Rough Multi-category Decision Theoretic Framework

Pawan Lingras[1], Min Chen[1,2], and Duoqian Miao[2]

[1] Department of Mathematics & Computing Science, Saint Mary's University,
Halifax, Nova Scotia, B3H 3C3, Canada
[2] School of Electronics and Information Engineering, Tongji University, Shanghai
201804, P.R. China
pawan.lingras@smu.ca

Abstract. Decision theoretic framework has been helpful in providing a better understanding of classification models. In particular, decision theoretic interpretations of different types of the binary rough set classification model have led to the refinement of these models. This study extends the decision theoretic rough set model to supervised and unsupervised multi-category problems. The proposed framework can be used to study the multi-classification and clustering problems within the context of rough set theory.

Keywords: Rough sets, Web usage mining, Rough approximation, k-means cluster algorithm.

1 Introduction

Probabilistic extensions have played a major role in the development of rough set theory since its inception. Recently, Yao [10] explained a list of probabilistic models under the decision theoretic framework. The models included in the overview were: rough set-based probabilistic classification [7], 0.5 probabilistic rough set model [4], decision-theoretic rough set models [8,9], variable precision rough set models [11], rough membership functions [4], parameterized rough set models [5], and Bayesian rough set models [6]. 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 [10] described how the decision theoretic framework exposed additional issues in probabilistic rough set models.

Rough set theory - like many other classification techniques - was originally developed for binary classification. That is, an object either belongs to a given class or does not. Many classification techniques are not easily extendible to a multi-class problem. The objective of a multi-class problem is to assign an object to any of the k possible classes. Whenever a technique cannot be easily extended

G. Wang et al. (Eds.): RSKT 2008, LNAI 5009, pp. 676–683, 2008.

to the multi-class problem, researchers have generally chosen two approaches, namely one-versus-one or one-versus-rest [1].

This paper describes how rough set theory does not need to use either the one-versus-one or one-versus-rest technique for extending the binary classification. The framework described in this paper uses the term *category* instead of class to emphasize the fact that it can be used in 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 multi-category framework is applicable to both classification and clustering problems.

The paper further extends the binary decision theoretic rough set framework for a multi-category problem. The extended framework is shown to reduce to Yao's binary classification approach when the number of categories is equal to two. Moreover, the framework is also shown to be applicable to rough clustering techniques. Finally, it is shown that the decision theoretic crisp categorization is a special case of the rough set based approach. The paper concludes with a discussion on the implications of introducing decision theoretic framework in further theoretical development, especially in the rough clustering area.

2 Literature Review

Due to space limitations, we assume familiarity with the rough set theory [5].

2.1 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 $\mathbf{x}$. Let $\lambda(a_i|\omega_j)$ denoted the loss, or cost for taking action a_i when the state is ω_j. For an object x with description $\mathbf{x}$, suppose action a_i is taken. Since $P(\omega_j|\mathbf{x})$ is the probability that the true state is ω_j given $\mathbf{x}$, the expected loss associated with taking action a_i is given by:

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

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

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

$$\mathbf{R} = \sum_{\mathbf{x}} R(\tau(\mathbf{x})|\mathbf{x})P(\mathbf{x}) \tag{2}$$

If the action $\tau(\mathbf{x})$ is chosen so that $R(\tau(\mathbf{x})|\mathbf{x})$ is as small as possible for every object $\mathbf{x}$. For every $\mathbf{x}$, compute the conditional risk $R(a_i|\mathbf{x})$ for $i = 1, \ldots, m$

defined by equation (1) 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.

Yao proposed probabilistic rough set approximations in [10], 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:

$$R(a_1|[x]) = \lambda_{11}P(A|[x]) + \lambda_{12}P(A^c|[x]), \tag{3}$$

$$R(a_2|[x]) = \lambda_{21}P(A|[x]) + \lambda_{22}P(A^c|[x]), \tag{4}$$

$$R(a_3|[x]) = \lambda_{31}P(A|[x]) + \lambda_{32}P(A^c|[x]), \tag{5}$$

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 rules:

If $R(a_1|[x]) \leq R(a_2|[x])$ and $R(a_1|[x]) \leq R(a_3|[x])$, decide $POS(A)$;
If $R(a_2|[x]) \leq R(a_1|[x])$ and $R(a_2|[x]) \leq R(a_3|[x])$, decide $NEG(A)$;
If $R(a_3|[x]) \leq R(a_1|[x])$ and $R(a_3|[x]) \leq R(a_2|[x])$, 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} $(i = 1, 2, 3\ ; j = 1, 2)$.

