based on the tolerance classes. After filling in the missing values, there are no missing labels in the decision information system, so an MSVDIS can be obtained.

Definition 10. [10] Let (U, At, d) be a p-MSVDIS, $C \subseteq At$, and $\lambda \in [0, 1]$. Given $x \in U$, $U^*/d = \{D_1, D_2, \cdots, D_r\}$, and $1 \le s \le t \le r$, $s, t \in \{1, 2, \cdots, r\}$. There exists a $k \in \{1, 2, \cdots, r\}$ such that

$$D_k = \arg\max\{p(D_j \mid S_\lambda^C(x)) : 1 \le j \le r\} \in U^*/d. \tag{2.5}$$

i.e.,

$$p(D_k \mid S_\lambda^C(x)) = \max\{p(D_j \mid S_\lambda^C(x)) : 1 \le j \le r\}. \tag{2.6}$$

If $s, t \in \{1, 2, \cdots, r\}$, $s < t$, and $p(D_s \mid S_\lambda^C(x)) = p(D_t \mid S_\lambda^C(x))$, then $D_k = D_s$. Let $D_k = [x']_d$ ($x' \in U^*$), then the predicted label of x is described as $\mathrm{Pr}_C^\lambda(x) = d(x')$.

Definition 11. [10] Let (U, At, d) be a p-MSVDIS, $C \subseteq At$, and $\lambda \in [0, 1]$. Assume that $\mathrm{Pr}_C^\lambda(x)$ is the predicted label of $x \in U^*$. Define the decision error set as

$$des_{C,\lambda}^*(d) = \{x \in U^* : \mathrm{Pr}_C^\lambda(x) \ne d(x)\}. \tag{2.7}$$

The decision error rate is described as

$$der^*_{C,\lambda}(d) = \frac{|des^*_{C,\lambda}(d)|}{|U^*|}. \tag{2.8}$$

Clearly, $0 \leq der^*_{C,\lambda}(d) \leq 1$.

In Definition 11, x, $d(x)$ and $\mathrm{Pr}^\lambda_C(x)$ can be regarded as the test object, the true label of the test object, and the predicted label of the test object, respectively. If $\mathrm{Pr}^\lambda_C(x)$ is obtained, it can be compared with $d(x)$. Therefore, $des^*_{C,\lambda}(d)$ and $der^*_{C,\lambda}(d)$ represent the number of misclassified objects and the percentage of misclassified objects, respectively.

Example 3. Continued from Example 2, by Definition 10,

$$p(D_1 \mid S^C_\lambda(x_3)) = \frac{|\{x_1, x_5\} \cap \{x_3, x_8\}|}{|\{x_3, x_8\}|} = 0,$$
$$p(D_2 \mid S^C_\lambda(x_3)) = \frac{|\{x_2, x_4, x_8\} \cap \{x_3, x_8\}|}{|\{x_3, x_8\}|} = \frac{1}{2},$$
$$p(D_3 \mid S^C_\lambda(x_3)) = \frac{|\{x_6\} \cap \{x_3, x_8\}|}{|\{x_3, x_8\}|} = 0.$$

Thus,

$$\max\{p(D_j \mid S^C_\lambda(x_3)) : 1 \leq j \leq 3\} = p(D_2 \mid S^C_\lambda(x_3)).$$

The predicted label of x_3 is $\mathrm{Pr}^\lambda_C(x_3) = d(x_2)$. Similarly, we can obtain

$$\mathrm{Pr}^\lambda_C(x_1) = \mathrm{Pr}^\lambda_C(x_5) = \mathrm{Pr}^\lambda_C(x_7) = d(x_1),$$
$$\mathrm{Pr}^\lambda_C(x_2) = \mathrm{Pr}^\lambda_C(x_3) = \mathrm{Pr}^\lambda_C(x_4) = \mathrm{Pr}^\lambda_C(x_8) = d(x_2),$$
$$\mathrm{Pr}^\lambda_C(x_6) = d(x_6).$$

According to Definition 11, $des^*_{C,\lambda}(d) = \varnothing$ and $der^*_{C,\lambda}(d) = 0$. Therefore, $D^C_1 = \{x_1, x_5, x_7\}$, $D^C_2 = \{x_2, x_3, x_4, x_8\}$, $D^C_3 = \{x_6\}$.

Definition 12. [10] Let (U, At, d) be a p-MSVDIS, $C \subseteq At$, and $\lambda \in [0, 1]$. Define

$$d^\lambda_C(x) = \begin{cases} d(x), & x \in U^*, \\ \mathrm{Pr}^\lambda_C(x), & x \in U_*. \end{cases} \tag{2.9}$$

Let $d^\lambda_{\{C\}} = d^\lambda_C$, $V_{d^\lambda_C} = \{d^\lambda_C(x) : x \in U\}$. Therefore, $V_{d^\lambda_C} = V^*_d$. In (U, At, d^λ_C), $R_{d^\lambda_C}$ is an equivalent relation on U, defined as $R_{d^\lambda_C} = \{(x, x') \in U \times U : d^\lambda_C(x) = d^\lambda_C(x')\}$. The decision class of x based on $R_{d^\lambda_C}$ is $[x]_{d^\lambda_C} = \{x' \in U : (x, x') \in R_{d^\lambda_C}\}$. Let $U/d^\lambda_C = \{[x]_{d^\lambda_C} : x \in U\} = \{D^C_1, D^C_2, \cdots, D^C_r\}$.

Definition 13. [10] Let (U, At, d) be a p-MSVDIS, $C \subseteq At$, and $\lambda \in [0, 1]$. Define the mapping $\partial^{\mathrm{Pr}}_C : U \to 2^{V^*_d}$ as

$$\forall x \in U,\ \partial^{\mathrm{Pr}}_C(x) = \{d^\lambda_C(y)\ :\ y \in S^C_\lambda(x)\}, \tag{2.10}$$

then ∂^{Pr}_C is referred to as a decision mapping in (U, At, d). If $\left|\partial^{\mathrm{Pr}}_C(x)\right| = 1$ for all $x \in U$, and (U, At, d) is called a coordinated predicted p-MSVDIS; otherwise, (U, At, d) is called an uncoordinated predicted p-MSVDIS.

Theorem 1. [10] Let (U, At, d) be a p-MSVDIS, $C \subseteq At$, and $\lambda \in [0, 1]$. Then, the following are equivalent:

1) (U, At, d) is coordinated predicted;

2) $S^C_\lambda \subseteq R_{d^\lambda_C}$.

3. Entropy measures on predicted p-MSVDISs

This section introduces entropy measures on p-MSVDISs and discusses the relationships between different entropies.

Li et al. introduced the definition of conditional entropy in p-MSVDISs for feature selection [10].

Definition 14. [10] Let (U, At, d) be a p-MSVDIS, $U = \{x_1, x_2, \cdots, x_n\}$, $C \subseteq At$, and $\lambda \in [0, 1]$. Denote $U/d_C^\lambda = \{D_1^C, D_2^C, \cdots, D_r^C\}$, and define the conditional entropy of C w.r.t. d based on predicted labels as

$$H_\lambda^{\Pr}(d \mid C) = -\sum_{j=1}^{n}\sum_{k=1}^{r}\frac{\left|S_\lambda^C(x_j) \cap D_k^C\right|}{n}\log_2\frac{\left|S_\lambda^C(x_j) \cap D_k^C\right|}{\left|S_\lambda^C(x_j)\right|}. \tag{3.1}$$

We present the definition of local conditional entropy.

Definition 15. Let (U, At, d) be a p-MSVDIS, $U = \{x_1, x_2, \cdots, x_n\}$, $C \subseteq At$, and $\lambda \in [0, 1]$. Denote $U/d_C^\lambda = \{D_1^C, D_2^C, \cdots, D_r^C\}$. For each $D_k^C \in U/d_C^\lambda$, the local conditional entropy of C w.r.t. D_k^C is defined as

$$LH_\lambda^{\Pr}(D_k^C \mid C) = -\sum_{x_j \in D_k^C}\frac{\left|S_\lambda^C(x_j) \cap D_k^C\right|}{\left|D_k^C\right|}\log_2\frac{\left|S_\lambda^C(x_j) \cap D_k^C\right|}{\left|S_\lambda^C(x_j)\right|}. \tag{3.2}$$

The total local conditional entropy of d is defined as

$$TLH_\lambda^{\Pr}(d \mid C) = \sum_{k=1}^{r} LH_\lambda^{\Pr}(D_k^C \mid C). \tag{3.3}$$

Next, we present the comparison between $LH_\lambda^{\Pr}(D_k^C \mid C)$ and $H_\lambda^{\Pr}(d \mid C)$ in Table 5.

Table 5. Comparison of differences between $LH_\lambda^{\Pr}(D_k^C \mid C)$ and $H_\lambda^{\Pr}(d \mid C)$ in a p-MSVDIS.

