J. P. Herbert, J.T. Yao, Rough Set Model Selection for Practical Decision Making, Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'07), Aug 24-27, 2007, Hainan, China, III, pp203-207.

Rough Set Model Selection for Practical Decision Making

Joseph P. Herbert JingTao Yao
Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada, S4S 0A2
{herbertj, jtyao}@cs.uregina.ca

Abstract

One of the challenges a decision maker faces is choosing a suitable rough set model to use for data analysis. The traditional algebraic rough set model classifies objects into three regions, namely, the positive, negative, and boundary regions. Two different probabilistic models, variable-precision and decision-theoretic, modify these regions via l,u user-defined thresholds and α, β values from loss functions respectively. A decision maker whom uses these models must know what type of decisions can be made within these regions. This will allow him or her to conclude which model is best for their decision needs. We present an outline that can be used to select a model and better analyze the consequences and outcomes of those decisions.

1. Introduction

Rough set theory is a way of representing and reasoning imprecision and uncertain information in data [4]. It deals with the approximation of sets constructed from descriptive data elements. This is helpful when trying to discover decision rules, important features, and minimization of conditional attributes. Rough sets creates three regions, namely, the positive, negative and boundary regions. These regions can be used for making decisions regarding "yes", "no", and "wait-and-see" cases. This method of data analysis is very useful for data mining [2, 6].

Researchers have extended the algebraic rough set model into probabilistic approaches. The variable-precision rough set model [14] and decision theoretic rough set model [12, 10] expand the POS and NEG regions by using l, u thresholds and α, β values respectively. The purpose of expanding the POS and NEG regions is to increase our certainty about the knowledge (rules) obtained through rough set analysis.

Decision makers that use these rough sets to aid in their decision making are now faced with the challenge of which model to choose from. An outline that details the kinds of decisions that can be made could be very beneficial.

The organization of this paper is as follows. Section 2 will discuss rough set theory and the extended probabilistic models that expand the positive and negative regions. Section 3 will state the outline based on all three models. We conclude this paper in Section 4.

2. Rough Set Models

We will review the algebraic, variable-precision, and decision-theoretic rough set models in this section.

2.1. Algebraic Rough Set Model

Discerning objects from each other is a major purpose in rough set theory. It may be impossible to precisely describe $A \subseteq U$. 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 [5]:

$$\begin{aligned} \underline{apr}(A) &= \{x \in U | [x] \subseteq A\}, \\ \overline{apr}(A) &= \{x \in U | [x] \cap X \neq \emptyset\}. \end{aligned} \quad (1)$$

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 [4] of A can be defined as:

$$\begin{aligned} POS(A) &= \underline{apr}(A), \\ NEG(A) &= U - \overline{apr}(A), \\ BND(A) &= \overline{apr}(A) - \underline{apr}(A). \end{aligned} \quad (2)$$

2.2. Variable-Precision Rough Set Model

The variable-precision rough set (VPRS) model aims at increasing the discriminatory capabilities of the rough set approach by using parameter grades of conditional probabilities. Two parameters, the lower-bound l and the upper-bound u, are provided by the user.

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

$$POS_u(A) = \{x \in A | P(A|[x]) \geq u\}. \quad (3)$$

Likewise, the l-negative region $NEG_l(A)$ is controlled by the lower-bound l, such that,

$$NEG_l(A) = \{x \in A | P(A|[x]) \leq l\}. \quad (4)$$

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

$$BND_{l,u}(A) = \{x \in A | l < P(A|[x]) l u\}. \quad (5)$$

Since the l and u parameters are given by the user, the quality is user-driven. 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 $l = 0$ results in this model performing exactly like the algebraic model. The VPRS model has been used in many areas [1, 15, 16]

2.3. Decision-Theoretic Rough Set Model

The decision-theoretic rough set (DTRS) model 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 [11, 13]:

$$R(a_i|\mathbf{x}) = \sum_{j=1}^{s} \lambda(a_i|w_j) P(w_j|\mathbf{x}). \quad (6)$$

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 $\Omega = \{A, A^c\}$. $a_\diamond$ 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.

