Knowledge-Based Systems 143 (2018) 179–191

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Maximum decision entropy-based attribute reduction in decision-theoretic rough set model

Can Gao[a,b,*], Zhihui Lai[a,b], Jie Zhou[a,b], Cairong Zhao[c,d], Duoqian Miao[c,d]

[a] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, PR China
[b] Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Kowloon, Hong Kong
[c] Department of Computer Science and Technology, Tongji University, Shanghai, 201804, PR China
[d] The Key Laboratory of "Embedded System and Service Computing", Ministry of Education, Shanghai, 201804, PR China

ARTICLE INFO

Article history:
Received 3 June 2017
Revised 9 December 2017
Accepted 11 December 2017
Available online 12 December 2017

Keywords:
Decision-theoretic rough set model
Attribute reduction
Maximum decision entropy
Decision monotonicity

ABSTRACT

Decision-theoretic rough set model, as a probabilistic generalization of the Pawlak rough set model, is an effective method for decision making from vague, uncertain or imprecise data. Attribute reduction is one of the most important problems in the decision-theoretic rough set model and several uncertainty measures for attribute reduction have been presented. However, the monotonicity of the uncertainty measures does not always hold. In this paper, a novel monotonic uncertainty measure is introduced for attribute reduction in the decision-theoretic rough set model. More specifically, based on the concepts of the maximum inclusion degree and maximum decision, a new uncertainty measure, named maximum decision entropy, is first proposed, and the definitions of the positive, boundary and negative region preservation reducts are then provided by using the proposed uncertainty measure. Theoretically, it is proved that the proposed uncertainty measure is monotonic when adding or deleting the condition attributes. Additionally, a heuristic attribute reduction algorithm based on the maximum decision entropy is developed, which maximizes the relevance of the reduct to the class attribute and also minimizes the redundancy of the condition attributes within the reduct. The experimental results on artificial as well as real data sets demonstrate the competitive performance of our proposal in comparison with the state-of-the-art algorithms.

1. Introduction

Since the initial work of Pawlak [1,2], rough set theory has witnessed the rapid development of theoretical research and also been extensively utilized in the fields of machine learning, pattern recognition and artificial intelligence [3]. In the Pawlak rough set model, the lower approximation is defined by all elementary granules each of which is fully contained by a concept. However, this requirement is too strict for some applications in the real world, especially the one with noisy data. By incorporating the probability theory into the Pawlak rough set model, several extended and generalized rough set models have been proposed [4]. Wong and Ziarko [5] introduced the notion of probabilistic approximation, and the lower and upper approximations are well formulated in the form of conditional probability. Pawlak et al. [6] proposed the 0.5 probabilistic rough set model, in which the lower approximation is defined by the elementary granules whose conditional probability is greater than 0.5 and the upper approximation by the elementary granules whose conditional probability is equal or greater than 0.5. Ziarko [7] presented the variable precision rough set model by introducing certain levels of errors into the lower approximation. Motivated by Bayesian risk decision procedure for classification, Yao [8] developed the decision-theoretic rough set model, and the threshold parameters in the probabilistic approximations are thus replaced by the real cost functions. Slezak and Ziarko [9] put forward the Bayesian rough set model by using the prior probability rather than the threshold parameters or cost functions to define the probabilistic approximations. Later, Yao [10] provided a general framework for the Pawlak rough set model and probabilistic rough set models. In addition, other probabilistic rough set models [10,11] have also attracted much attention and have been studied extensively.

The decision-theoretic rough set model [12,13] (referred to as the "DTRS model" hereinafter), as a probabilistic extension of the Pawlak rough set model, intrinsically simulates the decision

* Corresponding author.
E-mail addresses: david.gao@polyu.edu.hk, 2005gaocan@163.com (C. Gao), laizhihui@szu.edu.cn (Z. Lai), jjpuzhou@polyu.edu.hk (J. Zhou), zhaocairong@tongji.edu.cn (C. Zhao), dqmiao@tongji.edu.cn (D. Miao).

https://doi.org/10.1016/j.knosys.2017.12.014
0950-7051/

procedure of human beings under uncertainty and risk. Decision making within the DTRS model is not only based on the degree of confidence for making a decision but also the cost caused by the decision behavior. As a result, the commonly used binary decisions with the mutually exclusive options "yes" or "no" evolve into three-way decisions with three alternatives [14,15], namely decision with acceptance, rejection and noncommitment. Compared with other probabilistic rough set models, the DTRS model exhibits the salient characteristic and superiority in probabilistic reasoning and semantic interpretation [16]. Moreover, the DTRS model provides a unified and comprehensive framework for rough set theory, and several rough set models, such as the Pawlak rough set model, 0.5 probabilistic rough set model and variable precision rough set model, can be derived directly when the parameters calculated from the cost functions are set properly [17].

Attribute reduction [18–24] is one of the most important applications of rough set theory. A reduct is a jointly sufficient and individually necessary subset of condition attributes that has the same level of performance as the entire set of condition attributes. In the Pawlak rough set model, the universe is always divided into two mutually complementary sets, namely the positive and boundary regions. Therefore, the uncertainty measures that reflect only one of the two regions, such as the degree of dependence and quality of classification [2], are quite enough for attribute reduction. However, in the DTRS model, owing to the introduction of the threshold parameters in approximating the concept, all objects are grouped into the positive, boundary and negative regions. The uncertainty measures mentioned above are no longer a good choice for attribute reduction. Actually, the main reason for this problem is that the monotonicity does not hold when these uncertainty measures are applied to the DTRS model directly [25].

To tackle the aforementioned problem, several uncertainty measures with different objectives have been suggested. Roughly speaking, their objectives for attribution reduction can be classified into criterion preservation and criterion optimization [26]. In the former case, attribute reduction in the DTRS model can be treated as a problem of uncertainty measure preservation or improvement. Li et al. [27] argued that the positive region should be the same or even larger after attribute reduction, and an uncertainty measure was developed for the positive region extension reduct. Ma et al. [28,29] defined the uncertainty measures for the probabilistic positive, boundary, negative and non-negative region distribution preservation reducts, and the corresponding heuristic algorithms were also presented by using conditional information or information entropy. The heuristic algorithms with criterion preservation could quickly yield a reduct, while their reducts may still contain the redundant attributes. In the latter case, attribute reduction in the DTRS model can be intercepted as a problem of uncertainty measure optimization. Yao and Zhao [30] investigated the positive region, non-negative region, confidence of rules, coverage of rules, cost of rules and others as the uncertainty measures for attribute reduction, and a general definition for the optimal reduct was proposed. Zhao et al. [26] introduced the positive decision, positive region extension and non-negative region-based uncertainty measures for the optimal reduct and also developed a discernibility matrix-based approach for constructing the optimal reduct. Zhang and Miao [31,32] discussed four types of uncertainty measures for attribute reduction, namely knowledge, consistency, region and structure targets, and a general reduct and three kinds of the optimal reducts were consequently put forward. From the viewpoint of decision cost minimization, Jia et al. [33,34] provided a minimum cost attribute reduction algorithm using the technique of genetic, simulated annealing or particle swarm. Yu et al. [35] and Bi et al. [36] incorporated the significance of single attribute or joint attributes with cost minimization for attribute reduction, and the optimization algorithms with multi-objectives were thus proposed. Liao et al. [37] examined the problem of attribute reduction with decision cost and test cost and presented an attribute reduction algorithm to minimize these two kinds of costs. Although the attribute reduction algorithms with criterion optimization could generate an optimal reduct, their time complexity may be rather high. Additionally, under the DTRS model, some researchers also studied the problem of attribute reduction in quick method [38], multi-costs strategy [39], multigranulation [40–42], incomplete system [43,44], neighborhood system [45–47], etc.

However, the existing algorithms for attribute reduction still have some problems that need to be investigated further. On the one hand, some algorithms could obtain an optimal reduct with their optimization objectives, but for the inconsistent condition equivalence class or even the consistent one, its majority decision which has the maximum inclusion degree (called the maximum decision hereinafter) may be changed after the procedure of attribute reduction. The objective of attribute reduction is to remove the redundant attributes. The behavior of changing the decision essentially violates the principle of attribute reduction. On the other hand, some algorithms could yield an optimal reduct with criterion preservation, but their reducts, in a sense, still have the redundant attributes which could be further removed. A full explanation of these problems is given in Section 3. Usually, human beings mainly concern the decision that has the maximum inclusion degree when facing the problem with multiple uncertain (or indeterministic) alternatives. For example, the decision with conditional probability above 0.5 is usually taken by a decision maker for a binary decisions problem. Therefore, the uncertainty measure for attribute reduction should pay more attention to the information about the decision with the maximum inclusion degree. Actually, the reduct with preservation of the maximum inclusion degree and maximum decision is not only more consistent with human beings' decision procedure but also contains less redundant information.

To achieve the objective above, this paper proposes a monotonic uncertainty measure in which the maximum inclusion degree, maximum decision and cost functions are all taken into consideration. The main contributions of this paper are threefold. First, we examine the existing uncertainty measures for attribute reduction in the DTRS model and illustrate their potential problems by an artificial data set. Second, we introduce the concept of the maximum decision entropy, based on which a novel monotonic uncertainty measure is designed. Third, we develop a heuristic algorithm with the principle of maximum relevance and minimum redundancy, which could provide more concise, stable and accurate results.

The remainder of this paper is organized as follows. Section 2 outlines some concepts related to the Pawlak rough set model and DTRS model. Section 3 first indicates the problems of the existing uncertainty measures for attribute reduction using a toy data set. A novel uncertainty measure is then introduced, and its monotonicity is also proved. Finally, a heuristic algorithm is developed for attribute reduction in the DTRS model. Section 4 shows the results of applying the proposed algorithm to several UCI data sets. Section 5 concludes the paper and indicates the intended directions for further research.

2. Preliminary knowledge

This section will review some concepts in the Pawlak rough set model and DTRS model. A detailed description of the models can be found in [2,8,30].

Formally, an information system [2] is defined as $S = (U, A, V, f)$, where U is a non-empty and finite set of objects, called the universe, A is a non-empty and finite set of attributes, V is the union of the domains of all attributes, i.e., $V = \bigcup V_a$, where V_a denotes the domain of an attribute $a \in A$, and f is an information function which associates each attribute of

an object belonging to U with a unique value. If the attribute set A can be divided into condition attribute set C and decision attribute set D, the information system is also called as a decision information system or simply a decision table.

For an attribute subset B of A, it determines a binary relation $IND(B)$, which is called the indiscernibility relation and defined as follows [2]:

$$IND(B) = \{\langle x, y\rangle \in U \times U | \forall a \in B, f(x,a) = f(y,a)\}. \quad (1)$$

The binary relation $IND(B)$ is also an equivalence relation which satisfies reflexivity, symmetry and transitivity. The family of all equivalence classes of $IND(B)$, i.e., a partition of the universe U determined by B, is denoted by $U/IND(B)$ or simply by U/B. An equivalence class of $IND(B)$, i.e., a block of the partition U/B, is described as $[x]_B$ and referred to as B-elementary set or B-elementary granule [2].

The Pawlak rough set model is based on two basic notions, namely the lower and upper approximations of a set. Let X be a subset of the universe U, the lower approximation $\underline{B}(X)$ and upper approximation $\overline{B}(X)$ with respect to the attribute subset $B(B \subseteq A)$ are defined as [5]:

$$\underline{B}(X) = \{x \in U | \mu_B(x) = 1\},$$
$$\overline{B}(X) = \{x \in U | \mu_B(x) > 0\}, \quad (2)$$

where $\mu_B(x)$ denotes the inclusion degree that an object x belongs to X with respect to B, i.e., $\mu_B(x) = P(X|[x]_B) = |[x]_B \cap X|/|[x]_B|$, in which the symbol "$|\cdot|$" denotes the cardinality of a set.

