A STRAIGHTFORWARD SVM APPROACH FOR CLASSIFICATION WITH CONSTRAINTS

Abdenour BOUNSIAR, Pierre BEAUSEROY and Edith GRALL

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, 10010, Troyes Cedex, France
phone: + (33) 325718450, fax: + (33) 325715699, email: {abdenour.bounsiar, pierre.beauseroy, edith.grall}@utt.fr
web: www.utt.fr

ABSTRACT

This paper deals with constrained binary classification problems. First, new theoretical decision rules of two such problems are designed in a Bayesian framework. They are shown to be functions of the likelihood ratio and thresholds. Optimal performances of such classifiers can be obtained by varying only these thresholds. In order to implement such rules with sampled data, we tried to apply the same principle using SVMs. We show that varying only the intercept of the optimal SVM may lead to poor performances except for minimum error. Especially for first type error classification problems, an approach to learn SVM parameters (the *slope* and the *intercept*) that always improves the performance corresponding to the given constraint, is proposed and experimental results are discussed.

1. INTRODUCTION

Assuming known the a priori probabilities and probability density functions of classes, many decision rules have been developed (Bayes rule, Neyman-Pearson test, minimax...) giving optimal performances for certain performance criteria [1]. Some specific problems may introduce total error constraint with value smaller than Bayes error probability. In such a case, solutions including a concept of rejection have been developed. The improvement of classification is obtained at the expense of a decision discard [2] [3] [4].

This paper deals with constrained binary classification problems. First type error constrained classification is an example of such problems; it consists in minimizing second type error for a given first type one. In general, classification constraints can have more complex expressions, they can combine several error probabilities and be expressed by equality constraints, order constraints or both.

In order to introduce constrained binary classification problems, we consider the Bayesian framework in the first part of this paper and propose analytical study of two new problems in section 2. The obtained decision rules consist in comparing the likelihood ratio with thresholds. Optimal performances of such classifiers are obtained by varying only these thresholds.

However, in many real world problems described only by samples set, the probability density functions of classes are unknown and usually can not be correctly estimated. For such processes, constrained classification problems where the subject of little attention until now. In the second part of this paper, we will consider the particular case of first type error constrained classification and suggest to use Support Vector Machines, which are an elegant approach for high-dimensional classification problems and have good generalization ability, section 3 is devoted to a short review on SVMs. In section 4, we show experimentally that for such classifiers where ROC curves are usually constructed by varying only the intercept, such a procedure may lead to poor performances. In fact, the obtained ROC curves do not fit the optimal one except for the point corresponding to minimum error. A new approach for learning SVM parameters (*slope* and *intercept*) that improves performances on the point of the ROC curve corresponding to the given constraint, is proposed. A discussion is given in section 5 and conclusions are driven in section 6.

2. CONSTRAINED BINARY CLASSIFICATION

2.1 Introduction

Constrained binary classification in a Bayesian framework always refers to Neyman-Pearson test. This test introduces on the one hand a constraint (false alarm probability) and on the other hand a performance criterion to be optimized (non detection probability) [1]. In general, the constraints can have more complex expressions. They can combine simultaneously several error probabilities; they can be expressed by equality constraints, order constraints or both. Given a classification problem characterized by a set of constraints, two cases arise depending on wether the constraints can be jointly satisfied or not. If yes, the problem consists in optimizing a performance criterion based on an error probability or on a given cost function. In the other case, it is necessary to introduce rejection and the problem consists in minimizing the reject probability with respect to the constraints. To illustrate this type of constrained problems we consider the two following, 2 class ($_1$, $_2$) problems.

