Fundamenta Informaticae 94 (2009) 233–244
DOI 10.3233/FI-2009-128
IOS Press

Semi-supervised Rough Cost/Benefit Decisions

Pawan Lingras* [†]
Department of Mathematics & Computing Science, Saint Mary's University, Halifax
Nova Scotia, B3H 3C3, Canada
pawan@cs.smu.ca

Min Chen[‡], **Duoqian Miao**
School of Electronics and Information Engineering, Tongji University
Shanghai, 201804, P.R.China and
Key Laboratory of Embedded System & Service Computing
Ministry of Education of China, Tongji University, Shanghai 201804, P.R. China

Abstract. Most of the business decisions are based on cost and benefit considerations. Data mining techniques that make it possible for the businesses to incorporate financial considerations will be more meaningful to the decision makers. Decision theoretic framework has been helpful in providing a better understanding of classification models. This study describes a semi-supervised decision theoretic rough set model. The model is based on an extension of decision theoretic model proposed by Yao. The proposal is used to model financial cost/benefit scenarios for a promotional campaign in a real-world retail store.

Keywords: Rough sets, Rough approximation, Probability, Decision theory, Cost/benefit analysis, k-means clustering algorithm.

*Address for correspondence: Department of Mathematics & Computing Science, Saint Mary's University, Halifax, Nova Scotia, B3H 3C3, Canada

[†]The author would like to thank NSERC Canada
[‡]The author would like to thank China Scholarship Council. The work is partially supported by the National Natural Science Foundation of China (Serial No. 60775036 and Serial No. 60475019) and the Specialized Research Fund for the Doctoral Program of Higher Education of China (Serial No. 20060247039).

1. Introduction

Businesses are looking at data mining for improving their profits. Data mining can contribute towards profits by reducing costs as well as increasing revenues. For example, if a business is planning a marketing campaign, data mining can be used to target customers who are most likely to respond to the campaign. Such an analysis is usually based on expected loss in a probabilistic framework. Probabilistic extensions have played a major role in the development of rough set theory since its inception. Recently, Yao [15] provided a comprehensive overview of many of the probabilistic extensions of rough set theory. The models included in the overview were: rough set-based probabilistic classification [14], 0.5 probabilistic rough set model [7], decision-theoretic rough set models [16, 17], variable precision rough set models [3, 18], rough membership functions [7], parameterized rough set models [8, 10], and Bayesian rough set models [2, 11, 12]. The list of probabilistic models that could be explained under the decision theoretic framework is a testimony in itself to the usefulness of the approach. The study of such a variety of models under a common framework also helps understand the similarities and differences between the models. Such a comparison can help in choosing the right model for the application on hand. It can also help in creating a new model that combines desirable features of two or more models. Finally, it can also lead to a unified model that can be moulded to a given application requirement. Yao [15] described how the decision theoretic framework exposed additional issues in probabilistic rough set models. This paper builds on Yao's decision theoretic model to develop a framework that can be used for supervised, unsupervised, and semi-supervised learning. Moreover, the framework does not depend on the notions such as equivalence classes or decision tables. The contributions of this paper are as follows:

1. The framework described in this paper includes supervised and unsupervised learning. Conventionally, the classification techniques refer to only supervised learning. When the objects are categorized without the help of a supervisor, the categories are usually called clusters. The proposed framework is applicable to both classification and clustering problems. The term *category* emphasizes the fact that it can be based on supervised or unsupervised learning.

2. Another interesting feature of the proposal is that it does not rely on the traditional concepts in the rough set theory such as equivalence classes and decision tables.

3. While rough set classification is frequently applied to more than two classes, Yao's decision theoretic approach was illustrated for classification between positive and negative regions. This paper extends the framework to multiple categories that is applicable to both classification and clustering. The extended framework is shown to reduce to Yao's classification approach when the number of categories is equal to two. It is also shown that the decision theoretic crisp categorization is a special case of the rough set based approach.

4. The proposal is used to describe a semi-supervised learning based on cost/benefit considerations for a real world retail store. The experiment describes weighing financial implications of a targeted marketing campaign that is modelled using the proposed framework.

The paper concludes with a discussion on the advantages of introducing decision theoretic framework in further theoretical development, especially in the rough clustering area.

2. Literature review

This section provides a review of research that lays foundation for the proposed approach.

