Computers and Mathematics with Applications 57 (2009) 908–918

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

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

Criteria for choosing a rough set model

Joseph P. Herbert, JingTao Yao *

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

ARTICLE INFO

Keywords:
Rough sets
Decision making
Probabilistic rough sets
Decision-theoretic rough sets
Variable-precision rough sets

ABSTRACT

One of the challenges a decision maker faces in using rough sets is to choose a suitable rough set model for data analysis. We investigate how two rough set models, the Pawlak model and the probabilistic model, influence the decision goals of a user. Two approaches use probabilities to define regions in the probabilistic model. These approaches use either user-defined parameters or derive the probability thresholds from the cost associated with making a classification. By determining the implications of the results obtained from these models and approaches, we observe that the availability of information regarding the analysis data is crucial for selecting a suitable rough set approach. We present a list of decision types corresponding to the available information and user needs. These results may help a user match their decision requirements and expectations to the model which fulfills these needs.

1. Introduction

Rough set theory is a way of representing and reasoning imprecision and uncertain information in data [1]. It deals with the approximation of sets constructed from descriptive data elements. This is helpful for finding decision rules, important features, and the minimum number of conditional attributes. Rough sets create three regions when they analyze data, namely, the positive, negative and boundary regions. These regions can be used for making decisions during "yes", "no", and "wait-and-see" situations. The "yes" and "no" cases apply when the information provided leaves no reasonable doubt regarding the outcome of these decisions. The "wait-and-see" case applies when a user is unsure of the outcome of a possible decision and, thus, prefers to wait for more information to become available [2].

Expansion of the positive (POS) and negative (NEG) regions can be achieved by decreasing the classification precision in the knowledge (rules) obtained through rough set analysis. This increases the coverage of the decision rules describing these regions. Researchers have extended the traditional, or Pawlak, rough set model into probabilistic approaches. The variable-precision rough set approach [3] and decision-theoretic rough set approach [4,5] expand the POS and NEG regions by providing probabilities that define region boundaries.

Decision makers wishing to find a rough set model to aid in their decision making are now faced with the challenge of which rough set model to choose from. Criteria for matching the kinds of decisions that can be made in the two rough set models to the user's decision needs could be very beneficial. Depending on the level of the user's knowledge about the domain, certain models may be more suitable for the analysis of the data. For example, a domain expert may understand the acceptable levels of probability for the classification of objects in that domain, in which case user-defined regions can be used. However, if such information is unavailable, which is true in many cases, an alternative approach to ensuring the correct region boundaries using the expected cost of a classification can be used.

This article classifies the types of decisions that can be made from the use of the various rough set models. We classify the decisions into two high-level categories, namely, immediate decisions and delayed decisions. These categories are suitable for classifying the decisions made from Pawlak rough set analysis, as there are only three regions created.

* Corresponding author.
E-mail addresses: herbertj@cs.uregina.ca (J.P. Herbert), jtyao@cs.uregina.ca (J.T. Yao).

0898-1221/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2008.10.043

For the probabilistic rough set approaches, further divisions of the immediate decision type are made. Risk or cost associated with observed data can be used for rough set region divisions with the decision-theoretic approach. Immediate decisions are divided into accepted loss and rejected loss decision types, each corresponding to the measured boundaries between regions. In contrast, user-accepted and user-rejected decision types are defined to imply that the decisions made from these regions are influenced by a user.

This article is organized as follows: Section 2 discusses rough set theory and the extended probabilistic model that expand the positive and negative regions. Section 3 provides the classification of the decision types for each rough set model. Section 4 gives illustrative examples of how differently these two models analyze data, and thus, the types of decision which can be made. We conclude this article in Section 5.

2. A critical review of rough set models

We will take a critical look at the Pawlak and probabilistic rough set models in this section.

2.1. The Pawlak rough set model

Discerning objects from each other is a major purpose in rough set theory. We convey this by saying that for any subset $B \subseteq A$,

$$\mathrm{IND}_T(B) = \{(o, o') \in U \times U | \forall a \in B, I_a(o) = I_a(o')\}, \tag{1}$$

where $\mathrm{IND}_T(B)$ is called the B-indiscernibility relation. Approximation is used to characterize a set $A \subseteq U$. It may be impossible to precisely describe $A \subseteq U$. The notion of approximating a set further divides the area into an algebraic representation and a probabilistic representation of equivalence classes. An equivalence class is formed through an equivalence relation E that partitions a universe U. Thus, for the algebraic representation, an equivalence class containing an object x is given by $[x] = \{y | xEy\}$ with partitioning U/E. Equivalence classes are descriptions of objects in U. Approximations are formed around these equivalence classes. The regions, derived from the approximations, are used as a guiding principle in what decisions a user can make. Definitions of lower and upper approximations follow [6]:

$$\begin{aligned} \underline{apr}(A) &= \{x \in U | [x] \subseteq A\}, \\ \overline{apr}(A) &= \{x \in U | [x] \cap A \neq \emptyset\}. \end{aligned} \tag{2}$$

The lower approximation of A, $\underline{apr}(A)$, is the union of all elementary sets that are included in A. The upper approximation A, $\overline{apr}(A)$, is the union of all elementary sets that have a non-empty intersection with A. This approximates unknown sets with equivalence classes. The positive, negative, and boundary regions of A can be defined as [1]:

$$\begin{aligned} \mathrm{POS}(A) &= \underline{apr}(A), \\ \mathrm{NEG}(A) &= U - \overline{apr}(A), \\ \mathrm{BND}(A) &= \overline{apr}(A) - \underline{apr}(A). \end{aligned} \tag{3}$$

The positive region, $\mathrm{POS}(A)$, consists of all objects that are definitely contained in the set A. The negative region, $\mathrm{NEG}(A)$, consists of all objects that are definitely not contained in the set A. The boundary region, $\mathrm{BND}(A)$, consists of all objects that may be contained in A. Since approximations are formed from equivalence classes, inclusion into the boundary region reflects uncertainty about the classification of objects [7].

