MULTILABEL CLASSIFICATION RULE WITH PERFORMANCE CONSTRAINTS

Edith Grall-Maës, Pierre Beauseroy, Abdenour Bounsiar

Université de Technologie de Troyes
Institut des Sciences et Technologies de l'Information de Troyes (CNRS FRE 2732)
Équipe Modélisation et Sûreté des Systèmes
12, rue Marie Curie - BP 2060 -10010 Troyes cedex - FRANCE
email : {edith.grall, pierre.beauseroy, abdenour.bounsiar}@utt.fr

ABSTRACT

A formulation for multilabel and performance constraints classification problems is presented within the framework of statistical decision theory. The definition of the problem takes into account three concerns. The first is the cost function which defines the criterion to minimize ; the second is the decision options which are defined by the admissible assignment classes or subsets of classes and the third one is the constraints of performance. Assuming that the conditional probability density functions are known, the classification rule that is solution of the stated problem is expounded. Two examples are provided to illustrate the formulation and the decision rule obtained.

1. INTRODUCTION

Theoretical studies of binary classification have yielded different optimal rules, according to particular criteria (Bayes rule, Neyman-Pearson test, minimax test) [1]. In some systems, the performances obtained are unsatisfactory because the error rate is excessive. A reject option has been introduced as a mean to reduce the error rate through a rejection mechanism [2, 3, 4]. It consists in withholding a decision and directing the rejected pattern to an exceptional handling, using additional information. However the performance of the rule obtained can be unsatisfactory with respect to criteria other than the error rate. In general, the desired performances can be defined by several constraints, which can combine different total or conditional probabilities and which can be expressed using inequalities or order relationships. The classification rules for the case of two constraints stating that each of the two conditional errors is bounded and the case of one constraint stating that the ratio of the error probability to the non-rejection probability is bounded are studied in [5, 6].

Problems of classification with reject option in the case of more than two classes are more complex. The simplest decision rule with reject option was proposed by Chow [2]. It consists in rejecting a sample if its highest posterior probability is lower than some threshold. The optimality is based on a tradeoff between the error rate and the rejection rate. A more complex scheme called class-selective rejection was proposed by Ha [7]. The pattern is not rejected from all classes but only from those that are most unlikely to issue the pattern. The optimality is defined as the best tradeoff between the error rate and the average number of selected classes. It consists in assigning the pattern to all classes whose posterior probability is greater than a pre-specified threshold. Another rule consisting in minimizing the maximum distance between selected classes for a given average number of classes has been proposed in [8]. These rules are interesting since providing a list of classes, instead of making a simple rejection, can make the subsequent processing easier. However the proposed rules do not take into account performance constraints.

The aim of this paper is to expound the formulation of multilabel classification with constraints and to derive the classification rule for problems assuming that the conditional density functions and the a priori probabilities are known or correctly estimated.

The formulation considers the following concerns :

- decision options : they correspond to the assignment classes or subsets of classes that are deemed as admissible for the problem,
- constraints : they correspond to the performance constraints to be satisfied,
- cost function : it corresponds to the function to minimize.

These concerns are described in section 2. Section 3 considers how to elaborate the classification rule of the stated problem. In section 4, two simulated problems are provided to illustrate the formulation and the decision rule obtained. The paper is concluded in section 5.

2. PROBLEM FORMULATION

2.1. Decision options

Let us suppose that a pattern x belongs to a class j noted C_j, with $j = 1..N$ where N is the number of classes. The classification rule consists in assigning the pattern x to a label set

ω_i which is a class or a subset of classes. Assigning x to a subset of classes means that the element is considered as belonging to one of the classes in the subset without distinction of class. The decision options set Ω is defined by the label sets ω_i :

$$\Omega = \{\omega_1, \omega_2, \dots \omega_I\} \tag{1}$$

where I is the number of sets, whose maximum value is $2^N - 1$. Each ω_i is a subset of the N classes, containing at least one class, and specified by the numbers of the classes, for example $\omega_4 = \{1; 4; 5\}$.