An object x belongs to B-lower approximation of X if its equivalence class $[x]_B$ is a subset of X. An object x belongs to B-upper approximation of X if its equivalence class $[x]_B$ has a nonempty intersection with X.

Based on the lower and upper approximations of X, the universe U can be divided into three mutually disjoint regions with respect to B, namely the positive region $POS_B(X)$, boundary region $BND_B(X)$ and negative region $NEG_B(X)$ [2]:

$$POS_B(X) = \underline{B}(X),$$
$$BND_B(X) = \overline{B}(X) - \underline{B}(X),$$
$$NEG_B(X) = U - \overline{B}(X). \quad (3)$$

Let C and D be the sets of condition and decision attributes in a decision table, U/C and U/D be the partitions induced by the attribute sets C and D over U, respectively, the positive, boundary and negative regions of D with respect to C are defined as [30]:

$$POS_C(D) = \bigcup_{X \in U/D} \underline{C}(X),$$
$$BND_C(D) = \bigcup_{X \in U/D} \overline{C}(X) - \bigcup_{X \in U/D} \underline{C}(X),$$
$$NEG_C(D) = U - (POS_C(D) \cup BND_C(D)) = \emptyset. \quad (4)$$

The positive region is a set of C-elementary granules which completely belong to a block of the partition U/D, and the boundary region is the difference between the C-upper and C-lower approximations of all blocks within the partition U/D. A decision table is consistent if the formula $POS_C(D) = U$ holds, otherwise it is inconsistent.

Decision making in the Pawlak rough set model is only related to the data itself, not taking into consideration the cost of decision making. However, this is not in line with some applications in the real world. The DTRS model overcomes this limitation by introducing Bayesian decision theory.

Let $\Omega = \{X, X^c\}$ be a set of states that indicate an object x is in X or not in X, respectively, and $\Lambda = \{a_P, a_B, a_N\}$ be a set of actions that decide the object x to be $POS(X)$, $BND(X)$ or $NEG(X)$, respectively. The cost functions regarding different actions under the states X and X^c can be expressed as Table 1.

Table 1
Cost functions for different actions under the states X and X^c.

	a_P	a_B	a_N
X	λ_{PP}	λ_{BP}	λ_{NP}
X^c	λ_{PN}	λ_{BN}	λ_{NN}

In the table, λ_{PP}, λ_{BP} and λ_{NP} denote the costs incurred by taking the actions a_P, a_B and a_N, respectively, when the object x belongs to X, and λ_{PN}, λ_{BN} and λ_{NN} denote the costs incurred by taking the same actions when the object x does not belong to X.

Given an object x, the expected costs of taking different actions can be described as [30]:

$$R(a_P|[x]) = \lambda_{PP}P(X|[x]) + \lambda_{PN}P(X^c|[x]),$$
$$R(a_B|[x]) = \lambda_{BP}P(X|[x]) + \lambda_{BN}P(X^c|[x]),$$
$$R(a_N|[x]) = \lambda_{NP}P(X|[x]) + \lambda_{NN}P(X^c|[x]), \quad (5)$$

where $P(X|[x])$ and $P(X^c|[x])$ denote the probability that the object x belongs to X and X^c, respectively, and the formula $P(X|[x]) = 1 - P(X^c|[x])$ holds.

According to Bayesian decision theory, the following minimum-risk rules can be deduced [30]:

(P) if $R(a_P|[x]) \le \min\{R(a_B|[x]), R(a_N|[x])\}$, then decide $x \in POS(X)$;
(B) if $R(a_B|[x]) \le \min\{R(a_P|[x]), R(a_N|[x])\}$, then decide $x \in BND(X)$;
(N) if $R(a_N|[x]) \le \min\{R(a_P|[x]), R(a_B|[x])\}$, then decide $x \in NEG(X)$.

For any object, it takes the action that incurs the minimum cost. Tie-breaking criteria should be added so that each object is classified into only one region. Intuitionally, the cost for taking the right action is less than that for taking an improper one. Therefore, the formulae $\lambda_{PP} \le \lambda_{BP} < \lambda_{NP}$ and $\lambda_{NN} \le \lambda_{BN} < \lambda_{PN}$ hold. Additionally, the formula $P(X|[x]) = 1 - P(X^c|[x])$ holds for any object x under the states X and X^c. The decision rules can be simplified as [30]:

(P) if $P(X|[x]) \ge \alpha$ and $P(X|[x]) \ge \gamma$, then decide $x \in POS(X)$;
(B) if $P(X|[x]) < \alpha$ and $P(X|[x]) > \beta$, then decide $x \in BND(X)$;
(N) if $P(X|[x]) \le \beta$ and $P(X|[x]) \le \gamma$, then decide $x \in NEG(X)$,

where

$$\alpha = \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})},$$
$$\beta = \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})},$$
$$\gamma = \frac{\lambda_{PN} - \lambda_{NN}}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}. \quad (6)$$

If the constraint condition $(\lambda_{PN} - \lambda_{BN})(\lambda_{NP} - \lambda_{BP}) > (\lambda_{BP} - \lambda_{PP})(\lambda_{BN} - \lambda_{NN})$ is imposed on the cost functions, we have $0 \le \beta < \gamma < \alpha \le 1$. In this case, the following decision rules can be obtained [30]:

(P) if $P(X|[x]) \ge \alpha$, then decide $x \in POS(X)$;
(B) if $\beta < P(X|[x]) < \alpha$, then decide $x \in BND(X)$;
(N) if $P(X|[x]) \le \beta$, then decide $x \in NEG(X)$.

By introducing the parameters α and β, the probabilistic lower and upper approximations can be defined by Yao and Zhao [30]:

$$\underline{B}_{(\alpha,\beta)}(X) = \{x \in U | \mu_B(x) \ge \alpha\},$$
$$\overline{B}_{(\alpha,\beta)}(X) = \{x \in U | \mu_B(x) > \beta\}. \quad (7)$$

With different cost functions, the Pawlak rough set model and most probabilistic rough set models can be derived from the DTRS model [30]. For example, we can obtain the Pawlak rough set model if $\alpha = 1$ and $\beta = 0$, 0.5 probabilistic rough set

model if $\alpha = \beta = 0.5$, and variable precision rough set model if $\alpha = 1 - \beta \geq 0.5$. Based on probabilistic lower and upper approximations, the probabilistic positive, boundary and negative regions can be defined as [30]:

$$\begin{aligned} POS_C^{(\alpha,\beta)}(D) &= \{x \in U | P(D_{max}([x]_C)|[x]_C) \geq \alpha\}, \\ BND_C^{(\alpha,\beta)}(D) &= \{x \in U | \beta < P(D_{max}([x]_C)|[x]_C) < \alpha\}, \\ NEG_C^{(\alpha,\beta)}(D) &= \{x \in U | P(D_{max}([x]_C)|[x]_C) \leq \beta\}, \end{aligned} \tag{8}$$

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

Under the DTRS model, each object in the universe is classified into only one of the three regions. Let $p^C_{\max}(x)$ be the maximum inclusion degree of an object x, namely $p^C_{\max}(x) = P(D_{\max}([x]_C)|[x]_C)$. The cost for different kinds of decision rules can be described by Jia et al. [33]:

(1) Positive rule: $p^C_{\max}(x)\lambda_{PP} + (1 - p^C_{\max}(x))\lambda_{PN}$;
(2) Boundary rule: $p^C_{\max}(x)\lambda_{BP} + (1 - p^C_{\max}(x))\lambda_{BN}$;
(3) Negative rule: $p^C_{\max}(x)\lambda_{NP} + (1 - p^C_{\max}(x))\lambda_{NN}$.

Considering a special case in which the cost for a correct classification is 0, namely $\lambda_{PP} = \lambda_{NN} = 0$, the total decision cost for all objects in the universe is defined as [33]:

$$\begin{aligned} Cost_C^{(\alpha,\beta)}(D) = &\sum_{p^C_{max}(x) \geq \alpha} (1 - p^C_{max}(x))\lambda_{PN} \\ &+ \sum_{\beta < p^C_{max}(x) < \alpha} (p^C_{max}(x)\lambda_{BP} + (1 - p^C_{max}(x))\lambda_{BN}) \\ &+ \sum_{p^C_{max}(x) \leq \beta} p^C_{max}(x)\lambda_{NP}. \end{aligned} \tag{9}$$

For a decision table, the overall cost consists of three types of costs caused by the decision rules, namely the cost of the positive rules, the cost of the boundary rules and the cost of the negative rules. With this notion, an uncertainty measure for the minimum cost attribute reduction can be defined [33].

3. Maximum decision entropy-based attribute reduction in the DTRS model

In this section, we first illustrate some phenomena in the process of attribute reduction, and a novel monotonic uncertainty measure is then introduced for attribute reduction in the DTRS model. Finally, a heuristic attribute reduction algorithm based on the proposed uncertainty measure is developed.

3.1. Attribute reduction uncertainty measures within the DTRS model

At present, there are several uncertainty measures for attribute reduction in the DTRS model, which could be roughly classified into criterion preservation and criterion optimization. Their definitions can be formally described as follows.

Definition 1. [30] Let $S = (U, A = (C \cup D), V, f)$ be a decision table, and (α, β) be the parameters induced from the cost functions, a condition attribute set $R \subseteq C$ is a criterion preservation reduct of C with respect to D if the following two conditions are satisfied:

(1) $MES_R^{(\alpha,\beta)}(D) \succeq MES_C^{(\alpha,\beta)}(D)$;
(2) for any attribute $a \in R$, $MES_{R-\{a\}}^{(\alpha,\beta)}(D) \prec MES_C^{(\alpha,\beta)}(D)$,

where the symbol "$\succeq$" means that A is equal or superior to B, and attribute reduction uncertainty measure MES could be classification quality [7], region extension [27], distribution preservation (or entropy preservation) [28], etc.

Definition 2. [30] Let $S = (U, A = (C \cup D), V, f)$ be a decision table, and (α, β) be the parameters induced from the cost functions, a condition attribute set $R \subseteq C$ is a criterion optimization reduct of C with respect to D if the following two conditions are satisfied:

(1) $MES_R^{(\alpha,\beta)}(D) \succeq MES_C^{(\alpha,\beta)}(D)$;
(2) for any condition attribute subset $R' \subset R$, $MES_{R'}^{(\alpha,\beta)}(D) \prec MES_C^{(\alpha,\beta)}(D)$,

where attribute reduction uncertainty measure MES could be the minimum cost [33], optimal region preservation [26], optimal distribution preservation [48], etc.

In the Pawlak rough set model, an uncertainty measure with the preservation of the positive region is sufficient for attribute reduction because a positive region could implicitly deduce a unique boundary region. In other words, the positive and boundary regions are complementary to each other. Therefore, the uncertainty measures, such as the positive region and quality of classification, are intrinsically monotonic when adding or removing the condition attributes. However, in the DTRS model, this remarkable characteristic is not inherited by the uncertainty measures. As a result, some undesired phenomena or problems, such as decision non-monotonicity, confidence fluctuation, region transfer and cost bias, may happen in the process of attribute reduction [30]. In what follows, an example is given to illustrate these problems.

Table 2
A decision table.

	$a1$	$a2$	$a3$	$a4$	$a5$	d
x_1	0	0	0	0	0	d_1
x_2	0	0	0	0	1	d_2
x_3	0	0	0	1	1	d_1
x_4	0	0	0	1	1	d_2
x_5	0	0	0	1	1	d_2
x_6	0	0	0	1	1	d_2
x_7	0	0	1	1	1	d_2
x_8	0	0	1	1	1	d_2
x_9	0	0	1	1	1	d_2
x_{10}	0	0	1	1	1	d_3
x_{11}	0	1	1	1	1	d_2
x_{12}	0	1	1	1	1	d_3
x_{13}	1	1	1	1	1	d_1
x_{14}	1	1	1	1	1	d_2
x_{15}	1	1	1	1	1	d_3

Example 1. Let $S = (U, A = (C \cup D), V, f)$ be a decision table shown in Table 2, where $U = \{x_1, x_2, \ldots, x_{15}\}$, $C = \{a_1, a_2, a_3, a_4, a_5\}$ and $V_D = \{d_1, d_2, d_3\}$.