	$LH_\lambda^{\Pr}(D_k^C \mid C)$	$H_\lambda^{\Pr}(d \mid C)$
Definition	$-\sum_{x_j \in D_k^C}\frac{\left\|S_\lambda^C(x_j)\cap D_k^C\right\|}{\left\|D_k^C\right\|}\log_2\frac{\left\|S_\lambda^C(x_j)\cap D_k^C\right\|}{\left\|S_\lambda^C(x_j)\right\|}$	$-\sum_{j=1}^{n}\sum_{k=1}^{r}\frac{\left\|S_\lambda^C(x_j)\cap D_k^C\right\|}{n}\log_2\frac{\left\|S_\lambda^C(x_j)\cap D_k^C\right\|}{\left\|S_\lambda^C(x_j)\right\|}$
Calculation scope	$x_j \in D_k^C$	$x_j \in U$
Functional mechanism	$\frac{\left\|S_\lambda^C(x_j)\cap D_k^C\right\|}{\left\|D_k^C\right\|}$	$\frac{\left\|S_\lambda^C(x_j)\cap D_k^C\right\|}{n}$
Time complexity	$O(\left\|D_k^C\right\| \|U\|)$	$O(\left\|D_k^C\right\| \|U\|^2)$

Example 4. Continued from Example 3, by Definitions 14 and 15,

$$H_\lambda^{\Pr}(d \mid C) = -\left[4 \times \left(\frac{1}{8}\log_2\frac{1}{2} + \frac{1}{8}\log_2\frac{1}{2}\right)\right] = 1.$$

$$LH_\lambda^{\Pr}(D_1^C \mid C) = -2 \times \frac{1}{3}\log_2\frac{1}{2} = \frac{2}{3},\ LH_\lambda^{\Pr}(D_2^C \mid C) = -2 \times \frac{1}{4}\log_2\frac{1}{2} = \frac{1}{2},$$

$$LH_\lambda^{\Pr}(D_3^C \mid C) = 0.$$

$$TLH_\lambda^{\Pr}(d \mid C) = \frac{7}{6}.$$

Therefore, the local conditional entropy proposed in this paper is different from conditional entropy. Next, properties of $LH_\lambda^{\Pr}(D_k^C \mid C)$ and $TLH_\lambda^{\Pr}(d \mid C)$ are analyzed.

Proposition 1. Let (U, At, d) be a p-MSVDIS, $U = \{x_1, x_2, \cdots, x_n\}$, $C \subseteq At$, and $\lambda \in [0, 1]$, denote $U/d_C^\lambda = \{D_1^C, D_2^C, \cdots, D_r^C\}$. Then,
(1) $LH_\lambda^{\Pr}(D_k^C \mid C) \geq 0$ for all $D_k^C \in U/d_C^\lambda$;
(2) $TLH_\lambda^{\Pr}(d \mid C) \geq 0$;
(3) if $C_1 \subseteq C_2 \subseteq At$, then $LH_\lambda^{\Pr}(D_k^C \mid C_2) \leq LH_\lambda^{\Pr}(D_k^C \mid C_1)$ for all $k \in \{1, 2, \cdots, r\}$;
(4) if $C_1 \subseteq C_2 \subseteq At$, then $TLH_\lambda^{\Pr}(d \mid C_2) \leq TLH_\lambda^{\Pr}(d \mid C_1)$.

Proof. (1) Since $0 \leq \frac{|S_\lambda^C(x_j) \cap D_k^C|}{|D_k^C|} \leq 1$ and $S_\lambda^C(x_j) \cap D_k^C \subseteq S_\lambda^C(x_j)$ for all $x_j \in D_k^C$, we obtain $\log_2 \frac{|S_\lambda^C(x_j) \cap D_k^C|}{|S_\lambda^C(x_j)|} \leq 0$. Therefore, according to Definition 14, $LH_\lambda^{\Pr}(D_k^C \mid C) \geq 0$ for all $D_k^C \in U/d_C^\lambda$.

(2) It can be proven by Definition 14 and (1).

(3) Let $C_1 \subseteq C_2 \subseteq At$. It follows that $S_\lambda^{C_2}(x_j) \subseteq S_\lambda^{C_1}(x_j)$ and $\left|S_\lambda^{C_2}(x_j)\right| \leq \left|S_\lambda^{C_1}(x_j)\right|$ for all $x_j \in D_k^C$. Then, $\left|S_\lambda^{C_2}(x_j) \cap D_k^C\right| \leq \left|S_\lambda^{C_1}(x_j) \cap D_k^C\right|$ and $\left|S_\lambda^{C_2}(x_j) \cap (U - D_k^C)\right| \leq \left|S_\lambda^{C_1}(x_j) \cap (U - D_k^C)\right|$. Moreover,

$$LH_\lambda^{\Pr}(D_k^C \mid C_1) = -\frac{1}{\left|D_k^C\right|} \sum_{x_j \in D_k^C} \left|S_\lambda^{C_1}(x_j) \cap D_k^C\right| \log_2 \frac{\left|S_\lambda^{C_1}(x_j) \cap D_k^C\right|}{\left|S_\lambda^{C_1}(x_j) \cap D_k^C\right| + \left|S_\lambda^{C_1}(x_j) \cap (U - D_k^C)\right|},$$

$$LH_\lambda^{\Pr}(D_k^C \mid C_2) = -\frac{1}{\left|D_k^C\right|} \sum_{x_j \in D_k^C} \left|S_\lambda^{C_2}(x_j) \cap D_k^C\right| \log_2 \frac{\left|S_\lambda^{C_2}(x_j) \cap D_k^C\right|}{\left|S_\lambda^{C_2}(x_j) \cap D_k^C\right| + \left|S_\lambda^{C_2}(x_j) \cap (U - D_k^C)\right|}.$$

We obtain that the function $f(u, v) = -u \log_2 \frac{u}{(u+v)}$ is monotonically increasing w.r.t. u and v [25], and it follows that for each $x_j \in D_k^C$,

$$-\left|S_\lambda^{C_2}(x_j) \cap D_k^C\right| \log_2 \frac{\left|S_\lambda^{C_2}(x_j) \cap D_k^C\right|}{\left|S_\lambda^{C_2}(x_j) \cap D_k^C\right| + \left|S_\lambda^{C_2}(x_j) \cap (U - D_k^C)\right|}$$

$$\leq -\left|S_\lambda^{C_1}(x_j) \cap D_k^C\right| \log_2 \frac{\left|S_\lambda^{C_1}(x_j) \cap D_k^C\right|}{\left|S_\lambda^{C_1}(x_j) \cap D_k^C\right| + \left|S_\lambda^{C_1}(x_j) \cap (U - D_k^C)\right|}.$$

Then,

$$-\sum_{x_j \in D_k^C} \left|S_\lambda^{C_2}(x_j) \cap D_k^C\right| \log_2 \frac{\left|S_\lambda^{C_2}(x_j) \cap D_k^C\right|}{\left|S_\lambda^{C_2}(x_j) \cap D_k^C\right| + \left|S_\lambda^{C_2}(x_j) \cap (U - D_k^C)\right|}$$

$$\leq -\sum_{x_j \in D_k^C} \left|S_\lambda^{C_1}(x_j) \cap D_k^C\right| \log_2 \frac{\left|S_\lambda^{C_1}(x_j) \cap D_k^C\right|}{\left|S_\lambda^{C_1}(x_j) \cap D_k^C\right| + \left|S_\lambda^{C_1}(x_j) \cap (U - D_k^C)\right|}.$$

Thus, $LH_\lambda^{\Pr}(D_k^C \mid C_2) \leq LH_\lambda^{\Pr}(D_k^C \mid C_1)$.

(4) It is clear by Definition 14 and (3).

□

Theorem 2. [10] Let (U, At, d) be a p-MSVDIS, $U/d_C^\lambda = \{D_1^C, D_2^C, \cdots, D_r^C\}$, $C \subseteq At$, and $\lambda \in [0, 1]$. If $S_\lambda^C \subseteq R_{d_C^\lambda}$, then for each $x \in U$ and $k \in \{1, 2, \cdots, r\}$,

$$S_\lambda^C(x) \cap D_k^C = \begin{cases} S_\lambda^C(x), & x \in D_k^C, \\ \varnothing, & x \notin D_k^C. \end{cases}$$

Theorem 3. Let (U, At, d) be a p-MSVDIS, $C \subseteq At$, $U/d_C^\lambda = \{D_1^C, D_2^C, \cdots, D_r^C\}$, and $\lambda \in [0, 1]$. Then, the following are equivalent:
(1) $S_\lambda^C \subseteq R_{d_C^\lambda}$;
(2) for any $k \in \{1, 2, \cdots, r\}$, $LH_\lambda^{\Pr}(D_k^C \mid C) = 0$;
(3) $TLH_\lambda^{\Pr}(d \mid C) = 0$.

Proof. (1) $\Rightarrow$ (2). According to Theorem 2, $S_\lambda^C(x_j) \cap D_k^C = S_\lambda^C(x_j)$ for all $x_j \in D_k^C$ and $k \in \{1, 2, \cdots, r\}$. Therefore, for any $k \in \{1, 2, \cdots, r\}$, $\log_2 \frac{|S_\lambda^C(x_j) \cap D_k^C|}{|S_\lambda^C(x_j)|} = 0$. Thus, for any $k \in \{1, 2, \cdots, r\}$, $LH_\lambda^{\Pr}(D_k^C \mid C) = 0$.

(2) $\Rightarrow$ (1). By Definition 15, we obtain that for any $k \in \{1, 2, \cdots, r\}$, $x_j \in D_k^C$, $\log_2 \frac{|S_\lambda^C(x_j) \cap D_k^C|}{|S_\lambda^C(x_j)|} = 0$. According to Theorem 2, $S_\lambda^C \subseteq R_{d_C^\lambda}$.

