Information Sciences, 178(17), 3356-3373, Elsevier B.V., 2008.

Attribute Reduction in Decision-Theoretic Rough Set Models

Yiyu Yao, Yan Zhao

Department of Computer Science, University of Regina,
Regina, Saskatchewan, Canada S4S 0A2

Abstract

Rough set theory can be applied to rule induction. There are two different types of classification rules, positive and boundary rules, leading to different decisions and consequences. They can be distinguished not only from the syntax measures such as confidence, coverage and generality, but also the semantic measures such as decision-monotocity, cost and risk. The classification rules can be evaluated locally for each individual rule, or globally for a set of rules. Both the two types of classification rules can be generated from, and interpreted by, a decision-theoretic model, which is a probabilistic extension of the Pawlak rough set model.

As an important concept of rough set theory, an attribute reduct is a subset of attributes that are jointly sufficient and individually necessary for preserving a particular property of the given information table. This paper addresses attribute reduction in decision-theoretic rough set models regarding different classification properties, such as: decision-monotocity, confidence, coverage, generality and cost. It is important to note that many of these properties can be truthfully reflected by a single measure γ in the Pawlak rough set model. On the other hand, they need to be considered separately in probabilistic models. A straightforward extension of the γ measure is unable to evaluate these properties. This study provides a new insight into the problem of attribute reduction.

Key words: attribute reduction, decision-theoretic rough set model, Pawlak rough set model

1 Introduction

In recent years, researchers, motivated by a desire to represent information qualitatively, have proposed many models to incorporate probabilistic ap-

Email address: {yyao,yanzhao}@cs.uregina.ca (Yiyu Yao, Yan Zhao).

proaches into rough set theory, which was introduced by Pawlak in 1982 [22,24,26,27]. The proposals include probabilistic rough set models [12,20,28,40,44,45], decision-theoretic rough set models [43,47,48], variable precision rough set models [52], rough membership functions [25], parameterized rough set models [27,30], and Bayesian rough set models [8,9,32,33]. All these proposals share the common feature by introducing thresholds into the standard model. For example, the decision-theoretic rough set models was proposed in the early 1990s, in order to generalize the probabilistic rough set model [22]. The decision-theoretic models systematically calculate the parameters based on a set of loss functions based on the Bayesian decision procedure. The physical meaning of the loss functions can be interpreted based on more practical notions of costs and risks. The results of these studies increase our understanding of rough set theory and its domain of applications.

The results drawn from rough set based classification can be used for decision making. In the Pawlak model, one can have two types of rules, positive rules and boundary rules [45]. A positive rule indicates that an object or an object set for sure belongs to one decision class, which enables us to make a positive decision. A boundary rule indicates that an object or an object set partially belongs to the decision class, which leads to another type of decision. In a probabilistic rough set model, one can also have positive rules and boundary rules. A probabilistic positive rule expresses that an object or an object set belongs to one decision class beyond a certain confidence threshold. A probabilistic boundary rule expresses that an object or an object set belongs to one decision class beyond another weaker confidence threshold. Besides these two types of rules, there is another situation, such that one cannot indicate to which decision class the object or the object set belongs, since the confidence is too low to support any decision making. The probabilistic positive and boundary rules can be distinguished not only by the syntax measures, such as confidence, coverage and generality, but also the semantics measures, such as decision-monotocity, cost and risk. The syntax properties focus on the discovery of the rules, while the semantics properties focus on the utilization of the rules, and thus are more practical for the real applications. Measures regarding the semantics properties are less studied in the rough set literature.

The theory of rough sets has been applied to data analysis, data mining and knowledge discovery. A fundamental notion supporting such applications is the concept of attribute reduction [22]. The objective of reduct construction is to reduce the number of attributes, and at the same time, preserve a certain property that we want. Different algorithms, approaches and methodologies have been extensively studied [2,4,11,13,17,20,29,35,41,51]. Suppose we are interested in the property of concept classification. A reduct should be able to preserve the original classification power provided by the whole attribute set. This power may be interpreted by syntax properties and semantics properties for both positive and boundary rule sets.

For the Pawlak model, a single measure γ is suggested for evaluating the performance of classification and attribute reduction. For a probabilistic model, by introducing the probabilistic thresholds, the properties are not necessarily monotonic with respect to the set inclusion, and cannot be evaluated by a single measure. Instead, we need to consider multiple properties and multiple measures for evaluation. More specifically, this paper addresses different criteria, such as confidence, coverage, generality, cost, and decision-monotocity criteria based on the decision-theoretic rough set models.

2 The Pawlak Rough Set Model

In many data analysis applications, information and knowledge are stored and represented in an information table, where a set of objects are described by a set of attributes [22]. An information table represents all available information and knowledge. That is, objects are only perceived, observed, or measured by using a finite number of attributes.

Definition 1 *An information table is the following tuple:*

$$S = (U, At, \{V_a \mid a \in At\}, \{I_a \mid a \in At\}),$$

where U is a finite nonempty set of objects, At is a finite nonempty set of attributes, V_a is a nonempty set of values of $a \in At$, and $I_a : U \rightarrow V_a$ is an information function that maps an object in U to exactly one value in V_a.

In classification problems, we consider an information table of the form $S = (U, At = \mathbf{C} \cup \{D\}, \{V_a\}, \{I_a\})$, where $\mathbf{C}$ is a set of condition attributes describing the objects, and D is a decision attribute that indicates the classes of objects. In general, we may have a set of decision attributes. A table with multiple decision attributes can be easily transformed into a table with a single decision attribute by considering the Cartesian product of the original decision attributes.

An equivalence relation with respect to $A \subseteq At$ is denoted as E_A, or simply E. That is, $E_A = \{(x, y) \in U \times U \mid \forall a \in A(I_a(x) = I_a(y))\}$. Two objects in U satisfy E_A if and only if they have the same values on all attributes in A. An equivalence relation is reflexive, symmetric and transitive.

The pair (U, E_A) is called an approximation space defined by the attribute set A. The equivalence relation E_A induces a partition of U, denoted by U/E_A or π_A. The equivalence class of U/E_A containing x is given by $[x]_{E_A} = [x]_A = \{y \in U \mid (x, y) \in E_A\}$, or $[x]$ if E_A is understood.

Consider an equivalence relation E on U. The equivalence classes induced by the partition π (i.e., U/E) are the basic blocks to construct the Pawlak rough set approximations. For a subset $X \subseteq U$, the lower and upper approximations of X with respect to π are define by [22]:

$$\begin{aligned}
\underline{apr}_\pi(X) &= \{x \in U \mid [x] \subseteq X\}, \\
&= \{x \in U \mid P(X \mid [x]) = 1\}; \\
\overline{apr}_\pi(X) &= \{x \in U \mid [x] \cap X \neq \emptyset\}, \\
&= \{x \in U \mid P(X \mid [x]) > 0\},
\end{aligned} \tag{1}$$

where $P(X \mid [x])$ denotes the conditional probability that an object x belongs to X given that the object is in the equivalence class $[x]$, i.e., $P(X \mid [x]) = \frac{|[x] \cap X|}{|[x]|}$.

Based on the rough set approximations of X defined by π, one can divide the universe U into three disjoint regions: the positive region $\mathrm{POS}_\pi(X)$ indicating the union of all the equivalence classes defined by π that each for sure can induce the decision class X; the boundary region $\mathrm{BND}_\pi(X)$ indicating the union of all the equivalence classes defined by π that each can induce a partial decision of X; and the negative region $\mathrm{NEG}_\pi(X)$ which is the union of all equivalence classes that for sure cannot induce the decision class X [22]:

$$\begin{aligned}
\mathrm{POS}_\pi(X) &= \underline{apr}_\pi(X), \\
\mathrm{BND}_\pi(X) &= \overline{apr}_\pi(X) - \underline{apr}_\pi(X), \\
\mathrm{NEG}_\pi(X) &= U - \mathrm{POS}_\pi(X) \cup \mathrm{BND}_\pi(X) = U - \overline{apr}_\pi(X) = (\overline{apr}_\pi(X))^c
\end{aligned} \tag{2}$$

Let $\pi_D = \{D_1, D_2, \ldots, D_m\}$ be a partition of the universe U, defined by the decision attribute D, representing m classes, where $m = |V_D|$. The lower and upper approximations of the partition π_D with respect to π are the families of the lower and upper approximations of all the equivalence classes of π_D. That is [23],

$$\begin{aligned}
\underline{apr}_\pi(\pi_D) &= (\underline{apr}_\pi(D_1), \underline{apr}_\pi(D_2), \ldots, \underline{apr}_\pi(D_m)); \\
\overline{apr}_\pi(\pi_D) &= (\overline{apr}_\pi(D_1), \overline{apr}_\pi(D_2), \ldots, \overline{apr}_\pi(D_m)).
\end{aligned} \tag{3}$$

For this m-class problem, we can solve it in terms of m two-class problems. Then, $\mathrm{POS}_\pi(\pi_D)$ indicates the union of all the equivalence classes defined by π that each for sure can induce a decision. $\mathrm{BND}_\pi(\pi_D)$ indicates the union of all the equivalence classes defined by π that each can induce a partial decision. Formally, we have [45]:

$$\begin{aligned}
\mathrm{POS}_\pi(\pi_D) &= \bigcup_{1\le i\le m} \mathrm{POS}_\pi(D_i),\\
\mathrm{BND}_\pi(\pi_D) &= \bigcup_{1\le i\le m} \mathrm{BND}_\pi(D_i),\\
\mathrm{NEG}_\pi(\pi_D) &= U - \mathrm{POS}_\pi(\pi_D) \cup \mathrm{BND}_\pi(\pi_D).
\end{aligned} \tag{4}$$

We can easily verify the following properties of the three regions in the Pawlak model:

(1) The three regions are pairwise disjoint, and the union is a covering of U. Furthermore, $\mathrm{POS}_\pi(\pi_D) \cap \mathrm{BND}_\pi(\pi_D) = \emptyset$ and $\mathrm{POS}_\pi(\pi_D) \cup \mathrm{BND}_\pi(\pi_D) = U$. That means, for any equivalence class in π, it can either make a sure decision or a partial decision. Thus, $\mathrm{NEG}_\pi(\pi_D) = \emptyset$.
(2) For an equivalence class in $\mathrm{POS}_\pi(\pi_D)$, it associates with at most one decision class $D_i \in \pi_D$. The family of positive regions $\{\mathrm{POS}_\pi(D_i) \mid 1 \le i \le m\}$ contains pairwise disjoint sets, i.e., $\mathrm{POS}_\pi(D_i) \cap \mathrm{POS}_\pi(D_j) = \emptyset$, for any $i \ne j$.
(3) For an equivalence class in $\mathrm{BND}_\pi(\pi_D)$, it associates with at least two decision classes $D_i, D_j \in \pi_D$. The family of boundary regions $\{\mathrm{BND}_\pi(D_i) \mid 1 \le i \le m\}$ does not necessarily contain pairwise disjoint sets, i.e., it may happen that $\mathrm{BND}_\pi(D_i) \cap \mathrm{BND}_\pi(D_j) \ne \emptyset$, for some $i \ne j$.

An information table is consistent if each equivalence class defined by $\mathbf{C}$ decides a unique decision. In this case, $\mathrm{BND}_{\pi_\mathbf{C}}(\pi_D) = \mathrm{NEG}_{\pi_\mathbf{C}}(\pi_D) = \emptyset$ and $\mathrm{POS}_{\pi_\mathbf{C}}(\pi_D) = U$. An inconsistent information table contains at least one equivalence class $[x]_\mathbf{C} \in \pi_\mathbf{C}$, such that it associates with more than one decision.

3 Decision-Theoretic Rough Set Models

A decision-theoretic rough set model brings new insights into the probabilistic rough set approaches. The Bayesian decision procedure deals with making decisions with minimum risk based on observed evidence. We present a brief description of the procedure from the book by Duda and Hart [6]. Different probabilistic models can be easily derived from the decision-theoretic model.