In the Pawlak rough set model, we have

$U/C = \{\{x_1\}, \{x_2\}, \{x_3, x_4, x_5, x_6\}, \{x_7, x_8, x_9, x_{10}\}, \{x_{11}, x_{12}\}, \{x_{13}, x_{14}, x_{15}\}\}$,

$U/D = \{\{x_1, x_3, x_{13}\}, \{x_2, x_4, x_5, x_6, x_7, x_8, x_9, x_{11}, x_{14}\}, \{x_{10}, x_{12}, x_{15}\}\}$,

$POS_C(D) = \{x_1, x_2\}$,

$BND_C(D) = \{x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}$.

The decision rules induced by C are as follows:

$\{x_1\} \to d_1$(confidence $= \frac{|\{x_1\}|}{|\{x_1\}|} = \frac{1}{1} = 1$),

$\{x_2\} \to d_2$(confidence $= \frac{|\{x_2\}|}{|\{x_2\}|} = \frac{1}{1} = 1$),

$\{x_3, x_4, x_5, x_6\} \to d_2$(confidence $= \frac{|\{x_4,x_5,x_6\}|}{|\{x_3,x_4,x_5,x_6\}|} = \frac{3}{4} = 0.75$),

$\{x_7, x_8, x_9, x_{10}\} \to d_2$(confidence $= \frac{|\{x_7,x_8,x_9\}|}{|\{x_7,x_8,x_9,x_{10}\}|} = \frac{3}{4} = 0.75$),

$\{x_{11}, x_{12}\} \to d_2$ or d_3(confidence $= \frac{|\{x_{11}\}|}{|\{x_{11},x_{12}\}|} = \frac{1}{2} = 0.50$),

$\{x_{13}, x_{14}, x_{15}\} \to d_1$ or d_2 or d_3(confidence $= \frac{|\{x_{13}\}|}{|\{x_{13},x_{14},x_{15}\}|} = \frac{1}{3} = 0.33$),

Assume that the parameters $\alpha = 0.70$ and $\beta = 0.35$ are calculated from the cost functions in the DTRS model, we have

$POS_C^{(0.70,0.35)}(D) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$,

$BND_C^{(0.70,0.35)}(D) = \{x_{11}, x_{12}\}$,

$NEG_C^{(0.70,0.35)}(D) = \{x_{13}, x_{14}, x_{15}\}$.

According to the definition of the positive region preservation or extension [27], we have the reduct $\{a_2\}$, and the following decision rules could be induced:

$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\} \to d_2$(confidence $= \frac{|\{x_2,x_4,x_5,x_6,x_7,x_8,x_9\}|}{|\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\}|} = \frac{7}{10} = 0.70$),

$\{x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\} \to d_2$ or d_3(confidence $= \frac{|\{x_{11},x_{14}\}|}{|\{x_{11},x_{12},x_{13},x_{14},x_{15}\}|} = \frac{2}{5} = 0.40$).

It is noteworthy that the class of the decision rule relating to the object x_1 is changed from d_1 to d_2 after attribute reduction. Meanwhile, the confidence of the decision rule relating to the object x_1 is decreased from 1 to 0.70. In other words, the certain decision rule relating to the object x_1 becomes an uncertain one after attribute reduction. In fact, this kind of behavior violates the principle of attribute reduction and inevitably brings a negative effect on classification task.

The set of condition attributes $\{a_1, a_3, a_4, a_5\}$ is a reduct of the optimal decision preservation that keeps all possible decisions of each object unchanged [26], and the decision rules determined by the reduct are as follows:

$\{x_1\} \to d_1$(confidence $= \frac{|\{x_1\}|}{|\{x_1\}|} = \frac{1}{1} = 1$),

$\{x_2\} \to d_2$(confidence $= \frac{|\{x_2\}|}{|\{x_2\}|} = \frac{1}{1} = 1$),

$\{x_3, x_4, x_5, x_6\} \to d_2$(confidence $= \frac{|\{x_4,x_5,x_6\}|}{|\{x_3,x_4,x_5,x_6\}|} = \frac{3}{4} = 0.75$),

$\{x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\} \to d_2$(confidence $= \frac{|\{x_7,x_8,x_9,x_{11}\}|}{|\{x_7,x_8,x_9,x_{10},x_{11},x_{12}\}|} = \frac{4}{6} = 0.67$),

$\{x_{13}, x_{14}, x_{15}\} \to d_1$ or d_2 or d_3(confidence $= \frac{|\{x_{13}\}|}{|\{x_{13},x_{14},x_{15}\}|} = \frac{1}{3} = 0.33$).

However, compared with the original attribute set C, the objects x_7, x_8, x_9 and x_{10} under the optimal decision preservation reduct are transferred from the probabilistic positive region to probabilistic boundary region.

Assume that the cost functions in the DTRS model are $\lambda_{PP} = 0, \lambda_{BP} = 1, \lambda_{NP} = 4, \lambda_{PN} = 5, \lambda_{BN} = 2$ and $\lambda_{NN} = 0$, we have $\alpha = 0.75$ and $\beta = 0.40$. The probabilistic positive, boundary and negative regions are as follows:

$POS_C^{(0.75,0.40)}(D) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$,

$BND_C^{(0.75,0.40)}(D) = \{x_{11}, x_{12}\}$,

$NEG_C^{(0.75,0.40)}(D) = \{x_{13}, x_{14}, x_{15}\}$.

For different subsets of all condition attributes, the overall costs are as follows:

$Cost_C^{(0.75,0.40)}(D) = 2(1-1)\lambda_{PN} + 4(1-\frac{3}{4})\lambda_{PN} + 4(1-\frac{3}{4})\lambda_{PN} + 2(\frac{1}{2}\lambda_{BP} + (1-\frac{1}{2})\lambda_{BN}) + 3(\frac{1}{3}\lambda_{NP}) = 2\lambda_{PN} + (\lambda_{BP} + \lambda_{BN}) + \lambda_{NP} = 17$,

$Cost_{C-\{a_2\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_3\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_4\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_2,a_3\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_2,a_4\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_3,a_4\}}^{(0.75,0.40)}(D) = 17$,

$Cost_{C-\{a_1\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_1,a_2\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_1,a_3\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_1,a_4\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_1,a_2,a_3\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_1,a_2,a_4\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_1,a_3,a_4\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_2,a_3,a_4\}}^{(0.75,0.40)}(D) = 18$,

$Cost_{C-\{a_1,a_2,a_3,a_4\}}^{(0.75,0.40)}(D) = 19$,

$Cost_{C-\{a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_2,a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_3,a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_4,a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_2,a_3,a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_2,a_4,a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_3,a_4,a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_2,a_3,a_4,a_5\}}^{(0.75,0.40)}(D) = 20$,

$Cost_{C-\{a_1,a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_1,a_2,a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_1,a_3,a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_1,a_4,a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_1,a_2,a_3,a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_1,a_2,a_4,a_5\}}^{(0.75,0.40)}(D) = Cost_{C-\{a_1,a_3,a_4,a_5\}}^{(0.75,0.40)}(D) = 21$.

The attribute subset $C - \{a_2, a_3\}$ is one of the minimum cost reducts, we have the following probabilistic positive, boundary and negative regions:

$POS_C^{(0.75,0.40)}(D) = \{x_1, x_2\}$,

$BND_C^{(0.75,0.40)}(D) = \{x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\}$,

$NEG_C^{(0.75,0.40)}(D) = \{x_{13}, x_{14}, x_{15}\}$.

Obviously, the objects x_3, x_4, x_5, x_6, x_7, x_8, x_9 and x_{10} are changed from the probabilistic positive region into the probabilistic boundary region after attribute reduction. In an extreme case, an object within the probabilistic positive region could jump down to the probabilistic negative region without increasing the overall cost as long as the object does not change the maximum inclusion degree of each equivalence class within the original probabilistic negative region. In fact, the minimum cost reduct makes the object tend towards the action that incurs less cost.

As for the distribution preservation reduct [28], we could only obtain the entire set of condition attributes, namely $\{a_1, a_2, a_3, a_4, a_5\}$. That is to say, each condition attribute is necessary for classification.

Intuitively, a reduct should be able to preserve the classification power. Not only the maximum inclusion degree (or confidence) but also the decision should be consistent after attribute reduction. Moreover, human beings usually take the decision that has the maximum inclusion degree. Therefore, in a sense, a reduct that keeps the maximum inclusion degree and maximum decision unchanged is enough for classification. Actually, the set of condition attributes $\{a_1, a_2, a_4, a_5\}$ is a reduct that keeps the maximum inclusion degree as well the maximum decision unchanged. The classification power of this reduct is inherently the same as that of the whole condition attribute set.

To this end, we propose a novel uncertainty measure to reflect the information of the maximum inclusion degree and maximum decision. A detailed description of the uncertainty measure will be presented in the next section.

3.2. Monotonic attribute reduction uncertainty measure using the maximum decision entropy

Information entropy is an effective measure of the uncertainty of random variable. We first present the related concepts of information entropy and then introduce a definition of the maximum decision entropy for attribute reduction.

Definition 3. [6] Let $S = (U, A = (C \cup D), V, f)$ be a decision table, and $U/B = \{X_1, X_2, \ldots, X_{|U/B|}\}$ be the partition induced by a condition attribute subset B of C over U, the entropy of B is defined as:

$$H(B) = -\sum_{i=1}^{|U/B|} P(X_i) \log P(X_i), \tag{10}$$

where $P(X_i) = |X_i|/|U|$.

Definition 4. [6] Let $S = (U, A = (C \cup D), V, f)$ be a decision table, $U/C = \{C_1, C_2, \ldots, C_{|U/C|}\}$ and $U/D = \{D_1, D_2, \ldots, D_{|U/D|}\}$ be the partitions induced by the condition attribute set C and decision attribute set D over U, the conditional entropy of C with respect to D is defined as:

$$H(D|C) = -\sum_{i=1}^{|U/C|} P(C_i) \sum_{j=1}^{|U/D|} P(D_j|C_i) \log P(D_j|C_i), \tag{11}$$

where $P(C_i) = |C_i|/|U|$ and $P(D_j|C_i) = |C_i \cap D_j|/|C_i|$.

Definition 5. [30] Let $S = (U, A = (C \cup D), V, f)$ be a decision table, $U/C = \{C_1, C_2, \ldots, C_{|U/C|}\}$ and $U/D = \{D_1, D_2, \ldots, D_{|U/D|}\}$ be the partitions induced by the condition attribute set C and decision attribute set D over U, for a condition class $C_i \in U/C$, its maximum inclusion degree and maximum decision are denoted as $MP(D|C_i) = \max\{P(D_1|C_i), P(D_2|C_i), \ldots, P(D_{|U/D|}|C_i)\}$ and $MD(D|C_i) = \{f(y, D) | y \in D_j \wedge P(D_j|C_i) = MP(D|C_i)\}$, respectively.

For a condition class, the decision is unique only when all objects within the condition class take the same decision. However, in the case of data with noise, a condition class may have more than one decision, and its maximum decision may also not be unique. Tie-breaking information, such as prior knowledge and class distribution, should be added so that each condition class is assigned only one maximum decision. Formally, the objects within each condition class can be further divided into two groups, namely a set of the objects that take the maximum decision and the remaining objects. Actually, the maximum inclusion degree and maximum decision could fully reflect the information of the two groups.