(2) $\Leftrightarrow$ (3). It can be proven by Definition 15. □

In feature selection based on local conditional entropy, we need to select valuable and non-redundant conditional attributes. Thus, we have introduced the concept of attribute importance.

Definition 16. Let (U, At, d) be a p-MSVDIS, $U/d_C^\lambda = \{D_1^C, D_2^C, \cdots, D_r^C\}$, $C \subseteq At$, and $\lambda \in [0, 1]$. For any $a \in C$, define the importance of a w.r.t. D_k^C as

$$Sig(a, C, D_k^C) = LH_\lambda^{\Pr}(D_k^C \mid (C - a)) - LH_\lambda^{\Pr}(D_k^C \mid C), \tag{3.4}$$

and define the importance of a w.r.t. d as

$$TSig(a, C, d) = TLH_\lambda^{\Pr}(d \mid (C - a)) - TLH_\lambda^{\Pr}(d \mid C). \tag{3.5}$$

Proposition 2. Let (U, At, d) be a p-MSVDIS, $C \subseteq At$, and $\lambda \in [0, 1]$. Then,
(1) for each $a \in C$, $Sig(a, C, D_k^C) \geq 0$;
(2) for each $a \in C$, $TSig(a, C, d) \geq 0$.

Proof. It can be obtained from Proposition 1 and Definition 16. □

Example 5. Continued from Example 4, according to Definition 16,

$$Sig(a_3, C, D_1^C) = 1,\ Sig(a_3, C, D_2^C) = 0.25,\ Sig(a_3, C, D_3^C) = 2,$$
$$Sig(a_4, C, D_1^C) = 1.43,\ Sig(a_4, C, D_2^C) = 1.41,\ Sig(a_4, C, D_3^C) = 0.$$
$$TSig(a_3, C, d) = 3.25,\ TSig(a_4, C, d) = 2.84.$$

4. Feature selection of a p-MSVDIS based on local conditional entropy

In this section, feature selection for a p-MSVDIS is performed based on the proposed local conditional entropy.

Definition 17. Let (U, At, d) be a p-MSVDIS, $U/d_C^\lambda = \{D_1^C, D_2^C, \cdots, D_r^C\}$, $C \subseteq At$, and $\lambda \in [0, 1]$.

(1) If $LH_\lambda^{\Pr}(D_k^C \mid C) = LH_\lambda^{\Pr}(D_k^C \mid At)$ for each $k \in \{1, 2, \cdots, r\}$, then C is called a local conditional entropy consistent subset of At. All the local conditional entropy consistent subsets are defined as $lec_\lambda^{\Pr}(At)$;

(2) if $LH_\lambda^{\Pr}(D_k^C \mid C) = LH_\lambda^{\Pr}(D_k^C \mid At)$ for each $k \in \{1, 2, \cdots, r\}$, and for each $a \in C$, there exists a $k \in \{1, 2, \cdots, r\}$ such that $LH_\lambda^{\Pr}(D_k^C \mid C) \neq LH_\lambda^{\Pr}(D_k^C \mid (C - a))$, then C is a local conditional entropy reduct of At. All the local conditional entropy reducts are denoted as $ler_\lambda^{\Pr}(At)$;

(3) if $TLH_\lambda^{\Pr}(d \mid C) = TLH_\lambda^{\Pr}(d \mid At)$, then C is called a total local conditional entropy consistent subset of At. All the total local conditional entropy consistent subsets are defined as $tlec_\lambda^{\Pr}(At)$;

(4) if $TLH_\lambda^{\Pr}(d \mid C) = TLH_\lambda^{\Pr}(d \mid At)$ and for each $a \in C$, $TLH_\lambda^{\Pr}(d \mid C) \neq TLH_\lambda^{\Pr}(d \mid (C - a))$, then C is a total local conditional entropy reduct of At. All total local conditional entropy reducts are denoted as $tler_\lambda^{\Pr}(At)$.

Theorem 4. Let (U, At, d) be a p-MSVDIS, $U/d_C^\lambda = \{D_1^C, D_2^C, \cdots, D_r^C\}$, $C \subseteq At$, and $\lambda \in [0, 1]$. Then,
(1) $lec_\lambda^{\Pr}(At) = tlec_\lambda^{\Pr}(At)$;
(2) $ler_\lambda^{\Pr}(At) = tler_\lambda^{\Pr}(At)$.

Proof. (1) For any $C \in lec_\lambda^{\Pr}(At)$, $LH_\lambda^{\Pr}(D_k^C \mid C) = LH_\lambda^{\Pr}(D_k^C \mid At)$ for each $k \in \{1, 2, \cdots, r\}$. Then, $\sum_{k=1}^{r} LH_\lambda^{\Pr}(D_k^C \mid C) = \sum_{k=1}^{r} LH_\lambda^{\Pr}(D_k^C \mid At)$. It follows that $lec_\lambda^{\Pr}(At) \subseteq tlec_\lambda^{\Pr}(At)$. Conversely, for any $C \in tlec_\lambda^{\Pr}(At)$, we have $TLH_\lambda^{\Pr}(d \mid C) = TLH_\lambda^{\Pr}(d \mid At)$, that is, $\sum_{k=1}^{r} LH_\lambda^{\Pr}(D_k^C \mid C) = \sum_{k=1}^{r} LH_\lambda^{\Pr}(D_k^C \mid At)$. By Propositions 1 (1) and (3), $LH_\lambda^{\Pr}(D_k^C \mid C) \geq LH_\lambda^{\Pr}(D_k^C \mid At) \geq 0$ for all $k \in \{1, 2, \cdots, r\}$. Then, $LH_\lambda^{\Pr}(D_k^C \mid C) = LH_\lambda^{\Pr}(D_k^C \mid At)$, for all $k \in \{1, 2, \cdots, r\}$. Hence, $C \in lec_\lambda^{\Pr}(At)$. It follows that $lec_\lambda^{\Pr}(At) = tlec_\lambda^{\Pr}(At)$.

(2) It can be easily proven by Definition 17 and (1).

□

The two reduction definitions proposed in this paper are equivalent, as both satisfy the requirement for a reduction that requires preserving the core attributes of the original system and being non-redundant. In contrast, the proposed algorithms yield different reduction results; this is because there are differences in search paths and the order of redundant element elimination in their design logic, which leads to the selection of different specific subsets from the reduction sets corresponding to the equivalent definitions.

Theorem 5. Let (U, At, d) be a p-MSVDIS, $U/d_C^\lambda = \{D_1^C, D_2^C, \cdots, D_r^C\}$, $C \subseteq At$, and $\lambda \in [0, 1]$. Then,

(1) $C \in ler_\lambda^{\Pr}(At) \Leftrightarrow LH_\lambda^{\Pr}(D_k^C \mid C) = LH_\lambda^{\Pr}(D_k^C \mid At)$ for all $k \in \{1, 2, \cdots, r\}$, and for each $a \in C$, there exists a $k \in \{1, 2, \cdots, r\}$ such that $Sig(a, C, D_k^C) > 0$;

(2) $C \in tler_\lambda^{\Pr}(At) \Leftrightarrow TLH_\lambda^{\Pr}(d \mid C) = TLH_\lambda^{\Pr}(d \mid At)$, and for each $a \in C$, $TSig(a, C, d) > 0$.

Proof. It can be proven by Definition 17 and Proposition 2. □

Theorem 6. Let (U, At, d) be a coordinated p-MSVDIS, $U/d_C^\lambda = \{D_1^C, D_2^C, \cdots, D_r^C\}$, $C \subseteq At$, and $\lambda \in [0, 1]$. Then,

(1) $C \in ler_\lambda^{\Pr}(At) \Leftrightarrow LH_\lambda^{\Pr}(D_k^C \mid C) = 0$ for all $k \in \{1, 2, \cdots, r\}$, and for each $a \in C$ there exists a $k \in \{1, 2, \cdots, r\}$ such that $LH_\lambda^{\Pr}(D_k^C \mid (C - \{a\})) > 0$;

(2) $C \in tler_\lambda^{\Pr}(At) \Leftrightarrow TLH_\lambda^{\Pr}(d \mid C) = 0$, and $TLH_\lambda^{\Pr}(d \mid (C - \{a\})) > 0$ for all $a \in C$.

Proof. It is easily proved by Theorem 3. □

Based on the above theory, we propose and design two feature selection algorithms for a p-MSVDIS: feature selection algorithm based on local conditional entropy (p-LEFS) and feature selection algorithm based on total local conditional entropy (p-TLEFS).

Let $|U| = n$, $|At| = m$, and $|U/d| = r$. Algorithm 1 is designed to identify the reduction set based on local conditional entropy. Steps 2–5 calculate the Hellinger distance between every pair of objects for each attribute, thereby determining the tolerance classes with objects under different attributes. Steps 6–9 obtain the decision classes and predicted labels. Step 10 gets each decision class of the predicted label. Steps 11–13 calculate the local conditional entropy of each D_k^{At} of At on the predicted labels. Steps 14–33 compute the local conditional entropy of each D_k^{At} of $C \subseteq At$. Steps 34–41 calculate the importance of a of each D_k^{At}. The time complexity of Algorithm 1 is $O(n^2m + m^2r + nr)$.