Let λ_{P1} denote the loss function for classifying an object in A into the POS region, λ_{B1} denote the loss function for classifying an object in A into the BND region, and let λ_{N1} denote the loss function for classifying an object in A into the NEG region. A loss function $\lambda_{\diamond 2}$ 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:

$$\begin{aligned} R(a_P|[x]) &= \lambda_{P1} P(A|[x]) + \lambda_{P2} P(A^c|[x]), \\ R(a_N|[x]) &= \lambda_{N1} P(A|[x]) + \lambda_{N2} P(A^c|[x]), \\ R(a_B|[x]) &= \lambda_{B1} P(A|[x]) + \lambda_{B2} P(A^c|[x]), \end{aligned} \quad (7)$$

where $\lambda_{\diamond 1} = \lambda(a_\diamond|A)$, $\lambda_{\diamond 2} = \lambda(a_\diamond|A^c)$, and $\diamond = P$, N, or B. If we consider the loss functions $\lambda_{P1} \leq \lambda_{B1} < \lambda_{N1}$ and $\lambda_{N2} \leq \lambda_{B2} < \lambda_{P2}$, the following decision rules are formulated (P, N, B) [11]:

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,

$$\begin{aligned} \alpha &= \frac{\lambda_{P2} - \lambda_{B2}}{(\lambda_{B1} - \lambda_{B2}) - (\lambda_{P1} - \lambda_{P2})}, \\ \gamma &= \frac{\lambda_{P2} - \lambda_{N2}}{(\lambda_{N1} - \lambda_{N2}) - (\lambda_{P1} - \lambda_{P2})}, \\ \beta &= \frac{\lambda_{B2} - \lambda_{N2}}{(\lambda_{N1} - \lambda_{N2}) - (\lambda_{B1} - \lambda_{B2})}. \end{aligned} \quad (8)$$

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) [11]:

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($A$).

These minimum-risk decision rules offer us a basic foundation in which to build a rough set risk analysis component for a WSS [8]. This model has also been successfully used for data mining [7], feature selection [9], and information retrieval [3]. They give 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.

3. Practical Decision Making for Rough Sets

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. 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 the rough set components can offer for decision making:

1. **Immediate Decisions** (Unambiguous) - These types of decisions are based upon classification within the various POS and NEG regions. The user can interpret the findings as:
   (a) Classification to POS regions are a "yes" answer.
   (b) Classification to NEG regions are a "no" answer.
2. **Delayed Decisions** (Ambiguous) - These types of decisions are based on classification in the various BND regions. Proceeding 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 [4]. The user can insert more attributes to the information table. They may also conduct further studies to gain knowledge in order to make a immediate decision from the limited data sets.
   (b) A decreased tolerance for acceptable loss [11, 12] or user thresholds [14]. The probabilistic aspects of the rough set component allows the user to modify the loss functions or thresholds in order to increase certainty.

3.1. Decisions from the Algebraic Rough Set Model

1. **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 $POS(A)$.
   (b) If $P(A|[x]) = 0$, then x is in $NEG(A)$.
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 happens:

   If $0 < P(A|[x]) < 1$, then x is in $BND(A)$.

Region	Decision Type
$POS(A)$	Immediate
$BND(A)$	Delayed
$NEG(A)$	Immediate

Table 1. Decision types for the algebraic rough set model

The decision regions are given as follows:

$$\begin{aligned} POS(A) &= \underline{apr}(A) \\ &= \{x \in U | P(A|[x]) = 1\}, \quad (9) \\ BND(A) &= \overline{apr}(A) - \underline{apr}(A) \\ &= \{x \in U | 0 < P(A|[x]) < 1\}, \quad (10) \\ NEG(A) &= U - \overline{apr}(A) \\ &= \{x \in U | P(A|[x]) = 0\}. \quad (11) \end{aligned}$$

