Parameterized-maximum-distribution-entropy-based three-way approximate attribute reduction

Can Gao[a,b,c], Jie Zhou[a,b,c,*], Jinming Xing[a,b,c], Xiaodong Yue[d]

[a] *College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong, 518060, P.R. China*
[b] *Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen, Guangdong, 518060, P.R. China*
[c] *SZU Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, Guangdong, 518060, P.R. China*
[d] *School of Computer Engineering and Science, Shanghai University, Shanghai 200444, P.R. China*

Abstract

Three-way decision theory has emerged as an effective method for attribute reduction when dealing with vague, uncertain, or imprecise data. However, most existing attribute reduction measures in the three-way decision are non-monotonic and too strict, limiting the quality of attribute reduction. In this study, a monotonic measure called parameterized maximum distribution entropy (PMDE) is proposed for approximate attribute reduction. Specifically, considering that the classification ability under uncertainty is reflected by both the decision and the degree of confidence, a novel PMDE measure that attaches different levels of importance to the decision with the highest probability and other decisions is provided, and its monotonicity is theoretically proven. Furthermore, the idea of trisection in the three-way decision is introduced into the process of attribute reduction, and a heuristic algorithm based on the proposed measure is developed to generate an optimal three-way approximate reduct, which greatly improves the efficiency of attribute reduction. Several experiments conducted on UCI datasets show that the proposed method achieves a favorable performance with much fewer attributes in comparison with other representative methods.

Keywords: Three-way decision, attribute reduction, monotonicity, maximum distribution entropy, three-way approximate reduct

1. Introduction

Data collected from routine tasks, such as medical diagnosis, text classification, and gene analysis, usually contain thousands of attributes. Excessive attributes inevitably cause the problem of the curse of dimensionality, resulting in poor generalization performance, low learning efficiency, lack of interpretability, and difficulty in visualization [1]. Attribute reduction [2, 3, 4, 5, 6] aims to remove redundant and irrelevant attributes from data and has been proven to effectively alleviate the aforementioned problems. The theory of rough sets [7, 8] is a prominent method for attribute reduction, especially when data are entangled with vagueness, imprecision, or uncertainty. In essence, the indiscernibility

*Corresponding author.
Email addresses: 2005gaocan@163.com (Can Gao), jie_jpu@163.com (Jie Zhou), xjm0801@126.com (Jinming Xing), yswantfly@shu.edu.cn (Xiaodong Yue)

relation, that is, the equivalence relation, is the cornerstone of Pawlak's rough set model, from which the lower and upper approximations are derived to appropriately describe ambiguous concepts. However, the indiscernibility relation is too strict, limiting the application of rough sets. Fortunately, a wide variety of extension models [9, 10, 11, 12, 13, 14, 15, 16, 17] have been proposed to address this problem by relaxing the indiscernibility relation or incorporating other uncertainty theories.

The three-way decision [18, 19, 20, 21] has evolved from an effective and popular probabilistic rough set model, that is, the decision-theoretic rough set model [10]. It extends the two-way decision within the classic rough set model into three alternative decisions of acceptance, rejection, and noncommitment, and provides a unified and comprehensive framework for the rough set theory. Presently, the research and application of the three-way decision is far beyond the scope of the rough set theory and has become the methodology and philosophy for thinking or computing in threes [22, 23, 24, 25, 26, 27]. The three-way decision, due to the merit and superiority in interpretability and versatility [28], has been introduced in many research domains, such as clustering [29], active learning [30], cognitive computing [31], conflict analysis [32], and formal concept analysis [33].

Attribute reduction [34] has attracted considerable attention in the theory of three-way decision. The three-way decision-based attribute reduction methods can be classified into measure preservation and measure optimization [35]. In the methods of measure preservation, attribute reduction in the three-way decision is considered as a problem of preserving or improving uncertainty measures such as the positive region and information entropy. Yao and Zhao [34, 35] studied attribute reduction measures of the region, rule, and cost, and proposed a general definition for attribute reduction in the three-way decision. Li et al. [36] presented a region extension measure to maintain or even enlarge the positive region. Ma et al. [37] investigated decision region distribution preservation for attribute reduction and provided information-theoretic measures to maintain the decision region distribution. Zhang and Miao [38] introduced the concept of a three-way attribute reduct and used the measure of the relative dependency degree to yield quantitative reducts. Gao et al. [39] defined the measure of the maximum decision entropy and developed a heuristic attribute reduction algorithm with max-relevance and min-redundancy. Further, they [40] incorporated the information granularity and developed the measure of the granular maximum decision entropy. In the methods of measure optimization, attribute reduction in the three-way decision is treated as the process of optimizing uncertainty measures from the perspective of cost or risk using some optimization strategies. Jia et al. [41] investigated the minimum-cost attribute reduction in the decision-theoretic rough set model and utilized evolutionary computing algorithms to find the optimal reduct with minimum cost. Liao et al. [42] introduced decision and test costs into the measure of attribute reduction and provided a heuristic attribute reduction algorithm to minimize the two costs. Fang and Min [43] studied cost-sensitive attribute reduction under qualitative and quantitative measures and proposed two heuristic algorithms to yield cost-sensitive approximate reducts. Li et al. [44] combined the measures of the positive region, decision cost, and mutual information to propose a multi-objective attribute reduction algorithm. In addition, the three-way decision is introduced into the problem of attribute reduction in class-specific [45, 46], multigranulation [47, 48], neighborhood systems [49, 50, 51], and others [52, 53, 54, 55].

The aforementioned methods can generate optimal reducts under a specific

measure. However, some attribute reduction measures are non-monotonic, and their reducts may also suffer from the problems of low reduction rate and poor generalization. Moreover, human decision-making under uncertainty prefers the decision that has the highest probability (hereafter referred to as the maximum decision). Thus, attribute reduction measures should pay different attention to the information of the maximum decision and other non-maximum decisions. In this study, a monotonic measure and high-quality attribute reduction method are proposed based on the three-way decision. The main contributions of this study are threefold:

(1) To address the problem of the non-monotonicity in existing measures, a new attribute reduction measure is proposed, called parameterized maximum distribution entropy (PMDE), which assigns different importance levels to the maximum decision and other non-maximum decisions. To gain deeper insights, the properties of the proposed measure are analyzed and the monotonicity is theoretically proven.

(2) To weaken the negative effect of noisy examples and improve the efficiency of reducing redundant attributes, a three-way decision-based approximate attribute reduction method is presented to obtain more compact and informative reducts with relatively lower computation costs.

(3) To verify the effectiveness of the proposed method, extensive experiments on UCI datasets are performed and very promising results are achieved, indicating the potential of the proposed method.

The remainder of this paper is structured as follows. In Section 2, the theories of rough sets and three-way decision are briefly reviewed. In Section 3, the proposed PMDE and the three-way approximate attribute reduction method are described. In Section 4, the experimental results and the analysis are presented. Finally, conclusions are drawn in Section 5.

2. Preliminaries

2.1. Rough sets

In the theory of rough sets, the set of examples is called the universe of discourse U, and each example $x \in U$ is described by a set of attributes A. For each example x, a value is assigned from the domain of attributes V by an information function f, i.e., $f(x, a) \in V_a$ for every example x on each attribute $a \in A$ and $V = \bigcup V_a$. If there are decision attributes in A, the data of interest are called a decision information system or a decision table [8] which is formally denoted as $DS = (U, A = C \cup D, V, f)$, where C is a set of condition attributes and D is a set of decision attributes. In real-world applications, the decision information systems may have more than one decision attribute, i.e., $|D| \geq 1$. Unless otherwise stated, we assume the decision information system in the paper have one decision attribute, i.e., $D = \{d\}$.

The universe of discourse U can be partitioned into a set of equivalence classes U/B by any attribute subset $B \subseteq A$, which is called the indiscernibility relation and satisfies reflexivity, symmetry, and transitivity. Each equivalence class containing x, denoted by $[x]_B$, is referred to as a B-elementary granule [8]. For any subset $X \subseteq U$, the B-lower and B-upper approximations of X with respect to B are defined as follows [8]:

$$\begin{aligned} \underline{B}(X) &= \bigcup \{x \in U | [x]_B \subseteq X\}, \\ \overline{B}(X) &= \bigcup \{x \in U | [x]_B \cap X \neq \emptyset\}. \end{aligned} \tag{1}$$

Formally, the B-lower approximation $\underline{B}(X)$ is a set of B-elementary granules that are completely contained within X, whereas the B-upper approximation $\overline{B}(X)$ is a set of B-elementary granules that possibly belong to X. The difference between $\overline{B}(X)$ and $\underline{B}(X)$ is called the boundary region. X is called a rough set with respect to B if the boundary region of X is nonempty; otherwise, X is a crisp set.

Let $U/D = \{Y_1, Y_2, ..., Y_{|U/D|}\}$ be the partition induced by the decision attribute D, where Y_i is a set of examples with decision i. An equivalence class is regarded to be consistent if the examples in the equivalence class have the same class; otherwise, it is an inconsistent equivalence class and the examples in the equivalence class is considered as inconsistent examples. For any attribute subset $B \subseteq C$, the B-positive and B-boundary regions of D over U are defined as follows [8]:

$$\begin{aligned} POS_B(D) &= \bigcup_{Y_i \in U/D} \underline{B}(Y_i), \\ BND_B(D) &= \bigcup_{Y_i \in U/D} (\overline{B}(Y_i) - \underline{B}(Y_i)). \end{aligned} \tag{2}$$

The positive region is composed of the equivalence classes that can be certainly classified into one of the equivalence classes in U/D under a given subset $B \subseteq C$. In Pawlak's rough set model, there is no negative region when the concept of approximation contains all decision classes, that is, $NEG_B(D) = \emptyset$. The positive region can be quantified to measure the classification ability of an attribute or attribute set. For any attribute subset $B \subseteq C$, the classification quality of B with respect to D is defined as follows [8]:

$$\gamma_B(D) = \frac{|POS_B(D)|}{|U|}, \tag{3}$$

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

2.2. Three-way decision

In Pawlak's rough set model and its extensions, approximate reasoning and decision-making involve only data. However, decision-making in practical applications is closely related to costs or risks, that is, taking different decisions leads to different costs and risks. The theory of three-way decision aims to make decisions with the minimum cost by introducing the Bayesian decision theory.

More specifically, let a_P, a_B, and a_N be the actions to decide whether an example x belongs to the C-positive, C-boundary, or C-negative regions of X, respectively. The decision costs for taking different actions can be described as follows [18]:

$$\begin{aligned} R_C(a_P, x) &= \lambda_{PP} P(X|[x]_C) + \lambda_{PN}(1 - P(X|[x]_C)), \\ R_C(a_B, x) &= \lambda_{BP} P(X|[x]_C) + \lambda_{BN}(1 - P(X|[x]_C)), \\ R_C(a_N, x) &= \lambda_{NP} P(X|[x]_C) + \lambda_{NN}(1 - P(X|[x]_C)), \end{aligned} \tag{4}$$

where $P(X|[x]_C)$ denotes the probability that the example x belongs to X with respect to C, i.e., $P(X|[x]_C) = |X \cap [x]_C|/|[x]_C|$; $\lambda_{PP}, \lambda_{BP}$, and λ_{NP} denote the costs of taking the actions a_P, a_B, or a_N when x belongs to X, respectively, whereas λ_{PN}, λ_{BN}, and λ_{NN} denote the costs of taking the actions a_P, a_B, or a_N when x does not belong to X, respectively.