3.1 The Bayesian decision procedure

Given an object x, let $\mathbf{x}$ be a description of the object, $\Omega = \{w_1, \ldots, w_s\}$ be a finite set of s states that x is possibly in, and $\mathcal{A} = \{a_1, \ldots, a_t\}$ be a finite set of t possible actions. Let $P(w_j \mid \mathbf{x})$ be the conditional probability of x being

	a_1	a_2	...	a_i	...	a_t
w_1	$\lambda(a_1\|w_1)$	$\lambda(a_2\|w_1)$	...	$\lambda(a_i\|w_1)$	...	$\lambda(a_t\|w_1)$
w_2	$\lambda(a_1\|w_1)$	$\lambda(a_2\|w_2)$	...	$\lambda(a_i\|w_2)$	...	$\lambda(a_t\|w_2)$
...						
w_j	$\lambda(a_1\|w_j)$	$\lambda(a_2\|w_j)$	...	$\lambda(a_i\|w_j)$	...	$\lambda(a_t\|w_j)$
...						
w_s	$\lambda(a_1\|w_s)$	$\lambda(a_2\|w_s)$	...	$\lambda(a_i\|w_s)$	...	$\lambda(a_t\|w_s)$

Table 1
The $s \times t$ matrix for all the values of loss functions

in state w_j, and the loss function $\lambda(a_i|w_j)$ denote the loss (or cost) for taking the action a_i when the state is w_j.

All the values of loss functions can be conveniently expressed as an $s \times t$ matrix illustrated in Table 1, with the rows denoting the set Ω of s states and the columns the set $\mathcal{A}$ of t actions. Each cell denotes the cost $\lambda(a_i|w_j)$ for taking the action a_i in the state w_j. The cost $\lambda(a_i|w_j)$ can be written as $\lambda_{a_i w_j}$ for simplicity.

For an object x with description $\mathbf{x}$, suppose action a_i is taken. The expected cost associated with action a_i is given by:

$$\mathcal{R}(a_i \mid \mathbf{x}) = \sum_{j=1}^{s} \lambda(a_i|w_j) P(w_j \mid \mathbf{x}). \tag{5}$$

The quantity $\mathcal{R}(a_i \mid \mathbf{x})$ is called the conditional risk.

The $s \times t$ matrix has two important applications. First, given the loss functions and the probabilities, one can compute the expected cost of a certain action. Furthermore, comparing the expected costs of all the actions, one can decide a particular action with the minimum cost. Second, according to the loss functions, one can determine the condition or probability for taking a particular action.

Example 1 *The idea of the Bayesian decision procedure can be demonstrated by the following example. Suppose there are two states:* w_1 *indicates that a meeting will be over in less than or equal to 2 hours, and* w_2 *indicates that the meeting will be over in more than 2 hours. Two states are complement. Suppose the probability for having the state* w_1 *is* 0.80*, then the probability for having the state* w_2 *is* 0.20*, i.e.,* $P(w_1 \mid \mathbf{x}) = 0.80$ *and* $P(w_2 \mid \mathbf{x}) = 1 - 0.80 = 0.20$*. There are two actions:* a_1 *means to park the car on meter, and* a_2 *means to park the car in the parking lot. The loss functions for taking different actions in different states can be expressed as the following matrix:*

	a_1 *(park on meter)*	a_2 *(park in a parking lot)*
w_1 *(≤ 2 hours)*	*\$2.00*	*\$7.00*
w_2 *(> 2 hours)*	*\$12.00*	*\$7.00*

In this case, the cost for each action can be calculated as follows:

$$\mathcal{R}(a_1 \mid \mathbf{x}) = \$2.00 * 0.80 + \$12.00 * 0.20 = \$3.00,$$
$$\mathcal{R}(a_2 \mid \mathbf{x}) = \$7.00 * 0.80 + \$7.00 * 0.20 = \$7.00.$$

Since $\$3.00 < \$7.00$*, according to the minimum cost, one may decide to park the car on meter, instead of in a parking lot.*

Suppose a person wants to decide where to park the car. It is interesting to know if parking on meter is more suitable. According to Equation (5), one obtains:

$$\$2.00 * P(w_1 \mid \mathbf{x}) + \$12.00 * (1 - P(w_1 \mid \mathbf{x})) \leq \$7.00.$$

That is, $P(w_1 \mid \mathbf{x}) \geq 0.50$*. Thus, if the probability that the meeting is over within 2 hours is greater than or equal to 0.50, then it is more profitable to park the car on meter, otherwise, park in a parking lot.*

3.2 Decision-theoretic rough set models

In an approximation space (U, E), the equivalence relation E induces a partition $\pi = U/E$. Let $Des([x])$ denote the description of x. For simplicity, we write $Des([x])$ as $[x]$ in the subsequent discussions. The partition π is the set of all possible descriptions. The classification of objects can be easily fitted into the Bayesian decision framework. The set of states is given by $\Omega = \pi_D = \{X, X^c\}$, indicating that an object is in a decision class X and not in X, respectively. We use the same symbol to denote both a subset X and the corresponding state. The probabilities for these two complement states are denoted as $P(X \mid [x]) = \frac{|X \cap [x]|}{|[x]|}$ and $P(X^c \mid [x]) = 1 - P(X \mid [x])$.

With respect to the three regions defined by a partition π, the set of actions regarding the state X is given by $\mathcal{A} = \{a_P, a_N, a_B\}$, where a_P, a_N and a_B represent the three actions of deciding an object to be in the sets $\mathrm{POS}_\pi(X)$, $\mathrm{NEG}_\pi(X)$ and $\mathrm{BND}_\pi(X)$, respectively. When an object belongs to X, let λ_{PP}, λ_{BP} and λ_{NP} denote the costs of taking the actions a_P, a_B and a_N, respectively. When an object does not belong to X, let λ_{PN}, λ_{BN} and λ_{NN} denote the costs

of taking the same three actions. The loss functions regarding the states X and X^c can be expressed as a 2×3 matrix as follows:

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

The expected costs $\mathcal{R}(a_i \mid [x])$ of taking individual actions can be expressed as:

$$\begin{aligned}
\mathcal{R}(a_P \mid [x]) &= \lambda_{PP} P(X \mid [x]) + \lambda_{PN} P(X^c \mid [x]), \\
\mathcal{R}(a_N \mid [x]) &= \lambda_{NP} P(X \mid [x]) + \lambda_{NN} P(X^c \mid [x]), \\
\mathcal{R}(a_B \mid [x]) &= \lambda_{BP} P(X \mid [x]) + \lambda_{BN} P(X^c \mid [x]).
\end{aligned} \tag{6}$$

The Bayesian decision procedure leads to the following minimum-risk decision rules:

(P) If $\mathcal{R}(a_P \mid [x]) \leq \mathcal{R}(a_N \mid [x])$ and $\mathcal{R}(a_P \mid [x]) \leq \mathcal{R}(a_B \mid [x])$, decide $[x] \subseteq \mathrm{POS}_\pi(X)$;
(N) If $\mathcal{R}(a_N \mid [x]) \leq \mathcal{R}(a_P \mid [x])$ and $\mathcal{R}(a_N \mid [x]) \leq \mathcal{R}(a_B \mid [x])$, decide $[x] \subseteq \mathrm{NEG}_\pi(X)$;
(B) If $\mathcal{R}(a_B \mid [x]) \leq \mathcal{R}(a_P \mid [x])$ and $\mathcal{R}(a_B \mid [x]) \leq \mathcal{R}(a_N \mid [x])$, decide $[x] \subseteq \mathrm{BND}_\pi(X)$.

Tie-breaking criteria should be added so that each object is classified into only one region. Since for any state X, $P(X \mid [x]) + P(X^c \mid [x]) = 1$, we can simplify the rules to classify any object x based only on the probability $P(X \mid [x])$ and the loss functions.

Consider a special kind of loss functions with $\lambda_{PP} \leq \lambda_{BP} < \lambda_{NP}$ and $\lambda_{NN} \leq \lambda_{BN} < \lambda_{PN}$. That is, the cost of classifying an object x into the positive region $\mathrm{POS}_\pi(X)$ is less than or equal to the cost of classifying x into the boundary region $\mathrm{BND}_\pi(X)$, and both of these costs are strictly less than the cost of classifying x into the negative region $\mathrm{NEG}_\pi(X)$. The reverse order of costs is used for classifying an object that does not belong to X. This assumption implies that $\alpha \in (0, 1]$, $\gamma \in (0, 1)$, and $\beta \in [0, 1)$. In this case, the minimum-risk decision rules (P)-(B) can be written as:

(P) If $P(X \mid [x]) \geq \gamma$ and $P(X \mid [x]) \geq \alpha$, decide $[x] \subseteq \mathrm{POS}_\pi(X)$;
(N) If $P(X \mid [x]) \leq \beta$ and $P(X \mid [x]) \leq \gamma$, decide $[x] \subseteq \mathrm{NEG}_\pi(X)$;
(B) If $P(X \mid [x]) \geq \beta$ and $P(X \mid [x]) \leq \alpha$, decide $[x] \subseteq \mathrm{BND}_\pi(X)$,

where

$$\alpha = \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})},$$
$$\gamma = \frac{\lambda_{PN} - \lambda_{NN}}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})},$$
$$\beta = \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}. \tag{7}$$

When $(\lambda_{PN} - \lambda_{BN})(\lambda_{NP} - \lambda_{BP}) > (\lambda_{BP} - \lambda_{PP})(\lambda_{BN} - \lambda_{NN})$, we have $\alpha > \beta$, and thus $\alpha > \gamma > \beta$. After tie-breaking, we obtain:

(P1) If $P(X \mid [x]) \geq \alpha$, decide $[x] \subseteq \mathrm{POS}_{\pi}(X)$;
(N1) If $P(X \mid [x]) \leq \beta$, decide $[x] \subseteq \mathrm{NEG}_{\pi}(X)$;
(B1) If $\beta < P(X \mid [x]) < \alpha$, decide $[x] \subseteq \mathrm{BND}_{\pi}(X)$.

After computing the two parameters α and β from the loss functions, the probabilistic lower and upper approximations can be defined by:

$$\underline{apr}_{\pi_{(\alpha,\beta)}}(X) = \{x \in U \mid P(X \mid [x]) \geq \alpha\},$$
$$\overline{apr}_{\pi_{(\alpha,\beta)}}(X) = \{x \in U \mid P(X \mid [x]) > \beta\}. \tag{8}$$

The probabilistic positive, boundary and negative regions are defined by:

$$\begin{aligned}
\mathrm{POS}_{\pi_{(\alpha,\beta)}}(X) &= \underline{apr}_{\pi_{(\alpha,\beta)}}(X), \\
\mathrm{BND}_{\pi_{(\alpha,\beta)}}(X) &= \overline{apr}_{\pi_{(\alpha,\beta)}}(X) - \underline{apr}_{\pi_{(\alpha,\beta)}}(X), \\
\mathrm{NEG}_{\pi_{(\alpha,\beta)}}(X) &= U - \mathrm{POS}_{\pi_{(\alpha,\beta)}}(X) \cup \mathrm{BND}_{\pi_{(\alpha,\beta)}}(X) \\
&= U - \overline{apr}_{\pi_{(\alpha,\beta)}}(X) = (\overline{apr}_{\pi_{(\alpha,\beta)}}(X))^c.
\end{aligned} \tag{9}$$

Similar to the Pawlak rough set model, we can extend the concept of probabilistic approximations and regions of a single decision to a partition π_D. For simplicity, we assume that the same loss functions are used for all decisions. That is,

$$\underline{apr}_{\pi_{(\alpha,\beta)}(\pi_D)} = (\underline{apr}_{\pi_{(\alpha,\beta)}}(D_1), \underline{apr}_{\pi_{(\alpha,\beta)}}(D_2), \ldots, \underline{apr}_{\pi_{(\alpha,\beta)}}(D_m));$$
$$\overline{apr}_{\pi_{(\alpha,\beta)}(\pi_D)} = (\overline{apr}_{\pi_{(\alpha,\beta)}}(D_1), \overline{apr}_{\pi_{(\alpha,\beta)}}(D_2), \ldots, \overline{apr}_{\pi_{(\alpha,\beta)}}(D_m)). \tag{10}$$