Since the boundary region introduces uncertainty into the discernibility of objects, the major challenge in data analysis using rough sets is to minimize the size of this region. This is done by relaxing the definitions of the POS and NEG regions to include objects that would otherwise not have been included previously. That is, the expansion of the POS and NEG results in the reduction of the BND region. The probabilistic model discussed in the remainder of the section explains how this relaxation should be precisely defined.

2.2. The probabilistic rough set model

A probabilistic model for rough sets can change the size of the rough set regions in a universe. The first approach, variable-precision, makes use of user-provided probabilities as parameters to produce the different regions. The second approach, decision-theoretic, uses the cost of classifying an object either correctly or incorrectly to determine the different regions.

2.2.1. The variable-precision rough set approach

The variable-precision rough set (VPRS) approach has been used in many areas to support decision making [8–10]. The VPRS approach aims to increase the discriminatory capabilities of the rough set approach by using parameter grades of conditional probabilities [3]. Two parameters, the lower-bound ℓ and the upper-bound u, are provided by the user or domain expert.

The parameter u reflects the least acceptable degree of the conditional probability $P(A|[x])$ to include an object x with description $[x]$ into a set A. This is called the u-positive region,

$$\mathrm{POS}_u(A) = \{x \in A | P(A|[x]) \geq u\}. \tag{4}$$

That is, an object x is included in a set A if the probability, given its description, of it belonging to A is greater-than or equal-to an upper-bound probability specified by the user. The positive region of a set A consists of all objects that meet this criterion.

Likewise, the ℓ-negative region $\mathrm{NEG}_\ell(A)$ is controlled by the lower-bound ℓ, such that,

$$\mathrm{NEG}_\ell(A) = \{x \in A | P(A|[x]) \leq \ell\}. \tag{5}$$

An object x is included in a set A if the probability, given its description, of it belonging to A is less-than or equal-to a lower-bound probability specified by the user. The negative region of set A consists of all objects that, again, meet this criterion.

The boundary region is now potentially smaller in size since the u-positive and ℓ-negative regions increase the size of the positive and negative regions. The ℓ, u-boundary region, defined as,

$$\mathrm{BND}_{\ell,u}(A) = \{x \in A | \ell < P(A|[x]) < u\}, \tag{6}$$

classifies all remaining objects.

Since the ℓ and u parameters are given by the user, the quality is user-driven. That is, an expert must provide values for these parameters based on either their knowledge of the domain through empirical evidence or their intuition [11]. Precision, or accuracy of classification, is greatly effected by these values. An upper-bound u set too low decreases the certainty that any object is correctly classified. Likewise, a lower-bound ℓ that is set too high suffers from the same outcome. The special case $u = 1$ and $\ell = 0$ results in this approach performing exactly like the Pawlak model.

The VPRS approach has a limitation in its formulation of rough set regions: the user must provided the ℓ and u parameters to define region boundaries. Although the probabilities of each individual object are measurable from the data, correct inclusion of these objects into regions depends on the user's ability to correctly describe the limits of each region. Therefore, the performance of this approach is dependent on the user's previous knowledge of the domain or their intuition pertaining to what they think these limits should be. The decision-theoretic rough set approach calculates region boundaries based on the cost of a classification action and, thus, does not have this limitation.

2.2.2. The decision-theoretic rough set approach

The decision-theoretic rough set (DTRS) approach uses the Bayesian decision procedure which allows for minimum risk decision making based on observed evidence. Let $\mathcal{A} = \{a_1, \ldots, a_m\}$ be a finite set of m possible actions and let $\Omega = \{w_1, \ldots, w_s\}$ be a finite set of s states. $P(w_j|\mathbf{x})$ is calculated as the conditional probability of an object x being in state w_j given the object description $\mathbf{x}$. $\lambda(a_i|w_j)$ denotes the loss, or cost, for performing action a_i when the state is w_j. The expected loss (conditional risk) associated with taking action a_i is given by [12,13]:

$$R(a_i|\mathbf{x}) = \sum_{j=1}^{s} \lambda(a_i|w_j) P(w_j|\mathbf{x}). \tag{7}$$

Object classification with the approximation operators can be fitted into the Bayesian decision framework. The set of actions is given by $\mathcal{A} = \{a_P, a_N, a_B\}$, where a_P, a_N, and a_B represent the three actions in classifying an object into POS(A), NEG(A), and BND(A) respectively. To indicate whether an element is in A or not in A, the set of states is given by $\Omega = \{A, A^c\}$. Let $\lambda(a_\diamond|A)$ denote the loss incurred by taking action $a_\diamond$ when an object belongs to A, and let $\lambda(a_\diamond|A^c)$ denote the loss incurred by take the same action when the object belongs to A^c [14].

Let λ_{PP} denote the loss function for classifying an object in A into the POS region, λ_{BP} denote the loss function for classifying an object in A into the BND region, and let λ_{NP} denote the loss function for classifying an object in A into the NEG region. A loss function $\lambda_{\diamond N}$ denotes the loss of classifying an object that does not belong to A into the regions specified by $\diamond$.