Considering Bayesian decision theory, the following decision rules can be derived [18]:

(P) if $R_C(a_P, x) \leq \min\{R_C(a_B, x), R_C(a_N, x)\}$, then decide $x \in POS(X)$;

(B) if $R_C(a_B, x) \leq \min\{R_C(a_P, x), R_C(a_N, x)\}$, then decide $x \in BND(X)$;

(N) if $R_C(a_N, x) \leq \min\{R_C(a_P, x), R_C(a_B, x)\}$, then decide $x \in NEG(X)$.

When $(\lambda_{PN} - \lambda_{BN})(\lambda_{NP} - \lambda_{BP}) > (\lambda_{BP} - \lambda_{PP})(\lambda_{BN} - \lambda_{NN})$, the above Bayesian minimum-cost decision rules can be simplified as follows [18]:

(P) if $P(X|[x]_C) \geq \alpha$, then decide $x \in POS(X)$;

(B) if $\beta < P(X|[x]_C) < \alpha$, then decide $x \in BND(X)$;

(N) if $P(X|[x]_C) \leq \beta$, then decide $x \in NEG(X)$,

where

$$\begin{aligned} \alpha &= \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}, \\ \beta &= \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}. \end{aligned} \tag{5}$$

Formally, the above rules can be rewritten as follows [18]:

$$\begin{aligned} POS_C^{(\alpha,\beta)}(D) &= \{x \in U | P(D_{max}([x]_C)|[x]_C) \geq \alpha\}, \\ BND_C^{(\alpha,\beta)}(D) &= \{x \in U | \beta < P(D_{max}([x]_C)|[x]_C) < \alpha\}, \\ NEG_C^{(\alpha,\beta)}(D) &= \{x \in U | P(D_{max}([x]_C)|[x]_C) \leq \beta\}, \end{aligned} \tag{6}$$

where $D_{max}([x]_C) = argmax_{Y_i \in U/D}\{P(Y_i|[x]_C)\}$.

Accordingly, the (α, β) probabilistic classification quality of C with respect to D is defined as follows [18]:

$$\gamma_C^{(\alpha,\beta)}(D) = \frac{|POS_C^{(\alpha,\beta)}(D)|}{|U|}. \tag{7}$$

3. Parameterized-maximum-distribution entropy-based three-way approximate attribute reduction

In this section, we first elaborate on the proposed measure and then develop a heuristic approximate attribute reduction algorithm using the idea of trisection in the three-way decision.

3.1. Parameterized maximum distribution entropy

Information entropy is a useful measure for quantifying the degree of uncertainty, and many information-theoretic measures have thus been proposed to estimate the correlation or redundancy between attributes. In what follows, some concepts related to information entropy are first described.

Definition 1. *Given a decision information system $DS = (U, A = C \cup D, V, f)$ and the partition $U/B = \{X_1, X_2, ..., X_{|U/B|}\}$ induced by an attribute subset $B \subseteq A$, the entropy of B over U is defined as follows [56]:*

$$H(B) = -\sum_{i=1}^{|U/B|} P(X_i) \log P(X_i), \tag{8}$$

where $P(X_i) = |X_i|/|U|$, denoting the probability of occurrence of the equivalence class X_i.

Definition 2. *Given a decision information system* $DS = (U, A = C \cup D, V, f)$, *the partition* $U/D = \{Y_1, Y_2, ..., Y_{|U/D|}\}$ *induced by* D, *and the partition* $U/B = \{X_1, X_2, ..., X_{|U/B|}\}$ *induced by an attribute subset* $B \subseteq C$, *the conditional entropy of* D *given* B *is defined as follows [56]:*

$$H(D|B) = -\sum_{i=1}^{|U/B|} P(X_i) \sum_{j=1}^{|U/D|} P(Y_j|X_i) \log P(Y_j|X_i), \tag{9}$$

where $P(Y_j|X_i) = |X_i \cap Y_j|/|X_i|$.

In conditional entropy, uncertainty is measured by the entropy of each conditional class under different decisions, and each example in the condition class is considered equally, regardless of which decision the example belongs to. However, human decision-making under uncertainty tends to make the maximum decision because of the highest degree of confidence. The objective of attribute reduction is to preserve the classification ability of the data. Whereas the classification ability under uncertainty is reflected not only by the maximum decision, but also by the degree of confidence associated with the maximum decision. Therefore, in a sense, the preservation of the maximum decision and the corresponding confidence are sufficient for classification. The following example illustrates this problem.

Example 1. *Given a decision information system* $DS = (U, A = C \cup D, V, f)$ *as shown in Table 1, where* $U = \{x_1, x_2, ..., x_{17}\}$, $C = \{a_1, a_2, ..., a_5\}$, $D = \{d\}$, $V_{a_1} = V_{a_2} = V_{a_3} = V_{a_4} = V_{a_5} = \{0, 1\}$, *and* $V_d = \{d_1, d_2, d_3\}$.

Table 1: A decision information system

	a_1	a_2	a_3	a_4	a_5	d
x_1	0	1	0	0	0	d_1
x_2	0	1	0	0	0	d_1
x_3	0	1	0	0	1	d_2
x_4	0	1	0	0	1	d_2
x_5	0	0	0	1	1	d_1
x_6	0	0	0	1	1	d_1
x_7	0	0	0	1	1	d_2
x_8	0	0	0	1	1	d_3
x_9	0	0	0	1	1	d_3
x_{10}	0	0	0	1	1	d_3
x_{11}	0	0	1	1	1	d_1
x_{12}	0	0	1	1	1	d_2
x_{13}	0	0	1	1	1	d_2
x_{14}	0	0	1	1	1	d_3
x_{15}	0	0	1	1	1	d_3
x_{16}	0	0	1	1	1	d_3
x_{17}	1	0	1	1	1	d_3

In Table 1, the universe of discourse under all condition attributes is divided into three consistent equivalence classes $X_1 : \{x_1, x_2\}$, $X_2 : \{x_3, x_4\}$, and $X_3 : \{x_{17}\}$, and two inconsistent equivalence classes $X_4 : \{x_5, x_6, x_7, x_8, x_9, x_{10}\}$, and $X_5 : \{x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}\}$. Probabilistic classification quality is a non-monotonic measure. When the probabilistic classification quality is used as the attribute reduction measure for Table 1, and the parameter α is set to 0.55, the overall measure under all condition attributes is $\gamma_C^{0.55}(D) = 5/17$. The heuristic algorithm with forward adding strategy successively selects the attributes

a_3, a_5, a_2, a_4, and a_1, whereas the attribute reduction measures after adding these attributes are $\gamma^{0.55}_{\{a_3\}}(D) = 7/17$, $\gamma^{0.55}_{\{a_3,a_5\}}(D) = 9/17$, $\gamma^{0.55}_{\{a_3,a_5,a_2\}}(D) = 11/17$, $\gamma^{0.55}_{\{a_3,a_5,a_2,a_4\}}(D) = 11/17$, and $\gamma^{0.55}_{\{a_3,a_5,a_2,a_4,a_1\}}(D) = 5/17$, respectively. The probabilistic classification quality measure is non-monotonic as adding attributes, which brings great difficulties in evaluating the goodness of attribute subsets and determining the stopping conditions for the attribute reduction algorithm. Moreover, the resulting reduct may have redundant attributes so that its quality can not be ensured. Taking Table 1 as an example, the obtained reduct is $\{a_3, a_5, a_2, a_4, a_1\}$, which is exactly the same as the original condition attribute set. Whereas the attribute a_2 or a_4 is relatively redundant and can be removed without losing any classification information.

Conditional entropy is a commonly used monotonic measure, and the overall conditional entropy of D given C is $H(D|C) = 1.0300$. When the forward-adding heuristic search strategy is used, the conditional-entropy-based attribute reduction algorithm will add the attributes a_5, a_2, a_1, and a_3 to the reduct sequentially, and the condition entropies after adding these attributes are $H(D|\{a_5\}) = 1.3287$, $H(D|\{a_5, a_2\}) = 1.1144$, $H(D|\{a_5, a_2, a_1\}) = 1.0588$, and $H(D|\{a_5, a_2,\ a_1, a_3\}) = 1.0300$, respectively. In other words, the attribute a_4 is redundant under the measure of conditional entropy. Under the reduct $\{a_5, a_2, a_1, a_3\}$, the useful classification rules and their confidences are as follows.

$r_1 : x_1 \xrightarrow[\{a_5,a_2,a_1,a_3\}]{} d_1$(confidence: 1),

$r_2 : x_3 \xrightarrow[\{a_5,a_2,a_1,a_3\}]{} d_2$(confidence: 1),

$r_3 : x_8 \xrightarrow[\{a_5,a_2,a_1,a_3\}]{} d_3$(confidence: 0.5),

$r_4 : x_{14} \xrightarrow[\{a_5,a_2,a_1,a_3\}]{} d_3$(confidence: 0.5),

$r_5 : x_{17} \xrightarrow[\{a_5,a_2,a_1,a_3\}]{} d_3$(confidence: 1).

Among these rules, the rules r_3 and r_4 are related to the two inconsistent equivalence classes X_4 and X_5, respectively, and they are all probabilistic rules. To minimize the risk, the decisions of the two rules are determined by the maximum decisions that have the highest probability, i.e., the decision d_3. While their rule confidences are determined by the probability of the maximum decision (hereafter referred to as the maximum probability), i.e., the probability of 0.5. However, when removing the attribute a_3 from the reduct, the equivalence classes X_4 and X_5 are merged, and the useful classification rules and their confidences are

$r_1 : x_1 \xrightarrow[\{a_5,a_2,a_1\}]{} d_1$(confidence: 1),

$r_2 : x_3 \xrightarrow[\{a_5,a_2,a_1\}]{} d_2$(confidence: 1),

$r_3 : x_8 \xrightarrow[\{a_5,a_2,a_1\}]{} d_3$(confidence: 0.5),

$r_5 : x_{17} \xrightarrow[\{a_5,a_2,a_1\}]{} d_3$(confidence: 1).

After removing the attribute a_3, although the two inconsistent equivalence classes are merged into one inconsistent equivalence class, their maximum decisions and maximum probabilities are still d_3 and 0.5, respectively, and the useful classification rules derived by the attribute subsets $\{a_5, a_2, a_1, a_3\}$ and $\{a_5, a_2, a_1\}$ are almost the same. In other words, attribute subset $\{a_5, a_2, a_1\}$, in a sense, is sufficient for classification. Thus, when performing attribute reduction, it is better to treat the information of the maximum decision and other non-maximum decisions differently, so that the obtained reduct and decision rules can be more generalized. Furthermore, practical data may be contaminated by noise; thus,

the process of attribute reduction should capture the key attributes that reflect the inherent characteristics of data and avoid introducing attributes that overfit noisy examples. In Table 1, for example, the attribute reduction algorithm has to select the attribute a_1 to discern the example x_{17} from the others. If the example x_{17} is corrupted in attribute a_1, the algorithm has unintentionally added an unimportant attribute a_1 into the reduct. Moreover, after removing a_1, the corresponding decision rule related to x_{17} still has the decision d_3. That is, these kinds of attributes may limit the generalization of derived decision rules and thus can be further removed to make the reduct more concise. To this end, a measure of parameterized maximum distribution entropy is proposed.