2.2 Binary classification with local constraints

Consider the problem of a binary classifier verifying

$$P(D_1/\ \ _2) \leq e_{12} \quad \text{and} \quad P(D_2/\ \ _1) \leq e_{21}, \tag{1}$$

where $e_{12} \in [0,1]$, $e_{21} \in [0,1]$, $D_{i,i=1,2}$ the decision of class $_i$ and $(P(D_i/\ \ _j) \equiv P_{ij})$ the probability to make D_i when the class is $_j$. The solution of this problem can be seen the an intersection of the following two symmetrical Neyman-

Pearson tests [1]:

$$\begin{array}{l} \frac{P(x/\ _1)}{P(x/\ _2)} \underset{D_2}{\overset{D_1}{\gtrless}} \ ^*_1 \text{ with } \int_{Z_1^*} P(x/\ _2)dx = e_{12} \\ \frac{P(x/\ _1)}{P(x/\ _2)} \underset{D_2}{\overset{D_1}{\gtrless}} \ ^*_2 \text{ with } \int_{Z_2^*} P(x/\ _1)dx = e_{21}, \end{array} \quad (2)$$

where $P(x/\ _i)_{i=1,2}$ is the probability density function of a pattern $x \in \mathbb{R}^d$ in the class $\ _i$, and the decision areas Z_1^* and Z_2^* are determined by the first and the second tests (2) respectively. It is clear that if $\ ^*_1 \leq \ ^*_2$ the two constraints are jointly satisfied. An optimal partition (Z_1, Z_2) can then be found by minimizing a certain criterion such as P_e, P_{12} or P_{21}. Consider here the case of minimizing P_e, since a minimum is obtained for Bayes threshold P_2/P_1 and that P_e increases when the threshold is shifted, the decision rule is [5]:

$$\frac{P(x/\ _1)}{P(x/\ _2)} \underset{D_2}{\overset{D_1}{\gtrless}} \frac{P_2}{P_1} \text{ (Bayes decision rule)} \quad \text{if} \quad \ ^*_1 \leq \frac{P_2}{P_1} \leq \ ^*_2$$

$$\frac{P(x/\ _2)}{P(x/\ _1)} \underset{D_2}{\overset{D_1}{\gtrless}} \ ^*_1 (\text{resp. } \ ^*_2) \quad \text{if} \quad \frac{P_2}{P_1} \leq \ ^*_1 \Big(\text{resp. } \frac{P_2}{P_1} \geq \ ^*_2\Big),$$

where P_i, $i = 1, 2$, is the *a priori* probability of the class $\ _i$. On the contrary, if $\ ^*_1 > \ ^*_2$ the constraints can only be satisfied if rejection is introduced, in this case the criterion to minimize is the reject probability P_r. One may easily deduce the following decision rule [5] :

$$\frac{P(x/\ _1)}{P(x/\ _2)} \left\{ \begin{array}{ll} \geq \ ^*_1 & x \text{ is classified in } \ _1 \\ \leq \ ^*_2 & x \text{ is classified in } \ _2 \\ \in]\ ^*_2, \ ^*_1[& x \text{ is rejected} \end{array} \right.$$

2.3 Binary classification with a constraint on $P_e/(1-P_r)$

Let examine now a quite different problem, consider the problem of a binary classifier minimizing the rejection probability P_r subject to the constraint

$$\frac{P_e}{P_d} = \frac{P_e}{1-P_r} \leq \qquad \text{with} \quad 0 \leq \ \leq P_B,$$

where P_B is the Bayes error and P_d the decision probability. The corresponding Lagrangian is

$$Ł(Z_r, Z_d, \ ) \ = \ P_r + \ \big(P_e - \ (1-P_r)\big),$$

with $\ \geq 0$ is a Lagrange multiplier, Z_r the rejection area and Z_d the decision one. It can be shown [5] that the corresponding decision rule minimizing the Lagrangian is:

$$\frac{P(x/\ _1)}{P(x/\ _2)} \left\{ \begin{array}{ll} \geq (P_2/P_1) & x \text{ is classified in } \ _1 \\ \leq (P_2/P_1) \ ^{-1} & x \text{ is classified in } \ _2 \\ \in](P_2/P_1) \ ^{-1}, (P_2/P_1) \ [& x \text{ is rejected} \end{array} \right.$$