We can define the three regions of the partition π_D for the probabilistic rough set models [45]:

$$\begin{aligned}
\mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D) &= \bigcup_{1\leq i\leq m} \mathrm{POS}_{\pi_{(\alpha,\beta)}}(D_i),\\
\mathrm{BND}_{\pi_{(\alpha,\beta)}}(\pi_D) &= \bigcup_{1\leq i\leq m} \mathrm{BND}_{\pi_{(\alpha,\beta)}}(D_i),\\
\mathrm{NEG}_{\pi_{(\alpha,\beta)}}(\pi_D) &= U - \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D) \cup \mathrm{BND}_{\pi_{(\alpha,\beta)}}(\pi_D).
\end{aligned} \tag{11}$$

We can verify the following properties of the three regions in probabilistic models:

(1) The three regions are not necessarily pairwise disjoint. Nevertheless, the union is a covering of U, i.e., $\mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D) \cup \mathrm{BND}_{\pi_{(\alpha,\beta)}}(\pi_D) \cup \mathrm{NEG}_{\pi_{(\alpha,\beta)}}(\pi_D) = U$. Furthermore, it may happen that $\mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D) \cap \mathrm{BND}_{\pi_{(\alpha,\beta)}}(\pi_D) \neq \emptyset$, and $\mathrm{NEG}_{\pi_{(\alpha,\beta)}}(\pi_D)$ is not necessarily empty.
(2) The family of probabilistic positive regions $\{\mathrm{POS}_{\pi_{(\alpha,\beta)}}(D_i) \mid 1 \leq i \leq m\}$ does not necessarily contain pairwise disjoint sets, i.e., it may happen that $\mathrm{POS}_{\pi_{(\alpha,\beta)}}(D_i) \cap \mathrm{POS}_{\pi_{(\alpha,\beta)}}(D_j) \neq \emptyset$, for some $i \neq j$.
(3) The family of probabilistic boundary regions $\{\mathrm{BND}_{\pi_{(\alpha,\beta)}}(D_i) \mid 1 \leq i \leq m\}$ does not necessarily contain pairwise disjoint sets, i.e., it may happen that $\mathrm{BND}_{\pi_{(\alpha,\beta)}}(D_i) \cap \mathrm{BND}_{\pi_{(\alpha,\beta)}}(D_j) \neq \emptyset$, for some $i \neq j$.

The Pawlak model, as a special case, can be derived from the general probabilistic model by having $(\alpha = 1) > (\beta = 0)$, and $\alpha = 1 - \beta$ [28]. From decision rules (P1)-(B1), we can compute the approximations as $\underline{apr}_{\pi_{(1,0)}}(\pi_D) = \mathrm{POS}_{\pi_{(1,0)}}(\pi_D)$ and $\overline{apr}_{\pi_{(1,0)}}(\pi_D) = \mathrm{POS}_{\pi_{(1,0)}}(\pi_D) \cup \mathrm{BND}_{\pi_{(1,0)}}(\pi_D) = U$.

We can derive the 0.50 probabilistic model [28], the symmetric variable precision rough set model [52], and the asymmetric variable precision rough set model [14]. More specifically, we may have the following probabilistic rough set models [45]:

- If $\alpha > 0.50$, $\mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)$ contains pairwise disjoint sets.
- If $\beta > 0.50$, $\mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)$, $\mathrm{BND}_{\pi_{(\alpha,\beta)}}(\pi_D)$ and $\mathrm{NEG}_{\pi_{(\alpha,\beta)}}(\pi_D)$ contain pairwise disjoint sets.
- If $\beta = 0$, $\mathrm{NEG}_{\pi_{(\alpha,\beta)}}(\pi_D) = \emptyset$.

When generalizing results from the Pawlak rough set model to the probabilistic rough set models, it is necessary to consider the implications of those properties.

Example 2 *Consider an example discussed by Yao and Herbert [42]. Suppose there are two complementary states after a series of diagnoses for a certain*

type of cancer: w_C is a confirmed cancer state and w_H is a confirmed no-cancer state, thus $w_H = w_C^c$. There are three actions regarding the three regions of the decision: a_P is to take some cancer-treatments to a patient, a_B is to wait-and-see when the decision is pending, and a_N is to discharge the patient without any further treatment.

The loss function for taking an action should include the cost and risk of further testing, follow-up diagnoses, treatments, and the cost of the corresponding results. For example, the loss function $\lambda(a_P|w_C)$ indicates the cost of taking proper treatments for a confirmed patient and the price of postoperative effects. The loss function $\lambda(a_B|w_C)$ indicates the risk of the potential delay of the proper treatment to a cancer patient. The loss function $\lambda(a_N|w_H)$, indicating to discharge a no-cancer patient, contains very little cost. Suppose one can estimate all the values of the loss functions and express them in the following matrix:

	a_P *(treat)*	a_B *(wait-and-see)*	a_N *(discharge)*
w_C *(cancer)*	*\$1200.00*	*\$1500.00*	*\$3500.00*
w_H *(no-cancer)*	*\$2500.00*	*\$1000.00*	*\$0*

According to the given matrix loss functions, we can calculate the values of the two thresholds α and β according to Equation (7):

$$\alpha = \frac{\$2500.00 - \$1000.00}{(\$1500.00 - \$1000.00) - (\$1200.00 - \$2500.00)} = 0.83$$
$$\beta = \frac{\$1000.00 - \$0}{(\$3500.00 - \$0) - (\$1500.00 - \$1000.00)} = 0.33.$$

4 Rule Induction

One of the important applications of rough set theory is to induce decision or classification rules. In this section, we consider two related issues. The first issue is the form and interpretation of rules. Two different types of classification rules are introduced and examined. The second issue is the evaluation of a single rule and a set of rules. The evaluation is investigated by considering the local evaluation of each single rule and the global evaluation of a set of rules.

4.1 Two types of classification rules

Typically, a rule in rough set theory is expressed in the form of $[x] \longrightarrow D_i$, stating that an object with description $[x]$ would be in the decision class D_i. Based on the notions of positive and boundary regions, we may introduce two types of rules [45]. One type is called positive rules and the other is called boundary rules.

Consider a partition π defined by a subset of condition attributes and the partition $\pi_D = \{D_1, D_2, \ldots, D_m\}$ defined by the decision attribute. For any $[x] \in \pi$, one can induce one of the following two types of classification rules [45]:

- Positive rule: If $[x] \subseteq \text{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)$, the induced rule is a positive rule, denoted as:

$$[x] \longrightarrow_P D_i, \text{ where } D_i \in \pi_D \text{ and } [x] \subseteq \text{POS}_{\pi_{(\alpha,\beta)}}(D_i).$$

- Boundary rule: If $[x] \subseteq \text{BND}_{\pi_{(\alpha,\beta)}}(\pi_D)$, the induced rule is a boundary rule, denoted as:

$$[x] \longrightarrow_B D_i, \text{ where } D_i \in \pi_D \text{ and } [x] \subseteq \text{BND}_{\pi_{(\alpha,\beta)}}(D_i).$$

In the Pawlak model, we have $\alpha = 1$ and $\beta = 0$. In probabilistic models, we require $\alpha > \beta$. Both models can generate these two types of rules.

Although the two types of rules have the same form and are characterized by the same quantitative measures, they have different interpretations, and hence lead to different decisions and actions. For example, Yao and Herbert suggest that an "immediate positive decision" is made based on a positive rule, and a "delayed positive decision" is made based on a boundary rule [10,42]. Regarding the previous medical example, a positive rule $[x] \longrightarrow_P w_C$ means that treatment should be applied immediately to a patient with a high probability of having cancer. A positive rule $[x] \longrightarrow_P w_H$ results in the discharge of a patient with a high probability of not having cancer (i.e., low probability of having cancer). A boundary rule, in forms of $[x] \longrightarrow_B w_C$ or $[x] \longrightarrow_B w_H$, means that the doctor may put the patient in a wait-and-see status requiring further diagnoses and investigations.

Another example is the academic paper review process. A positive rule means a paper is accepted or rejected right away. A boundary rule means a paper requires minor or major revisions, and the final acceptance/rejection decision is pending.

In the induction and utilization of rules, we in fact consider two slightly different types of decisions or actions. One decision determines the region to which an equivalence class belongs. According to the decision-theoretic model, we

can determine if an equivalence class $[x]$ belongs to the positive region of a decision class D_i, i.e., $[x] \subseteq \text{POS}_\pi(D_i)$, or $[x]$ belongs to the boundary region of D_i, i.e., $[x] \subseteq \text{BND}_\pi(D_i)$. Another kind of decision is to determine the action resulted from a rule. According to a positive rule $[x] \longrightarrow_P D_i$, we can determine a positive action towards the decision class D_i; according to a boundary rule, we can determine a pending action towards D_i. Since the former decisions determine the latter decisions, we use these two types of decisions interchangeably in this paper.

4.2 Single rule evaluation

Many quantitative measures associated with rules have been studied [3,5,46,49]. We review some measures for single rule (local) evaluation.

Confidence: Given a rule $[x] \longrightarrow D_i$, the confidence measure is defined as the ratio of the number of objects in an equivalence class $[x]$ that are correctly classified as the decision class D_i and the number of objects in the equivalence class $[x]$:

$$\begin{aligned} \textit{confidence}([x] \longrightarrow D_i) &= \frac{\#\text{ of objects in } [x] \text{ correctly classified as } D_i}{\#\text{ of objects in } [x]} \\ &= \frac{|[x] \cap D_i|}{|[x]|} = P(D_i \mid [x]), \end{aligned} \tag{12}$$

where $|.|$ denotes the cardinality of the set. Confidence focuses on the classification of an equivalence class $[x]$. The higher the confidence, the stronger the rule is.

The confidence measure is directly associated with the thresholds α and β. That is, the confidence of a positive rule is greater than or equal to α. A positive rule can be a certain rule with confidence being 1, or a probabilistic rule with confidence in $[\alpha, 1)$. For an equivalence class $[x] \subseteq \text{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)$, if $\alpha > 0.50$, it induces only one positive rule, and if $\alpha \leq 0.50$, it may induce more than one positive rule. The confidence of a boundary rule is greater than β and less than α. For an equivalence class $[x] \subseteq \text{BND}_{\pi_{(\alpha,\beta)}}(\pi_D)$, it induces only one boundary rule if $\beta > 0.50$, and may induce more than one boundary rule if $\beta \leq 0.50$. If $[x] \subseteq \text{NEG}_{\pi_{(\alpha,\beta)}}(\pi_D)$, the rule with the confidence less than β is too weak to be meaningful, and does not support any action towards a decision class of π_D.

Coverage: The coverage measure of a rule is defined as the ratio of the number of correctly classified objects in the decision class D_i by an equivalence class $[x]$ and the number of objects in the decision class D_i:

$$\begin{aligned} \textit{coverage}([x] \longrightarrow D_i) &= \frac{\text{\# of objects in } [x] \text{ correctly classified as } D_i}{\text{\# of objects in } D_i} \\ &= \frac{|[x] \cap D_i|}{|D_i|} = P([x] \mid D_i). \end{aligned} \tag{13}$$

Coverage focuses on the recall of a decision class $D_i \in \pi_D$ by $[x]$. A rule with a higher coverage is more general with respect to the decision class D_i.

In general, a high confidence rule is not necessarily a low coverage rule, and a high coverage rule is not necessarily a low confidence rule. In many situations, however, there may exist an inverse relationship between confidence and coverage. A reduction of confidence may lead to an increase of coverage. Such a relationship in fact is one of the motivations for the study of probabilistic rough set models. By weakening the requirement of confidence being 1 in the Pawlak positive rules, one expects to increase the coverage of probabilistic positive rules.