Algorithm 2 is designed to identify the reduction set based on total local conditional entropy. The algorithm terminates when total local conditional entropy of C is equal to total local conditional entropy of At. The time complexity of Algorithm 2 is $O(n^2m + m^2r + nr)$.

We present the flowcharts of Algorithms 1 and 2 in Figures 2 and 3, respectively.

Example 6. Let $U = \{x_1, x_2, \cdots, x_6\}$, $At = \{b_1, b_2, \cdots, b_5\}$, $U/d_C^\lambda = \{D_1^C, D_2^C, D_3^C\}$ with $D_1^C = \{x_1, x_2\}$, $D_2^C = \{x_3, x_4\}$, and $D_3^C = \{x_5, x_6\}$. Let the tolerance class of x_j under b_i be shown in Table 6. By Definition 15,

$$LH_\lambda^{\Pr}(D_1^C \mid At) = 0,\ LH_\lambda^{\Pr}(D_2^C \mid At) = 0,\ LH_\lambda^{\Pr}(D_3^C \mid At) = 0.$$
$$LH_\lambda^{\Pr}(D_1^C \mid b_1) = 0,\ LH_\lambda^{\Pr}(D_2^C \mid b_1) = 0.5,\ LH_\lambda^{\Pr}(D_3^C \mid b_1) = 0.5.$$
$$LH_\lambda^{\Pr}(D_1^C \mid b_2) = 1.17,\ LH_\lambda^{\Pr}(D_2^C \mid b_2) = 0.79,\ LH_\lambda^{\Pr}(D_3^C \mid b_2) = 0.$$
$$LH_\lambda^{\Pr}(D_1^C \mid b_3) = 1,\ LH_\lambda^{\Pr}(D_2^C \mid b_3) = 0,\ LH_\lambda^{\Pr}(D_3^C \mid b_3) = 1.$$
$$LH_\lambda^{\Pr}(D_1^C \mid b_4) = 1.58,\ LH_\lambda^{\Pr}(D_2^C \mid b_4) = 1,\ LH_\lambda^{\Pr}(D_3^C \mid b_4) = 1.$$
$$LH_\lambda^{\Pr}(D_1^C \mid b_5) = 0,\ LH_\lambda^{\Pr}(D_2^C \mid b_5) = 0,\ LH_\lambda^{\Pr}(D_3^C \mid b_5) = 0.$$
$$TLH_\lambda^{\Pr}(d \mid At) = 0,\ TLH_\lambda^{\Pr}(d \mid b_1) = 1,\ TLH_\lambda^{\Pr}(d \mid b_2) = 1.96,$$
$$TLH_\lambda^{\Pr}(d \mid b_3) = 2,\ TLH_\lambda^{\Pr}(d \mid b_4) = 4.47,\ TLH_\lambda^{\Pr}(d \mid b_5) = 0.$$

According to Algorithms 1 and 2, $\{b_1, b_2\} \in ler_\lambda^{\Pr}(At)$, $\{b_5\} \in tler_\lambda^{\Pr}(At)$.

5. Experimental analysis

To validate the effectiveness and feasibility of the algorithms, the performance is evaluated by classification accuracy and analyzed in comparison with four existing feature selection methods.

Algorithm 1 A feature selection algorithm for a p-MSVDIS based on local conditional entropy (p-LEFS).

Require: A p-MSVDIS (U, At, d), $\lambda \in [0, 1]$.
Ensure: A reduct C.

$C \leftarrow \emptyset$, start = 1;
for $a \in At$ **do**
  Calculate $HD(a) = (HD(S_{a(x)}, S_{a(x')}))_{n\times n}$ and $S_\lambda^a(x)$ of each x;
end for
Calculate $S_\lambda^{At}(x)$ of each x;
Calculate $U^*/d = \{D_1, D_2, \cdots, D_r\}$, where $U^* = \{x \in U : d(x) \neq *\}$;
for $D_k \in U^*/d$ **do**
  Calculate $p(D_k \mid S_\lambda^{At}(x))$ and replace * to obtain a predicted p-MSVDIS;
end for
Calculate $U/d_{At}^\lambda = \{D_1^{At}, D_2^{At}, \cdots, D_r^{At}\}$;
for $D_k^{At} \in U/d_{At}^\lambda$ **do**
  Calculate $LH_\lambda^{\Pr}(D_k^{At} \mid At)$;
end for
Set the frequency of occurrence of all a to 0;
for $D_k^{At} \in U/d_{At}^\lambda$ **do**
  for $a \in At - C$ **do**
    Calculate $LH_\lambda^{\Pr}(D_k^{AT} \mid a)$;
  end for
  Select an $a_0 \in At - C$ such that $a_0 = argmin\{LH_\lambda^{\Pr}(D_k^{AT} \mid a) \mid a \in At - C\}$ and increment the frequency of occurrence of a by 1 (if there are two or more minimum values, select the a which ranks higher in the order of $At - C$);
end for
if two or more attributes have the maximal frequency of occurrences **then**
  Select an $a \in At - C$ which ranks higher in the order of $At - C$;
else
  Select an $a \in At - C$ with the maximal frequency of occurrences;
end if
$C = C \cup a$;
for $D_k^{At} \in U/d_{At}^\lambda$ **do**
  if $LH_\lambda^{\Pr}(D_k^{At} \mid C) = LH_\lambda^{\Pr}(D_k^{At} \mid At)$ **then**
    Go to step 34;
  else
    Go to step 14;
  end if
end for
for $a \in C$ **do**
  for $D_k^{At} \in U/d_{At}^\lambda$ **do**
    Calculate $Sig(a, C, D_k^{At})$;
  end for
  if $Sig(a, C, D_k^{At}) = 0$ for all $k \in \{1, 2, \cdots, r\}$ and $C \neq \emptyset$ **then**
    Remove a from C;
  end if
end for
Return C.

Algorithm 2 A feature selection algorithm for a p-MSVDIS based on total local conditional entropy (p-TLEFS).

Require: A p-MSVDIS (U, At, d), $\lambda \in [0, 1]$.

Ensure: A reduct C.

$C \leftarrow \varnothing$, start = 1;
for $a \in At$ **do**
  Calculate $HD(a) = (HD(S_{a(x)}, S_{a(x')}))_{n\times n}$ and $S_\lambda^a(x)$ of each x;
end for
Calculate $S_\lambda^{At}(x)$ of each x;
Calculate $U^*/d = \{D_1, D_2, \cdots, D_r\}$, where $U^* = \{x \in U : d(x) \neq *\}$;
for $D_k \in U^*/d$ **do**
  Calculate $p(D_k \mid S_\lambda^{At}(x))$ and replace * to obtain a predicted p-MSVDIS;
end for
Calculate $U/d_{At}^\lambda = \{D_1^{At}, D_2^{At}, \cdots, D_r^{At}\}$;
Calculate $TLH_\lambda^{\Pr}(d \mid At)$;
while start = 1 **do**
  for $a \in At - C$ **do**
    Calculate $TLH_\lambda^{\Pr}(d \mid a)$;
  end for
  Select an $a_0 \in At - C$ such that $a_0 = argmin\{TLH_\lambda^{\Pr}(d \mid a) \mid a \in At - C\}$ (if there are two or more minimum values, select the a which ranks higher in the order of $At - C$);
  Calculate $TLH_\lambda^{\Pr}(d \mid C \cup a)$;
  if $TLH_\lambda^{\Pr}(d \mid C \cup a^*) \neq TLH_\lambda^{\Pr}(d \mid At)$ **then**
    $C \leftarrow C \cup a$;
  else
    start = 0;
  end if
end while
for $a \in C$ **do**
  Calculate $TSig(a, C, d)$;
  if $TSig(a, C, d) = 0$ and $C \neq \varnothing$ **then**
    Remove a from C;
  end if
end for
Return C.