where $\ = \frac{-1-}{1+} > 1$ and $P_e(\ )/P_d(\ ) = \ .$

2.4 Conclusion

Optimal decision rules of these two problems and others [5], like those of the well known: Bayes rule, Neyman-Pearson test and the minimax test, are all expressed as likelihood ratio comparisons with thresholds. Such classifiers allow to obtain optimal performances just by choosing convenient values of thresholds. But unless we have the true or an estimate of what the likelihood ratio is, such decision rules can not be used, for example on sampled data. In next paragraphs, motivated by their good generalization ability, we try to use SVM for the specific problem of first type error constrained binary classification on sampled data.

3. SUPPORT VECTOR MACHINES

Suppose we have a labelled training data $\{x_i, y_i\}$, $i = 1, ..., l, y_i \in \{-1, +1\}, x_i \in \mathbb{R}^d$, a practical application of the principle of *Structural Risk Minimization* (SRM) [6] to the problem of pattern recognition leads to the definition of *Support Vector Machines* (SVM). Support Vector Machines realize the following idea: they map $x \in \mathbb{R}^n$ into a high (possibly infinite) dimensional space and construct an optimal separating hyperplane in this space [7]. The mapping $\ (.)$ is performed by a kernel function $K(.,.)$ such that $K(x,y) = \ (x) * \ (y)$, the kernel function represents the dot product of the data in that space. The kernels that have these properties satisfy the Mercer's condition [6], i.e. for any $g(x)$ with finite L_2 norm (3), equation 4) must hold. Any positif definite kernel satisfies this condition [8].

$$\int_{-}^{+} g^2(x)dx < \qquad (3)$$

$$\int_{-}^{+} \int_{-}^{+} K(u,v)g(u)g(v)dudv > 0. \qquad (4)$$

Here we consider a kernel K depending on a set of parameters $\ $. The decision function given by an SVM is thus:

$$\begin{array}{rcl} f \ (x) & = & \text{sign}\Big(w^T \ (x) + b\Big) \\ & = & \text{sign}\Big(\sum_{i=1}^{l} \ ^0_i y_i K \ (x_i, x) + b\Big), \end{array} \qquad (5)$$

where w and b are referred to *slope* and *intercept* respectively. The coefficients $\ ^0_i$ are obtained by maximizing the following functional [7] [9]:

$$W(\ ) = \ ^l_{i=1} \ _i - \tfrac{1}{2} \ ^l_{i,j=1} \ _i \ _j y_i y_j K \ (x_i, x_j), \quad \text{subject to}$$

$$^l_{i=1} \ _i y_i = 0 \quad \text{and} \quad \ _i \geq 0, \quad i = 1, ..., l.$$

The coefficients $\ ^0_i$ define the optimal hyperplane with the maximal distance (in the high dimensional space) to the closer image $\ (x_i)$ from the training data, called the *maximal margin*. For the non-separable case, one need to allow training errors which results in the so called *soft margin SVM* [10], in which the coefficients $\ ^0_i$ are obtained by maximizing the same functional [9]:

$$W(\ ) = \ ^l_{i=1} \ _i - \tfrac{1}{2} \ ^l_{i,j=1} \ _i \ _j y_i y_j K \ (x_i, x_j), \quad \text{subject to}$$

$$^l_{i=1} \ _i y_i = 0 \quad \text{and} \quad 0 \leq \ _i \leq C, \quad i = 1, ..., l,$$

where C is the training cost penalizing the training errors, and will be considered just as another parameter of the SVM:

$$f_{(\ ,C)}(x) = \text{sign}\Big(\sum_{i=1}^{l} \ ^0_i y_i K \ (x_i, x) + b\Big), \quad 0 \leq \ ^0_i \leq C. \ (6)$$