Definition 3. *Given a decision information system $DS = (U, A = C \cup D, V, f)$, the partition $U/D = \{Y_1, Y_2, ..., Y_{|U/D|}\}$ induced by D, and an attribute subset $B \subseteq C$, the maximum probability, the maximum decision, and the maximum distribution of an example $x \in U$ are denoted as $\mu_B^{max}(x) = \max\{P(Y_1|[x]_B), P(Y_2|[x]_B), ..., P(Y_{|U/D|}|[x]_B)\}$, $\nu_D^{max}(x) = \{f(y, D)|y \in Y_j \wedge \mu_B^{Y_j}(x) = \mu_B^{max}(x)\}$, and $\mathcal{D}(x) = (\mu_B^{max}(x), 1 - \mu_B^{max}(x))$, respectively.*

Definition 4. *Given a decision information system $DS = (U, A = C \cup D, V, f)$, an attribute subset $B \subseteq C$, and a trade-off parameter $\lambda \in [0, 1]$, the parameterized maximum distribution entropy (PMDE) of D given B is denoted as follows:*

$$\begin{aligned} MH_\lambda(D|B) = -\frac{1}{|U|}\sum_{x\in U}(&\lambda\mu_B^{max}(x)\log\mu_B^{max}(x) \\ &+ (1-\lambda)(1-\mu_B^{max}(x))\log(\frac{1-\mu_B^{max}(x)}{k-1})), \end{aligned} \tag{10}$$

where k is the number of decisions within the condition class containing x.

Formally, the PMDE degrades to conditional entropy when $|U/D| = 2$ and $\lambda = 0.5$. In the definition, the PMDE involves two parts of information embodied in the condition class of each example. That is, the certainty and uncertainty of making the maximum decision. The former is reflected by the examples that have the maximum decision. The larger the number of these examples, the higher the confidence in making the maximum decision, and the smaller the PMDE. Conversely, the latter is related to examples that do not belong to the maximum decision. According to the principle of maximum entropy, the probability distribution leading to the maximum entropy is the most objective and reasonable solution when the available knowledge about the unknown distribution is limited. Thus, the latter part is averaged to reflect the uncertainty of making the maximum decision. The greater the number of these examples, the lower the certainty to make the maximum decision, and the larger the PMDE. In decision-making under uncertainty, we usually pay different attention to these two parts. Particularly, we prefer to consider the decision that has the maximum probability. Thus, we introduce a trade-off parameter λ to flexibly adjust the importance of the two parts in the PMDE.

Proposition 1. *Given a decision information system $DS = (U, A = C \cup D, V, f)$ and a trade-off parameter $\lambda \in [0, 1]$, for an attribute subset $B \subseteq C$, $0 \leq MH_\lambda(D|B) < \log|U|$.*

Proof. For each example $x \in U$, on the one hand, when x is a consistent example with respect to D, namely $\mu_B^{max}(x) = 1$, its uncertainty is minimized to 0. Thus,

the overall PMDE is minimized to 0 when all the examples are consistent. On the other hand, when x is extremely inconsistent with D, that is, each example x has its own decision that differs from that of other examples, the overall uncertainty attains the maximum. In that case, the partition induced by B has only one equivalence class, namely $U/B = U$, and $\mu_B^{max}(x) = 1/|U|$, then $MH_\lambda(D|B) = -\frac{1}{|U|}|U|(\lambda\frac{1}{|U|}\log(\frac{1}{|U|}) - (1-\lambda)(1-\frac{1}{|U|})\log(\frac{|U|-1}{|U|(|U|-1)})) = \frac{1}{|U|}(\lambda\log(|U|) + (1-\lambda)(|U|-1)\log(|U|)) = \frac{(|U|-1)-\lambda(|U|-2)}{|U|}\log|U|$. When $\lambda = 0$, the overall PMDE is maximized at $MH_\lambda(D|B) = \frac{|U|-1}{|U|}\log|U| < \log|U|$. For all possible cases, $0 \leq MH_\lambda(D|B) < \log|U|$. This proposition has been proven.

Proposition 2. *Given a decision information system $DS = (U, A = C \cup D, V, f)$ and a trade-off parameter $\lambda \in [0,1]$, for two attribute subsets $P, Q \subseteq C$ with $P \subset Q$, then $MH_\lambda(D|P) \geq MH_\lambda(D|Q)$.*

Proof. Assume that only the equivalence class X_{ij} under P is partitioned into two equivalence classes X_i and X_j under Q. The proof is presented in Appendix.

Monotonicity is a highly desirable characteristic of attribute reduction measures, which is beneficial in designing heuristic attribute reduction algorithms. As shown in Proposition 2, the PMDE decreases monotonically when attributes are added; thus, it can be used as a quantitative measure for attribute reduction.

Similarly, the positive, boundary, and negative PMDE can be defined when the decision costs are given.

Definition 5. *Given a decision information system $DS = (U, A = C \cup D, V, f)$, a pair of threshold parameters (α, β) derived from the decision costs, and a trade-off parameter $\lambda \in [0,1]$ for an attribute subset $B \subseteq C$, the positive, boundary, and negative PMDE of D given B are denoted as follows:*

$$\begin{aligned} MH_\lambda(POS_\beta^\alpha(D|B)) &= \sum_{\mu_B^{max}(x) \geq \alpha} MH_\lambda(D, B, x), \\ MH_\lambda(BND_\beta^\alpha(D|B)) &= \sum_{\beta < \mu_B^{max}(x) < \alpha} MH_\lambda(D, B, x), \\ MH_\lambda(NEG_\beta^\alpha(D|B)) &= \sum_{\mu_B^{max}(x) \leq \beta} MH_\lambda(D, B, x), \end{aligned} \tag{11}$$

where

$$\begin{aligned} MH_\lambda(D, B, x) = -\frac{1}{|U|}(&\lambda\mu_B^{max}(x)\log\mu_B^{max}(x) \\ &+ (1-\lambda)(1-\mu_B^{max}(x))\log(\frac{1-\mu_B^{max}(x)}{k-1})). \end{aligned} \tag{12}$$

In Definition 5, the positive PMDE only considers the entropy of the examples whose maximum probability is not lower than α, and the negative PMDE is composed of the entropy of the examples whose maximum probability is not larger than β. Whereas the entropy of the examples with the maximum probability greater β but less than α constitutes the boundary PMDE. Formally, the non-negative PMDE can be defined by combining the positive PMDE and the boundary PMDE.

Definition 6. *Given a decision information system $DS = (U, A = C \cup D, V, f)$, a trade-off parameter $\lambda \in [0, 1]$, and an attribute subset $P \subset C$ for a condition attribute $a \in (C - P)$, the relative significance of D given P is defined as follows:*

$$Sig(a, P, D) = MH_\lambda(D|P) - MH_\lambda(D|(P \cup \{a\})), \tag{13}$$

Definition 7. *Given a decision information system $DS = (U, A = C \cup D, V, f)$ and a trade-off parameter $\lambda \in [0, 1]$, an attribute subset $P \subseteq C$ is called a measure preservation reduct of C with respect to D if the following conditions are satisfied:*

(I) $MH_\lambda(D|P) = MH_\lambda(D|C)$, *and*

(II) $\forall P^* \subset P$, $MH_\lambda(D|P^*) \neq MH_\lambda(D|C)$.

In general, finding a minimal reduct (the reduct with the fewest attributes) is NP-hard, and a heuristic search technique is an alternative to this problem. Based on the two principles in Definition 7, a heuristic attribute reduction algorithm can be designed to find the optimal reduct, and the procedure is outlined in Algorithm 1.

Algorithm 1 A heuristic attribute reduction algorithm based on parameterized maximum distribution entropy (PMDE)

Input:

A decision information system $DS = (U, A = C \cup D, V, f)$ and the trade-off parameter $\lambda \in [0, 1]$.

Output:

An optimal reduct of P;

1: Compute the overall parameterized maximum distribution entropy $MH_\lambda(D|C)$, and $P \leftarrow \emptyset$;
2: **while** $MH_\lambda(D|P) \neq MH_\lambda(D|C)$ **do**
3: For each a in $C - P$, calculate its relative significance $Sig(a, P, D)$;
4: Select an attribute a^* that has the maximum significance;
5: $P \leftarrow P \cup \{a^*\}$;
6: **end while**
7: **return** the optimal reduct P.

As in many existing heuristic algorithms, the proposed algorithm starts with an empty set and then uses a greedy strategy to select an attribute that achieves the largest information gain in each iteration until the selected attribute set has the same PMDE as the original condition attribute set. Assume that there are $|U|$ examples described by $|C|$ attributes. The algorithm takes $O(|C||U|)$ time to select each optimal attribute and terminates after $|C|$ rounds of selection. Thus, the time complexity of Algorithm 1 for finding an optimal reduct is at most $O(|C|^2|U|)$, and the space complexity is $O(|C||U|)$.

3.2. Three-way approximate reduct based on the proposed measure

Intuitively, the attribute reduction process determines whether an attribute is selected for the reduct. Formally, considering the correlation with the decision attribute, all attributes can be grouped into strongly related, weakly related, and irrelevant attributes, and a reduct should contain all strongly related attributes and some of the weakly related attributes, but exclude all irrelevant attributes. On the one hand, the existing heuristic attribute reduction algorithms greedily

select the attributes that maximize the correlation measure. Consequently, redundancy between the selected attributes is inevitably introduced into the reduct. Especially in the last few rounds of attribute selection, the relevance between the selected attributes and the decision attribute decreases gradually, whereas the redundancy of the selected attributes increases dramatically. On the other hand, examples collected from practical tasks may be contaminated by noise. Most attribute reduction algorithms terminate when their measures are strictly satisfied. To fit the noisy data completely and perfectly, the resulting reduct may select some unimportant attributes, which not only reduce the efficiency of attribute reduction, but also limit the generalization performance. Motivated by the theory of three-way decision, an approximate attribute reduction algorithm is proposed to further refine the resulting reduct obtained by the measure of criterion preservation. More specifically, each attribute is considered to be positive, uncertain, or negative, according to its role in attribute reduction. The attribute reduction algorithm can first exclude all negative and part of uncertain attributes by the measure of criterion preservation, and then try to relax the measure to some extent. As a result, some of the uncertain attributes in the criterion preservation reduct can be further removed. A more concise approximate reduct can be obtained through the process of adding and deleting attributes. In fact, a more efficient way of adding can be employed instead of adding and deleting, by imposing an extra constraint condition. The forward-adding search process for an approximate reduct is described in Algorithm 2.

Algorithm 2 A three-way approximate attribute reduction algorithm based on PMDE

Input:

A decision information system $DS = (U, A = C \cup D, V, f)$, the trade-off parameter λ, and the threshold parameter δ.

Output:

An optimal approximate reduct P;

Compute the overall parameterized maximum distribution entropy $MH_\lambda(D|C)$ and the number of consistent examples $N_C(D)$, and $P \leftarrow \emptyset$;
while $MH_\lambda(D|P) \neq MH_\lambda(D|C)$ **do**
  For each a in $C - P$, calculate its relative significance $Sig(a, P, D)$;
  Select an attribute a^* that has the maximum significance;
  $P \leftarrow P \cup \{a^*\}$, and update $MH_\lambda(D|P)$ and $N_P(D)$;
  if $N_P(D) \geq \delta N_C(D)$ **then**
    Break;
  end if
end while
return the optimal approximate reduct P.