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:

$$\lambda_{12} = \lambda_{21} = 1, \qquad \lambda_{31} = \lambda_{32} = 0.5, \qquad \lambda_{11} = \lambda_{22} = 0. \tag{6}$$

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 Extension to the Multi-category Problem

Many classification techniques are originally designed for binary classification. Examples include Decision trees, Perceptrons, and Support Vector Machines. These techniques tend to classify objects into two classes such as the positive or negative regions in rough set theory. Some of these techniques have natural extensions for multi-class problems. Others use either the one-versus-one or one-versus-rest technique [1]. Let $C = \{\boldsymbol{c_1}, \ldots, \boldsymbol{c_k}\}$ be a set of categories. We will use the terms category, classes, and clusters interchangeably whenever it is appropriate in the context. In the one-versus-one approach, a binary classification model is created for every pair of classes $(\boldsymbol{c_i}, \boldsymbol{c_j})$. The training of such a model uses only the subset of those objects, which were classified as either $\boldsymbol{c_i}$ or $\boldsymbol{c_j}$. It can be easily seen that there will be a total of $k \times (k-1)$ such models. Assuming uniform distribution, there will be $\frac{n}{k}$ objects belonging to each class, where n is the size of the complete training set. While it would require significant computational effort to train $k \times (k-1)$ models, on an average each model will have only $\frac{2\times n}{k}$ objects. The one-versus-rest technique, on the other hand, creates a binary model for each class $\boldsymbol{c_i}$ by classifying objects as either belonging to $\boldsymbol{c_i}$ or not belonging to $\boldsymbol{c_i}$. There are only k such models. However, the training set for each model is the same size as the complete training set, i.e. n. Moreover, the training set is biased towards objects not belonging to the class. For example, for any given class $\boldsymbol{c_i}$ there will be $\frac{n}{k}$ objects belonging to $\boldsymbol{c_i}$ and $\frac{(k-1)\times n}{k}$ objects not belonging to $\boldsymbol{c_i}$. Therefore, the chances of a classification model erring towards predicting that an object does not belong to $\boldsymbol{c_i}$ are higher. As a result, studies have shown that the one-versus-one approach tends to be more accurate than the one-versus-rest approach. However, one-versus-one multi-classification creates a large number of models and works with a small amount of training data for each model. Smaller training data can lead to over-fitting and may explain the relative accuracy of the one-versus-one approach.

Given the inadequacies of both one-versus-one and one-versus-rest models, a classification technique that has a natural multi-class extension is more desirable. Rough set theory has such a natural extension. In this section, the multi-class extension of rough set is described. It should be noted that many implementation of rough set theory use similar philosophy for multi-classification. This section provides a formal framework that can be used with both supervised and unsupervised rough categories. We will start with formal definitions for the proposed framework.

Objects: Let $X = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$ be a finite set of objects.

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

Object and category similarity: For every object, $\boldsymbol{x_l}$, we define a non-empty set T_l of all the categories that are similar to $\boldsymbol{x_l}$. Clearly, $T_l \subseteq C$. We will use $\boldsymbol{x_l} \rightarrow T_l$ to denote the fact that object $\boldsymbol{x_l}$ is similar to all the elements of set T_l. Let us further stipulate that object $\boldsymbol{x_l}$ 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 $\boldsymbol{x_l}$ is assigned to a set T_l, then the object belongs to the upper approximations of all categories $\boldsymbol{c_i} \in T_l$. If $\mid T_l \mid= 1$, then $\boldsymbol{x_l}$ belongs to the lower approximation of the only $\boldsymbol{c_i} \in T_l$. Please note that when $\mid T_l \mid= 1$, $\{\boldsymbol{c_i}\} = T_l$. Therefore, upper ($\overline{apr}$) and lower ($\underline{apr}$) approximation of each category $\boldsymbol{c_i}$ can be defined as follows:

$$\overline{apr}(\boldsymbol{c_i}) = \{\boldsymbol{x_l} | \boldsymbol{x_l} \rightarrow T_l, \boldsymbol{c_i} \in T_l\}, \tag{7}$$

$$\underline{apr}(\boldsymbol{c_i}) = \{\boldsymbol{x_l} | \boldsymbol{x_l} \rightarrow T_l, \{\boldsymbol{c_i}\} = T_l\}. \tag{8}$$

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 [2].

