14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

QUALITY ASSESSMENT OF A SUPERVISED 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 multilabel classification rule with performance constraints for supervised problems is presented. It takes into account three concerns : the loss function which defines the criterion to minimize, the decision options which are defined by the admissible assignment classes or subsets of classes, and the constraints of performance. The classification rule is determined using an estimation of the conditional probability density functions and by solving an optimization problem. A criterion for assessing the quality of the rule and taking into account the loss function and the issue of the constraints is proposed. An example is provided to illustrate the classification rule and the relevance of the criterion.

1. INTRODUCTION

Classification problems with performance constraints can arise in different real fields, like cancer diagnosis, currency verification, fraud detection or face identification. The specifications of such problems are given by the different classes whose number can be two or more, and by the desired performances. These ones, in general, can be defined by several constraints, which can combine different total or conditional probabilities and which can be expressed using inequalities or order relationships.

In statistical hypothesis testing, the Neyman-Pearson test [1] is the solution of a classification problem with one constraint and two classes : it minimizes the second type error subject to the first type error being equal to a constant. Another usual binary classification problem with a performance constraint is defined by a bound on the error rate. Its interest is to ensure a high reliability and avoid erroneous decisions. Because the rule which minimizes the error rate can lead to a larger rate than the desired bound, an ambiguity reject option has been introduced as a mean to reduce the error probability through a rejection mechanism [2, 3]. It consists in withholding a decision and directing the rejected pattern to an exceptional handling, using additional information. When rejection is needed, the optimal rule is the one which satisfies the error rate and minimizes the rejection rate. More generally, the classification rule researched is the one which minimizes a given loss function without rejection if the constraints are satisfied, and the one which minimizes the rejection rate otherwise. The rules for the two constraints case where each of the two conditional errors is bounded and for the one constraint case where the ratio of the error probability to the non-rejection probability is bounded are studied in [4, 5].

For solving classification problems with desired performances in the case of more than two classes, the reject option has also to be considered. However the reject option is more complex than in the case of two classes. The simplest rule of classification with rejection is similar to the one for two classes : it assigns the pattern to a class or it rejects it. It was proposed by Chow [2] for designing a classification rule that minimizes the reject rate for a given error rate. A more complex rejection scheme called class-selective rejection consists in not rejecting the pattern from all classes but only from those that are most unlikely to issue the pattern. It was introduced by Ha [6] for designing a classification rule that minimizes the average number of selected classes for a given error rate. Then it was used in [7] for designing a rule that minimizes the maximum distance between selected classes for a given average number of classes.

A general formulation has been introduced in [8] for multilabel classification problems with performance constraints. It considers three concerns : the first one deals with the decision options which correspond to the assignment classes or subsets of classes that are deemed as admissible for the problem, the second one deals with the performance constraints to be satisfied, and the third one deals with the loss function which defines the function to minimize. The classification rule in statistical hypothesis testing context, assuming that the conditional density functions and the a priori probabilities are known, was expounded.

This paper tackles the problem of classification adapted to this formulation when the process is described by a training sample set. A supervised learning rule and a criterion for assessing the quality of the rule are proposed. Indeed, to compare several rules, it is necessary to get a criterion taking into account the loss function and the issue of the constraints. In section 2, the problem of multilabel classification with performance constraints is presented and the classification rule in statistical hypothesis testing framework is expounded. In section 3, a supervised classification rule and a criterion for

assessing the quality of a rule are proposed. In section 4, simulations results are provided to illustrate the supervised rule and the relevance of the criterion. The paper is concluded in section 5.

2. MULTILABEL CLASSIFICATION

The three concerns of the formulation of the multilabel classification problem with performance constraints briefly introduced above are described more precisely in sections 2.1 to 2.3. The theoretical rule is exposed in section 2.4.

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 any class in the subset. The decision options set Ω is defined by the label sets ω_i :

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

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 :

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

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 $P(D_i/C_j)$:

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

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

2.2. Performance constraints

Any performance constraint $C^{(k)}$, where k refers to the constraint number, is defined by its expression $e^{(k)}$ and its threshold $\gamma^{(k)}$:

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

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.

2.3. Average expected loss

The average expected loss is general and enables to include simple problem formulations. It is given by :

$$c = \sum_{i=1}^{I}\sum_{j=1}^{N} c_{ij} P_j P(D_i/C_j) \qquad (2)$$

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.