The expected loss $R(a_\diamond|[x])$ associated with taking the individual actions can be expressed as:

$$\begin{aligned} R(a_P|[x]) &= \lambda_{PP} P(A|[x]) + \lambda_{PN} P(A^c|[x]), \\ R(a_N|[x]) &= \lambda_{NP} P(A|[x]) + \lambda_{NN} P(A^c|[x]), \\ R(a_B|[x]) &= \lambda_{BP} P(A|[x]) + \lambda_{BN} P(A^c|[x]), \end{aligned} \tag{8}$$

where $\lambda_{\diamond P} = \lambda(a_\diamond|A)$, $\lambda_{\diamond N} = \lambda(a_\diamond|A^c)$, and $\diamond = P$, N, or B. If we consider the loss functions $\lambda_{PP} \leq \lambda_{BP} < \lambda_{NP}$ and $\lambda_{NN} \leq \lambda_{BN} < \lambda_{PN}$, the following decision rules are formulated (P, N, B) [12]:

P: If $P(A|[x]) \geq \gamma$ and $P(A|[x]) \geq \alpha$, decide POS (A);
N: If $P(A|[x]) \leq \beta$ and $P(A|[x]) \leq \gamma$, decide NEG (A);
B: If $\beta \leq P(A|[x]) \leq \alpha$, decide BND ($A$);

where,

$$\alpha = \frac{\lambda_{PN} - \lambda_{BN}}{(\lambda_{BP} - \lambda_{BN}) - (\lambda_{PP} - \lambda_{PN})},$$
$$\gamma = \frac{\lambda_{PN} - \lambda_{NN}}{(\lambda_{NP} - \lambda_{NN}) - (\lambda_{PP} - \lambda_{PN})},$$
$$\beta = \frac{\lambda_{BN} - \lambda_{NN}}{(\lambda_{NP} - \lambda_{NN}) - (\lambda_{BP} - \lambda_{BN})}. \tag{9}$$

The α, β, and γ values define the different regions, giving us an associated risk for classifying an object. When $\alpha > \beta$, we get $\alpha > \gamma > \beta$ and can simplify (P, N, B) into (P1, N1, B1) [12,14]:

P1: If $P(A|[x]) \geq \alpha$, decide POS(A);
N1: If $P(A|[x]) \leq \beta$, decide NEG(A);
B1: If $\beta < P(A|[x]) < \alpha$, decide BND($A$).

When $\alpha = \beta = \gamma$, we can simplify the rules (P–B) into (P2–B2) [12]:

P2: If $P(A|[x]) > \alpha$, decide POS(A);
N2: If $P(A|[x]) < \alpha$, decide NEG(A);
B2: If $P(A|[x]) = \alpha$, decide BND(A).

The above rules divide the regions based solely on α. These minimum-risk decision rules offer us a basic foundation on which to build a rough set risk analysis component for a Web-based Support System [2]. The DTRS approach has also been successfully used for data mining [15], feature selection [16], and information retrieval [17]. It gives us the ability to not only collect decision rules from data, but also the calculated risk that is involved when discovering (or acting upon) those rules.

It has been shown [12] that the Pawlak and variable-precision approaches are derivable with the decision-theoretic model when the α and β parameters take on certain values. Thus, the variable-precision model can be considered as an intermediate step when using the decision-theoretic approach for rough analysis. A remaining challenge to be addressed in the decision-theoretic approach includes the question regarding what should occur when the loss functions cannot be provided. Determining how much the loss functions can change while maintaining sufficient classification abilities was recently formulated [14]. Methods for calculating the loss functions from the data itself still need to be discovered.

3. Decision types for choosing a rough set model

The basic approach to make decisions with a rough set model is to analyze a data set in order to acquire lower and upper approximations. The creation of regions are based upon this approximation of the universe. The objects of the universe contained in these regions can be used as a basis to make informed decisions. Yao showed that there are two kinds of rules when considering the positive and boundary regions for classification, namely, a positive rule ($\rightarrow_P$) and a boundary rule ($\rightarrow_B$) [18]. For objects x in the positive region and objects x' in the boundary region, he defines these rules as [12]:

$\rightarrow_P$: $[x] \xrightarrow{c>\alpha} A$,
$\rightarrow_B$: $[x'] \xrightarrow{\beta<c<\alpha} A$,

where c is a confidence measure of the rule. Fig. 1 shows the division of a universe of discourse and the appropriate types of decisions that can be made. In the situation where objects x'' lie in the negative region, a third type of rule, $\rightarrow_N$, is introduced [2]. These rules offer the same decision type as those of $\rightarrow_P$, since we can conclude, with certainty, the classification of an object (in this case, the negative region). Based on the regions from these approximations, rules can be gathered. These rules can then be used for guiding decisions.

With the three regions (POS, BND, and NEG), there are two types of decisions that a rough set component can offer for decision making [19]:

(1) *Immediate decisions* (Unambiguous) — These types of decisions are based upon classification within the various POS and NEG regions. We can classify into the NEG regions when no rules can be used to classify into the POS regions. These are labeled as POS, NEG, POS_1, and NEG_0 in Fig. 1. The user can interpret the findings as:
   (a) Classification to POS regions are a "yes" answer. (Regions POS and POS_1).
   (b) Classification to NEG regions are a "no" answer. (Regions NEG and NEG_0).
(2) *Delayed decisions* (Ambiguous) — These types of decisions are based on classification in the various BND regions. Decision regions BND, $BND_{l,u}$, and $BND_{\alpha,\beta}$ in Fig. 1. Users should proceed with a "wait-and-see" agenda since there is uncertainty present. Rough set theory may be meaningless when these cases are too large and unambiguous rules are scarce. Two approaches may be applied to decrease ambiguity:
   (a) Obtain more information [1]. The user can insert more attributes to the information table. They may also conduct further studies to gain knowledge in order to make an immediate decision from the limited data sets.

