A Multiple-category Classification Approach with Decision-theoretic Rough Sets

Dun Liu∗†
School of Economics and Management, Southwest Jiaotong University
Chengdu, 610031, P.R. China, newton83@163.com

Tianrui Li
School of Information Science and Technology, Southwest Jiaotong University
Chengdu, 610031, P.R. China, trli@swjtu.edu.cn

Huaxiong Li
School of Management and Engineering, Nanjing University
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, 210093, P.R. China, huaxiongli@nju.edu.cn

Abstract. By considering the levels of tolerance for errors and the cost of actions in real decision procedure, a new two-stage approach is proposed to solve the multiple-category classification problems with Decision-Theoretic Rough Sets (DTRS). The first stage is to change an m-category classification problem (m > 2) into an m two-category classification problem, and form three types of decision regions: positive region, boundary region and negative region with different states and actions by using DTRS. The positive region makes a decision of acceptance, the negative region makes a decision of rejection, and the boundary region makes a decision of abstaining. The second stage is to choose the best candidate classification in the positive region by using the minimum probability error criterion with Bayesian discriminant analysis approach. A case study of medical diagnosis demonstrates the proposed method.

Keywords: Decision-theoretic rough sets, probabilistic rough sets, bayesian decision procedure, three-way decisions, multiple-category.

∗This work is supported by the Major Program of National Natural Science Foundation of China (No. 71090402/G1), the National Science Foundation of China (Nos. 61175047, 60873108, 70971062), the Youth Social Science Foundation of the Chinese Education Commission (11YJC630127), the Natural Science Foundation of Jiangsu, China (BK2011564).
†Address for correspondence: School of Economics and Management, Southwest Jiaotong Univ., Chengdu, 610031, P.R. China
1. Introduction

In many decision-making problems, people are used to make a decision in two ways: do it, or not do it. Instead of the two-way decision, a three-way decision procedure seems more closer to human being’s behavior. In the three-way decision procedure, if it is considered as correct, then it should be done immediately; If it is regarded as incorrect then it should not be done; If it is hard to judge at the moment, it needs further investigation. Clearly, the three-way decision procedure comes closer to the philosophy of the real decision problems. There are lots of successful applications of the three-way decision procedure in many domains [17], e.g., medicinal clinics [24], products inspecting process [34], documents classification [16], model selection criteria [8], actions for an environmental manager [10], data packs chosen [31], E-learning system [1], Web-based support systems [35], multi-agent business [47], autonomous clustering [48], oil investment [19, 23, 49] and email spam filtering [50, 52].

Motivated by above arguments, it is necessary to introduce the three-way decision procedure to Rough Set Theory (RST) to solve decision problems. In RST, a pair of certain sets named lower approximation and upper approximation are used for describing a vague or imprecise set (It is called as the Pawlak rough set model) [25]. Clearly, two approximations can divide the universe into three pair-wise disjoint regions: positive region, boundary region and negative region. The rules generated by these three regions correspond to the results of a three-way decision that the situation is verified positively, negatively, or undecidedly based on the evidence [20, 42, 43].

Unfortunately, in the Pawlak rough set model, the set inclusion must be fully correct or certain, namely, it does not allow any tolerance of errors. In order to overcome the disadvantage, the probabilistic rough set model was proposed by allowing certain acceptable level of errors, and a pair of threshold parameters $\alpha$ and $\beta$ are used to redefine the lower and upper approximations [13, 29, 33, 36, 37, 40, 44, 55, 56]. Specially, Yao induced the three-way decision procedure into probabilistic rough set and proposed the Decision-Theoretic Rough Set (DTRS) model [36, 37]. In DTRS model, two thresholds can be directly calculated by minimizing the decision cost with Bayesian theory. We can use the two parameters to generate three pair-wise disjoint regions and obtain the corresponding decision rules automatically. However, when considering three-way decisions, most of researchers assume that there only exists a two-category classification, namely, the states are composed of an affair $X$ and it’s complement $\neg X$ [42, 43, 50]. The assumption needs to further investigate in real decision problems. In this paper, we propose a new approach in the three-way decision procedure, which extends a two-category classification into $m$-category classifications.

In the following, Bayesian theory is induced to solve $m$-category classification problems. As a fundamental statistical approach to deal with the problems of pattern classification, Bayesian criterion allows us to choose the best state or action with the overall minimum error (risk). In two-category classification problems, DTRS focuses on the actions for a certain classification, and the Bayesian decision procedure is used to choose the best action. However, $m$-category classification problems focuses the states for all classifications, and we should select the best one from $m$ classifications by using of Bayesian theory. With these observations, it seems DTRS with a two-category classification is not suitable for the practical problems, and a two-stage bayesian decision procedure which combines with action selection and state selection together, is proposed to solve multiple-category classification problems.

The remainder of this paper is organized as follows: Section 2 provides the basic concepts of the rough sets, probabilistic rough set model and it’s extensions. A new approach with two stages based on the granularity viewpoint is proposed to solve the multiple-category classification problems, and the
detailed modeling process is shown in Section 3. Then, a case study of medical diagnosis is given to explain our approach in Section 4. The paper ends with conclusions and further research topics in Section 5.

2. Preliminaries

Basic concepts, notations and results of rough sets as well as their extensions are briefly reviewed in this section [21, 25, 27, 29, 33, 36, 37, 55, 56]. We mainly introduce two categories of rough set models, one is Pawlak rough set model, and the others are probabilistic rough sets models.

