Thresholds Determination for Probabilistic Rough Sets with Genetic Algorithms

Babar Majeed, Nouman Azam, JingTao Yao

Department of Computer Science
University of Regina
{majeed2b,azam200n,jtyao}@cs.uregina.ca

September 29, 2014

University of Regina

Probabilistic Rough Sets

- Introducing probabilities to define the rough set based approximations with a pair of (α, β) thresholds (Yao, 2008).
  - The (α, β) probabilistic approximations are given by,

$$\begin{aligned} \underline{apr}_{(\alpha,\beta)}(C) &= \bigcup\{[x] \in U/E \mid Pr(C|[x]) \geq \alpha\}, \\ \overline{apr}_{(\alpha,\beta)}(C) &= \bigcup\{[x] \in U/E \mid Pr(C|[x]) > \beta\}. \end{aligned} \quad (1)$$

  - Probabilistic positive, negative and boundary regions:

$$\begin{aligned} \mathrm{POS}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C|[x]) \geq \alpha\}, \\ \mathrm{NEG}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C|[x]) \leq \beta\}, \\ \mathrm{BND}_{(\alpha,\beta)}(C) &= \{x \in U \mid \beta < Pr(C|[x]) < \alpha\}. \end{aligned} \quad (2)$$

Yao, Y. Y., (2008). Probabilistic rough set approximations, International Journal of Approximate Reasoning, 49.

University of Regina

Probabilistic Rough Sets: A Main Result and Key Issue

- A main result of probabilistic rough sets is that the rules for determining the three regions are given by,

$$\begin{aligned} &\text{Acceptance:} && \text{if } P(C|[x]) \geq \alpha, \\ &\text{Rejection:} && \text{if } P(C|[x]) \leq \beta, \text{ and} \\ &\text{Deferment:} && \text{if } \beta < P(C|[x]) < \alpha. \end{aligned} \tag{3}$$

- A major difficulty is the interpretation and determination of the (α, β) thresholds (Yao, 2011).

Yao, Y.Y., (2011). Two semantic issues in a probabilistic rough set model. Fundamenta Informaticae 108(3).

University of Regina

Determination of (α, β) Probabilistic Thresholds

- The determination of probabilistic thresholds can generally be approached as an optimization problem based on criterion C (Deng and Yao, 2012).

$$\arg \min_{(\alpha,\beta)} C(\alpha, \beta), \text{ where}$$

$$C(\alpha, \beta) = C_P(\alpha, \beta) + C_N(\alpha, \beta) + C_B(\alpha, \beta). \quad (4)$$

Deng, X. F., Yao, Y. Y., (2012). An information-theoretic interpretation of thresholds in PRS. In: (RSCTC'12).

University of Regina

Attempts for Determination of Probabilistic Thresholds

- Recent attempts.
  - Optimization viewpoint (Jia et al., 2011),
  - Multi-view model(Li and Zhou, 2011),
  - Method using probabilistic model criteria (Liu et al., 2011),
  - Information-theoretic interpretation (Deng and Yao, 2012),
  - Approach based on fuzzy function (Huang, et al., 2011)
  - Game-theoretic framework (Herbert and Yao, 2011).
- We consider the genetic algorithm based approach.

Jia, X. Y., Li, W. W., Shang, L., Chen, J. J., (2011). An optimization viewpoint of DTRS model. In: (RSKT'11).
Li, H.X., Zhou, X.Z., (2011). Risk decision making based on DTRS... IJCIS 4.
Liu, D., Li, T.R., Ruan, D., (2011). Probabilistic model criteria with DTRS. Information Science 181.
Deng, X. F., Yao, Y. Y., (2012). An information-theoretic interpretation of thresholds in PRS. In: (RSCTC'12).
Huang, K. Y., Chang, T. H. and Chang, T . C., (2011), Determination of the threshold value β of .. IJAR 52
Herbert, J.P., Yao, J.T., 2011. Game-theoretic rough sets. Fundamenta Informaticae 108(3-4).

University of Regina

Threshold Determination Using Genetic Algorithms

- Description of simple genetic algorithm.

Algorithm 1 Genetic algorithm

Input: Initial population
Output: Optimum solution
  Initialize population
  Evaluate population
  While termination criteria is not reached
    Selecting next population using a fitness function
    Perform crossover and mutation
  End

Holland, J.H.: Adaptation in Natural and Artificial Systems. MIT Press, Cam- bridge, MA, USA (1992)

University of Regina

Five Steps Approach using Genetic Algorithms

- Introducing a five step approach for determining thresholds using Genetic Algorithms.
- **Step 1**: Generating initial population.
- **Step 2**: Evaluating population.
- **Step 3**: Termination conditions or criteria.
- **Step 4**: Selecting new population.
- **Step 5**: Performing crossover and mutation.