2.1. Review of rough sets

The notion of rough set theory was proposed by Pawlak (1982, 1992). It can be approached as an extension of the classical set theory. Let U denote the universe (a finite ordinary set), and let $E \subseteq U \times U$ be an equivalence (indiscernibility) relation on U. The pair $apr = (U, E)$ is called an approximation space. The equivalence relation E partitions the set U into disjoint subsets, denoted by U/E. For $x \in U$, let $[x] = \{y \mid xEy\}$ denote the equivalence class of E containing x. For a subset $A \subseteq U$, A is characterized in the approximation space $apr = (U, E)$ by its lower and upper approximations, defined respectively as:

$$\begin{aligned} \underline{apr}(A) &= \{x \in U \mid [x] \subseteq A\}; \\ \overline{apr}(A) &= \{x \in U \mid [x] \cap A \neq \emptyset\}. \end{aligned} \tag{1}$$

The lower and upper approximations, $\underline{apr}$, $\overline{apr}$, are dual operators in the sense that $\underline{apr}(A) = (\overline{apr}(A^c))^c$ and $\overline{apr}(A) = (\underline{apr}(A^c))^c$, where A^c is set complement of A. Based on the rough set approximations of A, one can divide the universe U into three disjoint regions, the positive region $POS(A)$, the boundary region $BND(A)$, and the negative region $NEG(A)$:

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

Some of these regions may be empty. One can say with certainty that any element $x \in POS(A)$ belongs to A, and that any element $x \in NEG(A)$ does not belong to A. One cannot decide with certainty whether or not an element $x \in BND(A)$ belongs to A.

2.2. The Bayesian decision procedure

The Bayesian decision procedure deals with making decision with minimum risk based on observed evidence. Let $\Omega = \{\omega_1, \ldots, \omega_s\}$ be a finite set of s states, and let $A = \{a_1, \ldots, a_m\}$ be a finite set of possible m actions. Let $P(\omega_j|\mathbf{x})$ be the conditional probability of an object x being in state ω_j given that the object is described by **x**. Let $\lambda(a_i|\omega_j)$ denote the loss, or cost for taking action a_i when the state is ω_j. For an object x with description **x**, suppose action a_i is taken. Since $P(\omega_j|\mathbf{x})$ is the probability that the true state is ω_j given **x**, the expected loss associated with taking action a_i is given by:

$$R(a_i|\mathbf{x}) = \sum_{j=1}^{s} \lambda(a_i|\omega_j) P(\omega_j|\mathbf{x}) \tag{3}$$

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

Given a description **x**, a decision rule is a function $\tau(\mathbf{x})$ that specifies which action to take. That is, for every **x**, $\tau(\mathbf{x})$ takes one of the actions, $a_1, \ldots, a_m$. The overall risk **R** is the expected loss associated with a given decision rule, defined by:

$$\mathbf{R} = \sum_{\mathbf{x}} R(\tau(\mathbf{x})|\mathbf{x})P(\mathbf{x}) \tag{4}$$

If the action $\tau(\mathbf{x})$ is chosen so that $R(\tau(\mathbf{x})|\mathbf{x})$ is as small as possible for every object **x**. For every **x**, compute the conditional risk $R(a_i|\mathbf{x})$ for $i = 1, \ldots, m$ defined by equation (3) and select the action for which the conditional risk is minimum. If more than one action minimizes $R(a_i|\mathbf{x})$, a tie-breaking criterion can be used.

2.3. Yao's basic model

Yao proposed probabilistic rough set approximations in [15], which applies the Bayesian decision procedure for the construction of probabilistic approximations. The classification of objects according to approximation operators in rough set theory can be easily fitted into the Bayesian decision-theoretic framework. Let $\Omega = \{A, A^c\}$ denote the set of states indicating that an object is in A and not in A, respectively. Let $A = \{a_1, a_2, a_3\}$ be the set of actions, where a_1, a_2 and a_3 represent the three actions in classifying an object, deciding $POS(A)$, deciding $NEG(A)$, and deciding $BND(A)$, respectively. The probabilities $P(A|[x])$ and $P(A^c|[x])$ are the probabilities that an object in the equivalence class $[x]$ belongs to A and A^c, respectively. The expected loss $R(a_i|[x])$ associated with taking the individual actions can be expressed as:

$$R(a_1|[x]) = \lambda_{11}P(A|[x]) + \lambda_{12}P(A^c|[x]), \tag{5}$$
$$R(a_2|[x]) = \lambda_{21}P(A|[x]) + \lambda_{22}P(A^c|[x]), \tag{6}$$
$$R(a_3|[x]) = \lambda_{31}P(A|[x]) + \lambda_{32}P(A^c|[x]), \tag{7}$$

where $\lambda_{i1} = \lambda(a_i|A)$, $\lambda_{i2} = \lambda(a_i|A^c)$, and $i = 1, 2, 3$. The Bayesian decision procedure leads to the following minimum-risk decision risk:

If $R(a_1|[x]) \le R(a_2|[x])$ and $R(a_1|[x]) \le R(a_3|[x])$, decide $POS(A)$;
If $R(a_2|[x]) \le R(a_1|[x])$ and $R(a_2|[x]) \le R(a_3|[x])$, decide $NEG(A)$;
If $R(a_3|[x]) \le R(a_1|[x])$ and $R(a_3|[x]) \le R(a_2|[x])$, decide $BND(A)$.