2.1. Pawlak rough set model

Let $U$ be a finite and nonempty set and $R$ be an equivalence relation on $U$. The pair $apr = (U, R)$ is called an approximation space. The equivalence relation $R$ induces a partition of $U$, denoted by $[x]_R$ or $[x]$ [25, 27]. For a subset $X \subseteq U$, its lower and upper approximations are defined by:

$$apr(X) = \{ x \in U | [x] \subseteq X \};$$
$$\overline{apr}(X) = \{ x \in U | [x] \cap X \neq \emptyset \}. \quad (1)$$

A fundamental application of rough sets is to induce rules based on these two approximations. The rules induced from lower approximation are certain rules, and the rules induced from upper approximation are possible rules [27]. Unfortunately, the rules generated by the complement of the upper approximation are ignored. Based on the rough set approximations of $X$, one can divide the universe $U$ into three disjoint regions: the positive region $POS(X)$, the boundary region $BND(X)$, and the negative region $NEG(X)$.

$$POS(X) = apr(X);$$
$$BND(X) = \overline{apr}(X) - apr(X);$$
$$NEG(X) = U - \overline{apr}(X). \quad (2)$$

The rules generated by the element $x \in POS(X)$ stand for one can certainly do something; The rules generated by the element $x \in NEG(X)$ means one cannot do anything; One cannot decide with certainty when the element $x \in BND(X)$. The structure of three disjoint regions is the original motivation and basic idea of the three-way decision making.

2.2. Probabilistic rough set models

In Pawlak rough set model, the lower approximation is the union of those equivalence classes that are included in the set and the upper approximation is the union of those equivalence classes that have a nonempty overlap with the set. The rules induced by the lower approximation must be absolutely consistent or correct, namely, the classification must be completely correct or certain. However, the definitions of approximations does not allow any tolerance of errors. It rarely happens in practice and a generalized model of rough sets, the probabilistic rough set model, has been proposed to solve these problems.
Definition 2.1. [28] Let $S = (U, A, V, f)$ be an information system. $\forall x \in U, X \subseteq U$, let: $Pr(X|[x]) = \frac{|\{x \in X\}|}{|x|}$, where, $|\cdot|$ stands for the cardinal number of objects in sets. $Pr(X|[x])$ denotes the conditional probability of the classification.

Definition 2.2. Let $S = (U, A, V, f)$ be an information system. $\forall X \subseteq U$ and $0 \leq \beta < \alpha \leq 1$, the $(\alpha, \beta)$-lower approximation, $(\alpha, \beta)$-upper approximation are defined as follows.

\[
\text{apr}_{(\alpha, \beta)}(X) = \{x \in U | Pr(X|[x]) \geq \alpha\};
\]
\[
\text{apr}_{(\alpha, \beta)}(X) = \{x \in U | Pr(X|[x]) > \beta\}.
\]

From the $(\alpha, \beta)$-probabilistic lower and upper approximations, we can obtain the $(\alpha, \beta)$-probabilistic positive, boundary and negative regions.

\[
\text{POS}_{(\alpha, \beta)}(X) = \{x \in U | Pr(X|[x]) \geq \alpha\},
\]
\[
\text{BND}_{(\alpha, \beta)}(X) = \{x \in U | \beta < Pr(X|[x]) < \alpha\},
\]
\[
\text{NEG}_{(\alpha, \beta)}(X) = \{x \in U | Pr(X|[x]) \leq \beta\}.
\]

In most of probabilistic rough sets, the thresholds $\alpha$ and $\beta$ are given by experts with intuitive understanding or experiences. However, DTRS utilizes Bayesian decision procedure to acquire the optimal thresholds with minimum decision risk. In the following, we briefly review the concepts about DTRS.

Inspired by Duda’s discussions in [7], in a information system $S$, the set of states given by $\Omega = \{w_1, w_2, \cdots, w_m\}$ is a finite set of $m$ states and $A = \{a_1, a_2, \cdots, a_n\}$ is a finite set of $n$ possible actions. $P(w_i|x)$ is the conditional probability of an object $x$ being in state $w_i$ given that the object is described by $x$. Let $\lambda(a_j|w_i)$ denote the loss or cost for taking action $a_j$ when the state is $w_i$. For an object $x$, suppose action $a_j$ is taken. Since $Pr(w_i|x)$ is the probability that the true state is $w_i$ given $x$, the expected cost associated with taking action $a_j$ is given by the following formula:

\[
R(a_j|x) = \sum_{i=1}^{m} \lambda(a_j|w_i) Pr(w_i|x).
\]

In general, a decision rule can be conceived as a function $\tau(x)$ that specifies which action to take, and the overall risk $R$ of a decision rule is calculated by:

\[
R = \sum_{x \in U} R(\tau(x)|x) Pr(x),
\]

where the summation is over the set of all possible descriptions of objects. For every object $x$, we need compute the conditional risk $R(a_j|x)$ for every actions by equation (5) and select the action for which the conditional risk is minimum.

With respect to the equations (5) and (6), the DTRS model introduces Bayesian decision procedure to rough sets. It uses 2 states and 3 actions to describe the decision process for a given sample $x$ or an equivalent class $[x]$. The set of states is given by $\Omega = \{X, \neg X\}$ indicating that an element is in $X$ and not in $X$. The set of actions is given by $A = \{a_P, a_B, a_N\}$, where $a_P$, $a_B$, and $a_N$ represent the three actions in classifying an object $x$, namely, deciding $x \in \text{POS}(X)$, deciding $x \in \text{BND}(X)$, and deciding $x \in \text{NEG}(X)$, respectively. We utilize $\lambda_{PP}$, $\lambda_{BP}$, and $\lambda_{NP}$ to denote the cost incurred for taking actions
where the parameters $a_P$, $a_B$, and $a_N$, respectively, when an object belongs to $X$. Similarly, $\lambda_{PN}$, $\lambda_{BN}$ and $\lambda_{NN}$ denote the cost incurred for taking the same actions when the object does not belong to $X$. Specially, the loss functions $\lambda_{\bullet \bullet}$ have the detailed semantic explanations in a decision problem, i.e., the cost of money in investment decisions.