University of Regina

Step 1: Generating Initial Population

- An encoding mechanism is needed to obtain initial population.
- Encoding mechanism: representing values of variables in an optimization problem.
- Considering binary encoding for the sake of simplicity.
  - Example:
  - A possible encoding of α using two bits,
    - 00 = 0.7, 01 = 0.8, 10 = 0.9 and 11 = 1.0.
  - A possible encoding of β using two bits,
    - 00 = 0.0, 01 = 0.1, 10 = 0.2 and 11 = 0.3
- Increasing the number of bits in encoding may result in more accurate solution.

University of Regina

Correspondence Between Chromosomes and Thresholds

- A chromosome is represented by combining the encodings of one α value and one β value.
  - For instance, the encoding 00 for α and 00 for β are joined to obtained a chromosome 0000.
  - The chromosome 0000 corresponds to thresholds $(\alpha, \beta) = (0.7, 0.0)$, since 00 for α is 0.7 and for β is 0.0.
- Each chromosome represents a threshold pair.
- A set of chromosomes is known as population.
  - The population is a set or collection of threshold pairs.

University of Regina

Initial Population

- Approaches for generating initial population.
  - Random generation.
  - Using domain specific knowledge.
  - Inspection of data.
    - Minimum size boundary region is expected when data contains minimum level of noise and provides precise information.
    - An optimal solution is expected to be in the vicinity of threshold values $(\alpha, \beta) = (1, 0)$, i.e., the Pawlak model.
    - Initial population is therefore selected in the neighbourhood of Pawlak model.

University of Regina

Step 2: Evaluating Population

- Chromosome generated in **Step 1** are evaluated using a fitness function.
    - Fitness Functions: representing the chromosome ability to survive and reproduce.
- The fitness function is also used in selecting chromosomes in different iterations.
- The objective of the algorithm is to optimize the fitness function.

University of Regina

Fitness Function

- Utilizing the information-theoretic rough sets to define the Fitness function.
- A partition based on a concept C is given by $\pi_C = \{C, C^c\}$.
- Another partition based on the (α, β) thresholds is given by,

$$\pi_{(\alpha,\beta)} = \{\mathsf{POS}_{(\alpha,\beta)}(C), \mathsf{NEG}_{(\alpha,\beta)}(C), \mathsf{BND}_{(\alpha,\beta)}(C)\}. \quad (5)$$

- The uncertainty in $\pi_C = \{C, C^c\}$ with respect to the three probabilistic regions may be computed with Shannon entropy (Deng and Yao, 2012).

Deng, X. F., Yao, Y. Y., (2012). An information-theoretic interpretation of thresholds in PRS. In: (RSCTC'12).

University of Regina

Fitness Function

- Denoting the uncertainty in the positive, negative and boundary regions as $\Delta_P(\alpha, \beta), \Delta_N(\alpha, \beta)$ and $\Delta_B(\alpha, \beta)$,
  - The overall uncertainty is given by,

$$\Delta(\alpha, \beta) = \Delta_P(\alpha, \beta) + \Delta_N(\alpha, \beta) + \Delta_B(\alpha, \beta) \tag{6}$$

- Equation (6) represents a fitness function.
- Other fitness functions may be defined using other rough set models such as DTRS.

University of Regina

Step 3: Termination Conditions or Criteria

- Different approaches may be employed in defining stop conditions.
  - Bound on the number of iterations.
  - The evaluations using fitness functions reaching or crossing some limits, i.e., $\Delta(\alpha, \beta) < \tau$
  - Subsequent iterations does not provide any improvements in performance, i.e., $\Delta(\alpha, \beta)_i \approx \Delta(\alpha, \beta)_{i+1}$.

University of Regina

Step 4: Selecting New Population

- The selection mechanisms are used to provide more chances to chromosome with higher evaluations to be selected in the next iterations.
- Roulette-wheel selection mechanism.
  - The chances of selecting a chromosome is seen as spinning a roulette wheel.
  - The size of the slot for each chromosome as being proportional to its normalized fitness.
  - Selection probability is calculated based on normalized fitness.
  - The chromosomes with higher fitness (slot sizes) will have more chance of being chosen.