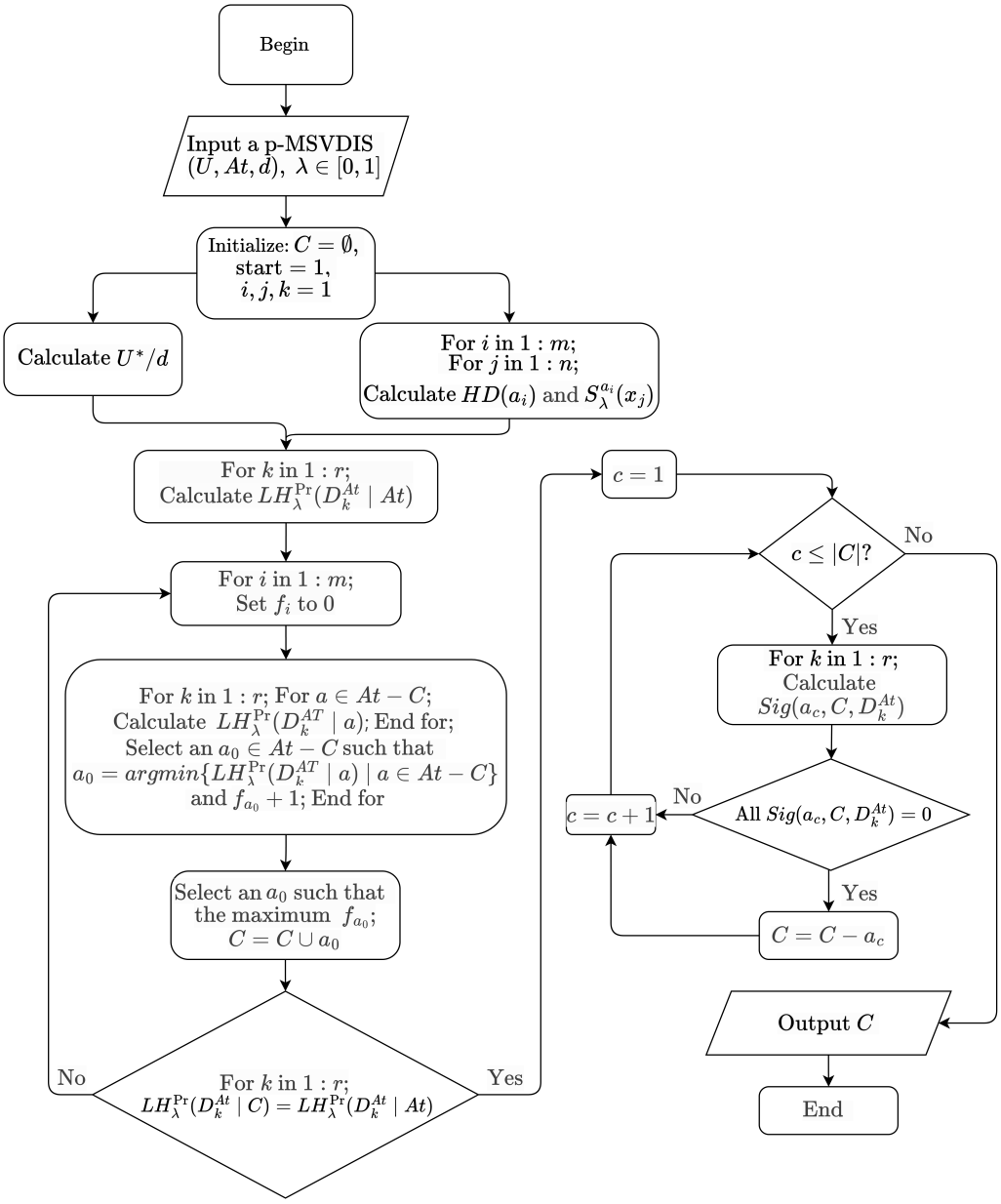


Figure 2. The framework of Algorithm 1.

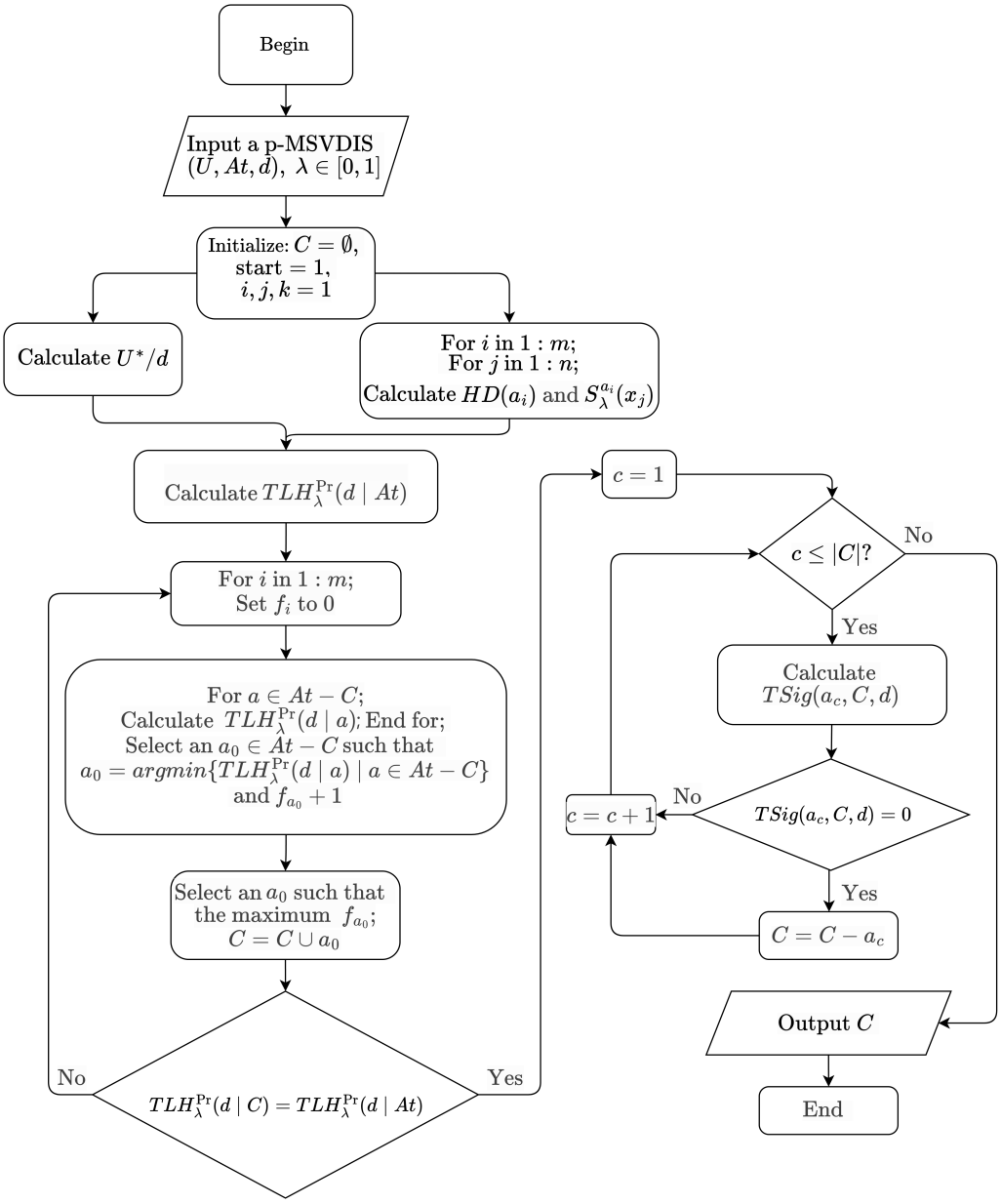


Figure 3. The framework of Algorithm 2.

Table 6. The tolerance class $S_{\lambda}^{a_j}(x_i)$ ($i = 1, 2, \cdots, 8$, $j = 1, 2, \cdots, 5$).

j	$S_{\lambda}^{b_1}(x_j)$	$S_{\lambda}^{b_2}(x_j)$	$S_{\lambda}^{b_3}(x_j)$	$S_{\lambda}^{b_4}(x_j)$	$S_{\lambda}^{b_5}(x_j)$
1	$\{x_1, x_2\}$	$\{x_1, x_2, x_3\}$	$\{x_1, x_5\}$	$\{x_1, x_2\}$	$\{x_1\}$
2	$\{x_1, x_2\}$	$\{x_1, x_2, x_3\}$	$\{x_2, x_6\}$	$\{x_1, x_2\}$	$\{x_2\}$
3	$\{x_3\}$	$\{x_1, x_2, x_3\}$	$\{x_3, x_4\}$	$\{x_3, x_5\}$	$\{x_3\}$
4	$\{x_4, x_5\}$	$\{x_4\}$	$\{x_3, x_4\}$	$\{x_4, x_6\}$	$\{x_4\}$
5	$\{x_4, x_5\}$	$\{x_5\}$	$\{x_1, x_5\}$	$\{x_3, x_5\}$	$\{x_5\}$
6	$\{x_6\}$	$\{x_6\}$	$\{x_2, x_6\}$	$\{x_4, x_6\}$	$\{x_6\}$

5.1. Datasets and experimental setup

For experimental validation, 12 UCI (dataset repository for machine learning research) datasets (Amphibians, Autistic, Cylinder bands, Hill, Horse, Monks, Music, Sports, Students, Urban, Zoo) are selected, as listed in Table 7. All experiments are implemented in a Python 3.0 simulation environment under Windows 10 OS, equipped with an Intel i7 CPU (the arithmetic and control core of a computer), and 16 GB RAM (the temporary data storage component of a computer). Classification accuracy is the main evaluation metric for analyzing the performance of the proposed classification algorithm on partially labeled datasets.

Table 7. Datasets.

NO.	Datasets	Objects	Attributes	Missing value	Classes
1	Amphibians	104	20	No	2
2	Autistic	396	20	No	2
3	Cylinder bands	541	38	Yes	3
4	Hill	606	100	No	2
5	Horse	300	27	Yes	2
6	Monks	432	6	No	2
7	Music	400	50	No	4
8	Sports	1000	59	No	2
9	Students	145	31	No	8
10	Tic	958	9	No	2
11	Urban	507	147	No	9
12	Zoo	101	16	No	7

5.2. Classification accuracy analysis

To evaluate the performance of the proposed algorithms, we compare p-LEFS and p-TLEFS (designed in this paper) with four existing rough set based attribute reduction algorithms: semrough-P (Sem-P) [17], attribute reduction via local conditional entropy (LCER) [14], semi-supervised attribute reduction methods for partially labeled multiset-valued data using predicted labels (p-DAR, p-CIEAR) [10], and semi-supervised attribute selection algorithms for partially labeled multiset-valued data (SARM1, SARM2) [12] focusing on classification accuracy.