The values of c_{ij} being relative since the aim is to minimize c, the values can be defined in the interval $[0; 1]$ without loss of generality. When the set ω_i contains only one class, c_{ij} will be generally chosen 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 chosen equal to 1 ; when class $j \in \omega_i$, c_{ij} defines an indistinctness cost then it will be generally chosen growing with the set size.

2.4. Theoretical classification rule

The optimal classification rule is defined by the partition Z^* so that the loss c is minimum and the K constraints given by (1) are satisfied. It is the solution of the following problem :

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

The solution to this optimization problem is given by the saddle point (Z^*, μ^*) of the Lagrangian associated with the problem :

$$L(Z, \mu) = c + \sum_{k=1}^{K} \mu_k \left(e^{(k)} - \gamma^{(k)}\right) \qquad (3)$$

in which $\mu = [\mu_1, \mu_2 \dots \mu_K]$, and $\mu_i \geq 0, i = 1 \dots K$ are the Lagrange multipliers associated with each of the constraints. Classical Lagrangian duality enables the primal problem to be transformed to its dual problem, which is easier to solve. The dual problem is given by :

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

It has been shown in [8] that the partition Z^* is defined by the Z_i^*, for $i = 1..I$, so that :

$$Z_i^* = Z_i^*(\mu^*) \qquad (4)$$

where $Z_i^*(\mu)$ is defined by :

$$Z_i^*(\mu) = \{x | \lambda_i(x, \mu) < \lambda_l(x, \mu), l = 1..I, \; l \neq i\} \qquad (5)$$

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

and μ^* is given by

$$\mu^* = \arg \max_{\mu \in \mathbb{R}^{K+}} w(\mu) \qquad (6)$$

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

3. SUPERVISED LEARNING RULE AND QUALITY ASSESSMENT OF THE RULE

3.1. Supervised classification rule

To design a supervised classification rule, a direct method is to estimate P_j and $P(x/C_j)$ from the training set and then to solve the optimization problem, as for the theoretical rule.

To estimate the probabilities P_j, an usual means is to use the ratio of samples in the different classes of the training set. To estimate the conditional distributions $P(x/C_j)$ a frequently used technique is the Parzen density estimate [1]. It allows to approximate the value of a density function at any location z knowing n independent samples z_i, $i = 1..n$ drawn from the same distribution. The expression of this estimate using a Gaussian kernel is

$$\hat{p}(z) = n^{\sigma-1}(2\pi)^{-d/2}|\Sigma|^{-1/2} \sum_{j=1}^{n} \exp\left(-\frac{1}{2}n^{2\sigma/d}(z-z_j)^t\Sigma^{-1}(z-z_j)\right) \quad (7)$$

where

- d is the vector dimension,
- Σ is an estimation of the covariance matrix of the data,
- $0 < \sigma < 1$ is a parameter that allows to adapt the smoothing of the estimated density.

3.2. Quality assessment of the rule

For measuring the quality of the classification rule, it is necessary to define a criterion. When the specifications of the problem involve no constraints, the criterion is simply given by the loss function. But when performance constraints are involved, the criterion has to take into account the constraints in addition to the loss function, and has to express that the constraints are satisfied while the loss function is minimum.

To define the criterion, we have considered the problem without constraints equivalent to the initial problem with constraints. It leads to the same classification rule and its formulation is based on a modified loss function, which depends on the loss function and the constraints of the original problem. Indeed, the classification rule of the initial problem with a loss function c and performance constraints is defined by $Z_i^*(\mu^*)$, where $Z_i^*(\mu)$ and μ^* are respectively given by (5) and (6). This is also the solution of the problem without constraints and with the loss function c' defined by :

$$c' = \sum_{i=1}^{I} \lambda_i(x,\mu^*) = \sum_{i=1}^{I}\sum_{j=1}^{N} c'_{ij}P_jP(D_i/C_j) \quad (8)$$