Compared with Algorithm 1, Algorithm 2 introduces another stopping condition, that is, the minimum number of consistent examples. Under the original attribute set, all examples are divided into consistent and inconsistent examples. Because the consistent examples have no uncertainty, their entropy is 0. To reduce the negative impact of noisy consistent examples, the algorithm terminates when the minimum number of consistent examples is met, which keeps most consistent examples unchanged and ignores the impact of noisy examples, thus resulting in high efficiency attribute reduction. The algorithm computes the entropy of the condition class of each example, so there is no extra computation for counting the number of consistent examples. Thus, the time and space complex-

ity of Algorithm 2 is almost the same as or even smaller than that of Algorithm 1 because the early stopping condition is embedded.

4. Empirical analysis

Extensive experiments were conducted to validate the effectiveness of the proposed measure for attribute reduction and to examine the quality of the three-way approximate reduct. The proposed methods were implemented in Python 3.8, and all experiments were performed on a Windows 10 PC with Intel i5-4590 CPU@3.30 GHz and 8 GB RAM.

4.1. Investigated datasets

Sixteen UCI datasets [1] are selected for the experiments, and their details are summarized in Table 2.

Table 2: The investigated datasets.

Dataset	$\|U\|$	$\|C\|$	$\|U/D\|$	Missing	Inconsistency
arrhythmia(arrhythmia)	452	279(206)	13	Yes	0
autos(autos)	205	25(15)	6	Yes	0
biodegradation(biodegrade)	1055	41(41)	2	No	15
cardiotocography-FHR pattern(cardio)	2126	21(21)	10	No	134
companies-bankruptcy-5year(company)	5910	64(64)	2	Yes	2
credit card clients(credit)	30000	23(23)	2	No	5999
hepatitis(hepatitis)	155	19(6)	2	Yes	0
hypothyroid(hypothyroid)	3772	29(7)	4	Yes	0
kr-vs-kp(krvskp)	3196	36(0)	2	No	0
mfeat-fourier(mfeat)	2000	76(76)	10	No	0
sick(sick)	3772	29(7)	2	Yes	0
sonar(sonar)	208	60(60)	2	No	0
splice(splice)	3175	60(60)	3	No	2
turkiye-student-evaluation-generic(turkiye)	5820	31(31)	3	No	3349
vote(vote)	435	16(0)	2	Yes	0
waveform-5000(waveform)	5000	40(40)	3	No	0

In Table 2, the second to fourth columns denote the numbers of examples, attributes, and classes in each dataset, respectively, where the number of continuous attributes is recorded in the brackets of the third column. The fifth column indicates whether the dataset has missing values. To facilitate the experiments, all missing values were completed with the attribute mean (or mode), and the continuous attributes in each dataset were discretized by the technique of equal frequency with three bins. The number of inconsistent examples within each preprocessed dataset is shown in the last column.

4.2. The effectiveness of the proposed measure

To validate the effectiveness of the PMDE, an attribute reduction was performed on the selected datasets. More specifically, all examples in each dataset were used for attribute reduction to ensure the uniqueness and repeatability of the obtained reduct, and the parameter λ was ranged from 0 to 1 with a step size of 0.05. The results are presented in Table 3.

Table 3 shows that the PMDE under different parameter values generates slightly different results. When the value of λ is small, the PMDE tends to select more attributes. This can be attributed to the fact that the proposed measure with a smaller value pays more attention to the uncertain information in making the maximum decision, whereas the objective of attribute reduction is

[1] http://archive.ics.uci.edu/ml/index.php.

Table 3: The number of attributes within the obtained reducts under different values λ.

	0.0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.0
arrhythmia	10	10	10	10	10	10	10	10	10	10	10	9	9	9	9	9	9	10	9	9	9
autos	11	11	11	11	11	11	11	11	11	11	11	10	10	10	11	11	11	11	11	11	11
biodegrade	21	20	20	20	20	20	20	20	18	18	18	18	18	18	17	19	19	19	19	19	19
cardio	19	19	19	19	19	19	19	19	19	19	19	19	19	20	20	20	20	20	20	19	19
company	22	22	22	22	22	22	22	22	22	22	22	22	21	21	21	21	21	21	21	21	21
credit	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23
hepatitis	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
hypothyroid	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23
krvskp	29	29	29	29	29	29	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30
mfeat	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	14	13	13	12
sick	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
sonar	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
splice	11	11	11	11	11	11	11	11	11	11	11	11	10	10	10	10	10	10	10	10	10
turkiye	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31
vote	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
waveform	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
Avg.	17.71	17.64	17.64	17.64	17.64	17.64	17.71	17.71	17.57	17.57	17.57	17.43	**17.29**	17.36	17.36	17.50	17.50	17.64	17.50	17.43	17.36

to preserve both certain and uncertain information. Accordingly, when the value of λ is large, the proposed measure also generates reducts with more attributes, but comparatively fewer than those with a smaller λ. In this case, the proposed measure prefers the attributes that make the examples confident in the decisions. Because uncertain information is also helpful in attribute reduction, the proposed measure with a higher value also yields undesirable results. On some datasets such as "credit" and "hepatitis", the proposed PMDE under different values of λ seems to obtain the reducts with the same number of attributes, but the attributes in the reducts are different, thus resulting in different performance. This fact will be confirmed in the following experiments. For all selected datasets, the proposed measure achieved the reducts with the lowest average number of attributes when the value of λ was 0.60, at which point certain and uncertain information seem to be well balanced.

To examine the quality of the reducts obtained by the proposed measure, a 10-fold cross-validation experiment under different values of λ was used for the performance evaluation. Specifically, the redundant and irrelevant attributes that are not included in the reduct were removed first. Then, the reduced dataset was randomly divided into 10 equal subsets, where nine subsets were used as the training data, and one subset was used as the testing data. The experiment was repeated 10 times, and the final performance was averaged. In the experiments, three commonly used classifiers were used, namely K-nearest neighbor ($K = 3$), support vector machine (LinearSVC and max-iter=10000), and Decision Tree (Default parameters) [57], and their error rates are displayed in Tables 4-6, respectively.

Table 4: The classification error rates of the obtained reducts under different values λ (KNN).

	0.0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.0
arrhythmia	**44.09**	44.09	44.09	47.60	47.60	47.60	47.60	47.60	47.60	47.60	47.60	48.52	48.52	48.52	47.92	47.92	47.88	46.29	47.60	47.60	47.14
autos	36.47	36.47	36.47	36.47	36.47	36.47	36.47	36.47	36.47	36.47	36.47	30.36	**30.36**	30.36	32.22	32.22	32.22	32.22	32.22	32.22	32.22
biodegrade	**16.24**	16.47	16.47	16.47	16.47	16.47	16.47	16.47	17.69	17.69	17.69	17.69	17.69	17.69	18.42	16.98	16.98	17.43	17.43	17.43	17.43
cardio	27.79	27.79	27.79	27.79	27.79	27.79	27.79	27.79	27.79	27.79	27.79	27.79	**27.79**	27.93	27.93	27.92	27.92	27.92	27.92	28.16	28.07
company	**5.84**	5.95	5.95	5.95	5.95	5.95	5.95	5.95	5.95	5.95	5.95	5.84	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99
credit	22.56	22.59	22.63	22.67	22.54	22.54	22.54	22.46	22.46	22.46	22.49	22.49	22.49	22.51	22.59	22.59	22.48	22.58	22.58	22.58	22.58
hepatitis	21.49	21.49	**16.11**	16.11	16.11	16.11	16.11	16.11	20.94	20.94	20.94	16.35	16.35	16.35	16.87	18.60	18.60	18.60	18.60	17.09	17.09
hypothyroid	9.01	9.01	9.01	9.01	9.01	9.01	9.01	9.01	9.01	9.01	9.01	9.01	**9.01**	9.01	9.01	9.01	9.01	9.01	9.01	9.01	9.01
krvskp	3.55	3.55	3.55	3.55	3.55	3.55	3.49	3.49	3.49	3.57	3.57	3.57	3.57	3.57	3.57	3.46	3.46	3.46	3.46	**3.39**	3.40
mfeat	31.34	31.34	31.34	31.34	31.34	31.34	31.34	31.34	28.67	28.27	28.27	28.27	**27.48**	27.48	27.48	27.48	30.58	29.79	33.13	36.52	43.16
sick	7.84	7.84	7.84	7.84	7.84	7.84	7.84	7.84	7.84	7.84	7.84	7.84	**7.84**	7.84	7.84	7.84	7.84	7.84	7.84	7.84	7.84
sonar	**20.97**	20.97	20.97	20.97	20.97	20.97	20.97	21.74	21.74	21.74	21.74	21.74	21.74	21.74	21.74	21.74	21.74	21.50	23.09	23.09	23.09
splice	15.41	15.41	15.41	15.41	15.41	15.41	15.41	15.41	15.43	15.43	15.43	15.43	12.26	12.26	12.26	12.10	12.10	12.10	12.10	12.10	12.14
turkiye	47.53	47.53	47.53	47.53	47.53	47.53	47.53	47.53	47.53	47.53	47.53	47.53	**47.51**	47.51	47.51	47.81	47.81	47.81	47.81	47.96	47.96
vote	5.93	5.93	5.93	5.93	5.93	5.93	5.93	5.93	5.93	5.93	5.93	5.93	**5.93**	5.93	5.93	5.93	5.93	5.93	5.93	5.93	5.93
waveform	28.40	28.40	28.40	28.16	28.16	28.16	28.16	28.16	28.22	28.22	28.22	28.33	28.33	29.27	29.27	29.27	29.27	29.27	28.43	28.43	**27.41**
Avg.	21.53	21.55	21.22	21.43	21.42	21.42	21.41	21.46	21.67	21.65	21.65	21.04	**20.80**	20.87	21.03	21.05	21.24	21.11	21.45	21.58	21.90

Table 5: The classification error rates of the obtained reducts under different values λ (SVM).