We define Z_i as the set of patterns x that are assigned to ω_i :

$$Z_i = \{x \in \mathbb{R}^n | x \text{ is assigned to } \omega_i\}. \tag{2}$$

Since each x has to be assigned to a unique ω_i, the sets Z_i build up a partition of $\mathbb{R}^n$, that we call Z.

The probability of deciding that an element of the class j belongs to the set ω_i is $P(D_i/C_j)$:

$$P(D_i/C_j) = \int_{Z_i} P(x/C_j)dx \tag{3}$$

where $P(x/C_j)$ are the conditional density functions.

2.2. Performance constraints

Any performance constraint $C^{(k)}$ is defined by its expression $e^{(k)}$ and its threshold $\gamma^{(k)}$:

$$e^{(k)} \leq \gamma^{(k)} \text{ with } e^{(k)} = \sum_{i=1}^{I} \sum_{j=1}^{N} \alpha_{i,j}^{(k)} P_j P(D_i/C_j) \tag{4}$$

where $\alpha_{i,j}^{(k)} \in \mathbb{R}$ and $P_j = P(C_j)$ are the a priori probabilities.

A large diversity of constraints can be defined using this formulation, and in particular the following ones :

- a constraint on the error associated with an assignment set, for example : $P(D_1/C_2) + P(D_1/C_3) \leq \gamma$,
- a constraint on the error associated with a class, for example : $P_1 P(D_2/C_1) + P_1 P(D_3/C_1) \leq \gamma$,
- a constraint on good decision associated with a class, for example : $P(D_1/C_1) \geq \gamma$,
- a constraint expressed by a ratio, for example : $P(D_1/C_1)/\left[P(D_1/C_1) + P(D_2/C_1)\right] \geq \gamma$,
- a constraint described by a order relation, for example : $P(D_2/C_1) \leq P(D_3/C_1)$.

2.3. Cost function

The cost function allows the inclusion of simple problem formulations. It is given by :

$$\overline{c} = \sum_{i=1}^{I} \sum_{j=1}^{N} c_{ij} P_j P(D_i/C_j). \tag{5}$$

where c_{ij} is the cost of deciding that an element x belongs to the set ω_i when it belongs to the class j.

This cost function is general and includes usual functions, in particular the ones described below.

First, let us consider the case where Ω is composed of $I = N$ sets $\omega_i = \{i\}$. Taking $c_{ij} = 1$ for $i \neq j$ and $c_{ij} = 0$ for $i = j$, $\overline{c}$ is the expression of the total error probability.

Second, let us consider the case of two classes where Ω is composed of sets $\omega_1 = \{1\}$, $\omega_2 = \{2\}$ and a set $\omega_3 = \{1; 2\}$ that represents a reject option. The case of the following two constraints problem has been developed in [5, 6] :

$$P(D_1/C_2) \leq e_{12} \text{ and } P(D_2/C_1) \leq e_{21} \tag{6}$$

with e_{12} and $e_{21} \in [0, 1]$. If these constraints can be satisfied without rejection, then the cost function to minimize is the error probability P_E. When the constraints can not be satisfied using only the sets $\{1\}$ and $\{2\}$, then rejection has to be used and the function to minimize is the rejection probability $P_R = \sum_{j=1}^{2} P_j P(D_3/C_j)$. Though it seems that two optimization problems arise, it can be shown that they can be expressed by the shared cost function $\overline{c} = P_E + P_R$ which corresponds to (5) with $c_{31} = 1$, $c_{32} = 1$, $c_{12} = 1$, $c_{21} = 1$, $c_{11} = 0$ and $c_{22} = 0$.

Finally, let us consider the cost function (5) with $c_{ij} = c_m \delta_{ij} + c_n |\omega_i|$ where δ_{ij} is equal to 1 if j is in the set ω_i and 0 otherwise ; $|\omega_i|$ is the cardinality of the set ω_i ; c_m and c_n are constant positive values. Then the obtained cost function is the same as the one in [7].