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

Then a mean to assess the quality of the rule is to measure the average expected loss of the equivalent problem. By subtracting the constant $\sum_{k=1}^{K} \mu_k^* \gamma^{(k)}$, the criterion remains equivalent and becomes equal to the Lagrangian function $L(Z,\mu^*)$, enabling thus comparison with this one. Since the theoretical values μ_k^* are unknown, they have to be replaced by estimated values $\widetilde{\mu_k}$. The proposed criterion κ is then given by :

$$\kappa = \sum_{i=1}^{I}\sum_{j=1}^{N}\left(c_{ij} + \sum_{k=1}^{K}\widetilde{\mu_k}\alpha_{ij}^{(k)}\right)P_jP(D_i/C_j) - \sum_{k=1}^{K}\widetilde{\mu_k}\gamma^{(k)}. \quad (9)$$

4. SIMULATION RESULTS

For a problem characterized by a loss function and constraints, the supervised learning classification rule depends on the sample set and the parameter σ used for estimating the density probability functions. The aim of the simulation was to determine if the proposed criterion is appropriate for selecting, on average for a large number of sample sets, the best classification rule with respect to the value of σ. Thus the optimal value of σ obtained using the proposed criterion has to be compared with the one which minimizes the loss function while respecting the constraints.

4.1. Problem description and theoretical rule

Each pattern x in $\mathbb{R}^2$ belongs to one of three equiprobable classes which have normal distributions. Their means and covariance matrices are given by : $m_1 = (-1.1;0), \Sigma_1 = I, m_2 = (1.1;0)$, $\Sigma_2 = I$, $m_3 = (0;2), \Sigma_1 = 0.5I$ where I is the identity matrix. The density probability functions are represented on figure 1.

The problem of classification is described by the following concerns :

- 7 label sets : $\omega_1 = \{1\}, \omega_2 = \{2\}, \omega_3 = \{3\}, \omega_4 = \{1;2\}, \omega_5 = \{1;3\}, \omega_6 = \{2;3\}, \omega_7 = \{1;2;3\}$,
- 2 constraints : $\begin{cases} P_E \leq 0.05 \\ P_I \leq 0.08 \end{cases}$ (10)

  where P_E is the probability of error :

$$P_E = P_2P(D_1/C2) + P_3P(D_1/C_3) + P_1P(D_2/C1) + P_3P(D_2/C_3) + P_1P(D_3/C1) + P_2P(D_3/C_2) + P_3P(D_4/C_3) + P_2P(D_5/C_2) + P_1P(D_6/C_1) \quad (11)$$

  and P_I is the probability of indistinctness :

$$P_I = P_1P(D_4/C_1) + P_2P(D_4/C_2) + P_1P(D_5/C_1) + P_3P(D_5/C_3) + P_2P(D_6/C_2) + P_3P(D_6/C_3), \quad (12)$$

- the average expected loss defined by the following costs :
  - for the sets ω_1, ω_2 and ω_3 containing one class, $c_{ij} = 1$ if $j \notin \omega_i$ (incorrect class) and $c_{ij} = 0$ otherwise (correct class),
  - for the sets ω_4, ω_5 and ω_6 containing two classes, $c_{ij} = 1$ if class $j \notin \omega_i$ (incorrect class) and $c_{ij} = 0.5$ otherwise (correct classification but with indistinctness),
  - for the set ω_7 containing the three classes, $c_{ij} = 1$ because it corresponds to total rejection ;

  then it can be written as : $c = P_E + 0.5P_I + P(D_7)$.

Given this problem, the proposed criterion defined by (9) rewrites as :