	0.0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.0
arrhythmia	42.61	42.61	42.61	42.83	42.83	42.83	42.83	42.83	42.83	42.83	42.83	43.16	43.16	43.16	42.10	42.10	42.50	41.99	43.03	43.03	41.63
autos	51.26	51.26	51.26	51.26	51.26	51.26	51.26	51.26	51.26	51.26	51.26	49.80	**49.80**	49.80	52.78	52.78	52.78	52.78	52.78	52.78	52.78
biodegrade	**14.36**	14.37	14.37	14.37	14.37	14.37	14.37	14.37	15.84	15.84	15.84	15.84	15.84	15.84	15.20	15.78	15.78	15.26	15.26	15.26	15.26
cardio	22.34	22.34	22.34	22.34	22.34	22.34	22.34	22.34	22.34	22.34	22.34	22.34	22.34	**22.14**	22.14	22.14	22.14	22.14	22.14	22.34	22.34
company	5.37	5.37	5.37	5.37	5.37	5.37	5.37	5.37	5.37	5.37	5.37	5.37	**5.27**	5.27	5.27	5.27	5.27	5.27	5.27	5.27	5.27
credit	18.57	18.57	18.57	18.57	18.57	18.57	18.57	18.57	18.57	18.57	18.57	18.57	**18.57**	18.57	18.57	18.57	18.57	18.57	18.57	18.57	18.57
hepatitis	18.85	18.85	15.50	15.50	15.50	15.50	15.50	15.50	17.37	17.37	17.37	15.00	**15.00**	15.00	15.50	16.92	16.92	16.92	16.92	15.00	15.00
hypothyroid	7.71	7.71	7.71	7.71	7.71	7.71	7.71	7.71	7.71	7.71	7.71	7.71	**7.71**	7.71	7.71	7.71	7.71	7.71	7.71	7.71	7.71
krvskp	2.60	2.60	2.60	2.60	2.60	2.60	2.68	2.68	2.68	2.40	2.4	2.40	2.40	2.40	2.40	2.40	2.40	2.40	2.40	**2.34**	2.34
mfeat	25.22	25.22	25.22	25.22	25.22	25.22	25.22	25.22	23.29	23.29	23.29	23.29	**21.78**	21.78	21.78	21.78	23.17	24.02	26.26	28.11	36.01
sick	6.12	6.12	6.12	6.12	6.12	6.12	6.12	6.12	6.12	6.12	6.12	6.12	**6.12**	6.12	6.12	6.12	6.12	6.12	6.12	6.12	6.12
sonar	18.85	18.85	18.85	18.85	18.85	18.85	18.85	17.89	17.89	17.89	17.89	17.89	**17.89**	17.89	17.89	17.89	17.89	20.02	21.17	21.17	21.17
splice	8.50	8.50	8.50	8.50	8.50	8.50	8.50	8.50	8.50	8.50	8.5	8.50	**7.19**	7.19	7.19	7.19	7.19	7.19	7.19	7.19	7.19
turkiye	37.78	37.78	37.78	37.78	37.78	37.78	37.78	37.78	37.78	37.78	37.78	37.78	**37.78**	37.78	37.78	37.78	37.78	37.78	37.78	37.78	37.78
vote	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14	**4.14**	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14
waveform	22.09	22.09	22.09	21.39	21.39	21.39	21.39	21.39	21.34	21.34	21.34	21.34	21.34	21.55	21.55	21.55	21.55	21.55	21.18	21.18	**20.85**
Avg.	19.15	19.15	18.94	18.91	18.91	18.91	18.92	18.86	18.94	18.92	18.92	18.70	**18.52**	18.52	18.63	18.76	18.87	18.99	19.25	19.25	19.64

Table 6: The classification error rates of the obtained reducts under different values λ (DT).

	0.0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.0
arrhythmia	53.18	53.07	53.18	53.01	52.26	53.01	**52.26**	53.01	52.26	53.01	52.26	54.45	54.14	54.45	55.27	55.04	55.43	52.61	53.27	53.41	52.40
autos	24.45	23.38	24.45	23.38	24.45	23.38	24.45	23.38	24.45	23.38	24.45	22.41	**22.12**	22.41	24.66	23.93	24.66	23.93	24.66	23.93	24.66
biodegrade	**18.21**	18.35	18.47	18.35	18.47	18.35	18.47	18.35	19.65	19.25	19.65	19.25	19.42	19.44	19.50	19.40	19.30	18.69	18.56	18.69	18.56
cardio	**27.22**	27.23	27.22	27.40	27.30	27.40	27.30	27.40	27.30	27.40	27.30	27.40	27.30	27.36	27.37	27.43	27.56	27.43	27.56	27.38	27.32
company	7.83	7.82	7.65	7.82	7.65	7.82	7.65	7.82	7.65	7.82	7.65	7.76	7.68	7.65	7.68	7.65	7.68	7.64	7.72	**7.64**	7.75
credit	26.58	26.64	**26.57**	26.64	26.61	26.67	26.61	26.62	26.60	26.62	26.59	26.62	26.59	26.60	26.58	26.64	26.63	26.65	26.63	26.65	26.63
hepatitis	22.42	22.06	23.42	23.23	23.42	23.23	23.42	23.23	22.90	22.28	22.90	20.78	**20.58**	20.72	23.61	23.03	23.05	23.03	23.05	20.98	20.70
hypothyroid	9.45	9.45	9.45	9.45	9.45	9.45	9.45	9.45	9.45	9.45	9.45	9.45	9.44	9.45	**9.43**	9.43	9.49	9.43	9.45	9.44	9.48
krvskp	0.58	0.56	0.57	0.58	0.55	0.58	0.48	0.48	0.48	**0.42**	0.47	0.44	0.44	0.47	0.49	0.46	0.43	0.45	0.43	0.54	0.55
mfeat	33.10	33.48	33.10	33.48	33.10	33.48	33.10	33.48	29.45	29.73	29.46	29.73	29.81	29.93	29.81	29.93	30.81	**29.38**	33.79	35.15	42.88
sick	8.32	8.38	8.30	8.34	8.32	8.33	8.32	8.36	8.31	8.36	8.31	8.36	8.31	8.32	8.31	8.30	**8.26**	8.30	8.26	8.27	8.33
sonar	22.47	22.41	22.47	22.41	22.47	22.41	22.47	**19.40**	19.84	19.40	19.84	19.40	19.84	19.40	19.84	19.40	19.84	23.98	25.84	25.55	25.84
splice	8.46	8.50	8.46	8.50	8.46	8.50	8.46	8.50	8.42	8.58	8.42	8.58	7.74	7.67	7.74	**7.65**	7.70	7.65	7.70	7.65	7.71
turkiye	43.08	**42.97**	43.13	43.08	43.13	43.08	43.10	43.10	43.10	43.16	43.08	43.21	43.12	43.14	43.04	43.08	43.02	43.07	43.01	43.08	43.05
vote	**5.61**	5.90	5.61	5.90	5.77	6.02	5.77	6.02	5.77	5.99	5.75	5.99	5.75	5.99	5.75	5.99	5.77	5.86	5.77	5.86	5.77
waveform	29.82	29.82	29.82	30.04	30.05	30.04	30.05	30.04	**29.68**	29.89	29.68	29.93	29.74	29.99	30.05	29.99	30.05	29.99	29.99	30.14	29.98
Avg.	21.30	21.25	21.37	21.35	21.34	21.36	21.33	21.16	20.96	20.92	20.95	20.86	**20.75**	20.81	21.19	21.08	21.23	21.13	21.61	21.52	21.98

Tables 4-6 present the classification error rates of the obtained reducts using the classifiers KNN, SVM, and DT. For each value λ, the performance is averaged from 10 runs of 10-fold cross-validation. The highest performance for different values λ is shown in bold, and the average performance over all data sets is recorded in the last row "Avg.".

In Tables 4-6, it is observed that the proposed measure achieves worse performance when the value of λ is relatively small or high. In the extreme case, $\lambda = 0$, the proposed measure does not consider the information related to the maximum decision. Compared with other decisions, the maximum decision is the dominant decision that has the highest probability and is more likely to be the certain decision after adding some attributes. Thus, the proposed measure yields a low-quality reduct with unsatisfactory performance. In contrast, when the value of λ is set to the highest value $\lambda = 1$, the proposed measure only considers the information about the maximum decision, whereas other uncertain decisions may become the maximum decision after selecting attributes. In light of the fact that the maximum decision has a relatively greater chance of being the deterministic decision, the proposed measure pays slightly more attention to the information of the maximum decision. For all selected datasets,the KNN, SVM, and DT classifiers achieved the highest performance when the value of λ approached 0.60, at which point different kinds of information were harmonized properly. In the following experiments, the value of λ was set to 0.60.

4.3. Performance comparison analysis

To further verify the effectiveness, the parameterized maximum distribution entropy (PMDE) and its extension with the three-way decision (PMDE-TWD)

were compared to the attribute reduction measures of classification quality (CQ) [8], conditional entropy (CE) [56], maximum decision entropy (MDE) [39] and granular maximum decision entropy (GMDE) [40], and the approximate attribute reduction measure of relative decision entropy (RDE) [58]. Note that the PMDE-TWD method requires a threshold to terminate the process of eliminating unnecessary boundary attributes. In the experiments, the parameter δ was set to 0.90 empirically. The results of the selected attribute reduction methods are presented in Table 7.

Table 7: The number of attributes within the reducts obtained by the selected methods.

	Raw	CQ	CE	MDE	GMDE	PMDE	RDE	PMDE-TWD
arrhythmia	279	15	9	12	9	9	14	8
autos	25	10	11	11	11	10	10	9
biodegrade	41	18	18	18	19	18	19	15
cardio	21	20	20	20	20	19	18	16
company	64	23	22	22	23	21	24	14
credit	23	23	23	23	23	23	20	20
hepatitis	19	8	8	8	8	8	8	8
hypothyroid	29	23	23	23	23	23	17	17
krvskp	36	30	30	30	29	30	31	24
mfeat	76	13	13	13	12	13	13	12
sick	29	20	20	20	20	20	16	12
sonar	60	7	7	7	7	7	7	6
splice	60	10	11	10	10	10	10	10
turkiye	31	31	31	31	31	31	17	25
vote	16	12	12	12	12	12	12	10
waveform	40	13	15	15	13	15	13	13
Avg.	53.06	17.25	17.06	17.19	16.88	16.81	15.56	**13.69**
Rank	7.31	2.75	3.13	2.94	2.69	2.44	2.81	**1.06**

In Table 7, the number of attributes reserved by each attribute reduction method is recorded. The attribute information in the raw data is also listed for comparison. The average results over all datasets and the mean rank among the selected methods are reported in the last two rows "Avg." and "Rank".

Table 7 shows that the selected methods achieve the objective of attribute reduction through removing some attributes, but yield subtly different results. CQ is a measure that keeps the number of positive examples unchanged, namely the certain information in the data. Whereas CE aims to retain the distribution of each example under different decisions. That is, both certain and uncertain information are preserved. Intuitively, CE is a stricter measure than CQ. On most datasets, CQ obtained reducts with the same or fewer attributes as CE. In the datasets of "arrhythmia" and "company", the reducts of CQ had more attributes than those of CE. These results can be explained by the fact that CQ adopts a greedy strategy to find the reduct, whereas the heuristic algorithm of CQ may be misguided in selecting more attributes. MDE and GMDE are extended from CE, and their obtained reducts were equal to or smaller than that of CE on most datasets except "arrhythmia" and "biodegrade". RDE employs the idea of approximate attribute reduction so that their reducts have fewer attributes, but their quality is relatively low, which will be confirmed by the following performance evaluation. The proposed PMDE attaches different and appropriate importance to the certain and uncertain information in the process of attribute reduction, whereas PMDE-TWD removes the attributes that fit the noisy examples, which usually limits the generalization performance. As a result, the two proposed methods achieve very promising attribute reduction results. On all datasets, the attribute reduction rates of CQ, CE, MDE, GMDE, and RDE were 67.49%, 67.84%, 67.61%, 68.20%, and 70.67%, respectively. PMDE gained an attribute reduction rate of 68.32%, whereas PMDE-TWD achieved an

attribute reduction rate of 74.20%, which is improved over the entropy-based CE and RDE by almost 9.37% and 5.00%. In terms of mean rank, PMDE-TWD and PMDE also gained the best results among the selected methods. These results demonstrate the effectiveness of the proposed methods for attribute reduction.