Generality: The generality of a rule is the ratio of the number of objects to which the rule can be applied and the total number of objects in the universe. It only tells us the degree of applicability of the rule, and does not say anything about its confidence nor its coverage. The generality measure can be denoted as:

$$\begin{aligned} \textit{generality}([x] \longrightarrow D_i) &= \frac{\text{\# of objects in } [x]}{\text{\# of objects in } U} \\ &= \frac{|[x]|}{|U|}. \end{aligned} \tag{14}$$

Cost: A positive rule $[x] \longrightarrow_P D_i$ decides that all the objects in $[x]$ are put into the positive region of the decision class D_i with confidence greater than or equal to α. If a positive action a_P of D_i is taken for $[x]$, the corresponding expected cost of applying the positive rule can be calculated as follows:

$$\begin{aligned} \mathcal{R}([x] \longrightarrow_P D_i) &= \mathcal{R}(a_P^{D_i} \mid [x]) \\ &= \lambda_{PP} P(D_i \mid [x]) + \lambda_{PN} P(D_i^c \mid [x]) \\ &= \textit{confidence}([x] \longrightarrow_P D_i)\lambda_{PP} + (1 - \textit{confidence}([x] \longrightarrow_P D_i))\lambda_{PN} \\ &= \lambda_{PN} + (\lambda_{PP} - \lambda_{PN})\textit{confidence}([x] \longrightarrow_P D_i). \end{aligned} \tag{15}$$

Generally, it is reasonable to assume that $\lambda_{PP} < \lambda_{PN}$. That is, the cost for putting an object with the decision D_i into the positive region of D_i is always lower than the cost of putting an object not with the decision D_i into the positive region of D_i. In this case, the cost measure of positive rules is de-

creasing with respect to the confidence measure. In decision-theoretic terms, the threshold α in fact imposes the following upper bound cost for each positive rule:

$$\mathcal{R}([x] \longrightarrow_P D_i) \leq \alpha\lambda_{PP} + (1-\alpha)\lambda_{PN}.$$

For the special case of $\lambda_{PP} = 0$ and $\lambda_{PN} = 1$, we have:

$$\mathcal{R}([x] \longrightarrow_P D_i) \leq 1 - \alpha.$$

The quantity $1 - \alpha$ becomes the error rate of a rule. In this special case, we in fact impose this upper bound on the error rate for positive rules.

A boundary rule $[x] \longrightarrow_B D_i$ decides that all the objects in $[x]$ are put into the boundary region of the decision class D_i with confidence greater than β and less than α. If a wait-to-see action a_B of D_i is taken for $[x]$, the corresponding expected cost of applying the boundary rule can be calculated as follows:

$$\begin{aligned}\mathcal{R}([x] \longrightarrow_B D_i) &= \mathcal{R}(a_B^{D_i} \mid [x]) \\ &= \lambda_{BP} P(D_i \mid [x]) + \lambda_{BN} P(D_i^c \mid [x]) \\ &= \mathit{confidence}([x] \longrightarrow_B D_i)\lambda_{BP} + (1 - \mathit{confidence}([x] \longrightarrow_B D_i))\lambda_{BN} \\ &= \lambda_{BN} + (\lambda_{BP} - \lambda_{BN})\mathit{confidence}([x] \longrightarrow_B D_i).\end{aligned} \tag{16}$$

From the Equations (15) and (16) we can see that two types of rules do lead to different decisions and have different costs and consequences. Such differences are explicitly shown by the cost measure, but cannot be differentiated by both the confidence and coverage measures.

Example 3 *In Example 2, we have calculated* $\alpha = 0.83$ *and* $\beta = 0.33$. *Suppose we have a patient* x, *whose symptoms are described by the description* $[x]$. *Based on the diagnoses, the probability for* x *getting cancer is* 0.90, *i.e., the confidence of the rule is written as confidence*$([x] \longrightarrow_P w_C) = 0.90$. *The cost of this positive rule is:*

$$\mathcal{R}(a_P^{w_C} \mid [x]) = 0.90 * \$1200.00 + 0.10 * \$2500.00 = \$1330.00.$$

Suppose we have another patient y. *Based on the diagnoses, the probability for* y *getting cancer is* 0.40. *In other words, the probability for* y *not getting cancer is* 0.60. *We pick the rule with a higher confidence, i.e. confidence*$([y] \longrightarrow_B w_H) = 0.60$. *The cost of the boundary rule is:*

$$\mathcal{R}(a_B^{w_H} \mid [y]) = 0.40 * \$1500.00 + 0.60 * \$1000.00 = \$1200.00.$$

4.3 Rule set evaluation

Given a partition π defined by a subset of condition attributes, we obtain two sets of rules about the decision classification: the sets of positive rules and the set of boundary rules. Let PRS and BRS be these two sets of rules, respectively.

In general, the evaluation of a rule set depends on the interaction of rules and rule conflict resolution for overlapping rules. These concepts are first discussed before introducing specific measures.

Definition 2 *Given a rule set* RS *induced from a partition* π*, if two rules in* RS *involving the same equivalence class* $[x]$ *and different decisions, that is,*

$$[x] \longrightarrow D_i \text{ and } [x] \longrightarrow D_j \text{ with } D_i \neq D_j,$$

they are called overlapping rules. The rule set RS *is called an overlapping rule set.*

The notion of overlapping rules is also known as conflicting rules or inconsistent rules. We want to emphasize the fact that the left-hand-sides of those rules have an overlap. If $\alpha \leq 0.50$, we may have two or more distinct positive rules for each equivalence class; if $\beta \leq 0.50$, we may have two or more distinct boundary rules for each equivalence class. In these cases, we may have conflict decisions for each equivalence class and an overlapping rule set. The condition $\alpha > \beta > 0.50$ is sufficient for obtaining non-overlapping rules and rule sets. For general cases, the values of α and β are calculated from the loss functions, and thus are not necessarily bounded by 0.50. The non-overlapping rule set is a special case of an overlapping rule set.

For a non-overlapping rule set, we can easily induce the unique rule $[x] \longrightarrow D_i$ for each equivalence class. For an overlapping rule set, rule conflict resolution is required. We adopt a simple maximum-confidence criterion for rule conflict resolution.

Definition 3 *For an overlapping rule set* RS*, the maximum-confidence criterion for each rule involving* $[x]$ *is denoted as*

$$[x] \longrightarrow D_{\max}([x]), \text{ where } D_{\max}([x]) = \arg \max_{([x] \longrightarrow D_i) \in \mathrm{RS}} \{\mathit{confidence}([x] \longrightarrow D_i)\}.$$

If $\lambda_{PP} < \lambda_{PN}$, the maximum-confidence criterion is equivalent to the minimum-

risk criterion, i.e.,

$$[x] \longrightarrow_P D_{\max}([x]), \text{ where } D_{\max}([x]) = \arg\min_{([x]\longrightarrow D_i)\in \mathrm{RS}}\{\mathcal{R}(a_P^{D_i} \mid [x])\};$$
$$[x] \longrightarrow_B D_{\max}([x]), \text{ where } D_{\max}([x]) = \arg\min_{([x]\longrightarrow D_i)\in \mathrm{RS}}\{\mathcal{R}(a_B^{D_i} \mid [x])\}.$$

Other conflict resolution methods can also be defined.

We are now ready to examine measures of a set of rules and their relationships with the measures of single rules.

Confidence: The confidence of the set of positive rules can be interpreted as the ratio of the number of correctly classified objects and the number of classified objects covered by all positive rules. We define the confidence measure as follows:

$$\begin{aligned} \mathit{confidence}(\mathrm{PRS}) &= \frac{\#\text{ of correctly classified objects by PRS}}{\#\text{ of classified objects by PRS}} \\ &= \frac{|\bigcup_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)}[x] \cap D_{\max}([x])|}{|\bigcup_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)}[x]|} \\ &= \frac{\sum_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)} |[x] \cap D_{\max}([x])|}{|\mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)|} \\ &= \sum_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)} \frac{|[x]|}{|\mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)|} \cdot \mathit{confidence}([x] \longrightarrow D_{\max}([x])). \end{aligned} \tag{17}$$

That is, for a set of positive rules, its confidence is the weighted sum of the confidence of individual rules in the set.

Coverage: The coverage of the set of positive rules is the ratio of the number of correctly classified objects in the set and the number of all objects in the universe. The coverage measure is defined as follows:

$$\begin{aligned} \mathit{coverage}(\mathrm{PRS}) &= \frac{\#\text{ of correctly classified objects by PRS}}{\#\text{ of objects in } U} \\ &= \frac{|\bigcup_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)}[x] \cap D_{\max}([x])|}{|U|} \\ &= \frac{\sum_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)} |[x] \cap D_{\max}([x])|}{|U|} \\ &= \sum_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)} \frac{|D_{\max}([x])|}{|U|} \cdot \mathit{coverage}([x] \longrightarrow D_{\max}([x])). \end{aligned}$$

(18)

That is, for a set of positive rules, its coverage is the weighted sum of the coverage of individual rules in the set.

Generality: For the set of positive rules, we can define the generality measure as follows:

$$\begin{aligned} \mathit{generality}(\mathrm{PRS}) &= \frac{\#\text{ of objects covered by PRS}}{\#\text{ of objects in } U} \\ &= \frac{|\bigcup_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)}[x]|}{|U|} \\ &= \sum_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)} \frac{|[x]|}{|U|} \\ &= \sum_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)} \mathit{generality}([x] \longrightarrow D_{\max}([x])) \\ &= \frac{|\mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)|}{|U|}. \end{aligned} \tag{19}$$

Again, the generality of a set of positive rules can be computed from the generality of individual rules in the set.

Cost: The cost of the set of positive rules is defined as:

$$\begin{aligned} \mathcal{R}(\mathrm{PRS}) &= \mathit{confidence}(\mathrm{PRS})\lambda_{PP} + (1 - \mathit{confidence}(\mathrm{PRS}))\lambda_{PN} \\ &= \left(\sum_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)} \frac{|[x]|}{|\mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)|} \cdot \mathit{confidence}([x] \longrightarrow D_{\max}([x])) \right) \cdot \lambda_{PP} + \\ &\quad \left(1 - \sum_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)} \frac{|[x]|}{|\mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)|} \cdot \mathit{confidence}([x] \longrightarrow D_{\max}([x])) \right) \cdot \lambda_{PN} \\ &= \sum_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)} \frac{|[x]|}{|\mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)|} \cdot [\mathit{confidence}([x] \longrightarrow D_{\max}([x]))\lambda_{PP} + \\ &\qquad\qquad (1 - \mathit{confidence}([x] \longrightarrow D_{\max}([x])))\lambda_{PN}] \\ &= \sum_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)} \frac{|[x]|}{|\mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)|} \cdot \mathcal{R}([x] \longrightarrow D_{\max}([x])). \end{aligned} \tag{20}$$

That is, for a set of positive rules, the cost equals to the weighted sum of the cost of individual positive rules in the set.

By following the same argument, the confidence, coverage, generality and cost

of the boundary rule set can be defined as follows:

$$\begin{aligned}
\mathit{confidence}(\mathrm{BRS}) &= \sum_{[x]\subseteq \mathrm{BND}_{\pi_{(\alpha,\beta)}}(\pi_D)} \frac{|[x]|}{|\mathrm{BND}_{\pi_{(\alpha,\beta)}}(\pi_D)|} \cdot \mathit{confidence}([x] \longrightarrow D_{\max}([x]));\\
\mathit{coverage}(\mathrm{BRS}) &= \sum_{[x]\subseteq \mathrm{BND}_{\pi_{(\alpha,\beta)}}(\pi_D)} \frac{|D_{\max}([x])|}{|U|} \cdot \mathit{coverage}([x] \longrightarrow D_{\max}([x]));\\
\mathit{generality}(\mathrm{BRS}) &= \sum_{[x]\subseteq \mathrm{BND}_{\pi_{(\alpha,\beta)}}(\pi_D)} \mathit{generality}([x] \longrightarrow D_{\max}([x]));\\
\mathcal{R}(\mathrm{BRS}) &= \sum_{[x]\subseteq \mathrm{BND}_{\pi_{(\alpha,\beta)}}(\pi_D)} \frac{|[x]|}{|\mathrm{BND}_{\pi_{(\alpha,\beta)}}(\pi_D)|} \cdot \mathcal{R}([x] \longrightarrow D_{\max}([x])).
\end{aligned}$$

This is, for the evaluation of the boundary rule set, we obtain the measures by replacing POS with BND in the corresponding positive measures.