According to the equation (5), the expected cost $R(a_i|\lambda |x])$ associated with the individual actions can be expressed as follows:

$$R(a_P|x]) = \lambda_{PP} Pr(X|x]) + \lambda_{PN} Pr(\neg X|x])$$

$$R(a_B|x]) = \lambda_{BP} Pr(X|x]) + \lambda_{BN} Pr(\neg X|x])$$

$$R(a_N|x]) = \lambda_{NP} Pr(X|x]) + \lambda_{NN} Pr(\neg X|x])$$

The Bayesian decision procedure in equation (6) suggests the following minimum-cost decision rules:

- (P) If $R(a_P|x]) \leq R(a_B|x])$ and $R(a_P|x]) \leq R(a_N|x])$, decide $x \in POS(X)$;
- (B) If $R(a_B|x]) \leq R(a_P|x])$ and $R(a_B|x]) \leq R(a_N|x])$, decide $x \in BND(X)$;
- (N) If $R(a_N|x]) \leq R(a_P|x])$ and $R(a_N|x]) \leq R(a_B|x])$, decide $x \in NEG(X)$.

Since $Pr(X|x]) + Pr(\neg X|x]) = 1$, we can simplify the rules based only on the probabilities $Pr(X|x])$ and the loss functions $\lambda_{\bullet \bullet}$. By considering a reasonable kind of loss functions with $\lambda_{PP} \leq \lambda_{BN} < \lambda_{NP} \leq \lambda_{NN}$, one can induce decision rules as follows:

- (P) If $Pr(X|x]) \geq \alpha$ and $Pr(X|x]) \geq \gamma$, decide $x \in POS(X)$;
- (B) If $Pr(X|x]) \leq \alpha$ and $Pr(X|x]) \geq \beta$, decide $x \in BND(X)$;
- (N) If $Pr(X|x]) \leq \beta$ and $Pr(X|x]) \leq \gamma$, decide $x \in NEG(X)$.

where the parameters $\alpha$, $\beta$, and $\gamma$ are defined as:

$$\alpha = \frac{\lambda_{PN} - \lambda_{BN}}{\lambda_{PN} - \lambda_{BN} + (\lambda_{BN} - \lambda_{PP})}$$

$$\beta = \frac{\lambda_{BN} - \lambda_{NN}}{\lambda_{BN} - \lambda_{NN} + (\lambda_{NP} - \lambda_{BP})}$$

$$\gamma = \frac{\lambda_{PN} - \lambda_{NN}}{\lambda_{PN} - \lambda_{NN} + (\lambda_{NP} - \lambda_{PP})}$$

In addition, as a well-defined boundary region, the conditions of rule (B) suggest that $\alpha > \beta$, that is,

$$\frac{\lambda_{PN} - \lambda_{BN}}{\lambda_{PN} - \lambda_{BN} + (\lambda_{BN} - \lambda_{PP})} > \frac{\lambda_{BN} - \lambda_{NN}}{\lambda_{BN} - \lambda_{NN} + (\lambda_{NP} - \lambda_{BP})}$$. So, we can easily obtain: $\frac{\lambda_{NP} - \lambda_{PP}}{\lambda_{BN} - \lambda_{NN}} > \frac{\lambda_{BP} - \lambda_{PP}}{\lambda_{BN} - \lambda_{NN}}$. Due to the inequality $\frac{b}{a} > \frac{c}{d} \implies \frac{b+d}{a+c} > \frac{d}{c}$, ($a, b, c, d > 0$), we have: $\frac{\lambda_{NP} - \lambda_{PP}}{\lambda_{BN} - \lambda_{NN}} > \frac{\lambda_{BP} - \lambda_{PP}}{\lambda_{BN} - \lambda_{NN}}$. Due to this, after tie-breaking, the following simplified rules are obtained. The parameter $\gamma$ is unnecessary [39, 42, 43, 44, 45].

- (P1) If $Pr(X|x]) \geq \alpha$, decide $x \in POS(X)$;
- (B1) If $\beta < Pr(X|x]) < \alpha$, decide $x \in BND(X)$;
- (N1) If $Pr(X|x]) \leq \beta$, decide $x \in NEG(X)$. 

D. Liu et al. / A Multiple-category Classification Approach with DTRS
The computing process can be easily found in [36, 37, 38, 39, 42, 43, 44, 45], which brings us a clear and intuitive way to acquire the optimal thresholds. In addition, Herbert and Yao introduced a game-theoretic approach to DTRS for learning optimal parameter values. Measures of classification ability were interpreted as players in a game, each associated with a goal of optimizing its value; Actions performed in this game consist of varying the size of the classification regions [11, 13]. Furthermore, Beynon and Griffiths designed an expert system to generate the $\beta$-reducts, and the threshold $\beta$ can be tuned to construct different decision rules with an expert system window [5]. They also did many successful applications on corporate failure prediction [4], UK monopolies and mergers commission [6] and bank credit ratings [9] by considering the tolerance of errors in rough sets.