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

$$\Updownarrow$$

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

4 Loss Functions for Multi-category Problem

Following Yao [10], 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 $C = \{\boldsymbol{c_1}, \ldots, \boldsymbol{c_k}\}$.

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

$$R(b_j|\boldsymbol{x_l}) = \sum_{i=1}^{k} \lambda_{\boldsymbol{x_l}}(b_j|\boldsymbol{c_i}) P(\boldsymbol{c_i}|\boldsymbol{x_l}) \tag{9}$$

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

We generalize the loss function for the 0.5 probabilistic model [3] given by Yao [10] as follows:

$$\begin{aligned} \lambda_{\boldsymbol{x_l}}(b_j|\boldsymbol{c_i}) &= \frac{|b_j - T_l|}{|b_j|} && if \quad \boldsymbol{c_i} \in b_j \ ; \\ \lambda_{\boldsymbol{x_l}}(b_j|\boldsymbol{c_i}) &= \frac{|b_j - \emptyset|}{|b_j|} && if \quad \boldsymbol{c_i} \notin b_j \ . \end{aligned} \tag{10}$$

When $\boldsymbol{c_i}$ belongs to b_j, the loss for taking action b_j corresponds to the fraction of b_j that is not related $\boldsymbol{x_l}$. 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, $C = \{\boldsymbol{c_1}, \boldsymbol{c_2}\}$. Therefore, $B = \{\{\boldsymbol{c_1}\}, \{\boldsymbol{c_2}\}, \{\boldsymbol{c_1}, \boldsymbol{c_2}\}\}$. Without loss of generality, we can designate $\boldsymbol{c_1}$ to be the positive class, $\boldsymbol{c_2}$ to be the negative class, and $\{\boldsymbol{c_1}, \boldsymbol{c_2}\}$ to be the boundary region. Then one can easily verify that $\lambda_{\boldsymbol{x_l}}(\{\boldsymbol{c_1}\}|\boldsymbol{c_1}) = 0$, $\lambda_{\boldsymbol{x_l}}(\{\boldsymbol{c_2}\}|\boldsymbol{c_1}) = 1$, and $\lambda_{\boldsymbol{x_l}}(\{\boldsymbol{c_1}, \boldsymbol{c_2}\}|\boldsymbol{c_1}) = \frac{1}{2}$, which corresponds to the loss function described by Yao [10] for the 0.5 probabilistic model [3].

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

Example 1. Let $C = \{\boldsymbol{c_1}, \boldsymbol{c_2}, \boldsymbol{c_3}, \boldsymbol{c_4}\}$ and $B = 2^C - \{\emptyset\}$ ($|B| = 2^4 - 1 = 15$). For an object $\boldsymbol{x_l}$, let $\{P(\boldsymbol{c_1}|\boldsymbol{x_l}), P(\boldsymbol{c_2}|\boldsymbol{x_l}), P(\boldsymbol{c_3}|\boldsymbol{x_l}), P(\boldsymbol{c_4}|\boldsymbol{x_l})\} = \{0.15, 0.2, 0.25, 0.4\}$. We will define the set T_l such that $\boldsymbol{x_l} \rightarrow T_l$ as: $T_l = \{\boldsymbol{c_h}|P(\boldsymbol{c_h}|\boldsymbol{x_l}) > 0.2\} = \{\boldsymbol{c_3}, \boldsymbol{c_4}\}$. 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 $T_l = \{\boldsymbol{c_3}, \boldsymbol{c_4}\}$. It is highest for the sets that do not contain either $\boldsymbol{c_3}$ or $\boldsymbol{c_4}$. Since the probability of $P(\boldsymbol{c_4}) > P(\boldsymbol{c_3})$, the sets containing $\boldsymbol{c_4}$ tend to have lower loss than those containing $\boldsymbol{c_3}$.

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 $\boldsymbol{x_l} \rightarrow T_l$ as: $T_l = \{\boldsymbol{c_h}\}$ such that $P(\boldsymbol{c_h}|\boldsymbol{x_l})$ is maximum. If more than one such c_h have the same (maximum) value, we arbitrarily choose the first c_h. This ensures that T_l is a singleton set. In our example, with $\{P(\boldsymbol{c_1}|\boldsymbol{x_l}), P(\boldsymbol{c_2}|\boldsymbol{x_l}), P(\boldsymbol{c_3}|\boldsymbol{x_l}), P(\boldsymbol{c_4}|\boldsymbol{x_l})\} = \{0.15, 0.2, 0.25, 0.4\}$, $T_l = \{\boldsymbol{c_4}\}$. The resulting expected loss function in this example is shown in Table 2.