Definition 6. Let $S=(U,A=(C\cup D),V,f)$ be a decision table, $U/C=\{C_1,C_2,\ldots,C_{|U/C|}\}$ and $U/D=\{D_1,D_2,\ldots,D_{|U/D|}\}$ be the partitions induced by the condition attribute set C and decision attribute set D over U, the probabilistic distribution of maximum inclusion degree of C with respect to D is denoted as $MS(D|C)=((MP(D|C_1),1-MP(D|C_1)),(MP(D|C_2),1-MP(D|C_2)),\ldots,(MP(D|C_{|U/C|}),1-MP(D|C_{|U/C|})))$.

Definition 7. Let $S=(U,A=(C\cup D),V,f)$ be a decision table, $MS(D|B)=((MP(D|B_1),1-MP(D|B_1)),(MP(D|B_2),1-MP(D|B_2)),\ldots,(MP(D|B_{|U/B|}),1-MP(D|B_{|U/B|})))$ be the probabilistic distribution of maximum inclusion degree of a condition attribute subset B of C, the maximum decision entropy of a condition class $B_i\in U/B$ and overall maximum decision entropy of B with respect to D are denoted as:

$$MH(D|B_i)=-P(B_i)\Big(MP(D|B_i)\log MP(D|B_i)+(m-1)\Big(\frac{1-MP(D|B_i)}{m-1}\Big)\log\Big(\frac{1-MP(D|B_i)}{m-1}\Big)\Big),\quad(12)$$

$$MH(D|B)=\sum_{i=1}^{|U/B|}MH(D|B_i),\quad(13)$$

where m is the number of classes, i.e., $m=|U/D|$.

In the definition, the maximum decision entropy degenerates into conditional entropy when $m=2$ and Shannon entropy when $m=1$.

Formally, the maximum decision entropy consists of two kinds of information, namely the certainty of the objects with the maximum decision and the uncertainty of the remaining objects. The former reflects the degree of certainty to make a decision. The higher the degree of certainty is, the lower the maximum decision entropy is and the better the decision is made. The latter embodies the degree of uncertainty to make the maximum decision. In practice, human beings mainly focus on the degree of uncertainty itself rather than the exact probability distribution. Therefore, it is reasonable to represent the overall uncertainty by averaging all decisions except the maximum one. The maximum decision entropy is minimized to 0 only when the probability distribution of all data is degenerated, namely each condition class takes only one decision, and reaches the maxima only when the probability distribution of all data is uniform (see Fig. 1).

Similar to the definition of the maximum decision entropy, the positive, boundary and negative maximum decision entropy could also be given.

Definition 8. Let $S=(U,A=(C\cup D),V,f)$ be a decision table, (α,β) be the parameters induced from the cost functions, and $MS(D|B)=((MP(D|B_1),1-MP(D|B_1)),(MP(D|B_2),1-MP(D|B_2)),\ldots,(MP(D|B_{|U/B|}),1-MP(D|B_{|U/B|})))$ be the probabilistic distribution of maximum inclusion degree of a condition attribute subset B of C, the (α,β) positive, boundary and negative maximum decision entropy of B with respect to D are denoted as

Table 3
Another decision table.

	$a1$	$a2$	$a3$	$a4$	$a5$	d
x_1	0	0	0	0	0	d_1
x_2	0	0	0	0	1	d_2
x_3	0	0	0	1	1	d_2
x_4	0	0	0	1	1	d_2
x_5	0	0	0	1	1	d_3
x_6	0	0	1	1	1	d_2
x_7	0	0	1	1	1	d_2
x_8	0	0	1	1	1	d_4
x_9	0	1	1	1	1	d_2
x_{10}	0	1	1	1	1	d_3
x_{11}	1	1	1	1	1	d_1
x_{12}	1	1	1	1	1	d_4
x_{13}	1	1	1	1	1	d_4

$$\begin{aligned}MH(POS_\beta^\alpha(D|B))&=\sum_{(B_i\in U/B)\wedge(MP(D|B_i)\geq\alpha)}MH(D|B_i),\\MH(BND_\beta^\alpha(D|B))&=\sum_{(B_i\in U/B)\wedge(\beta<MP(D|B_i)<\alpha)}MH(D|B_i),\\MH(NEG_\beta^\alpha(D|B))&=\sum_{(B_i\in U/B)\wedge(MP(D|B_i)\leq\beta)}MH(D|B_i).\end{aligned}\quad(14)$$

The definition of the non-negative maximum decision entropy could be defined similarly as $MH(\neg NEG_\beta^\alpha(D|B))$, which consists of the maximum decision entropy of the condition classes whose maximum inclusion degree is greater than β.

Proposition 1. *Let $S=(U,A=(C\cup D),V,f)$ be a decision table, (α,β) be the parameters induced from the cost functions, for any two attribute subsets P and Q of C, and P⊆Q, then*

(1) $MH(POS_\beta^\alpha(D|P))\geq MH(POS_\beta^\alpha(D|Q))$,
(2) $MH(BND_\beta^\alpha(D|P))\geq MH(BND_\beta^\alpha(D|Q))$,
(3) $MH(NEG_\beta^\alpha(D|P))\geq MH(NEG_\beta^\alpha(D|Q))$.

Without loss of generality, we assume that two condition classes C_i and C_j within $POS_\beta^\alpha(D|Q)$ will be merged when a condition attribute a is removed from the condition attribute set Q and $P=(Q-\{a\})$. The maximum inclusion degrees and maximum decisions of the condition classes C_i and C_j are denoted as $MP(D|C_i)$, $MP(D|C_j)$, $MD(D|C_i)$ and $MD(D|C_j)$, respectively. While the maximum inclusion degree and maximum decision of the merged condition class $C_i\cup C_j$ are denoted as $MP(D|(C_i\cup C_j))$ and $MD(D|(C_i\cup C_j))$, respectively. For simplicity, the maximum inclusion degrees $MP(D|C_i)$, $MP(D|C_j)$ and $MP(D|(C_i\cup C_j))$ are replaced by θ_i, θ_j and θ_{ij}, respectively. In what follows, we first give an example to demonstrate all cases of condition class mergence after an attribute is removed and then present the proof of Proposition 1.

Example 2. Let $S=(U,A=(C\cup D),V,f)$ be a decision table shown in Table 3, where $U=\{x_1,x_2,\ldots,x_{13}\}$, $C=\{a_1,a_2,a_3,a_4,a_5\}$ and $V_D=\{d_1,d_2,d_3,d_4\}$.

With all condition attributes, the universe in Example 2 is divided into 6 condition classes $U/C=\{C_1,C_2,C_3,C_4,C_5,C_6\}$, where $C_1=\{x_1\}$, $C_2=\{x_2\}$, $C_3=\{x_3,x_4,x_5\}$, $C_4=\{x_6,x_7,x_8\}$, $C_5=\{x_9,x_{10}\}$ and $C_6=\{x_{11},x_{12},x_{13}\}$, respectively. The maximum inclusion degrees of these condition classes are $MP(D|C_1)=1$, $MP(D|C_2)=1$, $MP(D|C_3)=0.67$, $MP(D|C_4)=0.67$, $MP(D|C_5)=0.50$ and $MP(D|C_6)=0.67$, respectively. While their maximum decisions are $MD(D|C_1)=d_1$, $MD(D|C_2)=d_2$, $MD(D|C_3)=d_2$, $MD(D|C_4)=d_2$, $MD(D|C_5)=d_2$ and $MD(D|C_6)=d_4$, respectively. When an attribute is removed from the whole condition attribute set, at least two condition classes in Example 2 will be merged. Table 4 tabulates different cases of condition class mergence.

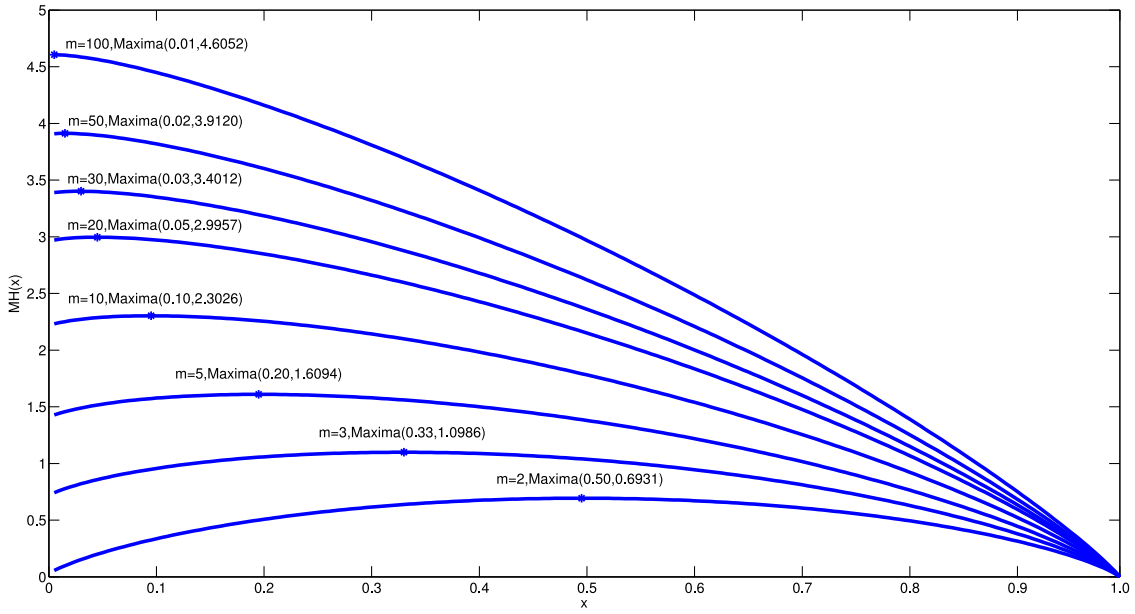


Fig. 1. Maximum decision entropy with different numbers of classes m (maxima marked by "*").

Table 4
Different cases of condition class mergence.

a	C_i, C_j	(θ_i, θ_j)	θ_{ij}		Generalized case	
a_3	C_3, C_4	(0.67,0.67)	0.67	$\theta_i = \theta_j$	$\theta_{ij} = \theta_i = \theta_j$	I
a_5	C_1, C_2	(1,1)	0.50		$\theta_{ij} < \min(\theta_i, \theta_j)$	II
N/A	N/A	N/A	N/A		$\theta_{ij} > \max(\theta_i, \theta_j)$	III
a_4	C_2, C_3	(1,0.67)	0.75	$\theta_i > \theta_j$	$\theta_j < \theta_{ij} < \theta_i$	I
a_1	C_6, C_5	(0.67,0.50)	0.40		$\theta_{ij} < \theta_j$	II
N/A	N/A	N/A	N/A		$\theta_{ij} > \theta_i$	III
a_2	C_5, C_4	(0.50,0.67)	0.60	$\theta_i < \theta_j$	$\theta_i < \theta_{ij} < \theta_j$	I
a_1	C_5, C_6	(0.50,0.67)	0.40		$\theta_{ij} < \theta_i$	II
N/A	N/A	N/A	N/A		$\theta_{ij} > \theta_j$	III

In the table, the attribute to be removed is listed in the first column. The condition classes that will be merged are indicated in the second column and their maximum inclusion degrees are given in the third column. The fourth column shows the maximum inclusion degree of the merged condition class. In addition, we also present the generalized cases for condition class mergence in the last column. In the table, the symbol "N/A" denotes that this situation could not happen in the real world application.

By observing Table 4, we could derive three generalized cases: (I) $\min\{\theta_i, \theta_j\} \le \theta_{ij} \le \max\{\theta_i, \theta_j\}$; (II)$\theta_{ij} < \min\{\theta_i, \theta_j\}$; and (III) $\theta_{ij} > \max\{\theta_i, \theta_j\}$. The proof for the three cases is presented in the Appendix.