To sum up, in Pawlak rough set model, the two parameters are defined by using the two extreme values, 0 and 1 as a qualitative nature for probabilities, but the magnitude of the value $\Pr(X|x)$ is not taken into account. By considering the real applications of a probabilistic rough set model, one may directly supply the parameters $\alpha$ and $\beta$ based on an intuitive understanding of the levels of tolerance for errors [55]. In 0.5 probabilistic rough set model [26], we set $\alpha = \beta = 0.5$, and this model corresponds to the application of the simple majority rule. In addition, two formulations probabilistic rough set models named DTRS model and variable precision rough set model (VPRS) are proposed based on the statistical information of membership function. The difference between DTRS and VPRS is the former can automatically calculate the two thresholds according to Bayesian theory while the latter can not do it. Furthermore, VPRS is just a special case of DTRS, and VPRS can be directly derived from DTRS when the decision costs are equal to some certain values [39, 46]. It seems DTRS brings a briefly semantics explanation with minimum decision risks, and it considers the decision risk issues in practical applications. Hence, DTRS can be regarded as a representative model for probabilistic rough sets and one can choose a suitable rough set approach to satisfy the user’s requirements and expectations which fulfill their needs [12, 21].

### 3. Multiple-category Classifications Decision Approach

In DTRS model, the two-category classification strategy is induced to solve the decision problem [42], and it uses two states $\Omega = \{X, \neg X\}$ to generate the probabilistic rules. For example, we utilize \{bankrupt, operation\} to describe the state of a corporation and \{illness, health\} to represent the state of a patient. However, the state space $\Omega$ may include many choices in a practical decision problem, e.g., the illness can be divided into many categories, such as cold, fever, influenza, chill, etc. Observed by this situation, we need to extend the two states $\Omega = \{X, \neg X\}$ to multiple states $\Omega = \{C_1, C_2, \ldots, C_m\}$ ($m \geq 2$) and construct a new DTRS model with multiple-category classifications. Like a bayes classifier in Bayesian discriminant analysis problem, the decision maker should select the best choice $C_l$ ($l \in m$) from all the categories. In order to achieve this goal, Yao suggested an alternative method to change an $m$-category classification problem into $m$ two-category classification problems [42]. However, the difficulty in Yao’s suggestion is that there may exist more than one classification enter into the position region. Then, we need more information to choose the best candidate. Moreover, Ślęzak presented an approach for defining the three probabilistic regions based on pair-wise comparisons of categories with bayesian discriminant analysis [29]. A matrix of threshold values is used, with a pair of different threshold values on each pair of categories. Although the approach is a natural way to solve the problem, one can hardly estimate all threshold values, especially when the database is massive. Although one
may simplify the model by using the same pair of threshold values for all pairs of categories [42], it is unreasonable in practical decision making.

With the insightful gain from the these opinions, we suggest to combine the above two approaches together to deal with the multiple-category classification problems and propose a two-stage method. In Section 3.1, we introduce the advantages, disadvantages and differences of two approaches in [29] and [42]. In Section 3.2, we utilize the method in [42] as the first stage, namely, we simply convert \( m \)-category classification problem into \( m \) two-category classification problems. Each classification can be divided into three regions by using DTRS. Obviously, the classifications which enter to the positive region may become the candidates for the best choice. In Section 3.3, we utilize the method in [29] as the second stage, namely, we choose the best candidate in positive region by using the minimum probability error criterion with bayesian discriminant analysis and find the best choice.

### 3.1. \( m \) two-category classification vs. \( m \)-category classification

The first approach proposed by Yao [43] focuses on choosing the proper action for a certain classification at one time. For an \( m \) classification problem, each classification may have three possible actions: acceptance, rejection or deferment. DTRS is used to calculate the thresholds by using equation (8) and select the proper action for every classification. After \( m \) times computing processes by DTRS, the proper action for \( m \) classifications can be divided into three decision groups: acceptance group, rejection group, and deferment group. The classifications in the acceptance group have higher probability to become the best choices in a decision problem. By considering the acceptance group may include more than one classification, the approach is ineffective to find a best decision classification. However, this approach allows multi-decision results for one thing, and there are no strongly independent requirement among different classifications. We can generate the decision rules with multi-decision classifications, e.g., “A patient may have disease \( A \) and disease \( B \), simultaneously.”

The second approach proposed by Ślezak [29] focuses on the states for all classifications. Like a Bayesian discriminant classifier, the classification corresponding to the largest discriminant is the best choice for the research object [7]. In fact, in this decision procedure, each classification may have two possible actions: acceptance or rejection. The classification which has the overall minimum error (risk) should be accepted, and the other classifications should be rejected. Unfortunately, the approach just works if all the misclassification parameters can be obtained, and the correlations among the classifications should be independent. The computational complexity generated by the high dimensions of \( m \) classification is the biggest challenge in practical problems.

Furthermore, the differences between the two approaches come from different angles. From the viewpoint of granularity [51], the former approach is based on a coarsen granular level and it roughly chooses the possible classifications as the candidates for final decision; the latter one is based on a refined granular level, which chooses the best classification with the overall minimum errors. As the viewpoint of research goal, both of them focus on a give sample \( x \) or an equivalent class \([x]\), but the former approach focuses on the proper action for \( x \) or \([x]\); the latter one focuses on the proper state for \( x \) or \([x]\). As the viewpoint of decision styles, the former approach is a three-way decision procedure. Three types of errors may happen in this situation: incorrect acceptance, incorrect rejection and deferment errors [43]; the latter one is two-way decision procedure. Since each classification has only two states: accept or reject, there are only two types of errors can happen: incorrect acceptance and incorrect rejection [7].
To sum up, a natural and intuitive method is to combine the two approaches together and form a new two-stage approach. In the first stage, the former approach is used for dimension reduction in classification, which simplifies a complex problem to an easy one. In the second stage, the latter approach is to find the best candidate from the classifications which are selected from the first stage. Compared with the original DTRS model, a simple flowchart of the difference between the multiple-category classification and the two-category classification with DTRS is shown in Figure 1.