Tie-breaking criteria should be added so that each object is classified into only one region. Since $P(A|[x]) + P(A^c|[x]) = 1$, the rules to classify any object in $[x]$ can be simplified based on the probability $P(A|[x])$ and the loss function λ_{ij} ($i = 1, 2, 3$; $j = 2, 2$).

2.4. Probabilistic rough set model

Based on the general decision-theoretic rough set model, it is possible to construct specific models by considering various classes of loss functions. In fact, many existing models can be explicitly derived from the general model. For example, the 0.5 probabilistic model can be derived when the loss function is defined as follows:

$$\lambda_{12} = \lambda_{21} = 1, \qquad \lambda_{31} = \lambda_{32} = 0.5, \qquad \lambda_{11} = \lambda_{22} = 0. \tag{8}$$

A unit cost is incurred if an object in A^c is classified into the positive region or an object in A is classified into the negative region; half of a unit cost is incurred if any object is classified into the boundary region. The 0.5 model corresponds to the application of the simple majority rule.

3. Proposed approach

Yao's basic model was described only for positive, negative, and boundary regions and applied to classification. This section provides a formal framework that can be used with both supervised and unsupervised multiple rough categories. Another interesting feature of the proposal is that it does not rely on the traditional concepts in the rough set theory such as equivalence classes and decision table. We will begin with formal definitions for the proposed framework that can be used for clustering and classification.

3.1. Framework for supervised and unsupervised learning

We will use the terms category, classes, and clusters interchangeably whenever it is appropriate in the context.
Objects: Let $X = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ be a finite set of objects.

Categories: Let $C = \{\mathbf{c_1}, \ldots, \mathbf{c_k}\}$ be a finite set of k states given that C is the set of categories and each category is represented by a vector $\mathbf{c_i}$ ($1 \leq i \leq k$). Furthermore, let C partition the set of objects X.

Object and category similarity: For every object, $\mathbf{x_l}$, we define a non-empty set T_l of all the categories that are similar to $\mathbf{x_l}$. Clearly, $T_l \subseteq C$. We will use $\mathbf{x_l} \rightarrow T_l$ to denote the fact that object $\mathbf{x_l}$ is similar to all the elements of set T_l. Let us further stipulate that object $\mathbf{x_l}$ can be similar to one and only one T_l. The definition of the similarity will depend on a given application. Later on we will see an example of how to calculate similarity using probability distribution.

Upper and lower approximations: If an object $\mathbf{x_l}$ is assigned to a set T_l, then the object belongs to the upper approximations of all categories $\mathbf{c_i} \in T_l$. If $\mid T_l \mid = 1$, then $\mathbf{x_l}$ belongs to the lower approximation of the only $\mathbf{c_i} \in T_l$. Please note that when $\mid T_l \mid = 1$, $\{\mathbf{c_i}\} = T_l$. Therefore, upper ($\overline{apr}$) and lower ($\underline{apr}$) approximation of each category $\mathbf{c_i}$ can be defined as follows:

$$\overline{apr}(\mathbf{c_i}) = \{\mathbf{x_l} | \mathbf{x_l} \rightarrow T_l, \mathbf{c_i} \in T_l\}, \tag{9}$$

$$\underline{apr}(\mathbf{c_i}) = \{\mathbf{x_l} | \mathbf{x_l} \rightarrow T_l, \{\mathbf{c_i}\} = T_l\}. \tag{10}$$

Since we do not define upper and lower approximations of all the subsets of X, we cannot test all the properties of rough set theory. However, it can be easily shown that the resulting upper and lower approximations in fact follow important rough set theoretic properties given the fact that C is a partition of X specified by Lingras and West [5].

- An object can be part of at most one lower approximation (P1)
- $\mathbf{x_l} \in \underline{apr}(\mathbf{c_i}) \Rightarrow \mathbf{x_l} \in \overline{apr}(\mathbf{c_i})$ (P2)
- An object $\mathbf{x_l}$ is not part of any lower approximation (P3)

$\Updownarrow$

$\mathbf{x_l}$ belongs to two or more upper approximations.

3.2. Loss functions for multi-category problem

Following Yao [15], we define a set of states and actions to describe the decision theoretic framework for multi-category rough sets.

States: The states are essentially the set of categories $C = \{\mathbf{c_1}, \ldots, \mathbf{c_k}\}$.