According to the relationships between measures on individual rules and measures on rule sets, we can easily obtain the following theorem.

Theorem 1 *For a set of rules,*

$$\begin{aligned}
&\forall([x] \longrightarrow D_{\max}([x])) \in \mathrm{PRS}\ (\mathit{confidence}([x] \longrightarrow_P D_{\max}([x])) \geq \alpha)\\
&\qquad\qquad \Longrightarrow \mathit{confidence}(\mathrm{PRS}) \geq \alpha;\\
&\forall([x] \longrightarrow D_{\max}([x])) \in \mathrm{BRS}\ (\mathit{confidence}([x] \longrightarrow_B D_{\max}([x])) > \beta)\\
&\qquad\qquad \Longrightarrow \mathit{confidence}(\mathrm{BRS}) > \beta.
\end{aligned}$$

Theorem 1 shows that the confidence bound of individual rules is the same as the confidence bound of the rule set. This implies that in a rule induction process one can ensure that the confidence of a rule set is above a certain threshold if one imposes the same bound on each individual rule. However, the reverse is not necessarily true. Thus, the requirement on the level of confidence of all individual rules is sufficient to guarantee the same level of performance of the rule set, but is not necessary.

5 Attribute Reduction in the Pawlak Model

The main results of rule induction in the last section can be summarized as follows. A subset of attributes defines an equivalence relation. Based on the corresponding partition, one can induce a set of positive rules and a set of boundary rules, respectively. An important issue not discussed yet is the choice of a suitable subset of attributes for rule induction. In machine learning, this is commonly referred to as the problem of feature selection. In rough set analysis, the problem is called attribute reduction, and a selected set of attributes for

rule induction is called a reduct [22]. Intuitively speaking, an attribute reduct is a minimal subset of attributes whose induced rule sets have the same level of performance as the entire set of attributes, or a lower but satisfied level of performance.

5.1 Pawlak reducts

A Pawlak reduct $R \subseteq \mathbf{C}$, more precisely a relative reduct with respect to the decision attribute D, is defined by requiring that the positive region of π_D is unchanged [22].

Definition 4 *Given an information table $S = (U, At = \mathbf{C} \cup \{D\}, \{V_a \mid a \in At\}, \{I_a \mid a \in At\})$, an attribute set $R \subseteq \mathbf{C}$ is a Pawlak reduct of $\mathbf{C}$ with respect to D if it satisfies the following two conditions:*

(s) *Jointly sufficient condition:*
$\mathrm{POS}_{\pi_R}(\pi_D) = \mathrm{POS}_{\pi_{\mathbf{C}}}(\pi_D)$;

(n) *Individually necessary condition:*
for any attribute $a \in R$, $\mathrm{POS}_{\pi_{R-\{a\}}}(\pi_D) \neq \mathrm{POS}_{\pi_{\mathbf{C}}}(\pi_D)$.

Based on this simple definition of a Pawlak reduct, we can make several important observations.

Two extreme cases of the confidence: In the definition of a Pawlak reduct, the positive region of the partition π_D is used. Recall that the definition of the positive region requires an equivalence class $[x]$ to be a subset of a decision class D_i. Thus, the definition of a reduct implicitly uses a condition that requires a Pawlak positive rule with a confidence of 1, which is the maximum value of confidence. On the other hand, the confidence of a Pawlak boundary rule must have a confidence greater than 0, which is the minimum value of confidence.

Implicit consideration of the boundary region: In the Pawlak model, for a reduct $R \subseteq \mathbf{C}$, we have $\mathrm{POS}_{\pi_R}(\pi_D) \cap \mathrm{BND}_{\pi_R}(\pi_D) = \emptyset$, and $\mathrm{POS}_{\pi_R}(\pi_D) \cup \mathrm{BND}_{\pi_R}(\pi_D) = U$. The condition $\mathrm{POS}_{\pi_R}(\pi_D) = \mathrm{POS}_{\pi_{\mathbf{C}}}(\pi_D)$ is equivalent to $\mathrm{BND}_{\pi_R}(\pi_D) = \mathrm{BND}_{\pi_{\mathbf{C}}}(\pi_D)$. Therefore, the requirement of the same boundary region is implicitly stated in the definition of a Pawlak reduct. It is sufficient to consider only the positive rule set in the Pawlak model.

Monotocity of positive regions and decision rules: The definition of a Pawlak reduct is based on the relationships between positive regions and, in turn, sets of positive rules, induced by different subsets of attributes. By introducing the concept of decision-monotocity of rules with respect to set inclusion of attributes, we can shed new lights on the notion of a reduct.

Consider any two subsets of attributes $A, B \subseteq \mathbf{C}$ with $A \subseteq B$. For any $x \in U$, we have $[x]_B \subseteq [x]_A$. We immediately obtain the monotocity of the positive regions with respect to set inclusion of attributes. That is,

$$A \subseteq B \Longrightarrow \forall D_i \in \pi_D(\mathrm{POS}_{\pi_A}(D_i) \subseteq \mathrm{POS}_{\pi_B}(D_i)), \text{ and thus}$$

$$A \subseteq B \Longrightarrow \mathrm{POS}_{\pi_A}(\pi_D) \subseteq \mathrm{POS}_{\pi_B}(\pi_D).$$

Therefore, if $[x]_A \subseteq D_i$ for some decision class $D_i \in \pi_D$, which implies $[x]_A \subseteq \mathrm{POS}_{\pi_A}(\pi_D)$, we can conclude that $[x]_B \subseteq D_i$ for the same decision class D_i, which implies $[x]_B \subseteq \mathrm{POS}_{\pi_B}(\pi_D)$. This suggests that the Pawlak positive rules induced by different subsets of attributes satisfy the following decision-monotocity with respect to set inclusion of attributes:

$$A \subseteq B \Longrightarrow (\forall x \in U([x]_A \longrightarrow_P D_i \Longrightarrow [x]_B \longrightarrow_P D_i)).$$

That is, if we can make a positive decision based on a smaller set of attributes, the decision must be consistent with the decision made by a larger set of attributes. However, the reverse is not necessarily true. By demanding that a reduct R produces the same positive region as the entire set $\mathbf{C}$, we in fact ensure the reverse is also true. In terms of rules, condition (s) of a reduct can be equivalently expressed by:

(s1) $\forall x \in U([x]_R \longrightarrow_P D_i \Longleftrightarrow [x]_{\mathbf{C}} \longrightarrow_P D_i)$,

or equivalently,

(s2) $\forall x \in U([x]_R \subseteq D_i \Longleftrightarrow [x]_{\mathbf{C}} \subseteq D_i)$.

Monotonicity of the quantitative measures: Many authors [3,11,22,29,38] use an equivalent quantitative definition of a Pawlak reduct. It is based on the following measure, called the *quality of classification* or the *degree of dependency of* D [24], on an attribute set $A \subseteq \mathbf{C}$:

$$\gamma(\pi_D \mid \pi_A) = \frac{|\mathrm{POS}_{\pi_A}(\pi_D)|}{|U|}. \tag{21}$$

Based on the monotocity of positive regions, we can obtain the monotocity of the γ measure. That is,

$$A \subseteq B \Longrightarrow \gamma(\pi_D \mid \pi_A) \leq \gamma(\pi_D \mid \pi_B).$$

By monotocity, the condition (s) of the definition can be re-expressed as:

(s3) $\gamma(\pi_D \mid \pi_R) = \gamma(\pi_D \mid \pi_{\mathbf{C}})$.

In other words, R and $\mathbf{C}$ are the same under the γ measure.

In general, any monotonic measure f can be used to define a Pawlak reduct if it satisfies the condition

$$(f(\pi_D \mid \pi_R) = f(\pi_D \mid \pi_{\mathbf{C}})) \Longleftrightarrow (\mathrm{POS}_{\pi_R}(\pi_D) = \mathrm{POS}_{\pi_{\mathbf{C}}}(\pi_D)).$$

For example, Shannon's entropy and many of its variations have been explored to measure the uncertainty in rough set theory [3,7,15,19,28,37,39], and thus can be understood as different forms of the f measure.

The equivalence of the two conditions $\gamma(\pi_D \mid \pi_R) = \gamma(\pi_D \mid \pi_{\mathbf{C}})$ and $\mathrm{POS}_{\pi_R}(\pi_D) = \mathrm{POS}_{\pi_{\mathbf{C}}}(\pi_D)$ is true under the condition $\alpha = 1$ used in defining the Pawlak positive region. They are no longer equivalent in the probabilistic rough set models when a different value of α is used in defining a probabilistic positive region.

5.2 Interpretations of the γ measure

In order to gain more insights into a reduct defined by the γ measure, we need to explicitly establish connections between γ and other measures of rule sets discussed in the last section.

Confidence: The Pawlak positive region is formed by the condition $\alpha = 1$. Therefore, each positive rule has a confidence of 1. By Equation (17), we have:

$$\mathit{confidence}(\mathrm{PRS}_A) = \frac{|\mathrm{POS}_{\pi_A}(\pi_D)|}{|\mathrm{POS}_{\pi_A}(\pi_D)|} = 1.$$

If $\mathrm{POS}_{\pi_A}(\pi_D) = \emptyset$, we assume $\mathit{confidence}(\mathrm{PRS}_A) = 1$. It can be observed that the confidence of a Pawlak positive rule set is always 1, independent of the set of attributes $A \subseteq \mathbf{C}$. Similarly, the confidence of a Pawlak boundary rule set is always less than 1 and greater than 0, and is independent of the set of attributes $A \subseteq \mathbf{C}$. The confidence of positive rules is imposed by the requirement of $\alpha = 1$, and it determines the positive region used in the γ measure. On the other hand, the γ measure does not determine the confidence of rules.

Coverage: According to Equation (18), the coverage of a Pawlak positive rule set can be computed by:

$$\mathit{coverage}(\mathrm{PRS}_A) = \frac{|\mathrm{POS}_{\pi_A}(\pi_D)|}{|U|} = \gamma(\pi_D \mid \pi_A).$$

Thus, $\gamma(\pi_D \mid \pi_A)$ is in fact the coverage of the Pawlak positive rule set.

Generality: According to Equation (19), the generality of a Pawlak positive rule set is given by:

$$\textit{generality}(\mathrm{PRS}_A) = \frac{|\mathrm{POS}_{\pi_A}(\pi_D)|}{|U|} = \gamma(\pi_D \mid \pi_A).$$

In the Pawlak model the coverage and generality of the positive rule set are the same as the γ measure.

Cost: According to Equation (20), the cost measure of a Pawlak positive rule set is given by:

$$\mathcal{R}(\mathrm{PRS}_A) = \textit{confidence}(\mathrm{PRS}_A)\lambda_{PP} + (1 - \textit{confidence}(\mathrm{PRS}_A))\lambda_{PN} = \lambda_{PP}.$$

This means that the cost of the Pawlak positive rule set is a constant and, moreover, the γ measure does not tell us anything about the cost.

The following theorem summarizes the main results developed so far:

Theorem 2 *For a reduct $R \subseteq \mathbf{C}$, the following conditions are equivalent in the Pawlak model:*

(i.) $\gamma(\pi_D \mid \pi_R) = \gamma(\pi_D \mid \pi_{\mathbf{C}})$;
(ii.) $\mathrm{POS}_{\pi_B}(\pi_R) = \mathrm{POS}_{\pi_{\mathbf{C}}}(\pi_D)$;
(iii.) $\textit{coverage}(\mathrm{PRS}_R) = \textit{coverage}(\mathrm{PRS}_{\mathbf{C}})$;
(iv.) $\textit{generality}(\mathrm{PRS}_R) = \textit{generality}(\mathrm{PRS}_{\mathbf{C}})$;
(v.) *for all* $x \in U$ $([x]_R \longrightarrow_P D_i) \Longleftrightarrow ([x]_{\mathbf{C}} \longrightarrow_P D_i)$.