Figure 1. The flowchart of the difference between the multiple-category classification and the two-category classification with DTRS

3.2. \textit{m}-category classification process

In an \textit{m}-category classification problem with DTRS, for a given sample \(x\) or an equivalent class \([x]\), the set of states is composed by \textit{m} classifications, which are given by \(C = \{C_1, C_2, \ldots, C_m\}\), where \(C_1, C_2, \ldots, C_m\) form a family of pair-wise disjoint subset of \(U\), namely, \(C_i \cap C_j = \emptyset\) for \(i \neq j\), and \(\cup C_i = U\). For each \(C_i\), we can define a two-category classification by \(\{C, \neg C\}\), where \(C = C_i\) and \(\neg C = \neg C_i = \cup_{j \neq i} C_j\) [43]. The descriptions of \([x]\) can be considered as the predecessor of a decision rule, and the classification \(C_i\) is the successor of the rule. The \textit{m}-category classification problem can be converted to \textit{m} two-category classification problems [43], and the results from DTRS model can be immediately applied in our research.

In general, different categories have different losses, and different threshold values should be used for different categories in a model [29]. So, as an \textit{m}-category classification problem, we need obtain
threshold value pairs \((\alpha_i, \beta_i)\) according to \(m\) different groups of loss functions. In detail, for each 
\(C_i \in C\), the set of states is given by 
\[\Omega_i = \{C_i, \neg C_i\}\] 
and the set of actions is given by 
\[A = \{a_{P}, a_{B}, a_{N}\}\] 
with respect to the three-way decisions. There are 6 loss functions regarding the cost: 
\(\lambda_{PC_i}, \lambda_{BC_i}, \lambda_{NC_i}, \lambda_{P-C_i}, \lambda_{B-C_i}\) and \(\lambda_{N-C_i}\), where 
\(\lambda_{PC_i}, \lambda_{BC_i}\) and \(\lambda_{NC_i}\) denote the cost incurred for taking actions 
\(a_{P}, a_{B}\) and \(a_{N}\), respectively, when an object belongs to \(C_i\). Similarly, \(\lambda_{P-C_i}, \lambda_{B-C_i}\) and \(\lambda_{N-C_i}\) denote the cost incurred for taking the same actions when the object does not belong to \(C_i\), respectively. The expected 
cost \(R(a_\bullet[x])\) associated with taking the individual actions \(a_\bullet\) can be expressed as:

\[
R(a_{P}[x]) = \lambda_{PC_i} Pr(C_i|[x]) + \lambda_{P-C_i} Pr(-C_i|[x]),
R(a_{B}[x]) = \lambda_{BC_i} Pr(C_i|[x]) + \lambda_{B-C_i} Pr(-C_i|[x]),
R(a_{N}[x]) = \lambda_{NC_i} Pr(C_i|[x]) + \lambda_{N-C_i} Pr(-C_i|[x]).
\]

(9)

Since \(Pr(C_i|[x]) + Pr(-C_i|[x]) = 1\), we can simplify the rules based only on the probabilities 
\(Pr(C_i|[x])\) and the loss functions \(\lambda_{\bullet\bullet}\). Then, we can easily induce the three-way decision rules when 
\(\lambda_{PC_i} \leq \lambda_{BC_i} < \lambda_{NC_i}\) and \(\lambda_{N-C_i} \leq \lambda_{B-C_i} < \lambda_{P-C_i}\), by using the minimum overall risk criterion:

- (P') If \(Pr(C_i|[x]) \geq \alpha_i\) and \(Pr(C_i|[x]) \geq \gamma_i\), decide \(x \in POS(C_i)\);
- (B') If \(Pr(C_i|[x]) \leq \alpha_i\) and \(Pr(C_i|[x]) \geq \beta_i\), decide \(x \in BND(C_i)\);
- (N') If \(Pr(C_i|[x]) \leq \beta_i\) and \(Pr(C_i|[x]) \leq \gamma_i\), decide \(x \in NEG(C_i)\);

According to the DTRS model [36, 37, 42, 43], the three-way decision rules generated by \(C_i\) are displayed as follows:

- If \(Pr(C_i|[x]) \geq \alpha_i\), decide \(x \in POS(C_i)\);
- If \(\beta_i < Pr(C_i|[x]) < \alpha_i\), decide \(x \in BND(C_i)\);
- If \(Pr(C_i|[x]) \leq \beta_i\), decide \(x \in NEG(C_i)\).

where, \(\alpha_i = \frac{(\lambda_{P-C_i} - \lambda_{B-C_i})}{(\lambda_{P-C_i} - \lambda_{B-C_i}) + (\lambda_{BC_i} - \lambda_{PC_i})}\) \(\beta_i = \frac{(\lambda_{B-C_i} - \lambda_{N-C_i})}{(\lambda_{B-C_i} - \lambda_{N-C_i}) + (\lambda_{BC_i} - \lambda_{NC_i})}\).