University of Regina

Step 5: Performing Crossover and Mutation

- A crossover represents the exchange of genetic material between two parents.
- One-point versus multi-point cross over.
  - One-point crossover: cutting and exchanging of parent chromosomes from one point.
  - Multi-point crossover: cutting and exchanging of parent chromosomes from more than one point.
- Mutation is applied to chromosomes by inverting a bit value.
  - Generally used when we have the same initial and final chromosomes.

University of Regina

Crossover

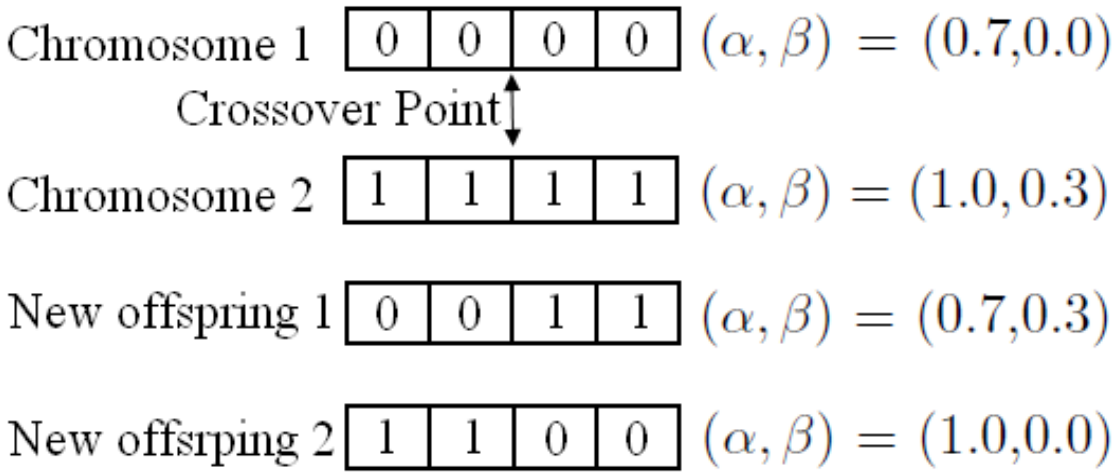

Figure : Crossover

University of Regina

Mutation

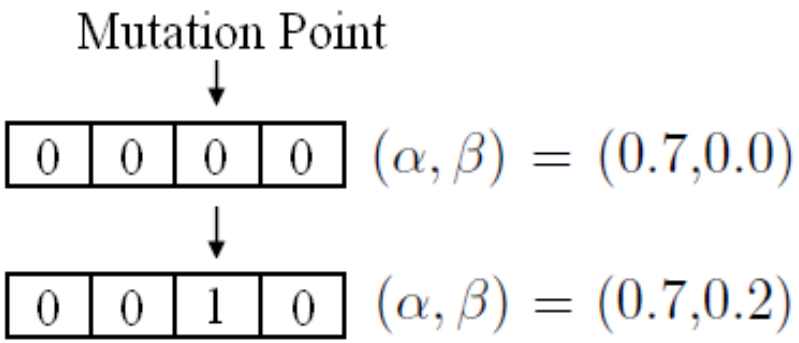

Figure : Mutation

University of Regina

A Demonstrative Example and Further Analysis

- Considering the following probabilistic information about a concept C with respect to 15 equivalence classes.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
$Pr(X_i)$	0.0177	0.1285	0.0137	0.1352	0.0580	0.0069	0.0498	0.1070
$Pr(C/X_i)$	1.0	1.0	1.0	1.0	0.9	0.8	0.8	0.6

	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
$Pr(X_i)$	0.1155	0.0792	0.0998	0.1299	0.0080	0.0441	0.0067
$Pr(C/X_i)$	0.5	0.4	0.4	0.2	0.1	0.0	0.0

University of Regina

Encoding Scheme

- Two-bit encoding for α and β.

α	α-Encoding	β	β-Encoding
0.6	00	0.0	00
0.8	01	0.1	01
0.9	10	0.2	10
1.0	11	0.4	11

University of Regina

Initial Population and their Evaluations: Steps 1 and 2

- Initial configuration.