Theorem 2 shows that a Pawlak reduct $R \subseteq \mathbf{C}$ produces a positive rule set with the same level of coverage and generality as the entire set $\mathbf{C}$. In addition, for any rules induced by R, it makes the same classification decision as the entire set $\mathbf{C}$. The same conclusions of (i.) - (iv.) are also true for the set of boundary rules induced by R. Therefore, the γ measure is a good choice for defining a reduct in the Pawlak model.

6 Attribute Reduction in Probabilistic Models

According to the analysis in the previous section, the γ measure truthfully reflects many properties of a reduct in the Pawlak model. We examine the possibility of defining a single measure in the probabilistic models and propose a general definition of an attribute reduct in probabilistic models.

6.1 A definition of a probabilistic attribute reduct

Being parallel to the study of a Pawlak reduct, a probabilistic attribute reduct can be defined by requiring that the probabilistic positive region of π_D is unchanged.

Definition 5 *Given an information table* $S = (U, At = \mathbf{C} \cup \{D\}, \{V_a \mid a \in At\}, \{I_a \mid a \in At\})$, *an attribute set* $R \subseteq \mathbf{C}$ *is a reduct of* $\mathbf{C}$ *with respect to* D *if it satisfies the following two conditions:*

(s) *Jointly sufficient condition:*
$$\mathrm{POS}_{\pi_{R_{(\alpha,\beta)}}}(\pi_D) = \mathrm{POS}_{\pi_{\mathbf{C}_{(\alpha,\beta)}}}(\pi_D);$$

(n) *Individually necessary condition:*
for any attribute $a \in R$, $\mathrm{POS}_{\pi_{R-\{a\}_{(\alpha,\beta)}}}(\pi_D) \neq \mathrm{POS}_{\pi_{\mathbf{C}_{(\alpha,\beta)}}}(\pi_D)$.

A similar definition has been proposed by Kryszkiewicz as a β-reduct for the variable precision rough set model [16]. Based on this definition, we can also make several observations.

Bounded confidence: In this definition, the probabilistic positive region of π_D is used. The definition of a probabilistic region indicates that the intersection of an equivalence class $[x]$ and a decision class is not empty, i.e., $[x] \cap D_{\max}([x]) \neq \emptyset$. More specifically, a positive rule $[x] \longrightarrow_P D_{\max}([x])$ is constrained by the confidence threshold α, and a boundary rule $[x] \longrightarrow_B D_{\max}([x])$ is constrained by the confidence threshold β. Note that α is not necessarily the maximum value 1, and β is not necessarily the minimum value 0.

Ignorance of the boundary region: In probabilistic models, for a reduct $R \subseteq \mathbf{C}$, we many have $\mathrm{POS}_{\pi_R}(\pi_D) \cup \mathrm{BND}_{\pi_R}(\pi_D) \neq U$. The $\gamma_{(\alpha,\beta)}$ measure only reflects the probabilistic positive region and does not evaluate the probabilistic boundary region. Attribute reduction in probabilistic rough set models needs to consider criteria for both the probabilistic positive region and the probabilistic boundary region.

Non-monotocity of probabilistic positive regions and decision rules: In a probabilistic model, we cannot obtain the monotocity of the probabilistic positive regions with respect to set inclusion of attributes. That is, for $A, B \subseteq \mathbf{C}$ with $A \subseteq B$, we may obtain $\mathrm{POS}_{\pi_{A_{(\alpha,\beta)}}}(D_i) \supseteq \mathrm{POS}_{\pi_{B_{(\alpha,\beta)}}}(D_i)$ for some $D_i \in \pi_D$, and thus $\mathrm{POS}_{\pi_{A_{(\alpha,\beta)}}}(\pi_D) \supseteq \mathrm{POS}_{\pi_{B_{(\alpha,\beta)}}}(\pi_D)$. These results have two consequences. First, for any $x \in U$, we may have two decision rules involving the equivalence classes $[x]_A$ and $[x]_B$, such that they are not connected by the decision-monotocity property. That is, we may make different decisions based on set A and its super set B of attributes, and the strength of such two decisions may be different. Second, the equality condition (s)

$\mathrm{POS}_{\pi_{R_{(\alpha,\beta)}}}(\pi_D) = \mathrm{POS}_{\pi_{\mathbf{C}_{(\alpha,\beta)}}}(\pi_D)$ is not enough for verifying a reduct, and may miss some reducts. Furthermore, the condition (n) should consider all the subsets of a reduct R, not only all the subsets $R - \{a\}$ for all $a \in R$.

Non-monotocity of the $\gamma_{(\alpha,\beta)}$ measure: In probabilistic models, many proposals have been made to extend the Pawlak attribute reduction by using extended and generalized measure of γ. For example, a straightforward transformation of the γ measure is denoted as follows [52]. For $A \subseteq \mathbf{C}$,

$$\gamma_{(\alpha,\beta)}(\pi_D \mid \pi_A) = \frac{|\mathrm{POS}_{\pi_{A_{(\alpha,\beta)}}}(\pi_D)|}{|U|}.$$

Based on the fact that the probabilistic positive regions are not monotonic with respect to set inclusion, the $\gamma_{(\alpha,\beta)}$ measure is also non-monotonic. That is, given $A \subseteq B$, we may have $\gamma_{(\alpha,\beta)}(\pi_D \mid \pi_A) \geq \gamma_{(\alpha,\beta)}(\pi_D \mid \pi_B)$.

Based on the condition $\gamma_{(\alpha,\beta)}(\pi_D \mid \pi_R) = \gamma_{(\alpha,\beta)}(\pi_D \mid \pi_{\mathbf{C}})$, we can obtain $|\mathrm{POS}_{\pi_{R_{(\alpha,\beta)}}}(\pi_D)| = |\mathrm{POS}_{\pi_{\mathbf{C}_{(\alpha,\beta)}}}(\pi_D)|$, but cannot guarantee $\mathrm{POS}_{\pi_{R_{(\alpha,\beta)}}}(\pi_D) = \mathrm{POS}_{\pi_{\mathbf{C}_{(\alpha,\beta)}}}(\pi_D)$. This means that the quantitative equivalence of the probabilistic positive regions does not imply the qualitative equivalence of the probabilistic positive regions.

6.2 Interpretations of the $\gamma_{(\alpha,\beta)}$ measure

Although the definition based on the extended $\gamma_{(\alpha,\beta)}$ measure is adopted by many researchers [4,11,35,52], the measure itself is inappropriate for attribute reduction in probabilistic models. Even if consider the evaluation of the probabilistic positive rule set only, we have the following observations and problems regarding the classification measures we have discussed so far.

Confidence: According to Theorem 1, $\alpha \leq \mathit{confidence}(\mathrm{PRS}) \leq 1$. The confidence of a probabilistic positive rule set is bounded by the value of α, and it determines the probabilistic positive regions used in $\gamma_{(\alpha,\beta)}$. The $\gamma_{(\alpha,\beta)}$ measure does not determine the confidence of the rules.

Coverage: According to Equation (18), the coverage of the positive rule set in a probabilistic model is computed as:

$$\mathit{coverage}(\mathrm{PRS}_A) \leq \frac{|\mathrm{POS}_{\pi_{A_{(\alpha,\beta)}}}(\pi_D)|}{|U|} = \gamma_{(\alpha,\beta)}(\pi_D \mid \pi_A).$$

Thus, $\gamma_{(\alpha,\beta)}(\pi_D \mid \pi_A)$ does not equal to the coverage measure of the probabilistic positive rule set.

Cost: According to Equation (20), the cost of the positive rule set in a probabilistic model is related to the confidence measure and the values of loss functions. Since the $\gamma_{(\alpha,\beta)}$ measure does not determine the confidence of the rules, it does not determine the cost of the rules.

Generality: According to Equation (19), the generality of the positive rule set in a probabilistic model is computed as:

$$\textit{generality}(\mathrm{PRS}_A) = \frac{|\mathrm{POS}_{\pi_{A_{(\alpha,\beta)}}}(\pi_D)|}{|U|} = \gamma_{(\alpha,\beta)}(\pi_D \mid \pi_A).$$

Thus, we can establish a two-way implication between $\gamma_{(\alpha,\beta)}$ and the generality of the positive rule set in a probabilistic model.

The consequence is that Theorem 2 does not hold in probabilistic models. Instead, we have the following theorem.

Theorem 3 *For $R \subseteq \mathbf{C}$, the following conditions are equivalent in a probabilistic model:*

(i.) $\gamma_{(\alpha,\beta)}(\pi_D \mid \pi_R) = \gamma_{(\alpha,\beta)}(\pi_D \mid \pi_{\mathbf{C}})$;
(ii.) $|\mathrm{POS}_{\pi_{R(\alpha,\beta)}}(\pi_D)| = |\mathrm{POS}_{\pi_{\mathbf{C}(\alpha,\beta)}}(\pi_D)|$;
(iii.) $\textit{generality}_{(\alpha,\beta)}(\mathrm{PRS}_R) = \textit{generality}_{(\alpha,\beta)}(\mathrm{PRS}_{\mathbf{C}})$.

Theorem 3 shows that a reduct $R \subseteq \mathbf{C}$ produces a probabilistic positive rule set with the same level of generality as the entire set $\mathbf{C}$. The same conclusion is not true for the set of boundary rules induced by R. The other properties, such as coverage and decision-monotocity, cannot be kept for both rule sets.

6.3 A general definition of a probabilistic attribute reduct

In light of the previous analysis, although the γ measure is suitable for attribute reduction in the Pawlak model by reflecting many properties of classification, the straightforward extension of the γ measure might not be suitable for attribute reduction in probabilistic rough set models. Instead, we need to consider multiple properties, such as confidence, coverage, generality, cost and decision-monotocity criteria.

For one certain property, we can use a particular measure as its indicator. A measure, roughly denoted as $e : 2^{\mathbf{C}} \longrightarrow (L, \succeq)$, maps a condition attribute set to an element of a poset L, which is equipped with the partial order relation

$\succeq$. That is, $\succeq$ is reflexive, anti-symmetric and transitive. Based on the partial order relation, we are able to pick the attribute set preserving the property. The evaluation of a reduct $R \subseteq \mathbf{C}$ with respect to e is the same or superior to $e(\mathbf{C})$, and the evaluation of any subset of R with respect to e is inferior to $e(\mathbf{C})$.

Given a certain property, a measure representing it is not unique. We have three basic forms. The first form, denoted as $e^{P,B}$, is to distinguish the positive regions from the boundary regions. This allows us emphasize the effectiveness of positive rules while keeping in mind the effectiveness of boundary rules. The second form, denoted as $e^{P\cup B}$, keeps tracking all the rules by combining boundary rules with positive rules. However, by doing so, the certainty of the positive rules are degraded. The third form, denoted as $e^{P,P\cup B}$, is to evaluate positive regions and non-negative regions separately. Inuiguch has commented that the third form should be a better choice for the definition [13]. We may also have distributed versions of the above three forms. For example, a distributive measure *distr-*$e^{P,B}$ is to evaluate the distribution of positive regions and boundary regions of individual decision classes.

By considering multiple criteria and multiple measures, a general definition of an attribute reduct can be described as follows.

Definition 6 *Given an information table* $S = (U, At = \mathbf{C} \cup \{D\}, \{V_a \mid a \in At\}, \{I_a \mid a \in At\})$. *Suppose we can evaluate the properties of* S *by a set of measures* $E = \{e_1, e_2, \ldots\}$. *An attribute set* $R \subseteq \mathbf{C}$ *is a reduct of* $\mathbf{C}$ *with respect to* D *if it satisfies the following two conditions:*

(s) *Jointly sufficient condition:*
$e(\pi_D \mid \pi_R) \succeq e(\pi_D \mid \pi_{\mathbf{C}})$ *for all* $e \in E$;

(n) *Individually necessary condition:*
for any subset $R' \subset R$, $e(\pi_D \mid \pi_{R'}) \prec e(\pi_D \mid \pi_{\mathbf{C}})$ *for all* $e \in E$.