An object is said to be in one of the categories. However, due to lack of information we are unable to specify the exact state of the object. Therefore, our actions are defined as follows.

Actions: Let $B = \{B_1, \ldots, B_s\} = 2^C - \{\emptyset\}$ be a family of non-empty subsets of C, where $s = 2^k - 1$. We will define a set of actions $b = \{b_1, \ldots, b_s\}$ corresponding to set B, where b_j represents the action in assigning an object $\mathbf{x_l}$ to the set B_j.

Note that some of the sets B_j's will be the same as the set T_l's defined in previous sections. The reason we choose to use a different notation is to emphasize the fact that we do not specify any similarity between $\mathbf{x_l}$ and B_j as we do in case of $\mathbf{x_l}$ and T_l. Note that there will be a total of n T_l's, one for each object, and they may not be distinctly different from each other. That is, two objects may be similar to the same subset of C. On the other hand, there will be exactly $s = 2^k - 1$ distinct B_j's.

Now we are ready to write the Bayesian decision procedure for our multi-category rough sets as follows.

Let $\lambda_{\mathbf{x_l}}(b_j|\mathbf{c_i})$ denote the loss, or cost, for taking action b_j when an object belongs to $\mathbf{c_i}$. Let $P(\mathbf{c_i}|\mathbf{x_l})$ be the conditional probability of an object $\mathbf{x_l}$ being in state $\mathbf{c_i}$. Therefore, the expected loss $R(b_j|\mathbf{x_l})$ associated with taking action b_j for an object $\mathbf{x_l}$ is given by:

$$R(b_j|\mathbf{x_l}) = \sum_{i=1}^{k} \lambda_{\mathbf{x_l}}(b_j|\mathbf{c_i}) P(\mathbf{c_i}|\mathbf{x_l}) \tag{11}$$

For an object $\mathbf{x_l}$, if $R(b_j|\mathbf{x_l}) \leq R(b_h|\mathbf{x_l}), \forall\ h = 1, \ldots, k$, then decide b_j.

We generalize the loss function for the 0.5 probabilistic model [10] given by Yao [15] as follows:

$$\begin{aligned} \lambda_{\mathbf{x_l}}(b_j|\mathbf{c_i}) &= \frac{|b_j - T_l|}{|b_j|} & \quad if \quad \mathbf{c_i} \in b_j \ ; \\ \lambda_{\mathbf{x_l}}(b_j|\mathbf{c_i}) &= \frac{|b_j - \emptyset|}{|b_j|} & \quad if \quad \mathbf{c_i} \notin b_j \ . \end{aligned} \tag{12}$$

When $\mathbf{c_i}$ belongs to b_j, the loss for taking action b_j corresponds to the fraction of b_j that is not related to $\mathbf{x_l}$. Otherwise, the loss for taking action b_j will have the maximum value of 1.

It can be easily seen that when k is equal to 2, $C = \{\mathbf{c_1}, \mathbf{c_2}\}$. Therefore, $B = \{\{\mathbf{c_1}\}, \{\mathbf{c_2}\}, \{\mathbf{c_1}, \mathbf{c_2}\}\}$. Without loss of generality, we can designate $\mathbf{c_1}$ to be the positive class, $\mathbf{c_2}$ to be the negative class, and $\{\mathbf{c_1}, \mathbf{c_2}\}$ to be the boundary region. Then one can easily verify that $\lambda_{\mathbf{x_l}}(\{\mathbf{c_1}\}|\mathbf{c_1}) = 0$, $\lambda_{\mathbf{x_l}}(\{\mathbf{c_2}\}|\mathbf{c_1}) = 1$, and $\lambda_{\mathbf{x_l}}(\{\mathbf{c_1}, \mathbf{c_2}\}|\mathbf{c_1}) = \frac{1}{2}$, which corresponds to the loss function described by Yao [15] for the 0.5 probabilistic model [9].

Let us illustrate the proposed rough multi-category expected loss function with the following example.

Table 1. Expected loss for all the actions from Example 1.