Additionally, by enumerating all possible values of the variables θ_i and θ_j, the maximum decision entropy after condition class mergence is shown in Fig. 2.

In Fig. 2, the left subfigure indicates the maximum decision entropy of the merged condition class $C_i \cup C_j$ (top surface) and the cumulative maximum decision entropy of the condition classes C_i and C_j (bottom surface). While the right subfigure presents the difference of the maximum decision entropy before and after condition class mergence, namely $\Delta MH = MH(D|(C_i \cup C_j)) - (MH(D|C_i) + MH(D|C_j))$. It is obvious that the maximum decision entropy becomes smaller when the values of the variables θ_i and θ_j are raised. Moreover, the value of ΔMH becomes larger when the difference between the variables θ_i and θ_j is increased. The maximum decision entropy after condition class mergence may not change only when the values of the variables θ_i and θ_j are the same.

Proposition 1 shows that the uncertainty measures of the (α, β) positive, boundary and negative maximum decision entropy are monotonic when adding or deleting a condition attribute.

Definition 9. Let $S = (U, A = (C \cup D), V, f)$ be a decision table and (α, β) be the parameters induced from the cost functions, a condition attribute $a \in C$ is indispensable with respect to the (α, β) positive, boundary and negative regions if the following conditions hold, respectively:

(1) $MH(POS^{\alpha}_{\beta}(D|(C - \{a\}))) > MH(POS^{\alpha}_{\beta}(D|C))$
(2) $MH(BND^{\alpha}_{\beta}(D|(C - \{a\}))) > MH(BND^{\alpha}_{\beta}(D|C))$
(3) $MH(NEG^{\alpha}_{\beta}(D|(C - \{a\}))) > MH(NEG^{\alpha}_{\beta}(D|C))$

The indispensable attribute is also called the core attribute. Conversely, a condition attribute is dispensable if the overall maximum decision entropy does not change after the attribute is removed.

Definition 10. Let $S = (U, A = (C \cup D), V, f)$ be a decision table, (α, β) be the parameters induced from the cost functions and $P \subseteq C$, for any condition attribute $a \in (C - P)$, its relative significance for the (α, β) positive, boundary and negative regions with respect to D are defined as follows, respectively:

(1) $SIG(a, P, POS^{\alpha}_{\beta}(D|P)) = MH(POS^{\alpha}_{\beta}(D|P)) - MH(POS^{\alpha}_{\beta}(D|(P \cup \{a\})))$
(2) $SIG(a, P, BND^{\alpha}_{\beta}(D|P)) = MH(BND^{\alpha}_{\beta}(D|P)) - MH(BND^{\alpha}_{\beta}(D|(P \cup \{a\})))$
(3) $SIG(a, P, NEG^{\alpha}_{\beta}(D|P)) = MH(NEG^{\alpha}_{\beta}(D|P)) - MH(NEG^{\alpha}_{\beta}(D|(P \cup \{a\})))$

Definition 11. Let $S = (U, A = (C \cup D), V, f)$ be a decision table and (α, β) be the parameters induced from the cost functions, for a condition attribute subset P of C, P is a reduct of the (α, β) positive, boundary and negative maximum decision entropy preservation iff:

(1) $MH(POS^{\alpha}_{\beta}(D|P)) = MH(POS^{\alpha}_{\beta}(D|C))$
and $\forall a \in P, MH(POS^{\alpha}_{\beta}(D|(P - \{a\}))) > MH(POS^{\alpha}_{\beta}(D|C))$

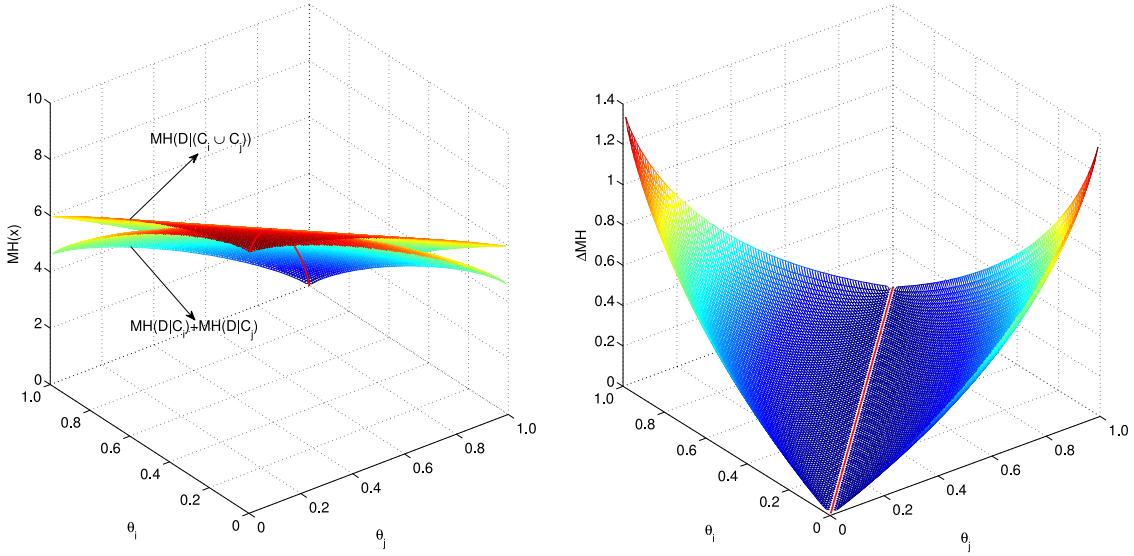


Fig. 2. Maximum decision entropy after mergence (equal values marked with red).(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(2) $MH(BND_{\beta}^{\alpha}(D|P)) = MH(BND_{\beta}^{\alpha}(D|C))$
and $\forall a \in P, MH(BND_{\beta}^{\alpha}(D|(P-\{a\}))) > MH(BND_{\beta}^{\alpha}(D|C))$
(3) $MH(NEG_{\beta}^{\alpha}(D|P)) = MH(NEG_{\beta}^{\alpha}(D|C))$
and $\forall a \in P, MH(NEG_{\beta}^{\alpha}(D|(P-\{a\}))) > MH(NEG_{\beta}^{\alpha}(D|C))$

Similarly, the (α, β) non-negative preservation reduct could be defined by using the non-negative maximum decision entropy.

3.3. Heuristic algorithm using the maximum decision entropy

It is well known that finding a minimal reduct of a decision table is NP-hard. The heuristic algorithm for an optimal reduct is an alternative way to achieve this goal. Roughly speaking, the strategies for attribute reduction could be classified into adding, deleting and add-deleting manners. In this paper, we use the strategy of adding-deleting with the objective of maximizing the relevance of the reduct to the class attribute and minimizing the redundancy of the condition attributes within the reduct [49]. The process could be depicted by Algorithm 1.

In the algorithm, the maximum decision entropy $MH(D|C)$ could be replaced by the (α, β) positive, boundary, negative or non-negative maximum decision entropy, and the corresponding reduct is then generated. Algorithm 1 includes three closely integrated stages: (1) Divides all condition attributes into the core and candidate attribute sets; (2) Recursively adds an optimal attribute into the reduct until the preset condition is met; and (3) Removes the redundant attributes from the reduct. At the first stage, from Step 1 to Step 3, the algorithm first splits the universe into several condition classes. The maximum decision entropy of each condition class and overall maximum decision entropy are then calculated. With the definition of the indispensable attribute (see Definition 9), all core attributes could be separated from the whole condition attribute set and be treated as an initial attribute set for the reduct. At the second stage, from Step 4 to Step 8, the algorithm repeatedly chooses an optimal attribute from the candidate attribute set and puts it into the reduct until the overall maximum decision entropy of all selected condition attributes is the same as that of the entire set of condition attributes. The criteria for finding an optimal attribute a_{opt} are based on two aspects: 1) The attribute should maximize the information gain of all selected condition attributes to the decision attribute, namely maximum relevance. In other words, the value of $MH(D|P) - MH(D|(P \cup \{a_{opt}\}))$ should

Algorithm 1: Maximum decision entropy-based heuristic algorithm for attribute reduction.

Input:
A decision table $S = (U, A = (C \cup D), V, f)$;
Output:
An optimal reduct P;
1: Partition the universe into a set of condition classes U/C;
2: Compute the maximum inclusion degree for each condition class within U/C and overall maximum decision entropy $MH(D|C)$;
3: Find the core attributes *Core* and put them into the reduct $P = Core$;
4: **while** $MH(D|P) \neq MH(D|C)$ **do**
5: For any condition attribute a_i within candidate attribute set $C - P$, calculate the relevance of the attribute set $P \cup \{a_i\}$ to D and the redundancy of the attribute a_i to P;
6: Select an attribute a_{opt} with maximum relevance and minimum redundancy;
7: $P = P \cup \{a_{opt}\}$
8: **end while**
9: Remove the superfluous attributes from P;
10: **return** the reduct P.

be as large as possible; 2)The attribute should minimize the correlation between the originally selected condition attributes and the attribute itself, namely minimum redundancy. That is to say, the value of $H(\{a_{opt}\}|P)$ should also be as large as possible. An attribute with the highest overall score is preferred by the algorithm. Tie-breaking information should be added when more than one optimal attribute is available. After the loops terminate, an unrefined reduct is generated by the algorithm. At the last stage, a backward deleting step is performed to further remove the redundant attributes, and an optimal reduct is finally yielded by the algorithm.

In Algorithm 1, the time complexity of forming all equivalence classes, computing the maximum decision entropy and finding all core attributes is at most $O(|C|^2|U|)$ with the technique of radix sort. The algorithm takes the time $O(|C|^2|U|)$ to find an optimal attribute. In the worst case, the set of candidate attributes is empty after $|C|$ rounds of selection, thus the time complexity of obtaining an unrefined reduct is $O(|C|^3|U|)$. The final stage for removing

Table 5
Investigated data sets.

Data sets	Categorical	Numeric	Objects	Classes	Missing	Consistency
annealing (anneal)	32	6	798	6	Y	(1.0,0.50)-24
wisconsin breast cancer(breast)	9	0	699	2	Y	(1.0,1.0)-0
credit rating(credit)	9	6	690	2	Y	(1.0,0.50)-20
kr-vs-kp(krvskp)	36	0	3196	2	N	(1.0, 1.0)-0
lung cancer(lung)	56	0	32	3	Y	(1.0,1.0)-0
lymphography(lymph)	15	3	148	4	N	(1.0,1.0)-0
sonar(sonar)	0	60	208	2	N	(1.0, 1.0)-0
table2(table)	5	0	15	3	N	(1.0, 0.33)-13
vehicle silhouettes(vehicle)	0	18	846	4	N	(1.0,0.26)-379
congressional voting(vote)	16	0	435	2	Y	(1.0,0.50)-5

the redundant attributes takes the time $O(|C|^2|U|)$. Therefore, the total time complexity of Algorithm 1 is $O(|C|^3|U|)$, and the space complexity is $O(|C||U|)$.

4. Empirical analysis

In the experiments, we consider three attribute reduction algorithms, namely the quantitative probabilistic region preservation algorithm (QPRP) presented in [27], the decision region distribution preservation algorithm (DRDP) introduced in [28] and the algorithm proposed in this study (PMDE). The purpose of the experiments is twofold. One is to verify the validity of our proposed uncertainty measure, namely the monotonicity of the maximum decision entropy. The other is to show the performance of the proposed algorithm compared with the selected algorithms.

4.1. Investigated data sets

The experimental evaluation is conducted over several data sets from UCI machine learning repository [50]. Table 5 tabulates the detailed information of the experimental data sets.

In the table, the sixth column indicates that whether the data set has missing value or not. The last column shows the consistency information of the data set, in which the maximum and minimum values over all maximum inclusion degrees of the condition classes are listed in brackets and the number of the inconsistent objects within the data set is also presented. Actually, a data set is consistent only if, in the last column of Table 5, the values in the brackets are all 1.0 and the number followed is 0.