We explain three criteria, decision-monotocity, generality and cost, in the following sub-sections. The confidence and coverage criteria can be explored in a similar manner.

6.3.1 The decision-monotocity criterion

For a particular object, it is desirable that the decision made with more attributes should stay the same with the decision made with less attributes. Let $R \subseteq \mathbf{C}$ be a reduct. The decision-monotocity property for a set of rules can be interpreted as:

For all $x \in U$,
$([x]_{\mathbf{C}} \longrightarrow_P D_{\max}([x]_{\mathbf{C}})) \Longrightarrow ([x]_R \longrightarrow_P D_{\max}([x]_{\mathbf{C}}))$, and

$$([x]_{\mathbf{C}} \longrightarrow_B D_{\max}([x]_{\mathbf{C}})) \Longrightarrow ([x]_R \longrightarrow_{B/P} D_{\max}([x]_{\mathbf{C}})).$$

The decision-monotocity criterion requires two things. First, the criterion requires that by reducing attributes a positive rule is still a positive rule of the same decision. That is, for any $x \in \mathrm{POS}_{\pi_{\mathbf{C}}}(D_i)$, we must have $x \in \mathrm{POS}_{\pi_R}(D_i)$. In other words, if $x \in \underline{apr}_{\pi_{\mathbf{C}}}(D_i)$, then $x \in \underline{apr}_{\pi_R}(D_i)$. Therefore, $\underline{apr}_{\pi_{\mathbf{C}}}(D_i) \subseteq \underline{apr}_{\pi_R}(D_i)$ for all $D_i \in \pi_D$.

The confidence of the positive rule $[x]_R \longrightarrow_P D_{\max}([x]_{\mathbf{C}})$ is not lower than the threshold α, but may be lower than the confidence of the positive rule $[x]_{\mathbf{C}} \longrightarrow_P D_{\max}([x]_{\mathbf{C}})$. In this case, the unique and same decision can be made for the equivalence class $[x]$ in the positive region. The decreasing confidence of positive rules causes two consequences: (a.) It increases the generality of the rule. In the domain of machine learning, this is exactly the idea of pruning an over-fitted rule to a more general rule by dropping off some descriptions of the rule. (b.) It increases the cost of the rule. This is because under the general assumption $\lambda_{PP} < \lambda_{PN}$, the cost of positive rules is monotonically increasing with respect to the decreasing confidence. Therefore, for positive rules, the decision-monotocity property normally means sacrificing the confidence and the cost for an increased generality.

Second, the criterion requires that by reducing attributes a boundary rule is still a boundary rule, or is upgraded to a positive rule with the same decision. That is, for any $x \in \mathrm{BND}_{\pi_{\mathbf{C}}}(D_i)$, we must have $x \in \mathrm{BND}_{\pi_R}(D_i)$ or $x \in \mathrm{POS}_{\pi_R}(D_i)$. In other words, if $x \in \overline{apr}_{\pi_{\mathbf{C}}}(D_i)$, then $x \in \overline{apr}_{\pi_R}(D_i)$. Therefore, $\overline{apr}_{\pi_{\mathbf{C}}}(D_i) \subseteq \overline{apr}_{\pi_R}(D_i)$ for all $D_i \in \pi_D$.

The confidence of the rule $[x]_R \longrightarrow_{B/P} D_{\max}([x]_{\mathbf{C}})$ is not lower than the threshold β, and may be higher than the confidence of the boundary rule $[x]_{\mathbf{C}} \longrightarrow_B D_{\max}([x]_{\mathbf{C}})$. In this case, the unique and same decision can be made for the equivalence class $[x]$ in the boundary region. The interpretation of the decision-monotocity criterion is only one-way upgrading. The degradation is not allowed by this interpretation. This is ensured by two conditions:

$$\underline{apr}_{\pi_{R_{(\alpha,\beta)}}}(D_i) \supseteq \underline{apr}_{\pi_{\mathbf{C}_{(\alpha,\beta)}}}(D_i) \text{ and}$$
$$\overline{apr}_{\pi_{R_{(\alpha,\beta)}}}(D_i) \supseteq \overline{apr}_{\pi_{\mathbf{C}_{(\alpha,\beta)}}}(D_i) \text{ for all } D_i \in \pi_D.$$

In this sense, the decision-monotocity criterion is consistent with the general definition of a reduct.

A criterion similar to decision-monotocity has been proposed by Slezak as the *majority decision* criterion [31] and by Zhang *et al.* as the *maximum distribution* criterion [50]. The majority decision criterion uses a binary information vector for each equivalence class to indicate to which decision class it be-

	$\mathbf{C}$						D
	c_1	c_2	c_3	c_4	c_5	c_6	
o_1	1	1	1	1	1	1	M
o_2	1	0	1	0	1	1	M
o_3	0	0	1	1	0	0	Q
o_4	1	1	1	0	0	1	Q
o_5	1	0	1	0	1	1	F
o_6	0	0	0	1	1	0	F
o_7	1	0	1	0	1	1	F

Table 2
An information table

longs. As Slezak suggested, there are many possibilities to modify, combine and generalize the majority decision function [31]. For example, instead of using a binary information vector, Kryszkiewicz defined a rough membership for each equivalence class with respect to all decision classes [17]. The partition based on the membership distribution vector is finer and more complex, and can preserve the quality of the decisions. Li *et al.* compare the differences of decision-monotocity criteria recently [18].

Example 4 *A simple information table* $S = (U, At = \mathbf{C} \cup \{D\}, \{V_a\}, \{I_a\})$ *shown in Table 2 is used for exemplifying the decision-monotocity criterion. Suppose the two thresholds* $\alpha = 0.81$ *and* $\beta = 0.58$ *are calculated from the loss functions for the three states regarding* M*,* Q *and* F*.*

The equivalence relation $E_{\mathbf{C}}$ *partitions the universe into five equivalence classes. The partition* $\pi_{\mathbf{C}}$ *induces the following five rules:*

$$\begin{aligned}
\{o_1\} &\longrightarrow_P M, (\mathit{confidence} = 1);\\
\{o_2, o_5, o_7\} &\longrightarrow_B F, (\mathit{confidence} = 0.67);\\
\{o_3\} &\longrightarrow_P Q, (\mathit{confidence} = 1);\\
\{o_4\} &\longrightarrow_P Q, (\mathit{confidence} = 1);\\
\{o_6\} &\longrightarrow_P F, (\mathit{confidence} = 1).
\end{aligned}$$

The equivalence relation $E_{\{c_2,c_5\}}$ *partitions the universe into four equivalence classes. The partition* $\pi_{\{c_2,c_5\}}$ *induces the following four rules:*

$$\begin{aligned}
\{o_1\} &\longrightarrow_P M, (\mathit{confidence} = 1);\\
\{o_2, o_5, o_6, o_7\} &\longrightarrow_B F, (\mathit{confidence} = 0.75);\\
\{o_3\} &\longrightarrow_P Q, (\mathit{confidence} = 1);
\end{aligned}$$

$$\{o_4\} \longrightarrow_P Q, (confidence = 1).$$

For the two equivalence classes of object o_6, we have a positive rule $[o_6]_{\mathbf{C}} \longrightarrow_P F$ and a boundary rule $[o_6]_{\{c_2,c_5\}} \longrightarrow_B F$ of the same decision class F. This result does not satisfy the decision-monotocity criterion of reducts. Thus, the attribute set $\{c_2, c_5\}$ is not a reduct according to the decision-monotocity criterion. It can be easily verified that $\{c_2, c_5\}$ satisfies the majority decision criterion, and thus is a majority decision reduct.

6.3.2 The generality criterion

It is reasonable to request that the generality of the new set of rules is kept or increased by the partition defined by a reduct. Let $R \subseteq \mathbf{C}$ be a reduct. The generality criterion means that the covered set derived from the partition π_R is more general than the covered set derived from the partition $\pi_{\mathbf{C}}$, i.e.,

$$generality(\pi_R \longrightarrow \pi_D) \geq generality(\pi_{\mathbf{C}} \longrightarrow \pi_D).$$

Although the generality criterion is used in many rough set models [4,24,52], it has crucial problems in the probabilistic rough set models. For example, suppose we have two positive rules $[x]_{\mathbf{C}} \longrightarrow_P D_i$ and $[x]_R \longrightarrow_P D_j$ with:

$$generality([x]_R \longrightarrow_P D_j) \geq generality([x]_{\mathbf{C}} \longrightarrow_P D_i).$$

In this case, even though we preserve the generality by the attribute set R, for these two particular rules, D_i and D_j may not be the same, and the rule $[x]_{\mathbf{C}} \longrightarrow_P D_i$ may not exist. Therefore, the generality criterion may conflict with the decision-monotocity criterion.

Example 5 *We can use a simple example to demonstrate the problem of the generality criterion. In information Table 2, suppose the thresholds $\alpha = 0.81$ and $\beta = 0.58$. The equivalence relation $E_{\{c_5\}}$ partitions the universe into two equivalence classes. The partition $\pi_{\{c_5\}}$ induces the following two rules:*

$$\{o_1, o_2, o_5, o_6, o_7\} \longrightarrow_B F, (confidence = 0.60);$$
$$\{o_3, o_4\} \longrightarrow_P Q, (confidence = 1).$$

For the two equivalence classes of object o_1, we have a positive rule of the decision class M, i.e., $[o_1]_{\mathbf{C}} \longrightarrow_P M$, with the generality being 1/7, and a boundary rule of the decision class F, i.e., $[o_1]_{\{c_5\}} \longrightarrow_B F$, with the generality being 5/7. Thus, the attribute set $\{c_5\}$ is a reduct according to the generality criterion. Although this result satisfies the generality criterion, it does not

satisfy the decision-monotocity criterion of reducts. Therefore, the generality criterion may disagree with the decision-monotocity criterion.

6.3.3 The cost criterion

Let $R \subseteq \mathbf{C}$ be a reduct. The cost criterion means that we need to make sure that the cost derived by the partition π_R does not increase, i.e.,

$$\mathcal{R}(\pi_R \longrightarrow \pi_D) \leq \mathcal{R}(\pi_{\mathbf{C}} \longrightarrow \pi_D).$$

The cost for the entire rule set can be defined as a distributed form for the cost of the positive rule set and the cost of the boundary rule set. It can be defined as the sum of the two costs. In a formal mathematical form:

$$\mathcal{R}^{P,B}(\pi \longrightarrow \pi_D) = (\sum_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)} \mathcal{R}(a_P \mid [x]), \sum_{[x]\subseteq \mathrm{BND}_{(\alpha,\beta)}(\pi_D)} \mathcal{R}(a_B \mid [x]));$$
$$\mathcal{R}^{P\cup B}(\pi \longrightarrow \pi_D) = \sum_{[x]\subseteq \mathrm{POS}_{\pi_{(\alpha,\beta)}}(\pi_D)} \mathcal{R}(a_P \mid [x]) + \sum_{[x]\subseteq \mathrm{BND}_{\pi_{(\alpha,\beta)}}(\pi_D)} \mathcal{R}(a_B \mid [x]).$$

It is important to note that the cost criterion should not be used alone. It should be used with decision-monotocity criterion and/or generality criterion. That is, the decrease of the cost should not change the original decision. Also, it should not sacrifice the generality of the rule set. This may not always be achievable.

Example 6 *The cost criterion can be illustrated by the same information Table 2. Suppose the two thresholds* $\alpha = 0.81$ *and* $\beta = 0.58$ *are calculated from the loss functions for the three states regarding* M, Q *and* F.