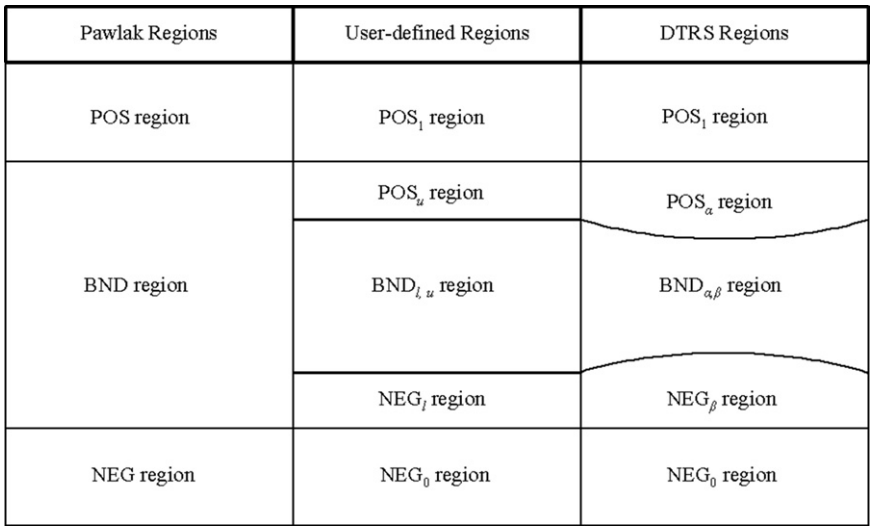

Fig. 1. A graphical depiction of the decisions made from rough regions.

Table 1
Decision types for the Pawlak rough set model.

Region	Decision type
POS(A)	Immediate
BND(A)	Delayed
NEG(A)	Immediate

(b) A decreased tolerance for acceptable loss [5,12] or user thresholds [3]. The probabilistic aspects of the rough set component allows the user to modify the loss functions or thresholds in order to increase certainty. Decision regions POS_u, NEG_l, POS_α, NEG_β in Fig. 1.

In the probabilistic approaches, the entire positive region can be derived from the union of the two sub-regions. The variable-precision approach calculates this with $\text{POS} = \text{POS}_1 \cup \text{POS}_u$ whereas the decision-theoretic approach calculates this with $\text{POS} = \text{POS}_1 \cup \text{POS}_\alpha$. Note that the new positive regions can no longer be thought of as regions that contain correct classifications at all times, since the additions of the POS_u and POS_α regions reduce the certainty in classification.

3.1. Decisions from the Pawlak rough set model

(1) *Immediate* — We can definitely classify an object x in this situation. According to the probability of an object x is in A given the description $[x]$, the following occurs:
   (a) If $P(A|[x]) = 1$, then x is in $\text{POS}(A)$.
   (b) If $P(A|[x]) = 0$, then x is in $\text{NEG}(A)$.

(2) *Delayed* — There is a level of uncertainty when classifying x in this situation. According to the probability of an object x is in A given the description $[x]$, the following occurs:
   If $0 < P(A|[x]) < 1$, then x is in $\text{BND}(A)$.

The decision regions are given as follows:

$$\begin{aligned}\text{POS}(A) &= \underline{apr}(A)\\ &= \{x \in U | P(A|[x]) = 1\},\end{aligned} \tag{10}$$

$$\begin{aligned}\text{BND}(A) &= \overline{apr}(A) - \underline{apr}(A)\\ &= \{x \in U | 0 < P(A|[x]) < 1\},\end{aligned} \tag{11}$$

$$\begin{aligned}\text{NEG}(A) &= U - \overline{apr}(A)\\ &= \{x \in U | P(A|[x]) = 0\}.\end{aligned} \tag{12}$$

The available decisions that can be made from the Pawlak model are summarized in Table 1. From this table, there are three types of decisions that can be made.

Table 2
Decision types for user-defined regions.

Region	Decision type
$\mathrm{POS}_1(A)$	Pure immediate
$\mathrm{POS}_u(A)$	User-accepted immediate
$\mathrm{BND}_{l,u}(A)$	Delayed
$\mathrm{NEG}_l(A)$	User-rejected immediate
$\mathrm{NEG}_0(A)$	Pure immediate

3.2. Decisions from user-defined regions

(1) *Pure immediate* — We can definitely classify x in this situation. According to the probability of an object x is in A given the description $[x]$, the following happens:
   (a) If $P(A|[x]) = 1$, then x is in $\mathrm{POS}_1(A)$.
   (b) If $P(A|[x]) = 0$, then x is in $\mathrm{NEG}_0(A)$.

(2) *User-accepted immediate* — The classification ability is greater than a user-defined upper-bound threshold. According to the probability of an object x is in A given the description $[x]$, the following happens:
   If $u \leq P(A|[x]) < 1$, then x is in $\mathrm{POS}_u(A)$.

(3) *User-rejected immediate* — The classification ability is less than a user-defined lower-bound threshold. According to the probability of an object x is in A given the description $[x]$, the following happens:
   If $0 < P(A|[x]) \leq \ell$, then x is in $\mathrm{NEG}_u(A)$.

(4) *Delayed* — There is a level of uncertainty when classifying x in this situation, between the user thresholds. According to the probability of an object x is in A given the description $[x]$, the following happens:
   If $\ell < P(A|[x]) < u$, then x is in $\mathrm{BND}_{l,u}(A)$.

The positive decision regions show an expanded sense of certainty. POS_1 contains those objects that have been classified with certainty. POS_u implies that the uncertainty of the object classification is within acceptable levels from the decision maker's perspective. They are given as follows:

$$\begin{aligned}\mathrm{POS}_1(A) &= \underline{apr_1}(A) \\ &= \{x \in U | P(A|[x]) = 1\}, \end{aligned} \tag{13}$$

$$\begin{aligned}\mathrm{POS}_u(A) &= \underline{apr_\alpha}(A) \\ &= \{x \in U | u \leq P(A|[x]) < 1\}. \end{aligned} \tag{14}$$

Classification into the boundary region occurs when the calculated probability lies between the user-defined thresholds l and u.

$$\begin{aligned}\mathrm{BND}_{l,u}(A) &= \overline{apr}(A) - (\underline{apr_u}(A) \cup \underline{apr_1}(A)) \\ &= \{x \in U | \ell < P(A|[x]) < u\}. \end{aligned} \tag{15}$$

NEG_0 implies that there is no uncertainty when an object x does not belong in A. NEG_l implies that there is no uncertainty from the decision maker's perspective when x does not belong to A. They are given as follows:

$$\begin{aligned}\mathrm{NEG}_0(A) &= U - (\mathrm{NEG}_l(A) - \overline{apr_1}(A) \\ &= \{x \in U | P(A|[x]) = 0\}, \end{aligned} \tag{16}$$

$$\begin{aligned}\mathrm{NEG}_l(A) &= U - (\mathrm{NEG}_0(A) - \overline{apr_1}(A) \\ &= \{x \in U | 0 < P(A|[x]) \leq l\}. \end{aligned} \tag{17}$$

We see that there are five types of decisions that can be made with the user-defined regions in Table 2. From a theoretical perspective, three decision types are apparent since the regions POS_1 and NEG_0 are special binary cases for POS_u and NEG_l for $u = 1$ and $\ell = 0$ respectively. However, from a practical decision perspective, the types of decisions that can be made from these special cases are distinct enough to warrant their own decision type, increasing this total to five types.