$The\ expected\ loss\ R(b_j\|\mathbf{x_l})$	$Action$
0.35	$\{\mathbf{c_3}, \mathbf{c_4}\}$
0.433	$\{\mathbf{c_2}, \mathbf{c_3}, \mathbf{c_4}\}$
0.467	$\{\mathbf{c_1}, \mathbf{c_3}, \mathbf{c_4}\}$
0.5	$\{\mathbf{c_1}, \mathbf{c_2}, \mathbf{c_3}, \mathbf{c_4}\}$
0.6	$\{\mathbf{c_4}\}$
0.7	$\{\mathbf{c_2}, \mathbf{c_4}\}$
0.725	$\{\mathbf{c_1}, \mathbf{c_4}\}$
0.75	$\{\mathbf{c}_3\}, \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_4\}$
0.775	$\{\mathbf{c_2}, \mathbf{c_3}\}$
0.8	$\{\mathbf{c_1}, \mathbf{c_3}\}, \{\mathbf{c_1}, \mathbf{c_2}, \mathbf{c_3}\}$
1	$\{\mathbf{c_1}\}, \{\mathbf{c_2}\}, \{\mathbf{c_1}, \mathbf{c_2}\}$

Table 2. Expected loss for all the actions from Example 2.

$The\ expected\ loss\ R(b_j\|\mathbf{x_l})$	$Action$
0.6	$\{\mathbf{c_4}\}$
0.75	$\{\mathbf{c_3}\}$
0.8	$\{\mathbf{c_2}\}$
0.85	$\{\mathbf{c_1}\}$

Example 1. Let $C = \{\mathbf{c_1}, \mathbf{c_2}, \mathbf{c_3}, \mathbf{c_4}\}$ and $B = 2^C - \{\emptyset\}$ $(|B| = 2^4 - 1 = 15)$. For an object $\mathbf{x_l}$, let $\{P(\mathbf{c_1}|\mathbf{x_l}), P(\mathbf{c_2}|\mathbf{x_l}), P(\mathbf{c_3}|\mathbf{x_l}), P(\mathbf{c_4}|\mathbf{x_l})\} = \{0.15, 0.2, 0.25, 0.4\}$. We will define the set T_l such that $\mathbf{x_l} \to T_l$ as: $T_l = \{\mathbf{c_h}|P(\mathbf{c_h}|\mathbf{x_l}) > 0.2\} = \{\mathbf{c_3}, \mathbf{c_4}\}$. The expected loss associated with taking action b_j is shown in Table 1. The values of the expected loss seem quite reasonable. The lowest value is obtained for the set $T_l = \{\mathbf{c_3}, \mathbf{c_4}\}$. It is highest for the sets that do not contain either $\mathbf{c_3}$ or $\mathbf{c_4}$. Since the probability of $P(\mathbf{c_4}) > P(\mathbf{c_3})$, the sets containing $\mathbf{c_4}$ tend to have lower loss than those containing $\mathbf{c_3}$.

Example 2. One can also obtain a crisp categorization from the proposed formulation by stipulating that all the T_l's in our formulation are singleton sets. We can demonstrate this by using the same probability function, but changing the criteria for defining the set T_l such that $\mathbf{x_l} \to T_l$ as: $T_l = \{\mathbf{c_h}\}$ such that $P(\mathbf{c_h}|\mathbf{x_l})$ is maximum. If more than one such $\mathbf{c}_h$ have the same (maximum) value, we arbitrarily choose the first $\mathbf{c}_h$. This ensures that T_l is a singleton set. In our example, with

$$\{P(\mathbf{c_1}|\mathbf{x_l}), P(\mathbf{c_2}|\mathbf{x_l}), P(\mathbf{c_3}|\mathbf{x_l}), P(\mathbf{c_4}|\mathbf{x_l})\} = \{0.15, 0.2, 0.25, 0.4\},$$

$T_l = \{\mathbf{c_4}\}$. The resulting expected loss function in this example is shown in Table 2.

4. Modeling promotional campaign for a real world retail store

As mentioned before, one of the major advantages of using the decision theoretic frameworks is that we can enhance our loss function using dollar amounts. Let us consider a real world retail store, which wants to increase its profits by classifying customers into those who could potentially help increase the profits, and those who may not. The profits will be represented by dollar amounts. Since the decision theoretic framework focuses on minimizing losses, we can look at profits as negative losses in our formulation. Minimizing the losses in the decision theoretic framework will translate to maximizing profits for the retailer.

The database we used for our experiment comes from a small specialty retail store. It has more than 257,000 transactions. The store has 9,080 distinct items and the database contains purchase history of more than 16,000 customers. All transactions are completed during the period of January 02, 2005 to September 30, 2007. Only a third of the identified customers are relatively frequent visitors. These 5,565 customers will be the potential targets of our marketing campaign.

Let S_l be the total annual sales for customer $\mathbf{x}_l$. Let $profitMargin$ be the ratio of profits and sales. The profit margin before fixed costs was found to be 30%. So the benefits from S_l amount of sales will be $S_l \times 0.3$. Loss will be negative of profits, i.e. $-S_l \times 0.3$.