Initial Population	Thresholds		Fitness Function
	α	β	
0000	0.6	0.0	0.6756
0101	0.8	0.1	0.6286
0010	0.6	0.2	0.6701
1111	1.0	0.4	0.6228

University of Regina

Results of the First Iteration: Steps 3 to 5

Initial Population	Thresholds		Fitness Function	Selection Probability	Crossover	Final Population
	α	β				
0000	0.6	0.0	0.6756	0.26	2	0100
0101	0.8	0.1	0.6286	0.24	1	0001
0010	0.6	0.2	0.6701	0.26	4	0110
1111	**1.0**	**0.4**	**0.6228**	0.24	3	1011

- The bold font represents the minimum fitness, i.e., **0.6228** and the corresponding thresholds $(\alpha, \beta) =$ **(1.0,0.4)**.

University of Regina

Results of the Second Iteration: Steps 3 to 5

Initial Population	Thresholds		Fitness Function	Selection Probability	Crossover	Final Population
	α	β				
0100	0.8	0.0	0.6290	0.25	4	0010
0001	0.6	0.1	0.6758	0.27	3	0111
0110	**0.8**	**0.2**	**0.6150**	0.24	2	0000
1011	0.9	0.4	0.6220	0.24	1	1101

- The minimum fitness is **0.6150** for the thresholds $(\alpha, \beta) =$ **(0.8,0.2)**.

University of Regina

Comparison with the Game-theoretic Rough Set Model

- The players: Immediate decision region I vs deferred decision region D (Azam and Yao, 2013).
  - Immediate decision region = positive and negative regions.
  - Deferred decision region = boundary region.
  - Decreasing the uncertainty of deferred decision region comes at cost of increasing uncertainty in the immediate decision region.
  - The game is to find a suitable tradeoff between the two decision regions.

Azam, N., Yao, J. T., (2013). Analyzing uncertainties of probabilistic rough set regions with GTRS. IJAR

University of Regina

GTRS based Game for Analyzing Uncertainty

- The strategies: Three types of strategies were formulated for each player.
  - s_1 ($\alpha_{\downarrow}$, decrease of α),
  - s_2 ($\beta_{\uparrow}$, increase of β) and
  - s_3 ($\alpha_{\downarrow}\beta_{\uparrow}$, decrease of α and increase of β).
- The payoffs or utility functions were defined as (1-uncertainty) of the respective regions.
  - For simplicity we refer to them as certainty.
  - $u_I(s_i, s_j) = \{(1 - \Delta_P(\alpha, \beta)) + (1 - \Delta_N(\alpha, \beta))\}/2$.
  - $u_D(s_i, s_j) = \{(1 - \Delta_B(\alpha, \beta))$.

University of Regina

GTRS Results

		D		
		$s_1 = \alpha_{\downarrow}$ = 10% dec. α	$s_2 = \beta_{\uparrow}$ = 10% inc. β	$s_3 = \alpha_{\downarrow}\beta_{\uparrow}$ = 10% (dec. α & inc. β)
I	$s_1 = \alpha_{\downarrow}$ = 10% dec. α	(0.949,0.474)	(0.977,0.416)	**(0.946,0.480)**
	$s_2 = \beta_{\uparrow}$ = 10% inc. β	(0.977,0.416)	(0.944,0.490)	(0.923,0.542)
	$s_3 = \alpha_{\downarrow}\beta_{\uparrow}$ = 10% (dec. α & inc. β)	(0.946,0.480)	(0.923,0.542)	(0.893,0.590)

- The cell with bold font represents the game solution.
- The corresponding thresholds are $(\alpha, \beta) = (0.8,0.1)$.

University of Regina

Comparison with GTRS

- Comparison of the two approaches.

Approach	Determined thresholds	Solution type	Fitness or payoffs associated with thresholds
Genetic Algo.	(0.8,0.2)	Optimization	0.6150
GTRS	(0.8.0.1)	Trade-off	(0.946,0.480)

University of Regina

Conclusion and Future Work

- The determination of thresholds is a key issue in probabilistic rough sets.
- We proposed and examined a genetic algorithm based approach for determining effective thresholds.
    - The approach is based on optimization of a fitness function defined using the information-theoretic rough set model.
- The proposed approach provides similar results to that of the GTRS.
- In future, further fitness functions may be investigated based on different rough set models.

University of Regina

Questions?

JingTao Yao, PhD

姚静涛

Professor

DEPARTMENT OF COMPUTER SCIENCE

University of Regina

Regina, Saskatchewan
Canada S4S 0A2
phone: 306.585.4071
fax: 306.585.4745
email: jtyao@cs.uregina.ca
http:/www.cs.uregina.ca/~jtyao

University of Regina