3.3. Decisions from the decision-theoretic rough set approach

The α and β thresholds in the DTRS approach are calculated through the combinations of λ loss functions. The following decisions types are therefore based on the expected cost of a classification action rather than arbitrary user-defined values.

(1) *Pure immediate* — We can definitely classify x in this situation. According to the probability of an object x is in A given the description $[x]$, the following occurs:
   (a) If $P(A|[x]) = 1$, then x is in $\mathrm{POS}_1(A)$.
   (b) If $P(A|[x]) = 0$, then x is in $\mathrm{NEG}_0(A)$.

Table 3
Decision types for the decision-theoretic rough set approach.

Region	Decision type
$\mathrm{POS}_1(A)$	Pure immediate
$\mathrm{POS}_\alpha(A)$	Accepted loss immediate
$\mathrm{BND}_{\alpha,\beta}(A)$	Delayed
$\mathrm{NEG}_\beta(A)$	Rejected loss immediate
$\mathrm{NEG}_0(A)$	Pure immediate

(2) *Accepted loss immediate* — Certain λ loss functions are utilized in order to derive the probability α, the minimum probability needed to be classified into the positive region, as defined in Eq. (9). In other words, the classification ability is greater than α. According to the probability of an object x is in A given the description $[x]$, the following occurs:
If $\alpha \leq P(A|[x]) < 1$, then x is in $\mathrm{POS}_\alpha(A)$.
(3) *Rejected loss immediate* — Other loss functions are adopted to find β, defined in Eq. (9). The classification ability is less than a β-based loss function. According to the probability of an object x is in A given the description $[x]$, the following occurs:
If $0 < P(A|[x]) \leq \beta$, then x is in $\mathrm{NEG}_\beta(A)$.
(4) *Delayed* — There is a level of uncertainty when classifying x in this situation, given by the fact that the probabilities lie between α and β derived from loss functions in Eq. (9). According to the probability of an object x is in A given the description $[x]$, the following occurs:
If $\beta < P(A|[x]) < \alpha$, then x is in $\mathrm{BND}_{\alpha,\beta}(A)$.

The positive decision regions show an expanded sense of certainty. Using the DTRS approach, two immediate decisions can arise from this classification (pure immediate and accepted loss immediate). The POS_1 region implies that there is no uncertainty when classifying object x. POS_α implies that there is an acceptable risk (loss) associated with classifying object x into A. They are given as follows:

$$\begin{aligned}\mathrm{POS}_1(A) &= \underline{apr_1}(A)\\ &= \{x \in U | P(A|[x]) = 1\},\end{aligned} \tag{18}$$

$$\begin{aligned}\mathrm{POS}_\alpha(A) &= \underline{apr_\alpha}(A)\\ &= \{x \in U | \alpha \leq P(A|[x]) < 1\}.\end{aligned} \tag{19}$$

Classification into the boundary region occurs when the calculated probability lies between the derived α and β loss values. That is, those objects that do not meet acceptable loss criteria are considered uncertain in their classification. Delayed decisions arise from the following situation in the DTRS approach:

$$\begin{aligned}\mathrm{BND}_{\alpha,\beta}(A) &= \overline{apr}(A) - (\underline{apr_\alpha}(A) \cup \underline{apr_1}(A))\\ &= \{x \in U | \beta < P(A|[x]) < \alpha\}.\end{aligned} \tag{20}$$

Again, the DTRS approach allows for two more immediate decisions to arise (pure immediate and rejected loss immediate). The NEG_0 region implies that there is no uncertainty when object x does not belong in A. The NEG_β region implies that there is an acceptable risk of not classifying object x into A. They are given as follows:

$$\begin{aligned}\mathrm{NEG}_0(A) &= U - (\mathrm{NEG}_\beta(A) - \overline{apr}(A)\\ &= \{x \in U | P(A|[x]) = 0\},\end{aligned} \tag{21}$$

$$\begin{aligned}\mathrm{NEG}_\beta(A) &= U - (\mathrm{NEG}_0(A) - \overline{apr}(A)\\ &= \{x \in U | 0 < P(A|[x]) \leq \beta\}.\end{aligned} \tag{22}$$

The regions POS_1 and NEG_0 are again special cases for POS_α and NEG_β for $\alpha = 1$ and $\beta = 0$ respectively, similar to that of the VPRS-based decisions. However, these decisions are inherently different from those of the VPRS approach because of the differences in the implication of the parameters and thresholds in both models. The l and u parameters are user-provided. Any decisions made from using these parameters should be considered as being influenced by the user. In contrast, when using the α and β thresholds, the decisions made by the user are only influenced by the scientific observations of the data.

The decision regions derived from the VPRS approach allow for the classification of objects from the decision maker's perspective. Although the VPRS approach and the DTRS approach look remarkably similar, they are fundamentally different in respect to the types of decisions that they can provide [19]. We can see that there are five types of decision that can be made with the DTRS approach in Table 3.