Let us consider a promotional campaign targeted at relatively high and regular spenders. Let us assume that it is a two tier campaign that will be aimed at the top two categories of customers. The customers that do not qualify for either of the two promotions will fall into the third category. While spending is the most important criteria, the regularity with which a customer visits the store is also an important consideration in determining the likelihood of the customer being attracted by the campaign. Therefore, in addition to the total spending, the number of months customers visited the store was used as an indication of their regular patronage.

For the two-tier promotional campaign we need to separate customers into three categories:

1. Most regular and high spenders: $\mathbf{c}_3$
2. Moderately regular and reasonable spenders: $\mathbf{c}_2$
3. Relatively infrequent low spenders: $\mathbf{c}_1$

Since we do not have an expert who can positively identify customers in each one of these categories, we first group the existing customers into three clusters using the K-means algorithm. Each customer $\mathbf{x}_l$ is represented by a two dimensional vector (x_{l1}, x_{l2}), where x_{l1} is the spending for the customer and x_{l2} is the number of months in which the customer visited the store. The range of sales for customers were as high as $10,000. On the other hand, the values of x_{l2} would be between $[1, 12]$. The sales amount would have completely dominated the number of months in our analysis. Therefore, x_{l1} was normalized to a value between $[0, 100]$. Similarity of a customer $\mathbf{x}_l$ with a centroid $\mathbf{c}_i$ generated by the K-means algorithm was used to calculate the probability $P(\mathbf{c}_i|\mathbf{x}_l)$ of the customer $\mathbf{x}_l$ belonging to the cluster $\mathbf{c}_i$ as:

$$P(\mathbf{c}_i|\mathbf{x}_l) = \frac{sim(\mathbf{c}_i, \mathbf{x}_l)}{\sum_{h=1}^{k} sim(\mathbf{c}_h, \mathbf{x}_l)} \tag{13}$$

We used inverse of the Euclidean distance between the $\mathbf{c}_i$ and $\mathbf{x}_l$ to determine $sim(\mathbf{c}_i, \mathbf{x}_l)$. We will define the set T_l such that $\mathbf{x_l} \rightarrow T_l$ as: $T_l = \{\mathbf{c}_i | P(\mathbf{c}_i|\mathbf{x}_l) > 0.3\}$. It should be noted that these T_l's

Table 3. Number of Customers in Upper and Lower Bounds.

Cluster	Lower bound	Upper bound
$\mathbf{c}_1$	3999	4413
$\mathbf{c}_2$	747	1331
$\mathbf{c}_3$	235	405

essentially give us rough set representations for the three groups of customers. For example, $T_l = \{\mathbf{c}_2\}$ means that the object $\mathbf{x}_l$ belongs to the lower bound of $\mathbf{c}_2$. On the other hand, $T_l = \{\mathbf{c}_2, \mathbf{c}_3\}$ implies that the object $\mathbf{x}_l$ belongs to the boundary region of $\mathbf{c}_2$ and $\mathbf{c}_3$. Table 3 shows the upper and lower bounds of the resulting clusters. There was an overlap of 414 customers between clusters $\mathbf{c}_1$ and $\mathbf{c}_2$. Similarly, 170 customers were in the intersection of $\mathbf{c}_2$ and $\mathbf{c}_3$. However, there were no customers in the intersection of $\mathbf{c}_1$ and $\mathbf{c}_3$.

Based on the probabilities and T_l's described above, we are now ready to apply the rough cost/benefit analysis to various possible promotional campaigns. The first tier promotion will be directed at the customers in the regularly visiting highest spending category $\mathbf{c}_3$. Each promotional campaign has an associated cost and possible benefits. Cost can be expressed in absolute dollar terms such as $75 or $50. The benefits usually vary depending on the customers. We will use possible percent increase in sales to calculate the benefits.

Let $cost_3$ be the cost for top tier promotion and $gain_3$ be the increase in sales for the top tier customers. That means the increase in sales for the top tier customers will be $S_l \times gain_3$. The nominal profit will increase by $S_l \times gain_3 \times profitMargin$. We have to subtract the cost of promotion in calculating the increase in profits, so the net profits will be $S_l \times gain_3 \times profitMargin - cost_3$. Since the dollar loss is the negative of profits, the loss will be $cost_3 - S_l \times gain_3 \times profitMargin$. We can now modify the cost for all the actions b_j such that $\mathbf{c}_3 \in b_j$, since these actions possibly assign a customer to the highest spending category $\mathbf{c}_3$. The modified loss function for such action will be given as:

$$\lambda_{\mathbf{x}_l}(b_j|\mathbf{c}_i) = (cost_3 - S_l \times gain_3 \times profitMargin) \times \frac{|b_j - T_l|}{|b_j|} \quad \text{if} \quad \mathbf{c}_i \in b_j \wedge \mathbf{c}_3 \in b_j \ ; \tag{14}$$

$$\lambda_{\mathbf{x}_l}(b_j|\mathbf{c}_i) = (cost_3 - S_l \times gain_3 \times profitMargin) \times \frac{|b_j - \emptyset|}{|b_j|} \quad \text{if} \quad \mathbf{c}_i \notin b_j \wedge \mathbf{c}_3 \in b_j \ . \tag{15}$$