To further verify the performance of the obtained reducts, the selected methods were compared in terms of the classification error rate using the KNN, SVM, and DT classifiers. The overall results are presented in Tables 8-10, where the mean rank information of the selected methods on all datasets is also listed.

Table 8: The performance of the reducts obtained by the selected methods (KNN).

	Raw	CQ	CE	MDE	GMDE	PMDE	RDE	PMDE-TWD
arrhythmia	**40.98±0.52**	48.07±0.81	49.42±1.07	47.56±0.77	49.76±0.76	48.52±0.62	46.71±1.20	49.10±0.88
autos	37.58±1.59	39.33±0.85	36.47±1.15	32.86±1.97	32.86±1.97	30.36±1.08	30.40±1.10	**28.38±1.06**
biodegrade	**15.14±0.70**	17.12±0.54	17.69±0.45	17.80±0.80	16.82±0.56	17.69±0.45	16.46±0.56	17.81±0.63
cardio	28.29±0.43	28.32±0.34	28.17±0.33	28.38±0.38	28.38±0.38	27.79±0.34	**27.60±0.36**	28.04±0.34
company	6.76±0.09	6.02±0.10	**5.95±0.08**	6.05±0.08	6.35±0.11	5.99±0.08	6.14±0.12	6.11±0.11
credit	22.67±0.17	22.75±0.15	22.49±0.09	22.59±0.12	22.59±0.12	**22.49±0.09**	22.63±0.14	22.65±0.13
hepatitis	17.55±1.55	18.53±1.44	20.94±1.59	20.07±1.10	20.85±1.95	16.35±1.10	18.51±1.71	**16.35±1.10**
hypothyroid	9.05±0.25	9.14±0.18	9.01±0.15	**8.81±0.19**	8.81±0.19	9.01±0.15	8.93±0.22	9.14±0.21
krvskp	4.27±0.35	3.67±0.39	3.57±0.28	4.14±0.27	4.03±0.29	3.57±0.28	3.86±0.27	**2.34±0.20**
mfeat	**21.39±0.44**	30.39±0.32	30.58±0.37	28.90±0.50	39.76±0.34	27.48±0.69	28.75±0.47	27.06±0.52
sick	7.88±0.34	7.95±0.23	7.84±0.24	7.68±0.22	7.68±0.22	7.84±0.24	7.64±0.30	**7.70±0.21**
sonar	**12.66±1.30**	28.91±1.10	21.74±0.86	21.41±1.51	24.40±1.96	21.74±0.86	24.78±1.06	19.72±1.50
splice	27.59±0.40	18.81±0.21	15.25±0.27	**12.11±0.22**	42.00±0.43	12.26±0.25	25.21±0.21	12.37±0.28
turkiye	47.81±1.63	48.30±0.35	48.22±0.59	47.75±1.99	47.75±1.99	47.51±1.00	**47.22±1.67**	48.16±0.87
vote	7.73±0.43	7.79±0.69	5.93±0.47	5.97±0.37	5.93±0.38	5.93±0.47	7.42±0.33	**4.97±0.36**
waveform	28.99±0.19	36.31±0.47	29.25±0.38	29.19±0.37	41.58±0.46	28.33±0.45	41.82±0.42	**27.03±0.34**
Avg.	21.02±0.65	23.21±0.51	22.03±0.52	21.33±0.68	24.97±0.76	20.80±0.51	22.76±0.63	**20.43±0.55**
Rank	4.81	6.13	4.56	4.06	5.13	**2.81**	4.00	3.63

Table 9: The performance of the reducts obtained by the selected methods (SVM).

	Raw	CQ	CE	MDE	GMDE	PMDE	RDE	PMDE-TWD
arrhythmia	**40.42±0.42**	41.66±0.40	46.22±0.23	43.95±0.45	45.03±0.39	43.16±0.47	42.37±0.40	43.79±0.37
autos	41.46±1.04	40.86±1.04	51.26±1.02	51.86±1.22	51.93±1.35	49.80±0.98	**34.56±1.23**	49.72±0.99
biodegrade	**13.30±0.35**	16.65±0.34	15.84±0.24	15.84±0.24	15.16±0.24	15.84±0.24	15.94±0.22	15.58±0.31
cardio	22.20±0.20	28.02±0.31	**22.14±0.21**	22.14±0.21	22.14±0.21	22.34±0.29	28.15±0.24	22.50±0.45
company	5.88±0.03	6.22±0.02	5.37±0.06	5.37±0.06	5.38±0.06	**5.27±0.04**	6.24±0.01	5.32±0.05
credit	18.57±0.02	20.16±0.02	18.57±0.02	18.57±0.02	18.57±0.02	18.57±0.02	20.21±0.01	**18.54±0.02**
hepatitis	16.01±0.84	18.35±1.15	17.37±0.94	15.77±1.00	17.12±1.09	15.00±1.17	18.35±1.15	**15.00±1.17**
hypothyroid	7.71±0.00	**7.09±0.05**	7.71±0.00	7.71±0.00	7.71±0.00	7.71±0.00	7.22±0.06	7.71±0.00
krvskp	2.53±0.06	3.81±0.05	2.40±0.08	2.40±0.08	3.39±0.09	2.40±0.08	3.79±0.07	**0.99±0.08**
mfeat	**19.22±0.32**	29.06±0.23	23.40±0.29	21.83±0.30	30.10±0.25	21.78±0.25	27.58±0.31	21.88±0.36
sick	6.12±0.00	6.12±0.00	6.12±0.00	6.12±0.00	6.12±0.00	6.12±0.00	6.12±0.01	**6.12±0.00**
sonar	**14.25±0.63**	27.90±1.01	17.89±1.29	17.89±1.29	17.86±0.68	17.89±1.29	27.50±0.72	17.50±0.75
splice	9.68±0.12	26.72±0.15	8.50±0.13	7.19±0.09	40.29±0.18	7.19±0.09	26.48±0.09	**7.17±0.13**
turkiye	37.78±0.04	38.17±0.06	37.78±0.04	37.78±0.04	37.78±0.04	**37.78±0.04**	38.18±0.05	37.79±0.07
vote	4.14±0.21	4.07±0.30	4.14±0.21	4.14±0.21	4.14±0.21	4.14±0.21	4.07±0.30	**3.47±0.26**
waveform	**17.13±0.11**	27.79±0.13	21.55±0.18	21.55±0.18	33.10±0.20	21.34±0.21	32.96±0.17	20.78±0.25
Avg.	**17.28±0.27**	21.42±0.33	19.14±0.31	18.76±0.34	22.24±0.31	18.52±0.34	21.23±0.32	18.37±0.33
Rank	2.69	5.44	3.56	3.06	4.50	2.86	5.56	**2.69**

Table 10: The performance of the reducts obtained by the selected methods (DT).

	Raw	CQ	CE	MDE	GMDE	PMDE	RDE	PMDE-TWD
arrhythmia	**47.66±0.94**	58.55±1.23	56.64±1.95	50.95±1.10	55.75±1.31	54.78±1.30	54.01±2.02	51.77±1.30
autos	24.10±1.77	23.73±1.51	23.38±1.68	24.90±1.24	24.46±1.43	22.51±1.38	21.98±1.53	**21.89±1.26**
biodegrade	18.41±0.74	**18.28±0.86**	19.25±0.50	19.36±0.87	19.40±0.48	19.72±0.49	18.80±0.76	19.06±0.63
cardio	27.66±0.34	27.55±0.53	27.56±0.63	27.41±0.38	**27.24±0.38**	27.36±0.45	27.58±0.56	28.03±0.53
company	7.62±0.28	7.64±0.19	7.82±0.21	7.68±0.23	7.97±0.23	7.71±0.24	7.53±0.18	**6.83±0.13**
credit	26.65±0.15	26.59±0.14	26.62±0.17	26.58±0.15	26.59±0.16	26.56±0.19	26.34±0.14	**25.78±0.10**
hepatitis	23.91±2.14	21.52±1.25	22.28±2.48	23.28±2.11	23.47±2.06	**20.89±2.19**	21.13±1.55	22.05±2.37
hypothyroid	9.46±0.26	9.47±0.25	9.51±0.23	9.47±0.28	9.42±0.23	**9.42±0.23**	9.62±0.17	10.05±0.17
krvskp	**0.44±0.14**	0.53±0.11	0.46±0.09	0.48±0.15	0.56±0.13	0.45±0.10	0.52±0.12	1.35±0.08
mfeat	**27.57±0.72**	31.87±0.71	32.00±0.57	29.50±0.46	40.83±0.32	29.83±0.88	31.05±0.49	31.73±0.71
sick	8.39±0.19	8.31±0.26	8.36±0.27	8.30±0.24	8.33±0.22	8.31±0.19	8.18±0.15	**7.67±0.16**
sonar	24.67±2.37	29.95±2.90	19.40±2.18	20.13±1.58	24.88±1.96	20.27±2.04	32.86±2.01	**18.93±1.85**
splice	8.80±0.33	17.51±0.43	8.57±0.38	7.56±0.26	38.66±0.51	7.64±0.28	18.02±0.33	**7.33±0.14**
turkiye	43.19±0.52	43.10±0.44	43.12±0.40	43.07±0.32	43.09±0.43	43.06±0.51	42.93±0.36	**42.81±0.15**
vote	6.07±0.43	5.77±0.65	5.99±0.68	5.72±0.44	5.88±0.48	5.84±0.55	5.74±0.53	**4.46±0.42**
waveform	**27.68±0.38**	38.22±0.33	30.10±0.31	30.13±0.37	42.89±0.44	29.92±0.21	43.11±0.40	28.36±0.45
Avg.	20.77±0.73	23.04±0.74	21.32±0.80	20.91±0.64	24.96±0.67	20.89±0.70	23.09±0.71	**20.51±0.65**
Rank	4.69	5.00	5.44	4.00	6.00	3.50	4.19	**3.13**

As shown in Tables 8-10, the quality of the reducts obtained by the selected methods is different. For most datasets, CQ achieved mediocre performance, especially on the datasets of "company", "splice", and "waveform". Compared

with CQ, the performance of CE was improved on most datasets, but failed on the datasets of "arrhythmia" and "autos". Both MDE and GMDE are evolved from CE, but GMDE yielded worse performance on the datasets of "mfeat", "splice", and "waveform". RDE combines the approximation information with the entropy. Nevertheless, the quality of the obtained reducts is undesired so that the performance is poor. The reducts generated by the proposed PMDE and PMDE-TWD have fewer attributes, but their performance was much better than that of the selected methods. When averaging the performance over all datasets, PMDE using the KNN, SVM, and DT classifiers gained an improvement on CE by 5.58%, 3.24%, and 2.00%, respectively. Interestingly, PMDE-TWD achieved a much better attribute reduction rate than CE, but its performance was improved over CE by 7.26%, 4.02%, and 3.80%, respectively, and improved over RDE by 10.24%, 13.47%, and 11.17%, respectively, when using the KNN, SVM, and DT classifiers. By viewing the rank information of all selected methods, the proposed methods PMDE-TWD and PMDE always obtained the best or comparable results. These results indicate the great potential of the proposed methods for attribute reduction, especially the PMDE with the three-way decision.