In choosing a probabilistic rough set model for decision making purposes, one should consider the amount of descriptive information that is available [19]. The DTRS approach can still be used even if the user's decision has no risk or cost consideration or the user is capable of providing meaningful thresholds for defining the decision regions (they can provide their own region boundaries). The strength of the DTRS approach emerges when the cost or risk elements are beneficial for the decisions and a decreased user involvement is desired in order to minimize user error.

Table 4
A financial time-series data set.

Date	MACD	MA5	MA12	PROC	RSI	Decision
1991-07-16	10.45598	0	−1	0.46997	67.29899	0
1991-07-18	8.66146	−1	−1	1.83377	75.16101	−1
1991-07-22	6.79574	−1	−1	0.38667	63.90746	0
1991-07-23	6.13677	−1	−1	0.57888	61.94267	−1
1991-07-24	5.44940	0	1	−0.60159	49.67291	−1

Table 5
A spam data set.

Email	Message Length	Word 1	Word 2	CAP Length	REC Length	Decision
E_1	212	0	0	1	2	0
E_2	15	4	0	14	2	1
E_3	365	0	0	3	1	0
E_4	291	0	11	32	0	1
E_5	48	3	0	9	0	1
E_6	107	0	1	2	3	0

4. Examples of rough set model use

To see how the two models discussed affect the types of decisions that can be made, let us consider an example for each. The Pawlak model can be used for financial time-series analysis, specifically, stock market data. The variable-precision rough set approach example will be spam filtering. The decision-theoretic rough set approach example will look at medical diagnosis.

4.1. Pawlak model example

As we have stated previously, the Pawlak model is capable of supporting three types of decisions: two immediate decision types corresponding to the POS and NEG regions, and one delayed decision type corresponding to the BND region. The traditional rough set approach has been used with some success for stock market timing decisions [20,21]. Table 4 shows a typical example of a financial time-series information table.

Each object is a particular date on the calendar year. The conditional attributes represent a set of financial measures derived from the actual opening, closing, high, and low price of a particular stock index. Referring to Table 1, the immediate decision types refer to those decision rules with accuracy measures of 100% or 0%. This is rarely, usually never, the case for forecasting stock prices. The delayed decision type corresponds to rules generated with soft values of accuracy with the range of 0 and 1.

The delayed decision type corresponding to a rule set that is derived from data implies that the user should be aware that there an amount of uncertainty for any actions undertaken. Using the rules to help invest their money better is always an uncertain undertaking.

The first option one can employ to decrease ambiguity is that of obtaining more information. This would include gathering more data and adding more attributes, most likely derived using computation finance techniques. Adding more attributes to the information table creates a finer partitioning of the universe, thus, potentially creating a finer boundary region. Application domains where domain experts are not available at the time or loss functions cannot be calculated, can use the Pawlak rough set model for the data analysis.

4.2. User-defined probabilities example

The variable-precision rough set approach is capable of making five types of decisions, as shown in Table 2. These include the user-accepted and user-rejected immediate decisions, a delayed decision, and two special cases of pure immediate decision types corresponding to the POS_1 and NEG_0 regions.

Spam filtering can seem to be unreliable at times, with important messages sometimes being flagged as spam and not receiving attention from the user,while other messages that should be classified as spam are presented as legitimate messages.

Spam filtering has been a recent application of the variable-precision rough set approach [22,23]. This application domain typically contains a set of records (emails) with measures corresponding to popular spam keyword information, capitalization, size of recipient list etc. Table 5 is an example of an information table pertaining to email data.

The decision attribute takes a value of 0 if the email is considered a legitimate correspondence and 1 if the email is spam. Based on previous data, an expert user can obtain a lower and upper bound corresponding to the probabilities that a particular message is spam. For example, let us say that the expert guessed the following values: $\ell = 0.1$ and $u = 0.9$. The

Table 6
An information table.

Patient	S_1	S_2	S_3	S_4	S_5	Decision
o_1	$s_{1,1}$	$s_{1,2}$	$s_{1,3}$	$s_{1,4}$	$s_{1,5}$	d_1
o_2	$s_{2,1}$	$s_{2,2}$	$s_{2,3}$	$s_{2,4}$	$s_{2,5}$	d_2
o_3	$s_{3,1}$	$s_{3,2}$	$s_{3,3}$	$s_{3,4}$	$s_{3,5}$	d_3
o_4	$s_{4,1}$	$s_{4,2}$	$s_{4,3}$	$s_{4,4}$	$s_{4,5}$	d_4
o_5	$s_{5,1}$	$s_{5,2}$	$s_{5,3}$	$s_{5,4}$	$s_{5,5}$	d_5
o_6	$s_{6,1}$	$s_{6,2}$	$s_{6,3}$	$s_{6,4}$	$s_{6,5}$	d_6

following regions are then calculated,

$$\begin{aligned}
&\mathrm{POS}_1(A) = \{x \in U | P(A|[x]) = 1\},\\
&\mathrm{POS}_{0.9}(A) = \{x \in U | 0.9 \leq P(A|[x]) < 1\},\\
&\mathrm{BND}_{0.1,0.9}(A) = \{x \in U | 0.1 < P(A|[x]) < 0.9\},\\
&\mathrm{NEG}_{0.1}(A) = \{x \in U | 0 < P(A|[x]) \leq 0.1\},\\
&\mathrm{NEG}_0(A) = \{x \in U | P(A|[x]) = 0\}.
\end{aligned} \tag{23}$$

Email messages with a probability $0.9 \leq P(A|[x]) < 1$ are considered spam in this system, as the domain expert has deemed that this will obtain the best region divisions. This is a *user-accepted immediate* decision type. Likewise, email messages with a probability $0 < P(A|[x]) \leq 0.1$ will be considered legitimate messages. This is a *user-rejected immediate* decision type. Of course, the special cases where $P(A|[x]) = 1$ or $P(A|[x]) = 0$ result in a *pure immediate* decision type. Email messages with a probability that lies between the thresholds, $0.1 \leq P(A|[x]) < 0.9$ are considered uncertain and require more information in order to classify correctly. This is a *delayed* decision type, reflecting a "wait-and-see" agenda suggestion to the user.