The partition $\pi_{\mathbf{C}}$ *determines three regions:* $\mathrm{POS}_{\pi_{\mathbf{C}_{(0.81,0.58)}}}(\pi_D) = \{o_1, o_3, o_4, o_6\}$, $\mathrm{BND}_{\pi_{\mathbf{C}_{(0.81,0.58)}}}(\pi_D) = \{o_2, o_5, o_7\}$ *and* $\mathrm{NEG}_{\pi_{\mathbf{C}_{(0.81,0.58)}}}(\pi_D) = \emptyset$. *The cost of the entire rule set is:*

$$\mathcal{R}^{P\cup B}(\pi_{\mathbf{C}} \longrightarrow \pi_D) = \lambda_{PP} + \frac{2}{3}\lambda_{BP} + \frac{1}{3}\lambda_{BN}.$$

The partition $\pi_{\{c_2,c_5\}}$ *defines another three regions:* $\mathrm{POS}_{U/\{c_2,c_5\}_{(0.81,0.58)}}(\pi_D) = \{o_1, o_3, o_4\}$, $\mathrm{BND}_{U/\{c_2,c_5\}_{(0.81,0.58)}}(\pi_D) = \{o_2, o_5, o_6, o_7\}$ *and* $\mathrm{NEG}_{U/\{c_2,c_5\}_{(0.81,0.58)}}(\pi_D) = \emptyset$. *The cost of the entire rule set is:*

$$\mathcal{R}^{P\cup B}(\pi_{\{c_2,c_5\}} \longrightarrow \pi_D) = \lambda_{PP} + \frac{3}{4}\lambda_{BP} + \frac{1}{4}\lambda_{BN}.$$

Comparing the costs of the two rule sets, if $\lambda_{BP} \leq \lambda_{BN}$ *then* $\{c_2, c_5\}$ *is a reduct regarding the cost criterion* $\mathcal{R}^{P \cup B}$*, otherwise, it is not.*

7 Conclusion

Regarding classification tasks, positive rules and boundary rules, derived from both the Pawlak model and probabilistic models have different confidence, coverage, costs and risks, and lead to different decisions and consequences. An attribute reduct should be able to preserve the classification power of both positive rules and boundary rules. This can be better understood and explained in the decision-theoretic rough set models.

We discuss various criteria for attribute reduction for probabilistic rough set models, such as decision-monotocity, generality and cost. It is noted that these criteria can be integrated as one simple quantitative measure in the Pawlak rough set model. However, for probabilistic models, these criteria have different expressive powers, and lead to different decision making and consequences. A systematic study of attribute reduction should consider one or more of these criteria, by using one or more corresponding measures, instead of using an oversimplified straightforward extension of the Pawlak γ measure.

This study provides a new insight into the problem of attribute reduction. It suggests that more semantics properties preserved by an attribute reduct should be carefully examined.

Acknowledgement

The authors thank the anonymous referees for the constructive comments and suggestions. This work is partially supported by a Discovery Grant from NSERC Canada.

References

[1] An, A., Shan, N., Chan, C., Cercone, N. and Ziarko, W. Discovering rules for water demand prediction: an enhanced rough-set approach, *Engineering Application and Artificial Intelligence*, 9, 645-653, 1996.

[2] Bazan, J.G., Nguyen, H.S., Nguyen, S.H., Synak, P. and Wroblewski, J. Rough set algorithms in classification problem, in: Polkowski, L., Tsumoto, S. and Lin, T.Y. (Eds.), *Rough Set Methods and Applications*, 49-88, 2000.

[3] Beaubouef, T., Petry, F.E. and Arora, G. Information-theoretic measures of uncertainty for rough sets and rough relational databases, *Information Sciences*, 109, 185-195, 1998.

[4] Beynon, M. Reducts within the variable precision rough sets model: a further investigation, European Journal of Operational Research, 134, 592-605, 2001.

[5] Clark, P. and Matwin, S., Using qualitative models to guide induction learning, *Proceedings of International Conference on Machine Learning*, 49-56, 1993.

[6] Duda, R.O. and Hart, P.E. *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[7] Düntsch, I. and Gediga, G. Uncertainty measures of rough set prediction, *Artificial Intelligence*, 106, 77107, 1998.

[8] Greco, S., Matarazzo, B. and Slowinski, R. Rough membership and Bayesian confirmation measures for parameterized rough sets, *LNAI 3641*, 314-324, 2005.

[9] Greco, S., Pawlak, Z. and Slowinski, R. Can Bayesian confirmation measures be useful for rough set decision rules? *Engineering Applications of Artificial Intelligence*, 17(4), 345-361, 2004.

[10] Herbert, J.P. and Yao, J.T. Rough set model selection for practical decision making, *Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery*, 203-207, 2007.

[11] Hu, Q., Yu D. and Xie, Z. Information-preserving hybrid data reduction based on fuzzy-rough techniques, *Pattern Recognition Letters*, 27, 414-423, 2006.

[12] Hu, Q., Yu D., Xie, Z. and Liu, J. Fuzzy probabilistic approximation spaces and their information measures, *Transactions on Fuzzy Systems*, 14, 191-201, 2006.

[13] Inuiguch, M. Several approaches to attribute reduction in variable precision rough set model, *Modeling Decisions for Artificial Intelligence*, 215-226, 2005.

[14] Katzberg, J.D. and Ziarko, W. Variable precision rough sets with asymmetric bounds, in: W. Ziarko (Ed.) *Rough Sets, Fuzzy Sets and Knowledge Discovery*, Springer, London, 167-177, 1994.

[15] Klir, J. and Wierman, M.J. *Uncertainty Based Information: Elements of Generalized Information Theory*, New York: Physica-Verlag, 1999.

[16] Kryszkiewicz, M. Maintenance of reducts in the variable precision rough sets model, *ICS Research Report* 31/94, Warsaw University of Technology, 1994.

[17] Kryszkiewicz, M. Certain, generalized decision, and membership distribution reducts versus functional dependencies in incomplete systems, *Proceedings of Rough Sets and Intelligent Systems Paradigms*, 162-174, 2007.

[18] Li, H., Zhou, X. and Huang, B. Attribute reduction in incomplete information systems based on connection degree rough set, *Computer Science (Ji Suan Ji Ke Xue)*, 34, 39-42, 2007.

[19] Liang, J.Y. and Shi, Z.Z. The information entropy, rough entropy and knowledge granulation in rough set theory, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 12, 3746, 2004.

[20] Liu, J., Hu, Q. and Yu, D. A weighted rough set based method developed for class imbalance learning, *Information Science*, 178(4), 1235-1256, 2008.

[21] Mi, J.S., Wu, W.Z. and Zhang, W.X. Approaches to knowledge reduction based on variable precision rough set model, *Information Sciences*, 159, 255-272, 2004.

[22] Pawlak, Z. Rough sets, *International Journal of Computer and Information Sciences*, 11, 341-356, 1982.

[23] Pawlak, Z. Rough classification, *International Journal of Man-Machine Studies*, 20, 469-483, 1984.

[24] Pawlak, Z. *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishers, Boston, 1991.

[25] Pawlak, Z. and Skowron, A. Rough membership functions, in: R.R. Yager and M. Fedrizzi and J. Kacprzyk (Eds.), *Advances in the Dempster-Shafer Theory of Evidence*, John Wiley and Sons, New York, 251-271, 1994.

[26] Pawlak, Z. and Skowron, A. Rudiments of rough sets, *Information Sciences*, 177, 3-27, 2007.

[27] Pawlak, Z. and Skowron, A. Rough sets: some extensions, *Information Sciences*, 177, 28-40, 2007.

[28] Pawlak, Z., Wong, S.K.M. and Ziarko, W. Rough sets: probabilistic versus deterministic approach, *International Journal of Man-Machine Studies*, 29, 81-95, 1988.

[29] Skowron, A. and Rauszer, C. The discernibility matrices and functions in information systems, in: Słowiński, R. (Ed.), *Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory*, Kluwer Academic Publishers, Dordrecht, 1992.

[30] Skowron, A. and Stepaniuk, J. Tolerance approximation spaces, *Fundamenta Informaticae*, 27, 245-253, 1996.

[31] Slezak, D. Normalized decision functions and measures for inconsistent decision tables analysis, *Fundamenta Informaticae*, 44, 291-319, 2000.

[32] Slezak, D. Rough sets and Bayes factor, *LNAI 3400*, 202-229, 2005.

[33] Slezak, D. and Ziarko, W. Attribute reduction in the Bayesian version of variable precision rough set model, *Electronic Notes in Theoretical Computer Science*, 82, 263-273, 2003.

[34] Su, C.T. and Hsu, J.H. Precision parameter in the variable precision rough sets model: an application, *Omega-International Journal of Management Science*, 34, 149-157, 2006.

[35] Swiniarski, R.W. Rough sets methods in feature reduction and classification, *International Journal of Applied Mathematics and Computer Science*, 11, 565-582, 2001.

[36] Swiniarski, R.W. and Skowron, A. Rough set methods in feature selection and recognition, *Pattern Recognition Letters*, 24, 833-849, 2003.

[37] Wang, G., Yu, H. and Yang, D. Decision table reduction based on conditional information entropy, *Chinese Journal of Computers*, 25, 759-766, 2002.

[38] Wang, G.Y., Zhao, J. and Wu, J. A comparitive study of algebra viewpoint and information viewpoint in attribute reduction, *Foundamenta Informaticae*, 68, 1-13, 2005.

[39] Wierman, M.J. Measuring uncertainty in rough set theory, *International Journal of General Systems*, 28, 283297, 1999.

[40] Wong, S.K.M. and Ziarko, W. Comparison of the probabilistic approximate classification and the fuzzy set model, *Fuzzy Sets and Systems*, 21, 357-362, 1987.

[41] Wu, W.Z., Zhang, M., Li, H.Z. and Mi, J.S. Knowledge reduction in random information systems via Dempster-Shafer theory of evidence, *Information Sciences*, 174, 143-164, 2005.

[42] Yao, J.T. and Herbert, J.P. Web-based support systems based on rough set analysis, *Proceedings of International Conference on Rough Sets and Emerging Intelligent System Paradigms*, 360-370, 2007.

[43] Yao, Y.Y. Information granulation and approximation in a decision-theoretical model of rough sets, in: Polkowski, L., Pal, S.K., and Skowron, A. (Eds), *Rough-neuro Computing: Techniques for Computing with Words*, Springer Berlin, 491-516, 2003.

[44] Yao, Y.Y. Probabilistic approaches to rough sets, *Expert Systems*, 20, 287-297, 2003.

[45] Yao, Y.Y. Decision-theoretic rough set models, *Proceedings of RSKT'07*, LNAI 4481, 1-12, 2007.

[46] Yao, Y.Y., Chen, Y.H. and Yang X.D., Measurement-theoretic foundation for rules interestingness evaluation, *Proceedings of ICDM'03 Workshop on Foundations of Data Mining*, 221-227, 2003.

[47] Yao, Y.Y. and Wong, S.K.M. A decision theoretic framework for approximating concepts, *International Journal of Man-machine Studies*, 37, 793-809, 1992.

[48] Yao, Y.Y., Wong, S.K.M. and Lingras, P. A decision-theoretic rough set model, in: Z.W. Ras, M. Zemankova and M.L. Emrich (Eds.), *Methodologies for Intelligent Systems*, North-Holland, New York, 5, 17-24, 1990.

[49] Yao, Y.Y. and Zhong, N., An analysis of quantitative measures associated with rules, *Proceedings Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 479-488, 1999.

[50] Zhang, W.X., Mi, J.S. and Wu, W.Z. Knowledge reduction in inconsistent information systems, *Chinese Journal of Computers*, 1, 12-18, 2003.

[51] Zhao, Y., Luo, F., Wong, S.K.M. and Yao, Y.Y. A general definition of an attribute reduction, *Proceedings of the Second Rough Sets and Knowledge Technology*, 101-108, 2007.

[52] Ziarko, W. Variable precision rough set model, *Journal of Computer and System Sciences*, 46, 39-59, 1993.

[53] Ziarko, W. Acquisition of hierarchy-structured probabilistic decision tables and rules from data, *Expert Systems*, 20, 305-310, 2003.

[54] Ziarko, W. Probabilistic rough sets, *LNAI 3641*, 283-293, 2005.