The choice strategy for the cost values is beyond the scope of this paper, however, some remarks can be made. Note that the values of c_{ij} are relative since the aim is to minimize $\overline{c}$, we suggest defining the values in the interval $[0; 1]$. When the set ω_i contains only one class, c_{ij} will be generally equal to 0 for j equal to the class in ω_i and 1 otherwise. When ω_i contains several classes, and class $j \notin \omega_i$, c_{ij} defines an error cost then this cost will be generally equal to 1 ; when class $j \in \omega_i$, c_{ij} defines an indistinctness cost then it will be generally growing with the set size.

3. CLASSIFICATION RULE

To derive the classification rule, it is necessary to find the partition Z^* so that the cost $\overline{c}$ is minimum and the constraints given by (4) satisfied. Then the problem to be solved is the following one :

$$\min_{Z} \ \overline{c} \quad \text{subject to } e^{(k)} \leq \gamma^{(k)} \ \forall k = 1..K \tag{7}$$

$\overline{c}$ and $e^{(k)}$ are functions of the partition Z through $P(D_i/C_j)$. A sufficient condition for Z^* to be solution of the problem (7) is that (Z^*, μ^*) is a saddle point of the Lagrangian associated to the problem, given by :

$$L(Z, \mu) = \overline{c} + \sum_{k=1}^{K} \mu_k \left(e^{(k)} - \gamma^{(k)}\right)$$

in which $\mu = [\mu_1, \mu_2 \dots \mu_K]$ and $\mu_i \geq 0, i = 1 \dots K$ are the Lagrange multipliers associated to each of the constraints. To determine the saddle point (Z^*, μ^*) it is possible to solve the dual problem given by :

$$\max_{\mu \in \mathbb{R}^{K+}} \left\{ \min_{Z} L(Z, \mu) \right\}. \quad (8)$$

Using (4) and (5), we can rewrite the Lagrangian $L(Z, \mu)$ as :

$$L(Z,\mu) = \sum_{i=1}^{I}\sum_{j=1}^{N} c_{ij} P_j P(D_i/C_j) + \sum_{k=1}^{K} \left(\mu_k \sum_{i=1}^{I}\sum_{j=1}^{N} \alpha_{ij}^{(k)} P_j P(D_i/C_j) - \mu_k \gamma^{(k)} \right)$$

which is equal to :

$$L(Z,\mu) = \sum_{i=1}^{I} \int_{Z_i} \lambda_i(x,\mu) dx - \sum_{k=1}^{K} \mu_k \gamma^{(k)}, \quad (9)$$

where $\lambda_i(x, \mu)$ is given by :

$$\lambda_i(x,\mu) = \sum_{j=1}^{N} P_j P(x/C_j) \left(c_{ij} + \sum_{k=1}^{K} \mu_k \alpha_{ij}^{(k)} \right). \quad (10)$$

For a given μ the minimum value of $L(Z, \mu)$ is obtained by choosing the Z_i so that the integrated expression is minimum. It follows that Z_i, for $i = 1..I$, is given by :

$$Z_i = \{x | \lambda_i(x,\mu) < \lambda_l(x,\mu), l = 1..I,\ l \neq i\} \quad (11)$$

The solution of the dual problem (8) is given by :

$$\mu^* = \arg \max_{\mu \in \mathbb{R}^{K+}} w(\mu) \quad (12)$$

$$\text{with } w(\mu) = \min_{Z} \sum_{i=1}^{I} \int_{Z_i} \lambda_i(x,\mu) dx - \sum_{k=1}^{K} \mu_k \gamma^{(k)} \quad (13)$$

in which the Z_i are defined by (11). Finally the optimal classification rule is defined by the partition Z^* so that the Z_i^* are given by (11) with $\mu = \mu^*$.

Since the dual function w is a concave function of μ, this function has an extremum. In order to determine the value μ^*, usual optimization algorithms can be used. It may occur that the problem has no solution, meaning that the constraints can not be simultaneously satisfied . In these cases, the extremum of the function w is obtained for infinite values of μ. Thus, if no maximum is found before reaching large values of μ, the problem has no solution.