In the experiments, the missing values (denoted as "?") within some data sets are completed by the mean (or mode) of corresponding attribute [51]. In addition, some data sets contain the numerical attribute. To increase the uncertainty of the data set, all numerical attributes are discretized into the categorical one using the principle of equal frequency binning with only two bins [51]. More specifically, all objects are first ranked by their values in ascending order and then divided into two bins, each of which has the same number of objects. For simplicity, in what follows, we use the abbreviation shown in the first column to represent the data set.

4.2. Experiment design and parameter settings

For each data set, all objects are used to yield a reduct, and the monotonicity of the proposed uncertainty measure is analyzed on these results. During the procedure of attribute reduction, especially the data set without the core attribute, there is more than one optimal attribute that could be picked up by the algorithm. The natural attribute order within the data set is employed as tie-breaking information to select an optimal attribute. For example, there are three eligible attributes a_5 a_2 and a_3. The attribute a_2 is preferred to the algorithm because its natural order is minimal.

To show the performance of the selected algorithms, two different classifiers are used as the base classifier, namely Naive Bayes and decision rules (PART classifier in Weka [51]). The performance evaluation is carried out with the technique of 10-fold cross validation. Specifically, the data set is randomly divided into ten equal subsets in which nine subsets of the data set are used as training data and one subset is used as testing data. As for the parameters used in the algorithm, they usually depend on the practical application itself. For simplicity, we mainly investigate the positive region preservation reduct in our experiments, and the threshold parameter α is calculated from the data set itself. Specifically, for each condition class of the data set, we first compute its maximum inclusion degree under the entire set of condition attributes and then select the minimal one as the threshold parameter, namely the minimum value in the brackets of Table 5. As regards the strategy for selecting an optimal attribute, the algorithm first ranks all candidate attributes by their relevance to the class attribute in descending order, and then the attributes with the same value are sorted again by their redundancy to all selected condition attributes in ascending order. Finally, the algorithm picks up the first attribute from the sorted queue of all candidate attributes.

4.3. Experimental results and analysis

4.3.1. The monotonicity of the proposed uncertainty measure

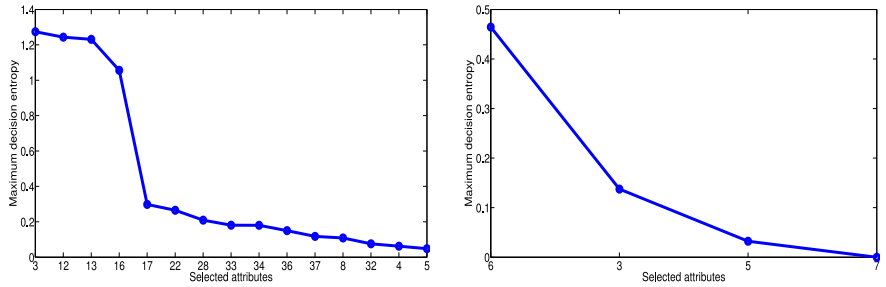
To show the monotonicity of the proposed uncertainty measure, we analyze the reduct generated by Algorithm 1 in adding manner. Specifically, an optimal reduct is first yielded by Algorithm 1. Each attribute within the optimal reduct is then one by one added into a condition attribute set which is initialized as empty, and the maximum decision entropy under the condition attribute set is recorded. The experimental results are shown in Fig. 3.

In Fig. 3, the x-coordinate denotes the selected condition attributes with ID information in the process of attribute reduction, and the y-coordinate indicates the maximum decision entropy under a condition attribute set. From all subfigures, it is quite obvious that the maximum decision entropy is strictly monotonic decreasing when adding the condition attributes, which validates the monotonicity of the proposed uncertainty measure.

4.3.2. The effectiveness of the proposed algorithm for attribute reduction

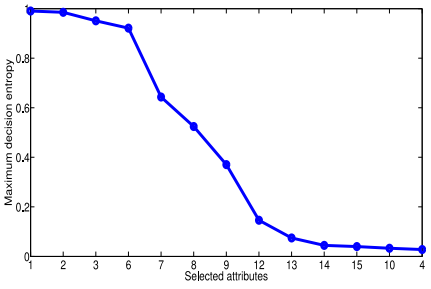
In the experiments, for each data set, the selected algorithms first generate an optimal reduct, based on which a classifier is then learned and tested. Table 6 lists the results of attribute reduction using the selected algorithms.

In Table 6, it is apparent that each algorithm achieves the objective of attribute reduction by removing some condition attributes, but in quite different ways. Compared with other algorithms, the quantitative probabilistic region preservation algorithm (QPRP) obtains a reduct with fewer attributes on most data sets. The objective of the QPRP algorithm is to keep the number of

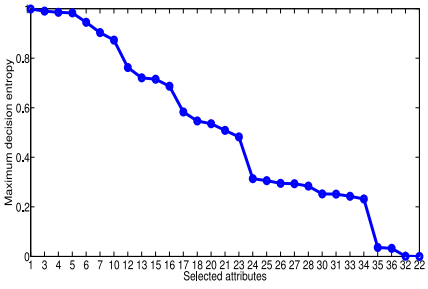


(a) anneal

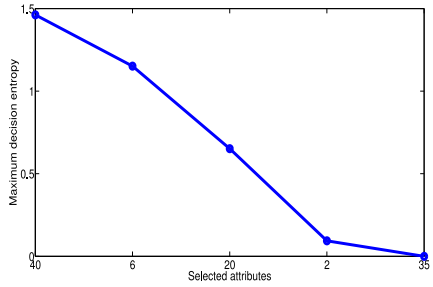
(b) breast



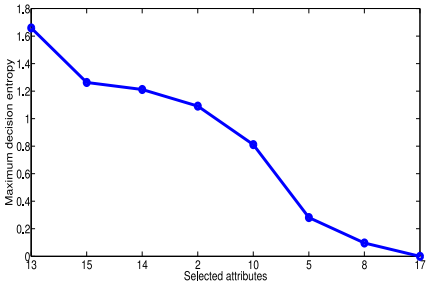
(c) credit



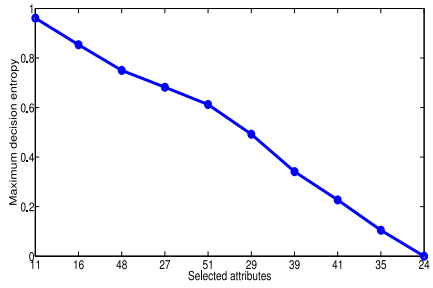
(d) krvskp



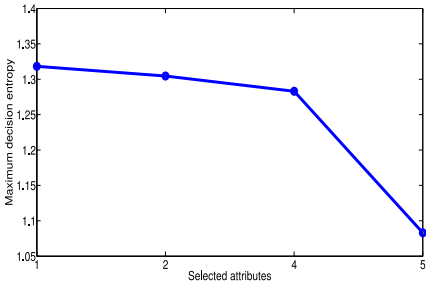
(e) lung



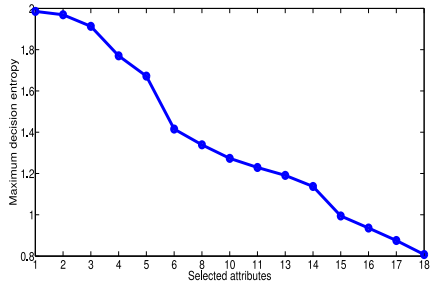
(f) lymph



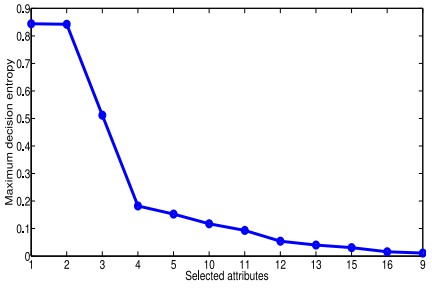
(g) sonar



(h) table



(i) vehicle



(j) vote

Fig. 3. Analysis of the monotonicity of the proposed uncertainty measure.

Table 6
Cardinality of the optimal reduct generated by the selected algorithms.

Data sets	Raw data	QPRP	DRDP	PMDE
anneal	38	1	15	15
breast	9	4	4	4
credit	15	1	13	13
krvskp	36	31	29	29
lung	56	5	4	5
lymph	18	7	8	8
sonar	60	11	10	10
table	5	1	5	4
vehicle	18	1	16	15
vote	16	1	12	12
avg.	27.1	6.3	11.6	11.5

Table 7
Classification accuracy of the selected algorithms using Naive Bayes classifier.

Data sets	Raw data	QPRP	DRDP	PMDE
anneal	0.9065	0.8363	0.9109	**0.9154**
breast	**0.9728**	0.9699	0.9657	0.9657
credit	0.8348	0.5551	0.8449	**0.8449**
krvskp	0.8789	0.8870	0.8827	**0.9030**
lung	0.5313	0.6250	0.6563	**0.6563**
lymph	**0.8581**	0.8311	0.8311	0.8513
sonar	0.7644	0.7308	0.7452	**0.7692**
table	0.4000	0.5333	0.4000	**0.5333**
vehicle	0.4527	0.2991	0.4929	**0.4964**
vote	0.9011	0.6736	0.9310	**0.9310**
avg.	0.7501	0.6941	0.7661	**0.7867**

Table 8
Classification accuracy of the selected algorithms using PART classifier.

Data sets	Raw data	QPRP	DRDP	PMDE
anneal	0.9588	0.8363	0.9566	**0.9621**
breast	0.9413	0.9371	0.9456	**0.9471**
credit	0.8377	0.5551	0.8420	**0.8420**
krvskp	0.9906	0.9906	**0.9915**	0.9912
lung	0.5625	0.6563	0.5938	**0.7813**
lymph	0.7905	0.7838	**0.7973**	0.7838
sonar	0.7356	0.7500	0.7500	**0.7740**
table	0.4667	**0.5333**	0.4667	0.4667
vehicle	0.6513	0.2991	0.6596	**0.6652**
vote	0.9540	0.6736	0.9563	**0.9563**
avg.	0.7889	0.7015	0.7959	**0.8170**

the positive objects unchanged or even enlarged. However, in the uncertain situation, the QPRP algorithm always terminates with a small set of condition attributes. We could see that, on all inconsistent data sets, the QPRP algorithm yields a reduct with only one attribute. Actually, this small attribute set is not enough for classification task. In other words, the reduct of the QPRP algorithm does not hold the same performance as the entire set of condition attributes. This conclusion will be further verified by the following experiments. The decision region distribution preservation algorithm (DRDP) and probabilistic maximum decision entropy preservation algorithm (PMDE) generate almost the same results but still have little difference. On data set "table", the reducts of the DRDP algorithm and our proposed algorithm PMDE are all made up of the core attributes, whereas the number of attributes within the reduct of our proposed algorithm PMDE is slightly less than that of the DRDP algorithm. In other words, the core attribute in the DRDP algorithm, in a sense, is not a real core attribute or even not the relevant attribute. Data set "vehicle" is also a good case. These observations show that our proposed algorithm compares favorably with the DRDP algorithm. In short, compared with other selected algorithms, our proposed algorithm could yield a better result with a reasonable number of condition attributes.

To fully evaluate the potentials of the proposed algorithm, two different base classifiers, namely Naive Bayes and PART, are utilized in the experiments, and the results are shown in Tables 7 and 8, respectively.

In Tables 7 and 8 we report the classification accuracy of the two selected algorithms and our proposed algorithm. Additionally, the classifier learned from the raw data, namely the performance without the procedure of attribute reduction, is also listed in the tables for comparison. The highest accuracy among these different algorithms is boldfaced, and the average accuracy over all data sets is also shown in the last row "avg.".