The second tier promotional campaign will be directed at the moderately regular and reasonable spenders from category $\mathbf{c}_2$. Let $cost_2$ be the cost for second tier promotion and $gain_2$ be the corresponding increase in sales. That means the increase in sales for the second tier customers will be $S_l \times gain_2$. The nominal profit will increase by $S_l \times gain_2 \times profitMargin$. We have to subtract the cost of promotion in calculating the increase in profits, so the net profits will be $S_l \times gain_2 \times profitMargin - cost_2$. Since the dollar loss is the negative of profits, the loss will be $cost_2 - S_l \times gain_2 \times profitMargin$. We will exclude all the customers who have already been a target of Tier 1 campaign. That means we need to modify the cost for all the actions b_j such that $\mathbf{c}_2 \in b_j$ and $\mathbf{c}_3 \notin b_j$. The modified loss function for

Table 4. Number of Customers in Each Group for Different Promotional Campaigns.

Scenario	Tier 1	Tier 2	No promotion
1	63	1095	4407
2	238	920	4407
3	115	1314	4136

such action will be given as:

$$\lambda_{\mathbf{x}_l}(b_j|\mathbf{c}_i) = (cost_2 - S_l \times gain_2 \times profitMargin) \times \frac{|b_j - T_l|}{|b_j|} \text{if } \mathbf{c}_i \in b_j \wedge \mathbf{c}_3 \notin b_j \wedge \mathbf{c}_2 \in b_j; \quad (16)$$

$$\lambda_{\mathbf{x}_l}(b_j|\mathbf{c}_i) = (cost_2 - S_l \times gain_2 \times profitMargin) \times \frac{|b_j - \emptyset|}{|b_j|} \text{ if } \mathbf{c}_i \notin b_j \wedge \mathbf{c}_3 \notin b_j \wedge \mathbf{c}_2 \in b_j. \quad (17)$$

The loss functions for the remaining actions b_j that do not assign customers to either $\mathbf{c}_3$ or $\mathbf{c}_2$ remain unchanged.

The loss function described above can be used with the multi-category decision theoretic framework outlined in the earlier section to generate rules for deciding which customers will be targeted for the two promotional campaigns.

The rough cost/benefit analysis allows us to test various scenarios based on different assumptions about the possible gains. We experimented with three scenarios:

1. $cost_3 = \$50, gain_3 = 10\%, cost_2 = \$10, gain_2 = 5\%.$
2. $cost_3 = \$75, gain_3 = 20\%, cost_2 = \$10, gain_2 = 5\%.$
3. $cost_3 = \$75, gain_3 = 20\%, cost_2 = \$25, gain_2 = 15\%.$

Table 4 shows the number of customers that qualify for Tier 1 and Tier 2 promotions, based on different values of costs and gains given by the three scenarios. First column in the table indicates the scenario. Second column describes the number of customers who qualify for our Tier 1 promotion. Third column represents the number of customers who should receive Tier 2 promotional offers. Number of customers who should not get any promotional offer is given in the last column.

If we compare numbers from Tables 3 and 4 we can draw a correspondence between the clusters obtained prior to cost/benefit analysis and the customers that qualify for different promotions. The customers qualifying for Tier 1 and Tier 2 promotions seem to come from the lower and upper bounds of $\mathbf{c}_3$ and $\mathbf{c}_2$. While the non-targeted customers seem to be from the lower and upper bounds of $\mathbf{c}_1$. The boundaries of these clusters will be affected by the costs and benefits given in the three scenarios.

It can be seen that when the values of $cost_3$ and $gain_3$ went up from scenario 1 to scenario 2, a number of customers moved from Tier 2 to Tier 1. This means that if we expect larger sales for Tier 1 promotion, more customers seem to qualify for the promotion. This movement from Tier 2 to Tier 1 is a result of shift in the boundary region of $\mathbf{c}_3$ and $\mathbf{c}_2$. The customers from this boundary region who originally qualified for Tier 2 promotion are now bumped up to Tier 1 promotional offer. However, non-targeted customers remained the same, because their sales were so low that the expected rise in sales did not justify increased cost of promotion. Moreover, the non-targeted customers are in the lower and upper bounds of $\mathbf{c}_1$. Since the boundary region of $\mathbf{c}_3$ and $\mathbf{c}_1$ was empty, the movement from non-targeted group to Tier 1 promotion is unlikely.