Although the spam detection application domain is important for reasons including information technology resource management and security, the risk or consequences of an incorrect email classification is less than the next example: medical diagnosis. We show that the decision-theoretic approach is more suitable for this next domain.

4.3. Expected cost probabilities example

The Pawlak rough set model has been used for medical diagnosing with wide success [24,25]. The decision-theoretic rough set approach can extend the application of rough sets in this domain by utilizing the loss functions based on the auditing of hospital performance [2]. In the case of medical support systems, the somewhat random or arbitrary selection of upper and lower probabilities by experts for inclusion into regions, as done by the variable-precision approach, is not applicable, and perhaps dangerous for this serious set of problems.

Table 6 is a generic information table consisting of six objects and six attributes.

Each object represents a particular patient record in a hospital patient database. The attributes $S = \{S_1, S_2, S_3, S_4, S_5\}$ correspond to symptoms that each patient is having. The decision attribute is the diagnosis for that particular patient.

The risk or cost is defined as consequences of the wrong diagnosis. Based on our common sense, the cost of the wrong diagnosis of a flu is lower than that of the wrong diagnosis of cancer. The cancer diagnosis tolerance levels of either a false-negative or false-positive are very low. Patients may sacrifice their lives when a false-negative level is high, as they may miss the best treatment time. They may suffer consequences of chemotherapy for non-existent cancer when a false-positive is high.

Using Table 6, let us form two hypothetical scenarios of patient diagnoses. First, a diagnosis of low severity with a low cost for a wrong diagnosis. This could be testing for a patient's minor allergies. An allergy test would be looking for positive indicators for symptoms $S = \{S_1, S_2, S_3\}$ and the diagnosis decision $D = \{\text{Decision}\}$. Below is a typical sample of loss functions for this situation:

$$\lambda_{PN} = \lambda_{NP} = 1u, \qquad \lambda_{PP} = \lambda_{NN} = \lambda_{BP} = \lambda_{BN} = 0, \tag{24}$$

where u is a unit cost determined by the individual administration. In this scenario, the administration has deemed that a false-positive (λ_{PN}) and false-negative (λ_{NP}) diagnosis has some form of cost whereas indeterminate diagnoses (λ_{BP}, λ_{BN}) and correct diagnoses (λ_{PP}, λ_{NN}) have no cost.

A diagnosis of high severity could have a high cost for a wrong diagnosis. This could be testing for whether a patient has a form of cancer. In Table 1, the cancer test would be looking for positive indicators for symptoms $S = \{S_1, S_4, S_5\}$ and the diagnosis decision $D = \{\text{Decision}\}$. Below is a typical sample of loss functions for this situation:

$$\lambda_{PN} = \lambda_{NP} = 2u, \qquad \lambda_{BP} = \lambda_{BN} = 1u, \qquad \lambda_{PP} = \lambda_{NN} = 0, \tag{25}$$

where u is a unit cost determined by the individual administration. In this scenario, the administration has deemed that a false-positive (λ_{PN}) and false-negative (λ_{NP}) diagnosis is twice as costly as indeterminate diagnoses (λ_{BP} and λ_{BN}). Correct diagnoses (λ_{PP} and λ_{NN}) have no cost.

Using the loss functions in (24) and calculating the parameters using the formulas in (9), we obtain $\alpha = 1$, $\gamma = 0.5$, and $\beta = 0$. When $\alpha > \beta$, we get $\alpha > \gamma > \beta$. We use the simplified decision rules (P1-B1) to obtain our lower and upper approximations:

$$\underline{apr}_{(1,0)}(A) = \{x \in U | P(A|[x]) = 1\},$$
$$\overline{apr}_{(1,0)}(A) = \{x \in U | P(A|[x]) > 0\}. \tag{26}$$

Using the loss functions in (25) and calculating the parameters using the formulas in (9), we obtain $\alpha = \beta = \gamma = 0.5$. When $\alpha = \beta = \gamma$, we use the simplified decision rules (P2–B2) to can obtain our new lower and upper approximations:

$$\underline{apr}_{(0.5,0.5)}(A) = \{x \in U | P(A|[x]) > 0.5\},$$
$$\overline{apr}_{(0.5,0.5)}(A) = \{x \in U | P(A|[x]) \geq 0.5\}. \tag{27}$$

The approximations in (26) mean that we can definitely class patient x into diagnosis class A if all similar patients are in diagnosis class A. The low loss functions (24) have indicated that users of the system can have high certainty when dealing with this class of patient. The approximations in (27) mean that we can definitely class patient x into diagnosis class A if strictly more than half of similar patients are in diagnosis class A. These examples use the loss functions to determine how high the level of certainty regarding a patient's symptoms needs to be in order to minimize cost.

5. Concluding remarks

When analysis is performed using the Pawlak model, the situation may arise where the boundary region is too large. If the user is unable to gather additional information within the table in order to decrease this ambiguity, he or she must decrease their acceptable levels of precision. This involves moving away from the Pawlak model and continuing on with the probabilistic model. The probabilistic model gives five regions, as shown in Fig. 1, in order to provide a more configurable and robust representation of data. The expected cost of performing a classification action, provided through loss functions, can be used to calculate probabilities that define region boundaries. The user may also provide probabilities based on their intuition to determine the region boundaries.

We present a list of decision types based on rough set regions created by two models: Pawlak and probabilistic. First, three types of decisions can be made when the Pawlak model is used. If additional information from the user or loss functions are not available, the user cannot use the probabilistic rough set models to their full potential. The two immediate and one delayed decision types correspond to the POS, NEG, and BND regions. We demonstrated these decision types through a financial time-series example.