By viewing the results, it is quite evident that, on most data sets, the performance of the QPRP algorithm becomes worse after attribute reduction, while the two entropy-based algorithms, namely the DRDP algorithm and our proposed one, both achieve better results. For the QPRP algorithm, its objective of attribute reduction is to keep or enlarge the number of the positive objects. Actually, two object sets with the same cardinality do not mean that the two sets have the same objects. Therefore, the QPRP algorithm could acquire comparable results on the data sets without any inconsistent object. Nevertheless, on some inconsistent data sets, such as "anneal", "credit", "vehicle" and "vote", the QPRP algorithm terminates with only one attribute. As a result, its performance is rather poor or even worse than that of the raw data. Both the DRDP algorithm and our proposed algorithm could deal with the inconsistent data. In the process of attribute reduction, the DRDP algorithm attaches the same importance to each inconsistent object, whereas our proposed algorithm classifies all inconsistent objects into two groups with different significance, namely a set of the inconsistent objects with the maximum inclusion degree and a set of the remaining objects. Actually, in the real world application, human beings usually take the decision that has the maximum inclusion degree. Therefore, during the procedure of attribute reduction, we should pay more attention to the information about the inconsistent objects with the maximum inclusion degree. Our proposed algorithm employs this idea to design the heuristic information, and its performance, as a result, is much better than that of the DRDP algorithm. By averaging the performance over all data sets, the DRDP algorithm using Naive Bayes and PART classifiers obtains an improvement over the raw data by 1.6% and 0.7%, respectively, whereas our proposed algorithm achieves an overall 3.7% and 2.8% improvement, respectively. These experimental results further validate the effectiveness of our proposed algorithm for attribute reduction in the context of the DTRS model.

5. Conclusions

The monotonicity is one of the most important properties for an effective uncertainty measure of attribute reduction. Most existing uncertainty measures within the DTRS model, however, do not have this salient property, and some undesired phenomena, such as decision non-monotonicity, confidence fluctuation, region transfer and cost bias, may happen in the process of attribute reduction. In this paper, we develop the concepts of the maximum inclusion degree and maximum decision for the condition class, based on which a novel uncertainty measure called the maximum decision entropy is presented. The monotonicity of the proposed uncertainty measure is not only proved in theory but also verified through the extensive experiments. Furthermore, we design a maximum decision entropy-based heuristic algorithm to yield an optimal reduct by using the criteria of maximum relevance and minimum redundancy. The experimental results on both artificial and UCI data sets demonstrate that the performance of

our proposed algorithm is better than that of the state-of-the-art algorithms. Currently, the proposed algorithm could only deal with the categorical data so that the numerical attribute must be discretized into the categorical one. In the future, it is desirable that an algorithm should be able to tackle with the hybrid data directly [52]. Another interesting research is to introduce our work to partially labeled data [53], which could enrich the theory as well as the application of the DTRS model.

Acknowledgement

The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions. The work was supported in part by the National Natural Science Foundation of China (No. 61573248, 61672358, 61703283, 61773328), in part by the China Postdoctoral Science Foundation (No. 2016M590812, 2017M612736, 2017T100645), and in part by the Guangdong Natural Science Foundation under Project 2017A030310067.

Appendix

Proof of Proposition 1:

The information gain ΔMH in Proposition 1.(1) is discussed as follows:

Proof. I.$\min\{\theta_i, \theta_j\} \le \theta_{ij} \le \max\{\theta_i, \theta_j\}$

$$\begin{aligned}
\Delta MH &= MH(POS_\beta^\alpha(D|P)) - MH(POS_\beta^\alpha(D|Q)) \\
&= MH(POS_\beta^\alpha(D|C_i \cup C_j)) - MH(POS_\beta^\alpha(D|C_i)) - MH(POS_\beta^\alpha(D|C_j)) \\
&= -P(C_i \cup C_j)\left(\theta_{ij}\log\theta_{ij} + (m-1)\left(\frac{1-\theta_{ij}}{m-1}\right)\log\left(\frac{1-\theta_{ij}}{m-1}\right)\right) \\
&\quad + P(C_i)\left(\theta_i\log\theta_i + (m-1)\left(\frac{1-\theta_i}{m-1}\right)\log\left(\frac{1-\theta_i}{m-1}\right)\right) \\
&\quad + P(C_j)\left(\theta_j\log\theta_j + (m-1)\left(\frac{1-\theta_j}{m-1}\right)\log\left(\frac{1-\theta_j}{m-1}\right)\right) \\
&= P(C_i)\theta_i\log\theta_i + P(C_j)\theta_j\log\theta_j - P(C_i \cup C_j)\theta_{ij}\log\theta_{ij} \\
&\quad + P(C_i)(1-\theta_i)\log\left(\frac{1-\theta_i}{m-1}\right) + P(C_j)(1-\theta_j)\log\left(\frac{1-\theta_j}{m-1}\right) \\
&\quad - P(C_i \cup C_j)(1-\theta_{ij})\log\left(\frac{1-\theta_{ij}}{m-1}\right)
\end{aligned}$$

Let $\Delta MH = \Delta MH_1 + \Delta MH_2$, where

$$\begin{aligned}
\Delta MH_1 &= P(C_i)\theta_i\log\theta_i + P(C_j)\theta_j\log\theta_j - P(C_i \cup C_j)\theta_{ij}\log\theta_{ij}, \\
\Delta MH_2 &= P(C_i)(1-\theta_i)\log\left(\frac{1-\theta_i}{m-1}\right) + P(C_j)(1-\theta_j)\log\left(\frac{1-\theta_j}{m-1}\right) \\
&\quad - P(C_i \cup C_j)(1-\theta_{ij})\log\left(\frac{1-\theta_{ij}}{m-1}\right).
\end{aligned}$$

Actually, in all cases, the certainty of the merged condition class $C_i \cup C_j$ is maximized when the maximum decisions of the condition classes C_i, C_j and $C_i \cup C_j$ are the same, namely $MD(D|(C_i \cup C_j)) = MD(D|C_i) = MD(D|C_j)$. Therefore, the information gain (the uncertainty) is minimized when $\theta_{ij} = (|C_i|\theta_i + |C_j|\theta_j)/(|C_i| + |C_j|)$.

$$\begin{aligned}
\Delta MH_1 &= \frac{1}{|U|}\Big(|C_i|\theta_i\log\theta_i + |C_j|\theta_j\log\theta_j - (|C_i|\theta_i + |C_j|\theta_j) \\
&\quad \times \log\left(\frac{|C_i|\theta_i + |C_j|\theta_j}{|C_i| + |C_j|}\right)\Big) \\
&= \frac{1}{|U|}\left(|C_i|\theta_i\left(\log\theta_i - \log\left(\frac{|C_i|\theta_i + |C_j|\theta_j}{|C_i| + |C_j|}\right)\right)\right. \\
&\quad \left. + |C_j|\theta_j\left(\log\theta_j - \log\left(\frac{|C_i|\theta_i + |C_j|\theta_j}{|C_i| + |C_j|}\right)\right)\right) \\
&= \frac{1}{|U|}\left(|C_i|\theta_i\log\left(\frac{|C_i|\theta_i + |C_j|\theta_i}{|C_i|\theta_i + |C_j|\theta_j}\right) + |C_j|\theta_j\log\left(\frac{|C_i|\theta_j + |C_j|\theta_j}{|C_i|\theta_i + |C_j|\theta_j}\right)\right)
\end{aligned}$$

Let $|C_i|\theta_i = \phi$ and $|C_j|\theta_j = \varphi$, we have

$$\Delta MH_1(\phi, \varphi) = \frac{1}{|U|}\left(\phi\log\left(\frac{\phi + \varphi\frac{\theta_i}{\theta_j}}{\phi + \varphi}\right) + \varphi\log\left(\frac{\phi\frac{\theta_j}{\theta_i} + \varphi}{\phi + \varphi}\right)\right)$$

Let $\theta_i/\theta_j = \lambda$, we have

$$\Delta MH_1(\phi, \varphi, \lambda) = \frac{1}{|U|}\left(\phi\log\left(\frac{\phi + \varphi\lambda}{\phi + \varphi}\right) + \varphi\log\left(\frac{\phi\frac{1}{\lambda} + \varphi}{\phi + \varphi}\right)\right).$$

ΔMH_1 is an explicit function of the variables ϕ, φ and λ. In fact, we are mainly concerned with the interaction between the parameters θ_i and θ_j rather than a single parameter. Therefore, we only consider the partial derivative of ΔMH_1 with respect to the variable λ.

$$\begin{aligned}
\frac{\partial(\Delta MH_1(\phi, \varphi, \lambda))}{\partial(\lambda)} &= \frac{\log e}{|U|}\left(\phi\varphi\frac{\phi + \varphi}{\phi + \varphi\lambda} - \phi\varphi\frac{\phi + \varphi}{\lambda(\phi + \lambda\varphi)}\right) \\
&= \frac{\log e}{|U|}\left(\phi\varphi(\phi + \varphi)\frac{(\lambda - 1)}{\lambda(\phi + \lambda\varphi)}\right)\begin{cases} < 0, & 0 < \lambda < 1 \\ = 0, & \lambda = 1 \\ > 0, & \lambda > 1 \end{cases}
\end{aligned}$$

The information gain ΔMH_1 reaches the minima 0 when $\lambda = 1$, namely $\theta_i = \theta_j$. Therefore, $\Delta MH_1 \ge 0$ holds for all possible values of θ_i and θ_j.

As for the information gain ΔMH_2, we have

$$\begin{aligned}
\Delta MH_2 &= P(C_i)(1-\theta_i)\log\left(\frac{1-\theta_i}{m-1}\right) + P(C_j)(1-\theta_j)\log\left(\frac{1-\theta_j}{m-1}\right) \\
&\quad - P(C_i \cup C_j)(1-\theta_{ij})\log\left(\frac{1-\theta_{ij}}{m-1}\right) \\
&= \frac{1}{|U|}\left(|C_i|(1-\theta_i)\log\left(\frac{1-\theta_i}{m-1}\right) + |C_j|(1-\theta_j)\log\left(\frac{1-\theta_j}{m-1}\right)\right. \\
&\quad \left. - (|C_i|(1-\theta_i) + |C_j|(1-\theta_j))\log\left(\frac{|C_i| + |C_j| - |C_i|\theta_i - |C_j|\theta_j}{(|C_i| + |C_j|)(m-1)}\right)\right) \\
&= \frac{1}{|U|}\left(|C_i|(1-\theta_i)\log\left(\frac{|C_i|(1-\theta_i) + |C_j|(1-\theta_i)}{|C_i|(1-\theta_i) + |C_j|(1-\theta_j)}\right)\right. \\
&\quad \left. + |C_j|(1-\theta_j)\log\left(\frac{|C_i|(1-\theta_j) + |C_j|(1-\theta_j)}{|C_i|(1-\theta_i) + |C_j|(1-\theta_j)}\right)\right).
\end{aligned}$$

Let $|C_i|(1-\theta_i) = \phi$ and $|C_j|(1-\theta_j) = \varphi$, we have

$$\Delta MH_2(\phi, \varphi) = \frac{1}{|U|}\left(\phi\log\left(\frac{\phi + \varphi\frac{1-\theta_i}{1-\theta_j}}{\phi + \varphi}\right) + \varphi\log\left(\frac{\phi\frac{1-\theta_j}{1-\theta_i} + \varphi}{\phi + \varphi}\right)\right).$$

Let $(1-\theta_i)/(1-\theta_j) = \lambda$, we have

$$\Delta MH_2(\phi, \varphi, \lambda) = \frac{1}{|U|}\left(\phi\log\left(\frac{\phi + \varphi\lambda}{\phi + \varphi}\right) + \varphi\log\left(\frac{\phi\frac{1}{\lambda} + \varphi}{\phi + \varphi}\right)\right).$$

We obtain the similar expression as ΔMH_1, thus $\Delta MH_2 \ge 0$ and $\Delta MH = \Delta MH_1 + \Delta MH_2 \ge 0$. Additionally, ΔMH is minimized to 0 when $\theta_i = \theta_j = \theta_{ij}$.