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

The expected loss $R(b_j \vert \boldsymbol{x_l})$	*Action*
0.35	$\{\boldsymbol{c_3}, \boldsymbol{c_4}\}$
0.433	$\{\boldsymbol{c_2}, \boldsymbol{c_3}, \boldsymbol{c_4}\}$
0.467	$\{\boldsymbol{c_1}, \boldsymbol{c_3}, \boldsymbol{c_4}\}$
0.5	$\{\boldsymbol{c_1}, \boldsymbol{c_2}, \boldsymbol{c_3}, \boldsymbol{c_4}\}$
0.6	$\{\boldsymbol{c_4}\}$
0.7	$\{\boldsymbol{c_2}, \boldsymbol{c_4}\}$
0.725	$\{\boldsymbol{c_1}, \boldsymbol{c_4}\}$
0.75	$\{\boldsymbol{c_3}\}, \{\boldsymbol{c_1}, \boldsymbol{c_2}, \boldsymbol{c_4}\}$
0.775	$\{\boldsymbol{c_2}, \boldsymbol{c_3}\}$
0.8	$\{\boldsymbol{c_1}, \boldsymbol{c_3}\}, \{\boldsymbol{c_1}, \boldsymbol{c_2}, \boldsymbol{c_3}\}$
1	$\{\boldsymbol{c_1}\}, \{\boldsymbol{c_2}\}, \{\boldsymbol{c_1}, \boldsymbol{c_2}\}$

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

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

5 Concluding Remarks

This paper describes an extension of the Bayesian decision procedure described by Yao [10] for multi-category rough sets. The proposal is a natural extension of the conventional binary rough set classification. Unlike some other classification techniques such as Perceptrons and Support Vector Machines, it is not necessary to create a multiple binary classifiers using either the one-versus-one or one-versus-rest approaches. This is a significant advantage of rough set theory as both one-versus-one and one-versus-rest approaches can be difficult to implement in practice. The one-versus-one approach can lead to large number of binary classifiers, which may overfit the training data. On the other hand, the one-versus-rest approach tends to have lower classification accuracy.

In addition to extending the Bayesian decision process from binary rough set classifiers to rough set multi-classifiers, the approach can easily be applied to unsupervised rough set classifiers. 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 [10]. 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. Results of such development will be reported in future publications.

Acknowledgments. The authors would like to thank China Scholarship Council and NSERC Canada for their financial support.

References

1. Lingras, P., Butz, C.J.: Rough set based 1-v-1 and 1-v-r approaches to support vector machine multi-classification. Information Sciences 177, 3298–3782 (2007)
2. Lingras, P., West, C.: Interval set clustering of web users with rough k-means. Journal of Intelligent Information System 23, 5–16 (2004)
3. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. International Journal of Man-Machine Studies 29, 81–95 (1988)
4. Pawlak, Z., Skowron, A.: Rough membership functions. In: Yager, R.R., Fedrizzi, M., Kacprzyk, J. (eds.) Advances in the Dempster-Shafer Theory of Evidence, pp. 251–271. John Wiley and Sons, New York (1994)
5. Pawlak, Z., Skowron, A.: Rough sets: some extensions. Information Sciences 177, 28–40 (2007)
6. Slezak, D.: Rough sets and Bayes factor. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets III. LNCS, vol. 3400, pp. 202–229. Springer, Heidelberg (2005)
7. Wong, S.K.M., Ziarko, W.: Comparison of the probabilistic approximate classification and the fuzzy set model. Fuzzy Sets and Systems 21, 357–362 (1987)
8. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximating concepts. International Journal of Man-machine Studies 37, 793–809 (1992)
9. Yao, Y.Y., Wong, S.K.M., Lingras, P.: A decision-theoretic rough set model. In: Ras, Z.W., Zemankova, M., Emrich, M.L. (eds.) Methodologies for Intelligent Systems, vol. 5, pp. 17–24. North-Holland, New York (1990)
10. Yao, Y.Y.: Decision-theoretic rough set models. In: Yao, J., Lingras, P., Wu, W.-Z., Szczuka, M., Cercone, N.J., Ślęzak, D. (eds.) RSKT 2007. LNCS (LNAI), vol. 4481, pp. 1–12. Springer, Heidelberg (2007)
11. Ziarko, W.: Variable precision rough set model. Journal of Computer and System Sciences 46, 39–59 (1993)