From equation (10), it can be noticed that the classification rule would be the same for the problem without constraints and with the costs c'_{ij} defined by :

$$c'_{ij} = c_{ij} + \sum_{k=1}^{K} \mu_k^* \alpha_{ij}^{(k)}. \quad (14)$$

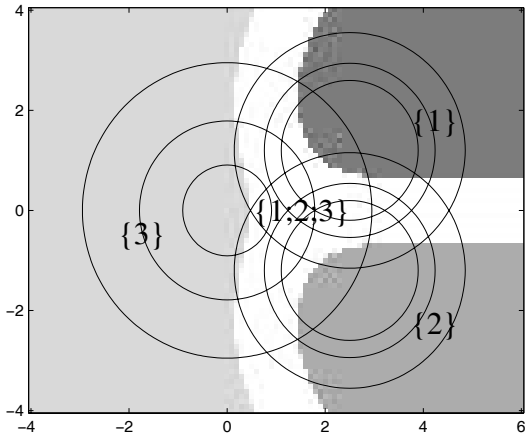

Fig. 1. Density probability functions and classification rule for the problem with one constraint

4. SIMULATION RESULTS

This section presents the results for a simulated problem. Each pattern $x \in \mathbb{R}^2$ belongs to one of three classes which are normal distributions. The means and covariance matrix are given by : $m_1 = (2.5; 1.2), \Sigma_1 = I, m_2 = (-2.5; 1.2), \Sigma_2 = I$, $m_3 = (0; 0), \Sigma_1 = 2.2I$ where I is the identity matrix. The density probability functions are represented on figure 1. Two classification rules corresponding to two different problems have been determined.

The first rule was determined according to the following components :

- 4 label sets : $\omega_1 = \{1\}, \omega_2 = \{2\}, \omega_3 = \{3\}, \omega_4 = \{1; 2; 3\}$
- 1 constraint : $P_E < 0.05$ with $P_E = P_2P(D_1/C_2) + P_3P(D_1/C_3) + P_1P(D_2/C1) + P_3P(D_2/C_3) + P_1P(D_3/C_1) + P_2P(D_3/C_2)$
- cost function : $\overline{c} = P_E + P_R$ with $P_R = P_1P(D_4/C_1) + P_2P(D_4/C_2) + P_3P(D_4/C_3)$.

The partition associated with the obtained classification rule is represented in figure 1. It was obtained for $\mu^* = 3.66$. This rule is the same as the one obtained without constraints and with the cost function :

$$\overline{c} = tP_E + P_R$$

for $t = 1 + \mu^*$. This cost function is the usual one when the rejection classification rule is determined using the trade-off between the error rate and the rejection rate. The usual way consists in testing different values for μ and choosing the value which gives the expected error rate. The proposed method allows us to quickly obtain the correct value of μ using an optimization algorithm.

The second classification rule was designed according to the following components :

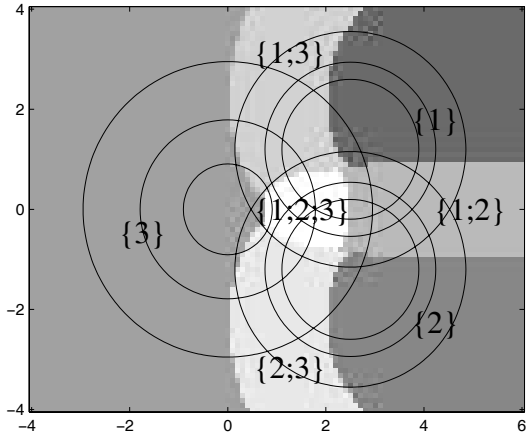

Fig. 2. Density probability functions and classification rule for the problem with six constraints

- 7 label sets : $\omega_1 = \{1\}, \omega_2 = \{2\}, \omega_3 = \{3\}, \omega_4 = \{1;2\}, \omega_5 = \{1;3\}, \omega_6 = \{2;3\}, \omega_7 = \{1;2;3\}$,
- 6 constraints :

$$\begin{cases} P_2P(D_1/C2) + P_3P(D_1/C_3) \leq 0.01 \\ P_1P(D_2/C1) + P_3P(D_2/C_3) \leq 0.01 \\ P_1P(D_3/C1) + P_2P(D_3/C_2) \leq 0.01 \\ P_3P(D_4/C_3) \leq 0.008 \\ P_2P(D_5/C_2) \leq 0.004 \\ P_1P(D_6/C_1) \leq 0.004 \end{cases} \quad (15)$$