II.$\theta_{ij} < \min\{\theta_i, \theta_j\}$

$$\begin{aligned}
\Delta MH &= -P(C_i \cup C_j)\left(\theta_{ij}\log\theta_{ij} + (m-1)\left(\frac{1-\theta_{ij}}{m-1}\right)\log\left(\frac{1-\theta_{ij}}{m-1}\right)\right) \\
&\quad + P(C_i)\left(\theta_i\log\theta_i + (m-1)\left(\frac{1-\theta_i}{m-1}\right)\log\left(\frac{1-\theta_i}{m-1}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&+P(C_j)\left(\theta_j\log\theta_j+(m-1)\left(\frac{1-\theta_j}{m-1}\right)\log\left(\frac{1-\theta_j}{m-1}\right)\right)\\
&=-\frac{1}{|U|}\left(\left(|C_i|+|C_j|\right)\left(\theta_{ij}\log\theta_{ij}+\left(1-\theta_{ij}\right)\log\left(\frac{1-\theta_{ij}}{m-1}\right)\right)\right.\\
&\quad-|C_i|\left(\theta_i\log\theta_i+(1-\theta_i)\log\left(\frac{1-\theta_i}{m-1}\right)\right)\\
&\quad\left.-|C_j|\left(\theta_j\log\theta_j+\left(1-\theta_j\right)\log\left(\frac{1-\theta_j}{m-1}\right)\right)\right)\\
&=-\frac{1}{|U|}\left(|C_i|\left(\left(\theta_{ij}\log\theta_{ij}+\left(1-\theta_{ij}\right)\log\left(\frac{1-\theta_{ij}}{m-1}\right)\right)\right.\right.\\
&\quad\left.-\left(\theta_i\log\theta_i+(1-\theta_i)\log\left(\frac{1-\theta_i}{m-1}\right)\right)\right)\\
&\quad+|C_j|\left(\left(\theta_{ij}\log\theta_{ij}+\left(1-\theta_{ij}\right)\log\left(\frac{1-\theta_{ij}}{m-1}\right)\right)\right.\\
&\quad\left.\left.-\left(\theta_j\log\theta_j+\left(1-\theta_j\right)\log\left(\frac{1-\theta_j}{m-1}\right)\right)\right)\right).
\end{aligned}
$$

The maximum decision entropy $MH(x)$ is monotonically decreasing within the range [maxima, 1](see Fig. 1), while the conditions $\theta_{ij}<\theta_i$ and $\theta_{ij}<\theta_j$ hold, thus $\Delta MH>0$.

III.$\theta_{ij}>\max\{\theta_i,\ \theta_j\}$

This case will not happen.

In all cases, the information gain is equal or greater than 0. Therefore, the maximum decision entropy is monotonically increasing when deleting a condition attribute.

The proofs of (2) and (3) are similar to that of (1). □

References

[1] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (5) (1982) 341–356.
[2] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992.
[3] Z. Pawlak, Rough sets and their applications, Physica-Verlag HD, Heidelberg, 2001, pp. 73–91.
[4] W. Ziarko, Probabilistic approach to rough sets, Int. J. Approximate Reasoning 49 (2) (2008) 272–284.
[5] S.K.M. Wong, W. Ziarko, Comparison of the probabilistic approximate classification and the fuzzy set model, Fuzzy Sets Syst. 21 (3) (1987) 357–362.
[6] Z. Pawlak, S.K.M. Wong, W. Ziarko, Rough sets: probabilistic versus deterministic approach, Int. J. Man Mach. Stud. 29 (1) (1988) 81–95.
[7] W. Ziarko, Variable precision rough set model, J. Comput. Syst. Sci. 46 (1) (1993) 39–59.
[8] Y.Y. Yao, S.K.M. Wong, A decision theoretic framework for approximating concepts, Int. J. Man Mach. Stud. 37 (6) (1992) 793–809.
[9] D. Slezak, W. Ziarko, The investigation of the bayesian rough set model, Int. J. Approximate Reasoning 40 (1) (2005) 81–91.
[10] Y.Y. Yao, Probabilistic rough set approximations, Int. J. Approximate Reasoning 49 (2) (2008) 255–271.
[11] J. Hu, T.R. Li, C. Luo, H. Fujita, S.Y. Li, Incremental fuzzy probabilistic rough sets over two universes, Int. J. Approximate Reasoning 81 (2017) 28–48.
[12] Y.Y. Yao, Probabilistic approaches to rough sets, Expert Syst. 20 (5) (2003) 287–297.
[13] Y.Y. Yao, Two semantic issues in a probabilistic rough set model, Fundam. Inf. 108 (3–4) (2011) 249–265.
[14] Y.Y. Yao, Three-way decisions and cognitive computing, Cognit. Comput. 8 (4) (2016) 543–554.
[15] Y.F. Chen, X.D. Yue, H. Fujita, S.Y. Fu, Three-way decision support for diagnosis on focal liver lesions, Knowl. Based Syst. 127 (2017) 85–99.
[16] Y.Y. Yao, The superiority of three-way decisions in probabilistic rough set models, Inf. Sci. 181 (6) (2011) 1080–1096.
[17] Y.Y. Yao, Three-way decisions with probabilistic rough sets, Inf. Sci. 180 (3) (2010) 341–353.
[18] K. Thangavel, A. Pethalakshmi, Dimensionality reduction based on rough set theory: a review, Appl. Soft Comput. 9 (1) (2009) 1–12.
[19] L.F. Li, J.K. Zhang, Attribute reduction in fuzzy concept lattices based on the t implication, Knowl. Based Syst. 23 (6) (2010) 497–503.
[20] H.M. Chen, T.R. Li, Y. Cai, C. Luo, H. Fujita, Parallel attribute reduction in dominance-based neighborhood rough set, Inf. Sci. 373 (2016) 351–368.
[21] Y.G. Jing, T.R. Li, C. Luo, S.J. Horng, G.Y. Wang, Z. Yu, An incremental approach for attribute reduction based on knowledge granularity, Knowl. Based Syst. 104 (2016) 24–38.
[22] Y.G. Jing, T.R. Li, H. Fujita, Z. Yu, B. Wang, An incremental attribute reduction approach based on knowledge granularity with a multi-granulation view, Inf. Sci. 411 (2017) 23–38.
[23] Z.H. Lai, W.K. Wong, Y. Xu, J. Yang, D. Zhang, Approximate orthogonal sparse embedding for dimensionality reduction, IEEE Trans. Neural Netw. Learn. Syst. 27 (4) (2016) 723–735.
[24] X.Y. Jia, L. Shang, B. Zhou, Y.Y. Yao, Generalized attribute reduct in rough set theory, Knowl. Based Syst. 91 (2016) 204–218.
[25] G.Y. Wang, X.A. Ma, H. Yu, Monotonic uncertainty measures for attribute reduction in probabilistic rough set model, Int. J. Approximate Reasoning 59 (2015) 41–67.
[26] 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.
[27] 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.
[28] X.A. Ma, G.Y. Wang, H. Yu, Heuristic method to attribute reduction for decision region distribution preservation, J. Software 25 (8) (2014) 1761–1780(inChinese).
[29] 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.
[30] Y.Y. Yao, Y. Zhao, Attribute reduction in decision-theoretic rough set models, Inf. Sci. 178 (17) (2008) 3356–3373.
[31] X.Y. Zhang, D.Q. Miao, Region-based quantitative and hierarchical attribute reduction in the two-category decision theoretic rough set model, Knowl. Based Syst. 71 (2014) 146–161.
[32] X.Y. Zhang, D.Q. Miao, Reduction target structure-based hierarchical attribute reduction for two-category decision-theoretic rough sets, Inf. Sci. 277 (2014) 755–776.
[33] 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.
[34] X.Y. Jia, Z.M. Tang, W.H. Liao, L. Shang, On an optimization representation of decision-theoretic rough set model, Int. J. Approximate Reasoning 55 (1, Part 2) (2014) 156–166.
[35] H. Yu, Y.Y. Yao, J. Zhao, An attribute reduction algorithm based on risk minimization, J. Nanjing Univ.(Natural Sciences) 49 (2) (2013) 210–216. (in Chinese)
[36] Z.Q. Bi, F.F. Xu, J.S. Lei, T. Jiang, Attribute reduction in decision-theoretic rough set model based on minimum decision cost, Concurrency Comput.: Pract. Exp. 28 (15) (2016) 4125–4143.
[37] S.J. Liao, Q.X. Zhu, F. Min, Cost-sensitive attribute reduction in decision-theoretic rough set models, Math. Prob. Eng. 35 (1) (2014) 1–9.
[38] Z.Q. Meng, Z.Z. Shi, On quick attribute reduction in decision-theoretic rough set models, Inf. Sci. 330 (2016) 226–244.
[39] H.L. Dou, X.B. Yang, X.N. Song, H.L. Yu, W.Z. Wu, J.Y. Yang, Decision-theoretic rough set: a multicost strategy, Knowl. Based Syst. 91 (2016) 71–83.
[40] G.P. Lin, J.Y. Liang, Y.H. Qian, J.J. Li, A fuzzy multigranulation decision-theoretic approach to multi-source fuzzy information systems, Knowl. Based Syst. 91 (2016) 102–113.
[41] W.H. Xu, Y.T. Guo, Generalized multigranulation double-quantitative decision-theoretic rough set, Knowl. Based Syst. 105 (2016) 190–205.
[42] J. Qian, C.Y. Dang, X.D. Yue, N. Zhang, Attribute reduction for sequential three--way decisions under dynamic granulation, Int. J. Approximate Reasoning 85 (2017) 196–216.
[43] Y.M. Zhang, X.Y. Jia, Z.M. Tang, Minimum cost attribute reduction in incomplete systems under decision-theoretic rough set model, in: 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, 2016, pp. 940–944.
[44] D. Liu, D.C. Liang, C.C. Wang, A novel three-way decision model based on incomplete information system, Knowl. Based Syst. 91 (2016) 32–45.
[45] Y.M. Chen, Z.Q. Zeng, Q.X. Zhu, C.H. Tang, Three-way decision reduction in neighborhood systems, Appl. Soft Comput. 38 (2016) 942–954.
[46] W.W. Li, Z.Q. Huang, X.Y. Jia, X.Y. Cai, Neighborhood based decision-theoretic rough set models, Int. J. Approximate Reasoning 69 (2016) 1–17.
[47] X.D. Yue, Y.F. Chen, D.Q. Miao, J. Qian, Tri-partition neighborhood covering reduction for robust classification, Int. J. Approximate Reasoning 83 (2017) 371–384.
[48] J.S. Mi, W.Z. Wu, W.X. Zhang, Approaches to knowledge reduction based on variable precision rough set model, Inf. Sci. 159 (3-4) (2004) 255–272.
[49] H.C. Peng, F.H. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[50] M. Lichman, UCI machine learning repository [http://archive.ics.uci.edu/ml], Technical Report, School of Information and Computer Science Irvine, University of California, 2013.
[51] E. Frank, M.A. Hall, I.H. Witten, The WEKA workbench, online appendix for "data mining: Practical machine learning tools and techniques", fourth ed., Morgan Kaufmann, 2016.
[52] Q.H. Hu, W. Pedrycz, D.R. Yu, J. Lang, Selecting discrete and continuous features based on neighborhood decision error minimization, IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) 40 (1) (2010) 137–150.
[53] D.Q. Miao, C. Gao, N. Zhang, Z.F. Zhang, Diverse reduct subspaces based co–training for partially labeled data, Int. J. Approximate Reasoning 52 (8) (2011) 1103–1117.