$$\kappa = P_E + 0.5P_I + P(D_7) + \mu_1(P_E - 0.05) + \mu_2(P_I - 0.08).$$

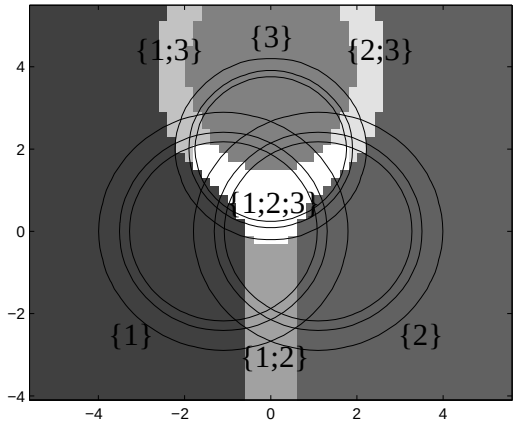

Fig. 1. Density probability functions and decision zones in function of x

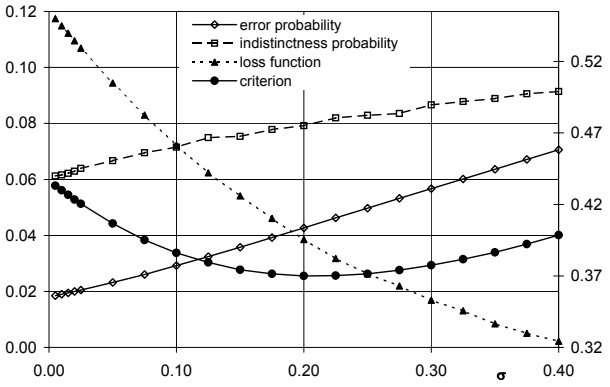

Fig. 2. Average of P_E and P_I (left vertical scale) and of c and κ (right vertical scale) of the classification rule, in function of σ.

The partition associated with the theoretical classification rule in function of x is represented in figure 1. It has been obtained for $\mu^* = [3.43; 0.46]$. Values of P_E, P_I, c and κ are given in table 1.

4.2. Experimental process

First, the classification rules were determined using the conditional probability functions estimated according to (7), with the same parameter σ, varying between 0.005 and 0.4, for all classes and with 200 samples in each class. The values of P_E, P_I, c and κ of the classification rule were computed using the theoretical density probability functions. To compute κ, the theoretical value for μ was used. The mean of these measures in function of σ were estimated from 100 sample sets. They are represented on figure 2. Since the loss decreases when σ grows and both constraints are verified for any σ smaller than 0.2, the best value is 0.2. The minimum criterion value is 0.370 and obtained for $\sigma = 0.2$. Thus the proposed criterion enables to select reliably the best value for σ.

To study the effect of the value of μ, the mean of the estimated values of μ_1 and μ_2 in function of σ was computed

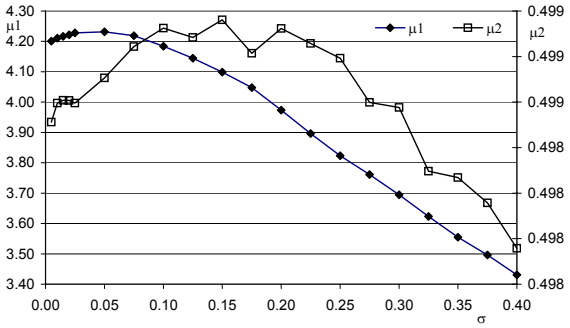

Fig. 3. Average of μ_1 and μ_2, in function of σ.

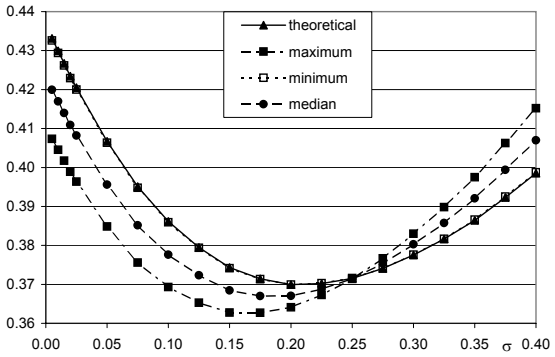

Fig. 4. Average of κ for different values of μ, in function of σ.

(it is represented on figure 3), and the value of the criterion κ was computed using an estimated value of μ instead of the theoretical one . For comparison, the same value was used for any value of σ. Results are given by figure 4. When the maximum values of μ_1 and μ_2 are used, i.e. $\mu = [4.23; 0.50]$, the minimum criterion value is obtained for $\sigma = 0.175$; when the median values are used, i.e. $\mu = [3.83; 0.50]$, the minimum is obtained for $\sigma = 0.2$, and when the minimum values are used, i.e. $\mu = [3.43; 0.50]$, the minimum is obtained for $\sigma = 0.2$. These results show that the selected value for σ is very close to the optimal one even if there is an error on the value used for μ.

Then, the classification rules were determined using the conditional probability functions estimated according to (7), with a parameter σ depending on the class. Due to similarity of classes 1 and 2, the same parameter σ_1 was used for

	50 samples	200 samples	theory
mean of PE	0.0457	0.0461	0.0505
mean of PI	0.0808	0.0783	0.0811
mean of c	0.3841	0.3609	0.3296
mean of κ	0.3698	0.3468	0.3317
standard deviation of PE	0.0107	0.0059	
standard deviation of PI	0.0148	0.0077	
standard deviation of c	0.0477	0.0235	
standard deviation of κ	0.0169	0.0050	

Table 1. Mean and standard deviation of different measures of the classification rule, in the cases of a set with 50 and 200 samples in each class, and in the case of the theoretical rule.