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

3.2. Decisions from the Variable-Precision Rough Set Model

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 $POS_1(A)$.
   (b) If $P(A|[x]) = 0$, then x is in $NEG_0(A)$.
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 $POS_u(A)$.
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 $0 < P(A|[x]) \leq l$, then x is in $NEG_u(A)$.
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:

Region	Decision Type
$POS_1(A)$	Pure Immediate
$POS_u(A)$	User-accepted Immediate
$BND_{l,u}(A)$	Delayed
$NEG_l(A)$	User-rejected Immediate
$NEG_0(A)$	Pure Immediate

Table 2. Decision types for the variable-precision rough set model

$$\text{If } l < P(A|[x]) < u, \text{ then } x \text{ is in } BND_{l,u}(A).$$

The positive decision regions show an expanded sense of certainty. POS_1 implies that there is no uncertainty when classifying object x. POS_u implies that there is no uncertainty from the decision maker perspective when classifying object x. They are given as follows:

$$\begin{aligned} POS_1(A) &= \underline{apr_1}(A) \\ &= \{x \in U | P(A|[x]) = 1\}, \quad (12) \\ POS_u(A) &= \underline{apr_\alpha}(A) \\ &= \{x \in U | u \leq P(A|[x]) < 1\}. \quad (13) \end{aligned}$$

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

$$\begin{aligned} BND_{l,u}(A) &= \overline{apr}(A) - (\underline{apr_u}(A) \cup \underline{apr_1}(A)) \\ &= \{x \in U | l < P(A|[x]) < u\}. \quad (14) \end{aligned}$$

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

$$\begin{aligned} NEG_0(A) &= U - (NEG_l(A) - \overline{apr_1}(A) \\ &= \{x \in U | P(A|[x]) = 0\}, \quad (15) \\ NEG_l(A) &= U - (NEG_0(A) - \overline{apr_1}(A) \\ &= \{x \in U | 0 < P(A|[x]) \leq l\}. \quad (16) \end{aligned}$$

We see that there are five types of decisions that can be made with the VPRS model 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 $l = 0$ 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 Model

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 $POS_1(A)$.
   (b) If $P(A|[x]) = 0$, then x is in $NEG_0(A)$.
2. **Accepted Loss Immediate** - The classification ability is greater than a α-based loss function. According to the probability of an object x is in A given the description $[x]$, the following happens:
   $$\text{If } \alpha \leq P(A|[x]) < 1, \text{ then } x \text{ is in } POS_\alpha(A).$$
3. **Rejected Loss Immediate** - 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 happens:
   $$\text{If } 0 < P(A|[x]) \leq \beta, \text{ then } x \text{ is in } NEG_\beta(A).$$
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:
   $$\text{If } \beta < P(A|[x]) < \alpha, \text{ then } x \text{ is in } BND_{\alpha,\beta}(A).$$

The positive decision regions show an expanded sense of certainty. Using the DTRS model, 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:

$$\begin{aligned} POS_1(A) &= \underline{apr_1}(A) \\ &= \{x \in U | P(A|[x]) = 1\}, \quad (17) \\ POS_\alpha(A) &= \underline{apr_\alpha}(A) \\ &= \{x \in U | \alpha \leq P(A|[x]) < 1\}. \quad (18) \end{aligned}$$

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 model:

$$\begin{aligned} BND_{\alpha,\beta}(A) &= \overline{apr}(A) - (\underline{apr_\alpha}(A) \cup \underline{apr_1}(A)) \\ &= \{x \in U | \beta < P(A|[x]) < \alpha\}. \quad (19) \end{aligned}$$

Region	Decision Type
$POS_1(A)$	Pure Immediate
$POS_\alpha(A)$	Accepted Loss Immediate
$BND_{\alpha,\beta}(A)$	Delayed
$NEG_\beta(A)$	Rejected Loss Immediate
$NEG_0(A)$	Pure Immediate