4.4. Discussion on the proposed methods

As the experimental analysis in Section 4.3, it can be found that the proposed methods PMDE and PMDE-TWD achieved better attribute reduction rates and classification performance in comparison with other representative methods. In particular, the PMDE-TWD obtained the reducts with much fewer attributes, but yielded the best or comparable classification performance. In general, based on the theoretical analysis and experimental results, the proposed methods have the following advantages over other methods.

(1) The proposed measure PMDE is monotonic, thus avoiding the problems of existing non-monotonic measures in the three-way decision and also facilitating the design of heuristic algorithms for attribute reduction.

(2) The proposed measure PMDE gives different levels of importance to the information of maximum decision and other non-maximum decisions, which makes the generated reduct more compact and informative. In addition, the parameter in the PMDE can be adjusted empirically, providing flexibility in dealing with practical tasks.

(3) The proposed PMDE-TWE method uses the idea of the three-way decision to remove attributes in the boundary region, which further improves the generalization ability of the resulting reduct and also provides a certain degree of noise tolerance.

5. Conclusions

In this study, to address the non-monotonicity of attribute reduction measures and the deficiency in attribute reduction rate, a measure of parameterized maximum distribution entropy is proposed to reconcile the importance of the certainty and uncertainty information of the maximum decision. The properties of the proposed measure are thoroughly investigated, and its monotonicity is theoretically proven. In addition, the idea of trisection is introduced to conduct the process of attribute reduction, and a heuristic algorithm is developed to yield the optimal approximate reduct by appropriately relaxing the proposed measure, thus resulting in a high attribute reduction rate and good generalization performance. Experiments conducted on several UCI datasets demonstrate that the proposed method is superior to other state-of-the-art methods in terms of

attribute reduction rate and classification performance. It should be noted that the proposed method is designed for data with discrete attributes. Therefore, it is worthwhile to improve and extend the proposed method to handle data with both discrete and continuous attributes.

6. Acknowledgement

The authors would like to thank the Editor-in-Chief, editors, and anonymous reviewers for their kind help and valuable comments. This work is supported in part by the National Natural Science Foundation of China (Nos. 61806127, 62076164), the Natural Science Foundation of Guangdong Province, China (No. 2021A1515011861), Shenzhen Science and Technology Program (No. JCYJ2021-0324094601005), and Shenzhen Institute of Artificial Intelligence and Robotics for Society.

7. Appendix

Proof of Proposition 2.

$$\begin{aligned}
&\Delta MH_\lambda = MH_\lambda(D|P) - MH_\lambda(D|Q) \\
&= -\frac{1}{|U|}\sum_{x_{ij}\in X_{ij}}(\lambda\mu_P^{max}(x_{ij})\log\mu_P^{max}(x_{ij}) + (1-\lambda)(1-\mu_P^{max}(x_{ij}))\log(\frac{1-\mu_P^{max}(x_{ij})}{k_{ij}-1})) \\
&+ \frac{1}{|U|}\sum_{x_i\in X_i}(\lambda\mu_Q^{max}(x_i)\log\mu_Q^{max}(x_i) + (1-\lambda)(1-\mu_Q^{max}(x_i))\log(\frac{1-\mu_Q^{max}(x_i)}{k_i-1})) \\
&+ \frac{1}{|U|}\sum_{x_j\in X_j}(\lambda\mu_Q^{max}(x_j)\log\mu_Q^{max}(x_j) + (1-\lambda)(1-\mu_Q^{max}(x_j))\log(\frac{1-\mu_Q^{max}(x_j)}{k_j-1}))
\end{aligned}$$

Let $\mu_Q^{max}(x_i) = \mu_i$, $\mu_Q^{max}(x_j) = \mu_j$, and $\mu_P^{max}(x_{ij}) = \mu_{ij}$. We have $\Delta MH_\lambda = \Delta MH_\lambda^1 + \Delta MH_\lambda^2$, where

$$\begin{aligned}
\Delta MH_\lambda^1 &= \frac{\lambda}{|U|}(|X_i|\mu_i\log\mu_i + |X_j|\mu_j\log\mu_j - |X_{ij}|\mu_{ij}\log\mu_{ij}) \\
\Delta MH_\lambda^2 &= \frac{(1-\lambda)}{|U|}(|X_i|(1-\mu_i)\log\frac{1-\mu_i}{k_i-1} + |X_j|(1-\mu_j)\log\frac{1-\mu_j}{k_j-1} \\
&\quad - |X_{ij}|(1-\mu_{ij})\log\frac{1-\mu_{ij}}{k_{ij}-1})
\end{aligned}$$

Assume the maximum decision of X_i, X_j, and X_{ij} are denoted as $\nu_D^{max}(x_{ij})$, $\nu_D^{max}(x_i)$, and $\nu_D^{max}(x_j)$, respectively. The maximum probability of X_{ij} can be expressed as $\mu_{ij} = (|X_i|\mu_i + |X_j|\mu_j)/(|X_i| + |X_j|)$ when $\nu_D^{max}(x_i) = \nu_D^{max}(x_j) = \nu_D^{max}(x_{ij})$. Since the examples in X_{ij} are all from X_i and X_j, the maximum probability of X_{ij} can be expressed as the inequality $\mu_{ij} \leq (|X_i|\mu_i + |X_j|\mu_j)/(|X_i| + |X_j|)$ if $\nu_D^{max}(x_i) \neq \nu_D^{max}(x_{ij})$ or $\nu_D^{max}(x_j) \neq \nu_D^{max}(x_{ij})$. Assume there are m classes in X_{ij}, then the inequality $1/m \leq \mu_{ij} \leq (|X_i|\mu_i + |X_j|\mu_j)/(|X_i| + |X_j|)$ holds. From the perspective of information entropy, the entropy is monotonically decreasing when the probability is larger than $1/m$. Therefore, the entropy of X_{ij} is minimized when the maximum decisions of X_i, X_j, and X_{ij} are the same. For the first part, we have

$$\Delta MH^1_\lambda = \frac{\lambda}{|U|}(|X_i|\mu_i \log \mu_i + |X_j|\mu_j \log \mu_j - (|X_i|\mu_i + |X_j|\mu_j) \log \frac{|X_i|\mu_i + |X_j|\mu_j}{|X_i| + |X_j|})$$
$$= \frac{\lambda}{|U|}(|X_i|\mu_i(\log \mu_i - \log \frac{|X_i|\mu_i + |X_j|\mu_j}{|X_i| + |X_j|}) + |X_j|\mu_j(\log \mu_j - \log \frac{|X_i|\mu_i + |X_j|\mu_j}{|X_i| + |X_j|}))$$
$$= \frac{\lambda}{|U|}(|X_i|\mu_i \log \frac{|X_i|\mu_i + |X_j|\mu_i}{|X_i|\mu_i + |X_j|\mu_j} + |X_j|\mu_j \log \frac{|X_i|\mu_j + |X_j|\mu_j}{|X_i|\mu_i + |X_j|\mu_j})$$

Let $|X_i|\mu_i = \zeta$, $|X_j|\mu_j = \eta$, and $\theta = \mu_i/\mu_j$. We have

$$\Delta MH^1_\lambda(\zeta, \eta, \theta) = \frac{\lambda}{|U|}(\zeta \log \frac{\zeta + \eta\theta}{\zeta + \eta} + \eta \log \frac{\zeta\frac{1}{\theta} + \eta}{\zeta + \eta})$$

Taking the partial derivative of $\Delta MH^1_\lambda(\zeta, \eta, \theta)$ with respect to the variable θ, we have

$$\frac{\partial \Delta MH^1_\lambda(\zeta, \eta, \theta)}{\partial \theta} = \frac{\lambda}{|U|}(\zeta \frac{\zeta + \eta}{\zeta + \eta\theta} \eta \log\ e - \eta \frac{\zeta + \eta}{\zeta\frac{1}{\theta} + \eta} \zeta \frac{1}{\theta^2} \log\ e)$$
$$= \frac{\lambda\zeta\eta(\zeta + \eta) \log\ e}{|U|}(\frac{1}{\zeta + \eta\theta} - \frac{1}{\theta(\zeta + \eta\theta)})$$
$$= \frac{\lambda\zeta\eta(\zeta + \eta) \log\ e}{|U|} \left(\frac{\theta - 1}{\theta(\zeta + \eta\theta)}\right) \begin{cases} < 0, & 0 < \theta < 1 \\ = 0, & \theta = 1 \\ > 0, & \theta > 1 \end{cases}.$$

589 $\Delta MH^1_\lambda(\zeta, \eta, \theta)$ is a concave function with respect to θ and is minimized to 0
590 when $\theta = 1$, namely $\mu_i = \mu_j$. Thus, in all possible cases, $\Delta MH^1_\lambda \geq 0$.

For ΔMH^2_λ, we have

$$\Delta MH^2_\lambda = \frac{(1 - \lambda)}{|U|}(|X_i|(1 - \mu_i) \log \frac{1 - \mu_i}{k_i - 1} + |X_j|(1 - \mu_j) \log \frac{1 - \mu_j}{k_j - 1}$$
$$- (|X_i|(1 - \mu_i) + |X_j|(1 - \mu_j)) \log \frac{|X_i| + |X_j| - |X_i|\mu_i - |X_j|\mu_j}{(|X_i| + |X_j|)(k_{ij} - 1)})$$
$$= \frac{(1 - \lambda)}{|U|}(|X_i|(1 - \mu_i) \log \frac{1 - \mu_i}{k_i - 1} + |X_j|(1 - \mu_j) \log \frac{1 - \mu_j}{k_j - 1}$$
$$- (|X_i|(1 - \mu_i) + |X_j|(1 - \mu_j)) \log \frac{|X_i|(1 - \mu_i) + |X_j|(1 - \mu_j)}{(|X_i| + |X_j|)(k_{ij} - 1)})$$
$$= \frac{(1 - \lambda)}{|U|}(|X_i|(1 - \mu_i) \log \frac{(|X_i|(1 - \mu_i) + |X_j|(1 - \mu_i))(k_{ij} - 1)}{(|X_i|(1 - \mu_i) + |X_j|(1 - \mu_j))(k_i - 1)}$$
$$+ |X_j|(1 - \mu_j) \log \frac{(|X_i|(1 - \mu_j) + |X_j|(1 - \mu_j))(k_{ij} - 1)}{(|X_i|(1 - \mu_i) + |X_j|(1 - \mu_j))(k_j - 1)})$$

Let $|X_i|(1 - \mu_i) = \zeta$, $|X_j|(1 - \mu_j) = \eta$, and $\theta = (1 - \mu_i)/(1 - \mu_j)$. We have

$$\Delta MH^2_\lambda = \frac{(1 - \lambda)}{|U|}(\zeta \log \frac{(\zeta + \eta\theta)(k_{ij} - 1)}{(\zeta + \eta)(k_i - 1)} + \eta \log \frac{(\zeta\frac{1}{\theta} + \eta)(k_{ij} - 1)}{(\zeta + \eta)(k_j - 1)})$$

In the formula, k_{ij}, k_i, and k_j denote the number of decisions within the condition classes X_{ij}, X_i, and X_j, respectively. Since X_i and X_j are divided from X_{ij}, $(k_{ij} - 1) \geq (k_i - 1)$ and $(k_{ij} - 1) \geq (k_j - 1)$ hold, and we have

$$\Delta MH^2_\lambda(\zeta, \eta, \theta) \geq \frac{(1 - \lambda)}{|U|}(\zeta \log \frac{(\zeta + \eta\theta)}{(\zeta + \eta)} + \eta \log \frac{(\zeta\frac{1}{\theta} + \eta)}{(\zeta + \eta)})$$

The right-side of the above inequality is almost the same as $\Delta MH^1_\lambda(\zeta,\eta,\theta)$, and $\Delta MH^2_\lambda(\zeta,\eta,\theta)$ is minimized to 0 when $\theta = 1$, namely $(1-\mu_i) = (1-\mu_j)$.