Intuitively, \((\alpha_i, \beta_i)\) depends on the loss functions \(\lambda_{\bullet\bullet}\). We should emphasize two necessary notes on the 
loss functions. The first one is the presentations of the cost. In some problems, we can directly utilize 
the objects (i.e., money, energy, time, etc) to describe the cost. But in some other problems, it’s very hard 
to use a detailed object to illuminate the cost. In this case, we need to do some experiments to acquire 
the cost. The second one is the performance of the individuals. Different people may have different 
attitudes on decision making. With respect to the personal performance, Zhou and Li introduced three 
decision attitudes into DTRS, namely, optimistic decisions, pessimistic decisions and equable decisions 
[15, 54]. They took a medical examination as an example. An optimistic person tends not to see 
the doctor because he or she believes the illness may be self-recover even without treatment. However, a 
pessimistic person tends to see the doctor immediately because he or she believes the illness may cause 
an aggravation without treatment [15, 54]. The optimistic persons pay a smaller costs when they make 
a wrong decisions, and the opposite situation happens with the pessimistic persons. Generally, one can 
summarize and integrate the expert’s suggestions to confirm the loss functions in an important decision 
problem. Some approaches in decision theory and data mining can be introduced to DTRS to get the cost.
As stated in equation (9), we can compute the threshold value pairs \((\alpha_i, \beta_i), \forall C_i \in C\), and \(m\) pairs of the threshold values can be obtained by repeating the calculational process with DTRS. By considering the fact that the rules generated by the positive region make the decision of acceptance, we choose the classifications which enter to the positive region as the candidates for the best choice. Clearly, only the classifications in positive region are considered to our following discussion. The classifications in boundary region and negative region are no longer discussed. It is a kind of dimension reduction which makes the decision procedure more easier. Furthermore, in our model, \((\alpha_i, \beta_i)\) is generated by a classification and its complement \(\{C_i, \neg C_i\}\). It’s easier to obtain the loss functions in one classification than in \(m\)-category classification. That is the advantage of the proposed approach.

Depending on the different values of \(\alpha_i\) and \(\beta_i\) for each \(C_i \in C\) computed by DTRS, the decision maker should choose a common criterion \(\alpha^*\) and \(\beta^*\) to form the three decision groups. With the inspiration by the risk management theory [2], the risk can be divided into three types: risk preference, risk averse and risk neutral. The risk-lovers search for high-risk and high-return, so they reduce the value of \(\alpha^*\), and choose \(\alpha^* < 0.5\); The risk-aversers hate the higher risk, so they increase the value of \(\alpha^*\), and choose \(\alpha^* > 0.5\); The risk neutrals are indifferent with risk, and they usually choose \(\alpha^* = 0.5\). Generally, a suitable criterion for choosing \(\alpha^*\) and \(\beta^*\) depends on decision maker’s risk attitude and the problem itself.

However, Yao also considered the case of the equivalence classes in an information system. By considering an equivalence class may produce more than one positive rule, he concluded that each equivalence class produces at most one positive rule when \(\alpha_i > 0.5\); each equivalence class produces at most one boundary rule when \(\beta_i > 0.5\) [43]. One has to consider the problem of rule conflict resolution in order to make effective acceptance, rejection, and abstaining decisions. Therefore, it is necessary to further investigate on the probabilistic regions of a classification, as well as the associated rules [43].

3.3. The process of choosing the best candidate classification

Suppose there are \(m'\) \((m' < m)\) classifications enter into the position region. We should choose the best one from these candidates. The strategy of our approach is to consider all the candidate classifications together and calculate the loss utility by using the Bayesian discriminant analysis approach for each candidate and finding the minimum one.

Furthermore, by thinking of a classifier as a device for partitioning the feature space into decision regions, we can obtain additional insight into the operation of a Bayesian discriminant classifier. Considering the \(m'\)-category cases, and suppose that the classification \(C_i\) has divided the space into two regions, \(a_P\) and \(a_N\). There are \((m' - 1)\) ways that a classification error can occur, that is, an observation \(x\) falls in \(C_j\) and the true state is \(C_i\) \((j \neq i)\). The probability of error is:

\[
P(error) = \sum_{i=1}^{m'} \sum_{j=1, j \neq i}^{m'} P(C_j, C_i) = \sum_{i=1}^{m'} \sum_{j=1, j \neq i}^{m'} P(C_j | C_i)P(C_i)
\]

By considering the cost of \(\lambda(C_i | C_i) = 0\) when doing a right classification, the overall error cost of
the misclassification of $C_i$ can be calculated by two parts with the cost of misclassification $\lambda(\bullet|\bullet)$:

$$\text{Er}(C_i) = \sum_{j=1, j \neq i}^{m'} P(C_j|C_i)P(C_i)\lambda(C_j|C_i) + \sum_{j=1, j \neq i}^{m'} P(C_i|C_j)P(C_j)\lambda(C_i|C_j)$$

(11)

where the first part stands for the losses of rejecting a right classification $C_i$, when the true state is $C_i$, which is called Type (I) errors (false positive) in statistics; the second part stands for the losses of accepting a wrong classification $C_i$, when the true state is not $C_i$, which is called Type (II) errors (false negative) in statistics [32]. In the viewpoint of a Bayesian discriminant classifier, the first part stands for the losses that one looks at information that should not substantially change one’s prior estimate of probability, but it actually do it; the second part stands for the losses that one looks at information which should change one’s estimate, but it actually not do it [7]. So, the following work is to select the best classification which has minimum overall conditional risk.