On the other hand, when we raised $cost_2$ and $gain_2$, the Tier 2 group gained customers from Tier 1 as well as previously non-targeted customers. The movement from Tier 1 to Tier 2 is a result of shift in the boundary region of $\mathbf{c}_3$ and $\mathbf{c}_2$. Some of these customers are moving down to Tier 2, as Tier 2's expected rise in sales has increased. The movement from non-targeted customers to Tier 2 promotion comes from the shift in the boundary region of $\mathbf{c}_2$ and $\mathbf{c}_1$. Some of these customers may have sufficient increase in sales to justify cost of promotion for Tier 2. As before, there is no movement of customers from non-targeted group to Tier 1 promotion, since the boundary region of $\mathbf{c}_3$ and $\mathbf{c}_1$ was empty.

This type of what-if scenarios can be used by a marketing manager to come up with the most suitable promotional campaign. Once a promotional campaign has been implemented in practice, the marketing managers will have better ideas for the values of $cost_3, gain_3, cost_2, gain_2$ for subsequent campaigns.

5. Concluding remarks

This paper builds on the Bayesian decision procedure described by Yao [15] by making it possible to use it for supervised and unsupervised learning. The proposed framework is shown to be useful even in the absence of explicit equivalence relations and decision tables. The paper also describes how such a framework can be used in cost/benefit analysis for a business.

The definition of probability used in this paper is abstract as opposed to the frequency based values used in various probabilistic rough set models, including the unified framework proposed by Yao [15]. By changing the definition of the probability one can easily adopt the Bayesian decision process to rough set based clustering. Such an adoption can be useful in further theoretical development in rough clustering.

Finally, the paper illustrates an application of the decision theoretic framework for monetary cost and benefit analysis in the business world with the help of a multi-tier promotional campaign for a real-world retail store. The approach described in this paper uses semi-supervised rough cost/benefit analysis to identify customers who should be targeted for different types of promotion. The customers are first clustered based on the regularity with which they visit a store and the amount of money they spend. The resulting profiles of clusters are used to calculate probabilities, and create rough set representations of these clusters. Finally, the proposed cost/benefit analysis is used to target customers with most beneficial promotional strategy. The experiments describe how the proposed approach can be used to do a what-if analysis based on different assumptions about the costs of promotions and expected gains in sales.

References

[1] Chang, F., Chou, C., Lin, C., Chen, C.: A prototype classification method and its application to handwritten character recognition, in: *IEEE International Conference on Systems and Man and Cybernetics*, 2004, 4738–4743.

[2] Greco, S., Matarazzo, B., Slowinski, R.: Rough membership and Bayesian confirmation measures for parameterized rough sets, in: *RSFDGrC 2005. LNCS and vol. 3641* (D. Slezak, G. Wang, M. Szczuka, I. Duentsch, Y. Yao, Eds.), Springer and Heidelberg, 2005, 314–324.

[3] Katzberg, J., Ziarko, W.: Variable precision rough sets with asymmetric bound, in: *Knowledge Discovery* (W. Ziarko, Ed.), Springer, London, 1994, 167–177.

[4] Lingras, P., Butz, C.: Rough set based 1-v-1 and 1-v-r approaches to support vector machine multi-classification, *Information Sciences*, **177**, 2007, 3782–3298.

[5] Lingras, P., West, C.: Interval set clustering of web users with rough k-means, *Journal of Intelligent Information System*, **23**, 2004, 5–16.

[6] Minsky, M., Papert, S.: *Perceptrons*, The MIT Press, Cambridge, MA, 1969.

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

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

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

[10] Skowron, A., Stepaniuk, J.: Tolerance approximation spaces, *Fundamenta Informaticae*, **27**, 1996, 245–253.

[11] Slezak, D.: Rough sets and Bayes factor, in: *Transactions of Rough Sets III, LNCS, vol. 3400* (D. v. A. J. F. Peters, A. Skowron, Ed.), Springer, Heidelberg, 2005, 202–229.

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

[13] Vapnik, V.: *Statistical learning theory*, Wiley, New York, 1998.

[14] Wong, S., Ziarko, W.: Comparison of the probabilistic approximate classification and the fuzzy set model, *Fuzzy Sets and Systems*, **21**, 1987, 357–362.

[15] Yao, Y.: Decision-theoretic rough set models, in: *RSKT 2007. LNCS, vol. 4481* (J. Yao, P. L. W. Wu, M. Szczuka, N. Cercone, D. Slezak, Eds.), Springer, Heidelberg, 2007, 1–12.

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

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

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