In the experimental design, for Sem-P, p-DAR, p-CIEAR, SARM1, SARM2, p-LEFS, and p-TLEFS, we adopted the technique from [26] to handle the missing values of conditional attributes in all datasets.

The missing rate for conditional attributes is set to 0.1 across all datasets. The parameter λ in p-DAR, p-CIEAR, SARM1, SARM2, p-LEFS, and p-TLEFS is set to 0.1. The parameter δ in LCER is also set to 0.1, and Sem-P does not need parameter setting because there are no parameters. The number of selected optimal attributes is shown in Table 8. The performance of the algorithm is evaluated using three classifiers: support vector machine (SVM), K-nearest neighbors (KNN), and naive bayes (NB). Detailed classification accuracy results are provided in Tables 9–11, where bold and underlined values indicate the highest accuracy.

Table 8. The number of optimal attributes selected by different algorithms.

Datasets	Raw	Sem-P	LCER	p-DAR	p-CIEAR	SARM1	SARM2	p-LEFS	p-TLEFS
Amp	20	3	1	8	5	1	1	6	6
Aut	20	4	1	9	5	5	5	6	6
Cyl	38	3	20	5	3	12	12	2	3
Hil	100	2	4	3	2	45	98	2	2
Hor	27	3	13	3	3	12	13	2	2
Mon	6	6	3	1	6	4	4	6	6
Mus	50	2	5	5	3	5	5	3	2
Spo	59	3	6	3	3	6	8	4	4
Stu	31	5	6	11	5	6	12	5	5
Tic	9	8	8	1	8	7	8	6	8
Urb	147	2	118	2	2	5	50	3	2
Zoo	16	5	5	11	8	6	10	5	5

Table 9. Classification accuracy of 12 UCI datasets under SVM.

Datasets	Sem-P	LCER	p-DAR	p-CIEAR	SARM1	SARM2	p-LEFS	p-TLEFS
Amp	66.00	66.00	81.55	65.00	66.00	66.00	**85.73**	**85.73**
Aut	61.29	55.19	79.62	80.01	80.11	79.51	**81.22**	**81.22**
Cyl	57.78	68.52	68.52	68.52	**68.70**	68.52	65.19	68.52
Hil	52.25	53.31	54.41	53.98	59.19	**64.30**	55.19	55.19
Hor	69.33	68.32	69.33	69.33	69.21	**70.21**	69.33	69.33
Mon	67.06	67.27	**68.69**	67.06	68.22	68.22	67.06	67.06
Mus	33.11	59.15	46.62	53.14	48.62	52.61	**76.92**	74.15
Spo	78.28	**80.58**	78.68	78.68	78.29	78.69	79.28	79.28
Stu	93.71	92.33	83.24	94.38	34.52	49.38	**97.19**	**97.19**
Tic	**65.31**	**65.31**	**65.31**	**65.31**	**65.31**	**65.31**	**65.31**	**65.31**
Urb	98.82	94.86	99.01	99.20	56.93	87.95	99.21	**99.31**
Zoo	88.00	88.00	96.00	90.00	89.00	96.00	**99.00**	**99.00**

Table 9 indicates the significant superiority of p-LEFS and p-TLEFS on most datasets under the SVM. For example, Sem-P, LCER, p-DAR, p-CIEAR, SARM1, SARM2, p-LEFS, and p-TLEFS achieve the highest classification accuracy in 1, 2, 2, 1, 2, 3, 6, 6 cases, respectively.

Table 10 shows the significant superiority of p-LEFS and p-TLEFS on most datasets under the KNN. Specifically, Sem-P, LCER, p-DAR, p-CIEAR, SARM1, SARM2, p-LEFS, and p-TLEFS attain the

Table 10. Classification accuracy of 12 UCI datasets under KNN.

Datasets	Sem-P	LCER	p-DAR	p-CIEAR	SARM1	SARM2	p-LEFS	p-TLEFS
Amp	96.00	98.09	98.09	97.09	98.09	98.09	**99.27**	**99.27**
Aut	95.22	95.92	92.92	91.63	94.45	90.88	**96.21**	**96.21**
Cyl	77.04	82.56	80.37	77.04	76.11	78.52	**82.59**	77.04
Hil	69.93	71.39	71.22	71.41	72.73	**73.39**	71.90	71.90
Hor	74.92	76.30	68.93	74.93	75.61	76.21	**76.61**	**76.61**
Mon	**88.65**	**88.65**	74.47	**88.65**	86.99	86.99	**88.65**	**88.65**
Mus	61.90	68.70	69.66	59.17	68.42	67.17	**75.93**	75.17
Spo	79.39	80.99	81.39	81.39	80.79	81.39	**81.49**	**81.49**
Stu	68.10	68.81	53.00	68.05	48.62	44.62	**72.76**	**72.76**
Tic	86.51	85.58	61.76	85.05	**87.57**	86.84	85.54	86.95
Urb	96.23	87.11	96.25	92.87	67.37	84.17	95.87	**98.25**
Zoo	85.00	95.00	**96.00**	91.00	95.00	**96.00**	**96.00**	**96.00**

Table 11. Classification accuracy of 12 UCI datasets under NB.

Datasets	Sem-P	LCER	p-DAR	p-CIEAR	SARM1	SARM2	p-LEFS	p-TLEFS
Amp	61.09	61.09	64.27	61.09	61.09	61.09	**71.00**	**71.00**
Aut	75.48	51.64	76.02	75.98	66.11	67.62	**76.74**	**76.74**
Cyl	57.78	62.40	57.78	57.78	56.93	**66.30**	57.78	57.78
Hil	50.58	50.50	50.58	50.58	51.40	**52.06**	50.58	50.58
Hor	**66.90**	66.23	**66.90**	66.23	66.23	67.23	**66.90**	**66.90**
Mon	**66.36**	65.90	50.12	**66.36**	64.73	64.73	**66.36**	**66.36**
Mus	32.83	54.12	38.83	39.10	43.38	42.40	**57.38**	56.61
Spo	63.46	63.46	63.46	63.46	**63.96**	63.66	63.46	63.46
Stu	31.24	39.52	33.95	33.33	23.62	28.29	**52.05**	**52.05**
Tic	65.62	65.62	65.31	65.62	65.31	65.31	65.41	**65.73**
Urb	62.83	67.23	65.43	47.45	38.92	**73.09**	66.83	52.22
Zoo	59.00	71.00	90.00	88.00	79.00	91.00	**94.00**	**94.00**

highest classification accuracy in 1, 1, 1, 1, 1, 2, 9, 8 cases, respectively.

Table 11 reveals that p-LEFS and p-TLEFS outperform most datasets under the NB. Here, Sem-P, LCER, p-DAR, p-CIEAR, SARM1, SARM2, p-LEFS, and p-TLEFS achieve the highest classification accuracy in 2, 0, 1, 1, 1, 3, 7, 7 cases, respectively.

The experimental results demonstrate that p-LEFS and p-TLEFS exhibit multiple advantages in SVM, KNN, and NB classifiers. First, they achieve high classification accuracy on most datasets across the different classifiers. Second, they are adaptable to complex data scenarios, showing strong robustness in cases such as the Cyl dataset with missing values, the Urb dataset with multiple classes, and the Zoo dataset with small samples. Third, they have excellent performance stability; even when tied with other methods for the optimal performance, they still maintain consistent accuracy.

5.3. Parameter analysis

This section analyzes the classification accuracy of the p-CIEAR, p-LEFS, and p-TLEFS algorithms across different datasets under the KNN as the key parameter λ varies. Specific details are provided in Table 12.

Based on the KNN classifier, we analyzed the classification accuracy of different algorithms as the parameter λ varies, and the conclusions are as follows: p-LEFS and p-TLEFS can significantly improve the classification accuracy and stability of the KNN classifier; compared with p-CIEAR, p-LEFS and p-TLEFS achieve better accuracy on most test datasets and also exhibit stronger resistance to parameter interference. This indicates that the two methods proposed in this paper have good applicability in high-dimensional classification problems. Reasonable adjustment of the parameter λ enables adaptation to different datasets and improves classification accuracy. In summary, during feature selection, it is necessary to select an appropriate λ value range in combination with specific datasets to achieve optimal classification performance.

5.4. Statistical analysis

To evaluate the classification performance of different algorithms, the Friedman test [27] and Bonferroni-Dunn test [28] are carried out in this subsection. The Friedman test is described as

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} \tag{5.1}$$

and

$$\chi_F^2 = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4}\right), \tag{5.2}$$

where N denotes the number of datasets, k represents the number of algorithms, r_i indicates the average ranking of a given algorithm across all datasets, and F_F is the F distribution satisfying $F(k-1,(k-1)(N-1))$ degrees of freedom.

Table 12. The variation of classification accuracy with λ under different algorithms.