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

Again, the DTRS model 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:

$$\begin{aligned} NEG_0(A) &= U - (NEG_\beta(A) - \overline{apr}(A) \\ &= \{x \in U | P(A|[x]) = 0\}, \quad (20) \\ NEG_\beta(A) &= U - (NEG_0(A) - \overline{apr}(A) \\ &= \{x \in U | 0 < P(A|[x]) \leq \beta\}. \quad (21) \end{aligned}$$

From a theoretical perspective, the regions POS_1 and NEG_0 are again special binary cases for POS_α and NEG_β for $\alpha = 1$ and $\beta = 0$ respectively, similar to that of the VPRS-based decisions. A practical decision perspective shows that there are five distinct types of decisions.

The decision regions derived from the VPRS model allow for the classification of objects from the decision maker's perspective. Although the VPRS model and the DTRS model look remarkably similar, they are fundamentally different in respect to the types of decisions that they can provide. We can see that there are five types of decisions that can be made with the DTRS model 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. If the user's decision has no risk or cost consideration and if the user is capable of providing meaningful thresholds for defining the decision regions, the VPRS model is suitable. Otherwise, the DTRS model is useful if cost or risk elements are beneficial for the decisions as well as the decreased user involvement.

4. Conclusions

We present a decision outline based on rough set regions created by three models: algebraic, variable-precision, and decision-theoretic. First, three types of decisions can be made when the algebraic model is used. Second, five types of decisions can be made using the VPRS model, with two of these decisions user-driven. Third, five types of decisions can be made using the DTRS model. Two of these decisions are based on an acceptable loss and rejected loss derived from loss functions. In total, our outline details thirteen types of decisions that can be made using rough sets. The outline helps decision makers to choose which particular rough set model is best for their decision goals.

References

[1] J. D. Katzberg and W. Ziarko. Variable precision rough sets with asymmetric bounds. In W. Ziarko, editor, *Rough Sets, Fuzzy Sets and Knowledge Discovery*, pages 167–177, London, 1994. Springer.

[2] A. Kusiak, J. A. Kern, K. H. Kernstine, and B. T. L. Tseng. Autonomous decision-making: A data mining approach. *IEEE Transactions on Information Technology In Biomedicine*, 4(4):274–284, 2000.

[3] Y. Li, C. Zhang, and J. R. Swanb. Rough set based model in information retrieval and filtering. In *Proceedings of the 5th Internation Conference on Information Systems Analysis and Synthesis*, pages 398–403, 1999.

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

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

[6] J. F. Peters and A. Skowron. A rough set approach to knowledge discovery. *International Journal of Intelligent Systems*, 17(2):109–112, 2002.

[7] S. Tsumoto. Accuracy and coverage in rough set rule induction. In *LNAI 2475*, pages 373–380, 2002.

[8] J. T. Yao and J. P. Herbert. Web-based support systems based on rough set analysis. In *Proceedings of RSEISP'07, LNAI*, 2007.

[9] J. T. Yao and M. Zhang. Feature selection with adjustable criteria. In *LNAI 3641*, pages 204–213, 2005.

[10] Y. Y. Yao. *Information granulation and approximation in a decision-theoretical model of rough sets*, pages 491–516. Springer, Berlin, 2003.

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

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

[13] Y. Y. Yao, S. K. M. Wong, and P. Lingras. A decision-theoretic rough set model. In Z. W. Ras, M. Zemankova, and M. L. Emrich, editors, *Methodologies for Intelligent Systems*, volume 5, pages 17–24. North-Holland, New York, 1990.

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

[15] W. Ziarko. Acquisition of heirarchy-structured probabilitic decision tables and rules from data. *Expert Systems*, 20:305–310, 2003.

[16] W. Ziarko and X. Fei. Vprsm approach to web searching. In *Lecture Notes In Artificial Intelligence*, 2475, pages 514–521, 2002.