Let $C' = \{C_1, C_2, \ldots, C_{m'}\}$ ($m' < m$) be the candidate classifications in position region, and the best one $C_l$ satisfies:

$$\text{Decide } C_l : \text{if } \text{Er}(C_l) \leq \text{Er}(C_i), \ i = 1, 2, \ldots, m'; \ i \neq l.$$ 

(12)

The classification $C_l$ which has the minimum overall risk is the best performance that can be achieved. The decision maker can choose it as the best choice. In addition, if there are more than one classification satisfy the equation (12), a tie-breaking criterion can be used [37]. However, we don’t consider the reducts of decision rule in our discussion. The attribute reduction strategies in probabilistic rough sets and DTRS can be found in [3, 41, 53]. We can simply introduce these reduction algorithms to our approach when we deal with a real problem.

4. An Illustration

Let us illustrate the above concepts based on a didactic example in medical diagnosis. The symptoms of a patient are described as $\{\text{fever, cough, nausea, headache, nose snivel, dysentery}\}$ according to a series of carefully diagnoses. Furthermore, the results of the symptom may lead to 5 possible diseases named H1N1, SARS, Viral influenza, Wind chill and Common cold, respectively. The doctors want to choose the best candidate from the 5 diseases.

According to the analysis process in Section 3, we denote the 5 possible classifications as $C = \{C_1, C_2, C_3, C_4, C_5\}$ to describe the 5 diseases. Firstly, the five-category classification problem can convert to 5 two-category classification problems $\{C_i, \neg C_i\} \ (i = 1, 2, 3, 4, 5)$. There are 6 parameters in each model, in which $\lambda_{FC_i}$, $\lambda_{BC_i}$, $\lambda_{NC_i}$ denote the cost incurred for taking actions of accepting $C_i$, need further examination and rejecting $C_i$ when the state is $C_i$, respectively. $\lambda_{P-C_i}$, $\lambda_{B-C_i}$, $\lambda_{N-C_i}$ denote the cost incurred for taking actions of accepting $C_i$, need further examination and rejecting $C_i$ when the state is not $C_i$, respectively. By considering the fact that there is no cost when doing a right decision, we set $\lambda_{FC_i} = \lambda_{N-C_i} = 0$. Furthermore, due to the cost of classifying an object $x$ in $C_i$ to the boundary region $BND(C_i)$ is no more than the cost of classifying an object $x$ into the negative region $NEG(C_i)$, the cost of classifying an object $x$ not belonging to $C_i$ in the boundary region $BND(C_i)$ is no more than the cost of classifying an object $x$ into the positive region $POS(C_i)$. We have $\lambda_{BC_i} \leq \lambda_{NC_i}$ and
executed. That is, the candidates which in Table 3 denotes the cost for misclassification in formula (11).

\[ \lambda_{B-C_i} \leq \lambda_{P-C_i} \] The detailed loss functions for the 5 diseases are presented in Table 1, where \( u \) is a unit cost determined by the individual administration [12].

By using the formula (8), we can directly compute the values of \( \alpha_i \) and \( \beta_i \) for the 5 diseases in Table 1, which are shown in Table 2.

### Table 1. The loss functions for the 5 diseases

<table>
<thead>
<tr>
<th>( C_i )</th>
<th>( \lambda_{PC_i} )</th>
<th>( \lambda_{EC_i} )</th>
<th>( \lambda_{NC_i} )</th>
<th>( \lambda_{P-C_i} )</th>
<th>( \lambda_{B-C_i} )</th>
<th>( \lambda_{N-C_i} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>0</td>
<td>4( u )</td>
<td>10( u )</td>
<td>9( u )</td>
<td>4( u )</td>
<td>( \lambda_{N-C_1} )</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>0</td>
<td>5( u )</td>
<td>10( u )</td>
<td>5( u )</td>
<td>4( u )</td>
<td>( \lambda_{N-C_2} )</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>0</td>
<td>2( u )</td>
<td>7( u )</td>
<td>6( u )</td>
<td>1( u )</td>
<td>( \lambda_{N-C_3} )</td>
</tr>
<tr>
<td>( C_4 )</td>
<td>0</td>
<td>3( u )</td>
<td>6( u )</td>
<td>4( u )</td>
<td>1.5( u )</td>
<td>( \lambda_{N-C_4} )</td>
</tr>
<tr>
<td>( C_5 )</td>
<td>0</td>
<td>1( u )</td>
<td>3( u )</td>
<td>1( u )</td>
<td>0.5( u )</td>
<td>( \lambda_{N-C_5} )</td>
</tr>
</tbody>
</table>

### Table 2. The \( \alpha_i \) and \( \beta_i \) value of the 5 diseases

<table>
<thead>
<tr>
<th>Parameters</th>
<th>( \alpha_i )</th>
<th>( \beta_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>0.5556</td>
<td>0.4</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>0.375</td>
<td>0.2875</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>0.7143</td>
<td>0.1667</td>
</tr>
<tr>
<td>( C_4 )</td>
<td>0.4545</td>
<td>0.3333</td>
</tr>
<tr>
<td>( C_5 )</td>
<td>0.3333</td>
<td>0.2</td>
</tr>
</tbody>
</table>

As stated in Section 3.2, we should choose a common criterion for the 5 diseases. A risk-preference doctor may choose a lower value of \( \alpha^* \) (i.e., \( \alpha^* = 0.4 \)), and the diseases \( C_1 \) (H1N1), \( C_3 \) (Viral influenza), \( C_4 \) (Wind chill) enter into the positive region in this case. A risk-neutral doctor may choose a higher value of \( \alpha^* \) (i.e., \( \alpha^* = 0.6 \)), and only \( C_3 \) (Viral influenza) enters into the positive region in this case. Here, inspired by the simple majority rule, the project which has more than half the votes should be executed. That is, the candidates which \( \alpha_i > 0.5 \) may enter into the positive region. Hence, the diseases \( C_1 \) (H1N1) and \( C_3 \) (Viral Influenza) may be considered as the positive choices according to primary diagnosis. However, we only need to consider 2 diseases instead of all 5 diseases in the next step, and our approach achieves the goal of dimension reduction.