Datasets	Algorithms	Parameter (λ)								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Amp	p-CIEAR	97.09	97.09	97.09	97.09	93.27	93.27	90.18	89.27	83.36
	p-LEFS	99.27	99.27	99.27	99.27	95.45	93.18	91.18	90.45	91.27
	p-TLEFS	99.27	99.27	99.27	99.27	96.18	93.18	91.18	88.73	87.45
Aut	p-CIEAR	91.63	91.63	96.21	96.21	95.92	95.68	95.68	91.71	92.69
	p-LEFS	96.21	96.21	96.21	96.21	96.69	96.16	96.16	90.91	90.67
	p-TLEFS	96.21	96.21	96.21	96.21	96.69	96.16	96.16	92.95	90.67
Cyl	p-CIEAR	77.04	77.04	77.04	77.04	77.04	77.04	77.04	77.04	78.89
	p-LEFS	82.59	82.59	82.59	82.59	82.59	82.59	82.59	82.59	78.89
	p-TLEFS	77.04	77.04	77.04	77.04	77.04	77.04	77.04	77.04	78.89
Hil	p-CIEAR	71.41	71.41	71.41	71.41	71.41	71.41	71.41	71.41	71.41
	p-LEFS	71.90	71.90	71.90	71.90	71.90	71.90	71.90	71.90	71.90
	p-TLEFS	71.90	71.90	71.90	71.90	71.90	71.90	71.90	71.90	71.90
Hor	p-CIEAR	74.93	74.93	74.93	73.63	73.63	73.63	73.63	73.63	68.93
	p-LEFS	76.61	76.61	76.61	76.61	76.61	76.61	76.61	63.91	68.93
	p-TLEFS	76.61	76.61	76.61	76.61	76.61	76.61	76.61	63.91	68.93
Mon	p-CIEAR	88.65	88.65	88.65	88.65	88.65	88.65	88.65	88.65	88.65
	p-LEFS	88.65	88.65	88.65	88.65	88.65	88.65	88.65	88.65	88.65
	p-TLEFS	88.65	88.65	88.65	88.65	88.65	88.65	88.65	88.65	88.65
Mus	p-CIEAR	59.17	59.17	59.17	59.17	59.17	59.17	59.17	59.17	59.17
	p-LEFS	75.93	75.93	75.93	75.93	75.93	75.93	75.93	75.93	75.93
	p-TLEFS	75.17	75.17	75.17	75.17	75.17	75.17	75.17	75.17	75.17
Spo	p-CIEAR	81.39	81.39	81.39	81.39	81.39	81.39	81.39	81.39	81.39
	p-LEFS	81.49	81.49	81.49	81.49	81.49	81.49	81.49	81.49	81.49
	p-TLEFS	81.49	81.49	81.49	81.49	81.49	81.49	81.49	81.49	81.49
Stu	p-CIEAR	68.05	68.05	68.05	68.05	68.05	64.48	58.19	66.71	63.90
	p-LEFS	72.76	72.76	72.76	72.76	72.76	64.48	61.05	84.05	86.71
	p-TLEFS	72.76	72.76	72.76	72.76	72.76	64.48	61.05	94.48	90.95
Tic	p-CIEAR	85.05	85.05	85.05	85.05	85.05	87.35	87.35	87.35	87.35
	p-LEFS	85.54	85.54	85.54	85.54	85.54	87.35	87.35	87.35	87.35
	p-TLEFS	86.95	86.95	86.95	86.95	86.95	87.35	87.35	87.35	87.35
Urb	p-CIEAR	92.87	92.87	92.87	92.87	92.87	92.87	92.87	92.87	92.87
	p-LEFS	95.87	95.87	95.87	95.87	95.87	95.87	95.87	95.87	95.87
	p-TLEFS	98.25	98.25	98.25	98.25	98.25	98.25	98.25	98.25	98.25
Zoo	p-CIEAR	91.00	91.00	96.00	96.00	95.00	96.00	96.00	96.00	96.00
	p-LEFS	96.00	96.00	98.00	98.00	92.00	87.00	93.00	93.00	95.00
	p-TLEFS	96.00	96.00	98.00	98.00	92.00	87.00	93.00	93.00	95.00

Based on the classification accuracy presented in Tables 9–11, each algorithm was ranked according to its performance under individual datasets. Subsequently, the average ranking of each algorithm across all 12 datasets was computed to reflect its overall performance. Table 13 shows the Friedman test results of eight algorithms under the SVM, KNN, and NB classifiers. At a significance level of $\alpha = 0.05$, the F_F distribution with $F(7, 77)$ had a critical value of 2.131. As shown in Table 13, the corresponding F_F of the three classifiers are all greater than 2.131; thus, further post-hoc tests are required. Furthermore, the Bonferroni-Dunn test is used to distinguish the eight algorithms. The critical value is described as

$$CD_\alpha = q_\alpha \sqrt{\frac{k(k+1)}{6N}}. \tag{5.3}$$

This method compares the average ranks of any two algorithms with the critical value. For a significance level of $\alpha = 0.05$ and $q_\alpha = 2.69$, the corresponding critical value $CD_\alpha = 2.69$ can be calculated according to the above formula. Figure 4 presents the results of the Bonferroni-Dunn test for different feature selection algorithms under three classifiers. The figure plots the average ranks of the eight algorithms and uses horizontal lines to connect two algorithms whose difference in average ranks is less than 2.69.

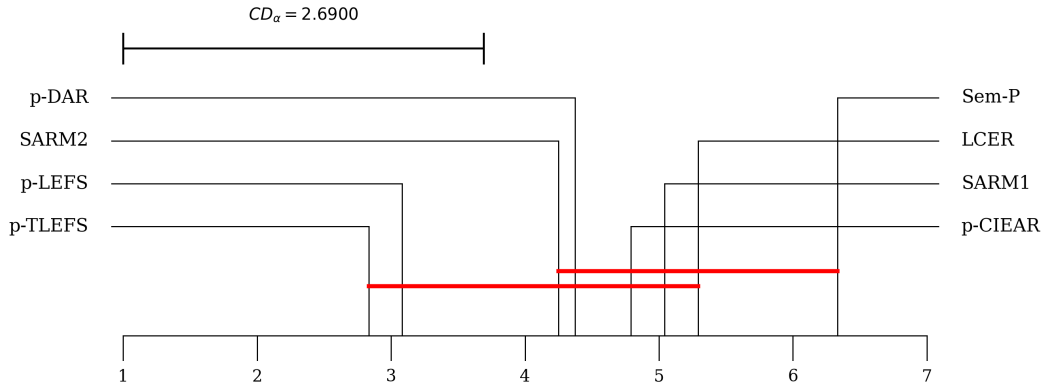
From Figure 4, several conclusions can be drawn regarding algorithm performance. On the SVM classifier, the proposed p-TLEFS and p-LEFS algorithms rank highest among all methods, demonstrating significantly superior performance compared to the four algorithms clustered on the right (Sem-P, LCER, SARM1, and p-CIEAR), and showing advantages over p-DAR and SARM2. From the statistical results, there is no significant difference between p-TLEFS and p-LEFS. On the KNN classifier, p-TLEFS and p-LEFS achieve the best performance among all feature selection methods, exhibiting statistically significant advantages over algorithms including p-CIEAR, Sem-P, SARM1, and p-DAR. On the NB classifier, both algorithms maintain competitive performance with an average rank of approximately 3. In conclusion, the proposed p-TLEFS and p-LEFS algorithms demonstrate consistent and superior performance across all three classifiers, particularly excelling with KNN and ranking among the top performers with SVM. This consistency indicates strong generalization capability and robustness across different classifiers, confirming their effectiveness as feature selection methods.

Table 13. Friedman test of eight algorithms with three classifiers.

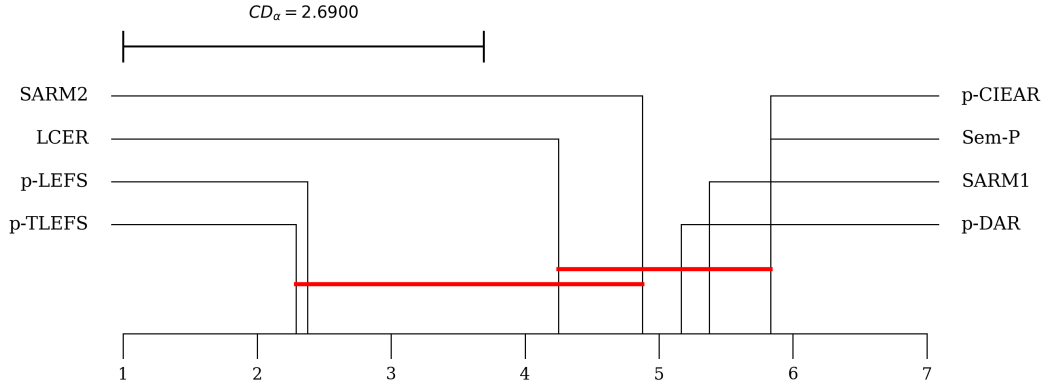
Classifiers	Average ranking								χ_F^2	F_F
	Sem-P	LCER	p-DAR	p-CIEAR	SARM1	SARM2	p-LEFS	p-TLEFS		
SVM	6.33	5.29	4.38	4.79	5.04	4.25	3.08	2.83	18.46	3.10
KNN	5.83	4.25	5.17	5.83	5.38	4.88	2.38	2.29	28.72	5.72
NB	5.21	4.96	4.92	5.08	5.88	3.88	3.04	3.04	15.52	2.49