- cost function :

$$\begin{aligned} \overline{c} = {} & P_2P(D_1/C_2) + P_3P(D_1/C_3) + P_1P(D_2/C1) \\ & + P_3P(D_2/C_3) + P_1P(D_3/C_1) + P_2P(D_3/C_2) \\ & + 0.5\,(P_1P(D_4/C_1) + P_2P(D_4/C_2)) + P_3P(D_4/C_3) \\ & + 0.5\,(P_1P(D_5/C_1) + P_3P(D_5/C_3)) + P_2P(D_5/C_2) \\ & + 0.5\,(P_2P(D_6/C_2) + P_3P(D_6/C_3)) + P_1P(D_6/C_1) \\ & + P_1P(D_7/C_1) + P_2P(D_7/C_2) + P_3P(D_7/C_3). \end{aligned}$$

  For the sets ω_1, ω_2 and ω_3 containing one class, the cost c_{ij} is equal to 1 if $i \neq j$ (incorrect class) and 0 otherwise (correct class). For the sets ω_4, ω_5 and ω_6 containing two classes, the cost c_{ij} is equal to 1 if class $j \notin \omega_i$ (incorrect class) and 0.5 otherwise (correct classification among classes subsets but without distinction of class). For the set ω_7 containing the three classes, the cost c_{ij} is equal to 1 because it corresponds to total rejection.

The partition associated with the obtained classification rule is represented in figure 2. It was obtained for $\mu^* = [4.04; 4.04; 2.04; 4.39; 3.92; 3.92]$.

5. CONCLUSION

A formulation is proposed for multilabel and performance constraints classification problems. It considers three concerns defining the problem : the label sets, the performance constraints and the cost function. This formulation embodies the usual framework of statistical decision theory and common rejection rules.

The classification rule for the stated problem, assuming that the probability density functions are known, is found. It is obtained using standard optimization algorithms. The approach also enables us to decide about the existence of solutions for a given problem. It is shown that any problem defined by a cost function and performance constraints can be transformed into an equivalent problem, defined by another cost function but without constraints, yet the classification rule remains the same. Whereas different approaches for determining a rejection rule with a constraint on the error probability consist in testing several costs so that the given performance is achieved, the proposed approach is more suitable since the performance is directly achieved by solving an optimization problem.

Future work will focus on designing classification rules for such problems where the process is learned using a training sample set.

6. REFERENCES

[1] K. Fukunaga, *Introduction to statistical pattern recognition*, Boston, 1990.

[2] C.K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. IT-16, no. 1, pp. 41–46, 1970.

[3] B. Dubuisson and M. Masson, "A statistical decision rule with incomplete knowledge about classes," *Pattern recognition*, vol. 26, no. 1, pp. 155–165, 1993.

[4] G. Fumera, F. Roli, and Giacinto G., "Reject option with multiple thresholds," *Pattern recognition*, vol. 33, no. 12, pp. 2099–2101, 2000.

[5] A. Bounsiar, P. Beauseroy, and E. Grall-Maës, "A straightforward SVM approach for classification with constraint," in *Proceedings of the EUSIPCO'05*, Antalya, Turkey, 2005.

[6] E. Grall-Maës, P. Beauseroy, and A. Bounsiar, "Classification avec contraintes : problématique et apprentissage d'une règle de décision," in *Proceedings of GRETSI'05*, Louvain-la-Neuve, Belgique, 2005, pp. 1145–1148.

[7] T. Ha, "The optimum class-selective rejection rule," *Transactions on Pattern Analysis ans Machine Intelligence*, vol. 19, no. 6, pp. 608–615, 1997.

[8] T. Horiuchi, "Class-selective rejection rule to minimize the maximum distance between selected classes," *Pattern recognition*, vol. 31, no. 10, pp. 579–1588, 1998.