Then, we choose the best one and make the final decision from the two positive candidates. According to the discussion in Section 3.2, it need further observation to get some other parameters. After an intensive examination, the symptoms of the patient also include sore throat, body aches, chills and fatigue, the correlative parameters are shown in Table 3. Different with the loss functions \( \lambda_{\bullet\bullet} \) in Table 1, the \( \lambda(\bullet\bullet) \) in Table 3 denotes the cost for misclassification in formula (11).

In addition, we suppose the prior probability of 5 diseases is \( P(C_1) = 0.1 \), \( P(C_2) = 0.05 \), \( P(C_3) = 0.25 \), \( P(C_4) = 0.2 \), \( P(C_5) = 0.4 \) according to the historical database. So, we can compute the overall conditional risk of the misclassification of \( C_1 \) and \( C_3 \) with formula (11), respectively.

\[
\text{Er}(C_1) = \sum_{j=1, j \neq 1}^{5} P(C_j|C_1)P(C_1)\lambda(C_j|C_1) + \sum_{j=1, j \neq 1}^{5} P(C_1|C_j)P(C_j)\lambda(C_1|C_j) = 1.355u
\]

\[
\text{Er}(C_3) = \sum_{j=1, j \neq 3}^{5} P(C_j|C_3)P(C_3)\lambda(C_j|C_3) + \sum_{j=1, j \neq 3}^{5} P(C_3|C_j)P(C_j)\lambda(C_3|C_j) = 0.63u
\]
Table 3. The misclassification correlative parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$C_1$</th>
<th>$C_3$</th>
<th>Parameters</th>
<th>$C_1$</th>
<th>$C_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(C_1</td>
<td>C_i)$</td>
<td>0.3</td>
<td>0.2</td>
<td>$\lambda(C_1</td>
<td>C_i)$</td>
</tr>
<tr>
<td>$P(C_2</td>
<td>C_i)$</td>
<td>0.1</td>
<td>0.2</td>
<td>$\lambda(C_2</td>
<td>C_i)$</td>
</tr>
<tr>
<td>$P(C_3</td>
<td>C_i)$</td>
<td>0.3</td>
<td>0.3</td>
<td>$\lambda(C_3</td>
<td>C_i)$</td>
</tr>
<tr>
<td>$P(C_4</td>
<td>C_i)$</td>
<td>0.1</td>
<td>0.1</td>
<td>$\lambda(C_4</td>
<td>C_i)$</td>
</tr>
<tr>
<td>$P(C_5</td>
<td>C_i)$</td>
<td>0.2</td>
<td>0.2</td>
<td>$\lambda(C_5</td>
<td>C_i)$</td>
</tr>
<tr>
<td>$P(C_1</td>
<td>C_i)$</td>
<td>0.3</td>
<td>0.1</td>
<td>$\lambda(C_1</td>
<td>C_1)$</td>
</tr>
<tr>
<td>$P(C_1</td>
<td>C_2)$</td>
<td>0.05</td>
<td>0.1</td>
<td>$\lambda(C_1</td>
<td>C_2)$</td>
</tr>
<tr>
<td>$P(C_1</td>
<td>C_3)$</td>
<td>0.15</td>
<td>0.3</td>
<td>$\lambda(C_1</td>
<td>C_3)$</td>
</tr>
<tr>
<td>$P(C_1</td>
<td>C_4)$</td>
<td>0.1</td>
<td>0.4</td>
<td>$\lambda(C_1</td>
<td>C_4)$</td>
</tr>
<tr>
<td>$P(C_1</td>
<td>C_5)$</td>
<td>0.05</td>
<td>0.25</td>
<td>$\lambda(C_1</td>
<td>C_5)$</td>
</tr>
</tbody>
</table>

Due to $\text{Er}(C_3) < \text{Er}(C_1)$, we conclude that the patient gets influenza according to equation (12).

5. Conclusion

In order to strike the right balance between the simplicity and flexibility, and between mathematical generality and practical applicability, a two-stage approach was introduced into our research to solve the multiple-category classification problems. The first stage was to change an $m$-category classification problem into $m$ two-category classification problems based on DTRS, and the second stage was to choose the best candidate by using the Bayesian discriminant analysis approach. The process of our approach provides a basic and naive thought on the multiple-category classification problems with DTRS. In addition, a case study of medicine diagnosis validates the feasibility of our method. However, our discussion in this paper is based on a given sample $x$ or an equivalent class $[x]$. The extension of our approach to the universe of all objects (including the unseen so far objects) seem important in real life situations. Some behavior strategies or psychological methods will also be introduced to deal with this issue in our future work.

Acknowledgements

This paper is an extended version of the paper published in the proceedings of RSKT2010. The authors appreciate Professor Y.Y. Yao for his insightful suggestions. The authors also thank all the members in Web Intelligence Lab, Department of Computer Science, University of Regina, Canada. The authors are grateful to constructive comments from the reviewers.

References