6. Conclusions

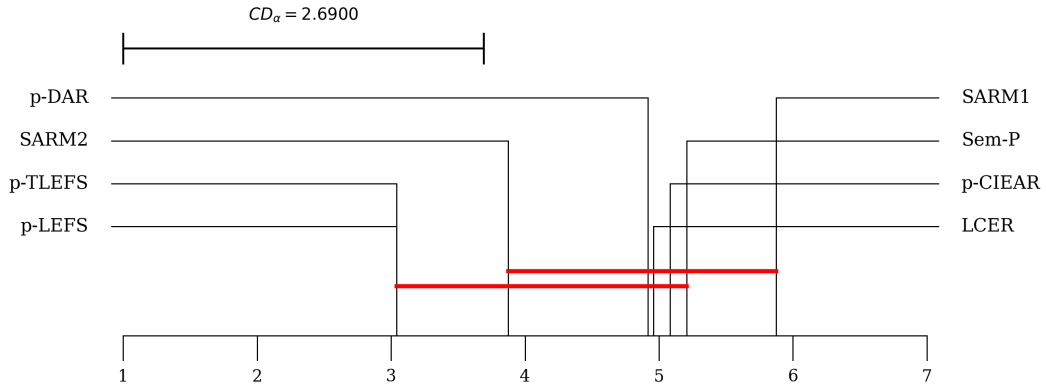
The p-MSVDIS and the predicted label strategy provided an effective method for addressing missing labels in datasets. This paper has proposed two algorithms for feature selection of the p-MSVDIS, based on local conditional entropy and total local conditional entropy, respectively. The introduction of local conditional entropy has alleviated the problems inherent in information entropy for feature



(a) SVM



(b) KNN



(c) NB

Figure 4. Bonferroni-Dunn test on three classifiers.

selection in datasets, which include computational inefficiency caused by the need to traverse all objects and decision classes, and overfitting that tends to occur with large-scale datasets. By leveraging the advantages of the p-MSVDIS and the predicted label strategy in feature selection, as shown by the classification experimental results, the classification accuracy on most datasets has been improved. After feature selection via these two algorithms, the classification accuracy has outperformed that of the four existing methods. The method proposed in this paper is more suitable for large-scale datasets, in which scenarios it can significantly improve computational efficiency and classification accuracy. However, when the degree of missing labels is relatively high, algorithms based on local conditional entropy may become unstable. In our future work, we will consider introducing an adaptive neighborhood strategy to enhance its robustness. Additionally, we will attempt to incorporate adaptive parameter adjustment and hybrid entropy measurement mechanisms to further enhance the applicability and stability of the method in diverse data scenarios.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Natural Science Foundation of Fujian Province (No. 2022J01912, No. 2024J01803).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. A. Bargiela, W. Pedrycz, Granular computing, in: *Handbook on Computational Intelligence* (ed. P. P. Angelov), World Scientific, (2016), 43–66. https://doi.org/10.1142/9548
2. C. Romero, S. Ventura, Data mining in education, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, **3** (2013), 12–27. https://doi.org/10.1002/widm.1075
3. J. Dai, Q. Xu, Approximations and uncertainty measures in incomplete information systems, *Inf. Sci.*, **198** (2012), 62–80. https://doi.org/10.1016/j.ins.2012.02.032
4. S. U. Amin, A. Hussain, B. Kim, S. Seo, Deep learning based active learning technique for data annotation and improve the overall performance of classification models, *Expert Syst. Appl.*, **228** (2023), 120391. https://doi.org/10.1016/j.eswa.2023.120391
5. D. Huang, H. Lin, Z. Li, Information structures in a multiset-valued information system with application to uncertainty measurement, *J. Intell. Fuzzy Syst.*, **43** (2022), 7447–7469. https://doi.org/10.3233/jifs-220652
6. J. Liang, Z. Shi, The information entropy, rough entropy and knowledge granulation in rough set theory, *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, **12** (2004), 37–46. https://doi.org/10.1142/S0218488504002631

7. C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, *Electron. Mark.*, **31** (2021), 685–695. https://doi.org/10.1007/s12525-021-00475-2
8. D. Huang, Y. Chen, F. Liu, Z. Li, Feature selection for multiset-valued data based on fuzzy conditional information entropy using iterative model and matrix operation, *Appl. Soft Comput.*, **142** (2023), 110345. https://doi.org/10.1016/j.asoc.2023.110345
9. P. Cunningham, M. Cord, S. J. Delany, Supervised learning, in: *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval* (eds. M. Cord and P. Cunningham), Springer Berlin Heidelberg, (2008), 21–49. https://doi.org/10.1007/978-3-540-75171-7_2
10. Z. Li, T. Yang, J. Li, Semi-supervised attribute reduction for partially labelled multiset-valued data via a prediction label strategy, *Inf. Sci.*, **634** (2023), 477–504. https://doi.org/10.1016/j.ins.2023.03.127
11. X. Guo, Y. Peng, Y. Li, H. Lin, Unsupervised attribute reduction algorithms for multiset-valued data based on uncertainty measurement, *Mathematics*, **13** (2025), 1718. https://doi.org/10.3390/math13111718
12. Y. He, J. He, H. Liu, Z. Li, Semi-supervised attribute selection algorithms for partially labeled multiset-valued data, *Mathematics*, **13** (2025), 1318. https://doi.org/10.3390/math13081318
13. Y. Qian, X. Liang, Q. Wang, J. Liang, B. Liu, A. Skowron, et al., Local rough set: a solution to rough data analysis in big data, *Int. J. Approx. Reason.*, **97** (2018), 38–63. https://doi.org/10.1016/j.ijar.2018.01.008
14. Y. Wang, X. Chen, K. Dong, Attribute reduction via local conditional entropy, *Int. J. Mach. Learn. Cybern.*, **10** (2019), 3619–3634. https://doi.org/10.1007/s13042-019-00948-z
15. L. Xie, G. Lin, J. Li, Y. Lin, A novel fuzzy-rough attribute reduction approach via local information entropy, *Fuzzy Sets Syst.*, **473** (2023), 108733. https://doi.org/10.1016/j.fss.2023.108733
16. J. Chen, Z. Li, X. Wang, J. Zhai, Adaptive feature selection framework using local fuzzy dominance neighborhood composite entropy for imbalanced ordered decision systems, *Inf. Sci.*, **720** (2025), 122533. https://doi.org/10.1016/j.ins.2025.122533
17. J. Dai, Q. Hu, J. Zhang, H. Hu, N. Zheng, Attribute selection for partially labeled categorical data by rough set approach, *IEEE Trans. Cybern.*, **47** (2016), 2460–2471. https://doi.org/10.1109/TCYB.2016.2636339
18. H. Zhang, Q. Sun, K. Dong, Information-theoretic partially labeled heterogeneous feature selection based on neighborhood rough sets, *Int. J. Approximate Reasoning*, **154** (2023), 200–217. https://doi.org/10.1016/j.ijar.2022.12.010
19. Z. Luo, C. Gao, J. Zhou, Rough sets-based tri-trade for partially labeled data, *Appl. Intell.*, **53** (2023), 17708–17726. https://doi.org/10.1007/s10489-022-04405-3
20. N. Zhou, S. Liao, H. Chen, W. Ding, Y. Lu, Semi-supervised feature selection with multi-scale fuzzy information fusion: from both global and local perspectives, *IEEE Trans. Fuzzy Syst.*, **33** (2025), 1825–1839. https://doi.org/10.1109/TFUZZ.2025.3540884
21. Y. Song, H. Lin, Z. Li, Outlier detection in a multiset-valued information system based on rough set theory and granular computing, *Inf. Sci.*, **657** (2024), 119950. https://doi.org/10.1016/j.ins.2023.119950

22. X. Xie, Z. Li, P. Zhang, G. Zhang, Information structures and uncertainty measures in an incomplete probabilistic set-valued information system, *IEEE Access*, **7** (2019), 27501–27514. https://doi.org/10.1109/ACCESS.2019.2897752

23. V. B. Gisin, E. S. Volkova, Equivalence relations on multisets, in *2024 XXVII International Conference on Soft Computing and Measurements*, IEEE, (2024), 258–261. https://doi.org/10.1109/SCM62608.2024.10554263

24. Z. Pawlak, Rough set theory and its applications, *J. Telecommun. Inf. Technol.*, **9** (2002), 7–10. https://doi.org/10.26636/jtit.2002.140

25. W. Shu, W. Qian, Y. Xie, Incremental feature selection for dynamic hybrid data using neighborhood rough set, *Knowl. Based Syst.*, **194** (2020), 105516. https://doi.org/10.1016/j.knosys.2020.105516

26. J. W. Grzymala-Busse, M. Hu, A comparison of several approaches to missing attribute values in data mining, in *Rough Sets and Current Trends in Computing* (eds. Z. Wojciech and Y. Yiyu), Springer Berlin Heidelberg, (2001), 378–385. https://doi.org/10.1007/3-540-45554-X_46

27. M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, *Ann. Math. Statist.*, **11** (1940), 86–92. https://doi.org/10.1214/aoms/1177731944

28. O. J. Dunn, Multiple comparisons among means, *J. Am. Stat. Assoc.*, **56** (1961), 52–64. https://doi.org/10.1080/01621459.1961.10482090

AIMS Press

© 2025 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)