Secondly, ten types of decisions can be made using the probabilistic model. These consist of five types of decisions pertaining to user-specified region boundaries and five types pertaining to the probability derivation by the expected cost of a classification action. The filtering of spam from an email system is a suitable application for user-defined region boundaries. Medical diagnosis is an excellent example of utilizing decision-theoretic rough set decision types as loss functions should be readily available.

In total, our procedure details thirteen types of decisions that can be made using rough sets. It is hoped that the outline helps decision makers to choose which particular rough set model is best for their decision goals. Although we examined three separate examples in parallel to demonstrate the decision types, we believe that there could be a sequential application of moving from one rough set model to another. In the medical domain, it may be impossible to gain additional information regarding patients and, thus, the Pawlak model is unsuitable. The second option is to move onto a probabilistic model. However, a domain expert may be unable to give accurate lower and upper bound estimates. Ultimately, one could examine the hospital's tolerance for mistreatment and construct loss functions, therefore enabling the use of the decision-theoretic approach. This sequence of refining the decision types is an avenue of research we wish to explore in the future.

Acknowledgments

The authors wish to sincerely thank their colleague, Dr. Yiyu Yao, for his discussion and suggestions. This work is partially supported by a Discovery Grant from NSERC Canada and the University of Regina FGSR Dean's Scholarship Program.

References

[1] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (1982) 341–356.
[2] J.T. Yao, J.P. Herbert, Web-based support systems based on rough set analysis, in: Proceedings of Rough Sets and Emerging Intelligent Systems Paradigms, RSEISP'07, in: Lecture Notes in Artificial Intelligence, vol. 4585, 2007, pp. 360–370.
[3] W. Ziarko, Variable precision rough set model, Journal of Computer and System Sciences 46 (1993) 39–59.
[4] Y.Y. Yao, Information granulation and approximation in a decision-theoretical model of rough sets, in: L. Polkowski, S.K. Pal, A. Skowron (Eds.), Rough-neuro Computing: A Way to Computing with Words, Springer, Berlin, 2003, pp. 491–516.
[5] Y.Y. Yao, S.K.M. Wong, A decision theoretic framework for approximating concepts, International Journal of Man-machine Studies 37 (1992) 793–809.
[6] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academatic Publishers, Boston, 1991.
[7] J.F. Peters, A. Skowron, A rough set approach to knowledge discovery, International Journal of Intelligent Systems 17 (2) (2002) 109–112.

[8] W. Ziarko, Acquisition of hierarchy-structured probabilistic decision tables and rules from data, Expert Systems 20 (2003) 305–310.
[9] W. Ziarko, X. Fei, VPRSM approach to web searching, in: Lecture Notes In Artificial Intelligence, vol. 2475, 2002, pp. 514–521.
[10] J.D. Katzberg, W. Ziarko, Variable precision rough sets with asymmetric bounds, in: W. Ziarko (Ed.), Rough Sets, Fuzzy Sets and Knowledge Discovery, Springer, London, 1994, pp. 167–177.
[11] Y.Y. Yao, Probabilistic rough set approximations, International Journal of Approximate Reasoning 49 (2) (2008) 255–271.
[12] Y.Y. Yao, Decision-theoretic rough set models, in: Proceedings of Rough Sets and Knowledge Technology, RSKT'07, in: Lecture Notes in Artificial Intelligence, vol. 4481, 2007, pp. 1–12.
[13] Y.Y. Yao, S.K.M. Wong, P. Lingras, A decision-theoretic rough set model, in: Z.W. Ras, M. Zemankova, M.L. Emrich (Eds.), in: Methodologies for Intelligent Systems, vol. 5, North-Holland, New York, 1990, pp. 17–24.
[14] J.P. Herbert, J.T. Yao, Game-theoretic risk analysis in decision-theoretic rough sets, in: Proceedings of Rough sets and Knowledge Technology, RSKT'08, in: Lecture Notes in Artificial Intelligence, vol. 5009, 2008, pp. 132–139.
[15] S. Tsumoto, Accuracy and coverage in rough set rule induction, in: Lecture Notes in Artificial Intelligence, vol. 2475, 2002, pp. 373–380.
[16] J.T. Yao, M. Zhang, Feature selection with adjustable criteria, in: Lecture Notes in Artificial Intelligence, vol. 3641, 2005, pp. 204–213.
[17] Y. Li, C. Zhang, J.R. Swanb, Rough set based model in information retrieval and filtering, in: Proceedings of the 5th International Conference on Information Systems Analysis and Synthesis, 1999, pp. 398–403.
[18] Y.Y. Yao, Probabilistic approaches to rough sets, Expert Systems 20 (5) (2003) 287–297.
[19] J.P. Herbert, J.T. Yao, Rough set model selection for practical decision making, in: Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD'07, vol. III, 2007, pp. 203–207.
[20] J. Herbert, J.T. Yao, Time-series data analysis with rough sets, in: Proceedings of the 4th International Conference on Computational Intelligence in Economics and Finance, CIEF'05, 2005, pp. 908–911.
[21] L. Shen, H.T. Loh, Applying rough sets to market timing decisions, Decision Support Systems 37 (4) (2004) 583–597.
[22] M. Glymin, W. Ziarko, Rough set approach to spam filter learning, in: Proceedings of Rough Sets and Emerging Intelligent Systems Paradigms, RSEISP'07, in: Lecture Notes in Artificial Intelligence, vol. 4585, 2007, pp. 350–359.
[23] W.Q. Zhao, Y.L. Zhu, Classifying email using variable precision rough set approach, in: Lecture Notes In Artificial Intelligence, vol. 4062, 2006, pp. 766–771.
[24] P. Mitra, S. Mitra, S.K. Pal, Staging of cervical cancer with soft computing, IEEE Transactions on Biomedical Engineering 47 (7) (2000) 934–940.
[25] S. Tsumoto, Automated extraction of medical expert system rules from clinical databases based on rough set theory, Information Sciences 112 (1–4) (1998) 67–84.