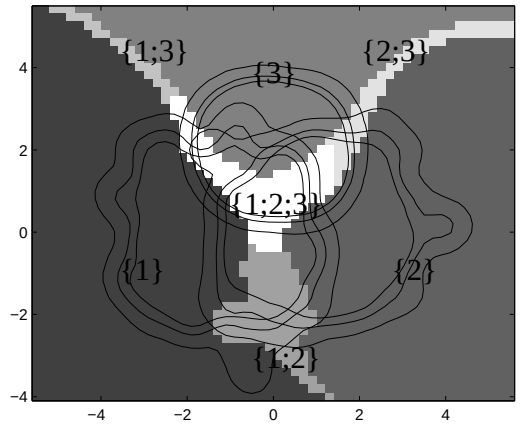

Fig. 5. Estimated density probability functions and decision zones in function of x using a set with 200 samples in each class

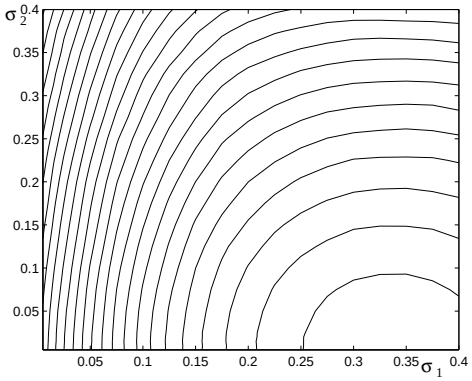

Fig. 6. Mean criterion in function of parameters σ_1 and σ_2 in the case of $N = 200$.

both classes and a parameter σ_2 was used for class 3. The experiments were carried out with sample sets containing N samples in each class, with $N = 50$ and $N = 200$. An example of a rule when $N = 200$ is given in figure 5. The mean of the criterion κ in function of σ_1 and σ_2 is given by figure 6 for $N = 200$. The parameters leading to the minimum criterion value are, for $N = 200 : \sigma_1 = 0.35$ and $\sigma_2 = 0.015$, and for $N = 50 : \sigma_1 = 0.3$ and $\sigma_2 = 0.01$. The mean and standard deviation of P_E, P_I, C and κ, estimated from 100 sample sets, are given in table 1. These results show that the criterion allows to select values of σ_1 and σ_2 so that on average the constraints are verified and the loss function is small. When the number of samples is equal to 200, the criterion value is close to the theoretical one.

5. CONCLUSION

Multilabel classification problems with constraints of performance take into account three concerns : the loss function which defines the criterion to minimize, the decision options which are defined by the admissible assignment classes or subsets of classes, and the performance constraints. A supervised learning rule for such problems is proposed. It consists in estimating the conditional density probability functions using a Parzen estimate and in solving an optimization problem as for determining the rule when the theoretical probability functions are known.

A criterion for assessing the quality of a supervised learning classification rule that takes into account the loss function and the issue of the constraints is introduced. It corresponds to the Lagrangian function of the optimization problem, and provides a measure of the loss function of the unconstrained problem leading to the same classification rule than the initial one.

Simulations on a problem with three classes and two constraints were carried out. Different values for the number of patterns in the sample set and for the parameter setting the smoothing of the estimated density functions were considered. The constraints, the loss function and the proposed criterion of the classification rule were measured using the theoretical density functions. It has been shown that the proposed criterion is appropriate for assessing the quality of a rule : it allows to choose the value of the smoothing parameter which is the best one from the point of vue of the issue of the constraints and of the minimization of the loss function.

Future work will focus on determining the rule and choosing the value of the smoothing parameter entirely from a 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] G. Fumera, F. Roli, and Giacinto G., "Reject option with multiple thresholds," *Pattern recognition*, vol. 33, no. 12, pp. 2099–2101, 2000.

[4] 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.

[5] 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.

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

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

[8] E. Grall-Maës, P. Beauseroy, and A. Bounsiar, "Multilabel classification rule with performance constraints," in *ICASSP'06*, Toulouse, France, 2006, accepted.