In all possible cases, $\Delta MH_\lambda(\zeta,\eta,\theta) = \Delta MH^1_\lambda(\zeta,\eta,\theta) + \Delta MH^2_\lambda(\zeta,\eta,\theta) \geq 0$. The Proposition is proven.

References

[1] C. Bishop, Pattern recognition and machine learning, Springer, New York, NY, USA, 2007.

[2] K. Thangavel, A. Pethalakshmi, Dimensionality reduction based on rough set theory: A review, Appl. Soft Comput. 9 (2009) 1-12.

[3] Y.H. Qian, J.Y. Liang, W. Pedrycz, C.Y. Dang, Positive approximation: An accelerator for attribute reduction in rough set theory, Artif. Intell. 174 (2010) 597-618.

[4] G.Y. Wang, X.A Ma, H. Yu, Monotonic uncertainty measures for attribute reduction in probabilistic rough set model, Int. J. Approx. Reason. 59 (2015) 41-67.

[5] C.Z. Wang, Y. Huang, M.W. Shao, Q.H. Hu, D.G. Chen, Feature selection based on neighborhood self-information, IEEE Trans. Cybern. 50 (9) (2020) 4031-4042.

[6] L. Sun, L.Y. Wang, W.P. Ding, Y.H. Qian, J.C. Xu, Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy neighborhood multigranulation rough sets, IEEE Trans. Fuzzy Syst. 29 (1) (2021) 19-33.

[7] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982) 341-356.

[8] Z. Pawlak, Rough sets: Theoretical aspects of reasoning about data, Kluwer Academic Publishers, Dordrecht, Netherlands, 1991.

[9] J.T. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: Perspectives and challenges, IEEE Trans. Cybern. 43 (6) (2013) 1977-1989.

[10] Y.Y. Yao, S.K.M. Wong, A decision theoretic framework for approximating concepts, Int. J. Man-Mach. Stud. 37 (1992) 793-809.

[11] J.P. Herbert, J.T. Yao, Game-theoretic rough sets, Fundam. Informaticae, 108 (3-4) (2011) 267-286.

[12] N. Azam, J.T. Yao, Analyzing uncertainties of probabilistic rough set regions with game-theoretic rough sets, Int. J. Approx. Reason. 55 (1) (2014) 142-155.

[13] Y. Zhang, J.T. Yao, Game theoretic approach to shadowed sets: A three-way tradeoff perspective, Inf. Sci. 507 (2020) 540-552.

[14] C. Cornelis, J. Medina, N. Verbiest, Multi-adjoint fuzzy rough sets: Definition, properties and attribute selection, Int. J. Approx. Reason. 55 (1) (2014) 412-426.

[15] O.U. Lenz, D. Peralta, C. Cornelis, Scalable approximate FRNN-OWA classification, IEEE Trans. Fuzzy Syst. 28 (5) (2020) 929-938.

[16] Q.H. Hu, L. Zhang, D. Zhang, W. Pan, S. An, W. Pedrycz, Measuring relevance between discrete and continuous features based on neighborhood mutual information, Expert Syst. Appl. 38 (2011) 10737-10750.

[17] Q.H. Zhang, M. Gao, F. Zhao, G.Y. Wang, Fuzzy-entropy-based game theoretic shadowed sets: A novel game perspective from uncertainty, IEEE Trans. Fuzzy Syst. 30 (3) (2022) 597-609.

[18] Y.Y. Yao, Three-way decisions with probabilistic rough sets, Inf. Sci. 180 (2010) 341-353.

[19] Y.Y. Yao, Three-way decision and granular computing, Int. J. Approx. Reason. 103 (2018) 107-123.

[20] Y.Y. Yao, Tri-level thinking: Models of three-way decision, Int. J. Mach. Learn. Cybern. 11 (2020) 947-959.

[21] Y.Y. Yao, The geometry of three-way decision. Appl. Intell. 51 (9) (2021) 6298-6325.

[22] F. Min, Z.H. Zhang, W.J. Zhai, R.P. Shen, Frequent pattern discovery with tri-partition alphabets, Inf. Sci. 507 (2020) 715-732.

[23] C. Gao, J. Zhou, D. Miao, J.J. Wen, X.D. Yue, Three-way decision with co-training for partially labeled data, Inf. Sci. 544 (2021) 500-518.

[24] D.C. Liang, W. Cao, Z.S. Xu, M.W. Wang, A novel approach of two-stage three-way co-opetition decision for crowdsourcing task allocation scheme, Inf. Sci. 559 (2021) 191-211.

[25] J.L. Yang, Y.Y. Yao, A three-way decision based construction of shadowed sets from Atanassov intuitionistic fuzzy sets, Inf. Sci. 577 (2021) 1-21.

[26] Y.Y. Yao, Symbols-Meaning-Value (SMV) space as a basis for a conceptual model of data science, Int. J. Approx. Reason. 144 (2022) 113-128.

[27] Y.Y Yao, J.L. Yang, Granular rough sets and granular shadowed sets: Three-way approximations in Pawlak approximation spaces, Int. J. Approx. Reason. 142 (2022) 231-247.

[28] Y.Y. Yao, The superiority of three-way decisions in probabilistic rough set models, Inf. Sci. 181 (2011) 1080-1096.

[29] H. Yu, Y. Chen, P. Lingras, G.Y. Wang, A three-way cluster ensemble approach for large-scale data, Int. J. Approx. Reason. 115 (2019) 32-49.
[30] F. Min, S.M. Zhang, D. Ciucci, M. Wang, Three-way active learning through clustering selection, Int. J. Mach. Learn. Cybern. 11 (5) (2020) 1033-1046.
[31] Y.Y. Yao, Three-way decisions and cognitive computing, Cogn. Comput. 8 (2016) 543-554.
[32] Y.Y. Yao, Three-way conflict analysis: Reformulations and extensions of the Pawlak model, Knowl.-Based Syst. 180 (2019) 26-37.
[33] Y.Y. Yao, Three-way granular computing, rough sets, and formal concept analysis, Int. J. Approx. Reason. 116 (2020) 106-125.
[34] Y.Y. Yao, Y. Zhao, Attribute reduction in decision-theoretic rough set models, Inf. Sci. 178 (2008) 3356-3373.
[35] Y. Zhao, S.K.M. Wong, Y.Y. Yao, A note on attribute reduction in the decision-theoretic rough set model, in: Lecture Notes in Computer Science, 2015, pp. 260-275.
[36] H.X. Li, X.Z. Zhou, J.B. Zhao, D. Liu, Non-monotonic attribute reduction in decision-theoretic rough sets, Fundam. Inform. 126 (2013) 415-432.
[37] X.A. Ma, G.Y. Wang, H. Yu, T.R. Li, Decision region distribution preservation reduction in decision-theoretic rough set model, Inf. Sci. 278 (2014) 614-640.
[38] X.Y. Zhang, D.Q. Miao, Three-way attribute reducts, Int. J. Approx. Reason. 88 (2017) 401-434.
[39] C. Gao, Z.H. Lai, J. Zhou, C.R. Zhao, D.Q. Miao, Maximum decision entropy-based attribute reduction in decision-theoretic rough set model, Knowl. Based Syst. 143 (2018) 179-191.
[40] C. Gao, Z.H Lai, J. Zhou, J.J. Wen, W.K. Wong, Granular maximum decision entropy-based monotonic uncertainty measure for attribute reduction, Int. J. Approx. Reason. 104 (2019) 9-24.
[41] X.Y. Jia, W.H. Liao, Z.M. Tang, L. Shang, Minimum cost attribute reduction in decision-theoretic rough set models, Inf. Sci. 219 (2013) 151-167.
[42] S.J. Liao, Q.X. Zhu, F. Min, Cost-sensitive attribute reduction in decision-theoretic rough set models, Math. Prob. Eng. 35 (2014) 1-9.
[43] Y. Fang, F. Min, Cost-sensitive approximate attribute reduction with three-way decisions, Int. J. Approx. Reason. 104 (2019) 148-165.
[44] W.W. Li, X.Y. Jia, L. Wang, B. Zhou, Multi-objective attribute reduction in three-way decision-theoretic rough set model, Int. J. Approx. Reason. 105 (2019) 327-341.
[45] X.A. Ma, Y.Y. Yao, Three-way decision perspectives on class-specific attribute reducts. Inf. Sci. 450 (2018) 227-245.
[46] X.Y. Zhang, X. Tang, J.L. Yang, Z.Y. Lv, Quantitative three-way class-specific attribute reducts based on region preservations, Int. J. Approx. Reason. 117 (2020) 96-121.
[47] W.H. Xu, Y.T. Guo, Generalized multigranulation double-quantitative decision-theoretic rough set, Knowl. Based Syst. 105 (2016) 190-205.
[48] C. Zhang, J.J. Ding, D.Y. Li, J.M. Zhan, A novel multi-granularity three-way decision making approach in q-rung orthopair fuzzy information systems, Int. J. Approx. Reason. 138 (2021) 161-187.
[49] W.W. Li, Z.Q. Huang, X.Y. Jia, X.Y. Cai, Neighborhood based decision-theoretic rough set models, Int. J. Approx. Reason. 69 (2016) 1-17.
[50] X.D. Yue, J. Zhou, Y.Y. Yao, D.Q. Miao, Shadowed neighborhoods based on fuzzy rough transformation for three-way classification, IEEE Trans. Fuzzy Syst. 28 (5) (2020) 978-991.
[51] X.Y. Zhang, H.Y.Gou, Z.Y. Lv, D.Q. Miao, Double-quantitative distance measurement and classification learning based on the tri-level granular structure of neighborhood system, Knowl.-Based Syst. 217 (2021) 106799.
[52] D. Liu, D.C. Liang, C.C. Wang, A novel three-way decision model based on in-complete information system, Knowl. Based Syst. 91 (2016) 32-45.
[53] M. Gao, Q.H. Zhang, F. Zhao, G.Y. Wang, Mean-entropy-based shadowed sets: A novel three-way approximation of fuzzy sets, Int. J. Approx. Reason. 120 (2020) 102-124.
[54] G.M. Lang, Y.Y. Yao, New measures of alliance and conflict for three-way conflict analysis, Int. J. Approx. Reason. 132 (2021) 49-69.
[55] J. Qian, D.W. Tang, Y. Yu, X.B. Yang, S. Gao, Hierarchical sequential three-way decision model, Int. J. Approx. Reason. 140 (2022) 156-172.
[56] Z. Pawlak, S.K.M. Wong, W. Ziarko, Rough sets: Probabilistic versus deterministic approach, Int. J. Man-Mach. Stud. 29 (1988) 81-95.
[57] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825-2830.
[58] F. Jiang, X. Yu, J. Du, D. Gong, Y. Zhang, Y. Peng, Ensemble learning based on approximate reducts and bootstrap sampling, Inf. Sci. 547 (2021) 797-813.