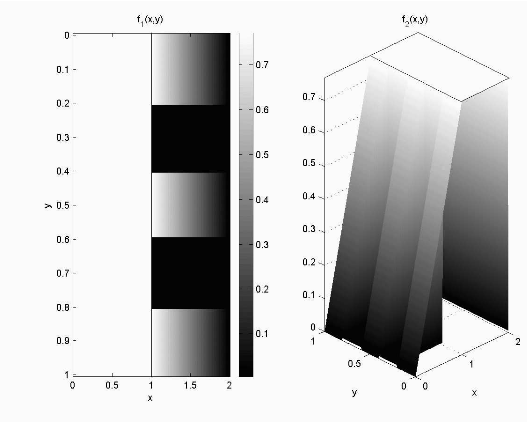

Figure 1: Data densities, the right one is a top view of $f_1(x,y)$, the left one is a lateral view of $f_2(x,y)$.

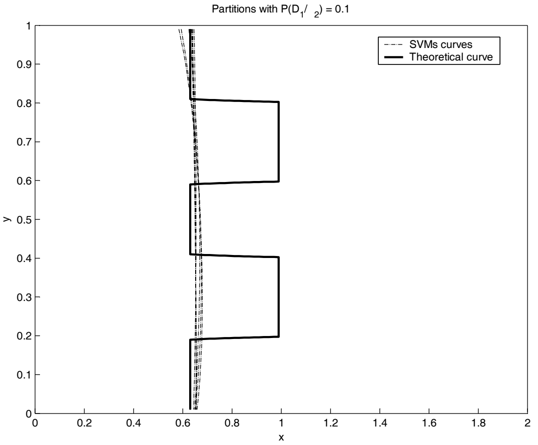

Figure 3: Partitions corresponding to $P_{21} = 0.1$.

4. EXPERIMENTAL ANALYSIS

In order to study binary classification with first type error constraint (P_{21}), we considered two classes of sampled data, with the same probabilities and probability density functions (see Figure 1). They are symmetric one to the other according to the plane $x = 1$, and took the form of letter 'E'. Experiments were driven using 7 training sets of 300 samples in each class. To learn SVMs we considered a RBF kernel (characterized by a width) $K\ (x,y) = \exp\Big(-\ \ |x-y|^2\Big)$. We first searched the optimal couple of parameters $(C,\ )$ of the SVM minimizing the generalization error [11]. This error was computed using the real probability density functions on decision areas (for more accuracy). Figure 2 shows the decision boundaries obtained by both exact theoretical rule (Bayes rule) and SVMs, they are very close. To obtain boundaries for the situation where P_e is minimized with respect to $P_{21} = 0.1$, the intercept of optimal hyperplanes were shifted (by analogy of shifting the decision threshold of Neyman-Pearson test). The resulting boundaries are reported on Figure 3. We can clearly see that these boundaries are very different from the theoretical one. Figure 4 shows ROC curves ($P_e = f(P_{21})$) obtained by varying decision thresholds for theoretical rule (2) and the intercept b of optimal hyperplanes for SVMs (5). We can see that the two curves (theoretical and SVM ones) are very close one to the other at the point corresponding to minimum P_e, and become distant when moving away. That mean that we can not obtain an optimal classifier for our problem only by moving the intercept of SVM. This is clearly seen with the decision boundaries on figure 3 (the intercepts of optimal hyperplanes were shifted in order to obtain $P_{21} = 0.1$). It is very clear that the partitions obtained by SVMs in that case are not optimal and are very far from the optimal one. In order to improve classification performances for this problem, we considered a new strategy to select SVM parameters: C, and b are tuned in order to optimize P_e subject to $P_{21} = 0.1$, i.e. the three parameters are jointly optimized such that the decision function:

$$f_{(\ ,C,b)}(x) = \text{sign}\Big(\overset{l}{\underset{i=1}{}}\ {}^{0}_{i}y_iK\ (x_i,x)+b\Big),\quad 0\le\ {}^{0}_{i}\le C \quad (7)$$

minimizes P_e verified that $P_{21} = 0.1$, where $\ =\ $ here. Figure 5 represents the ROC curves of SVMs trained on the pre-

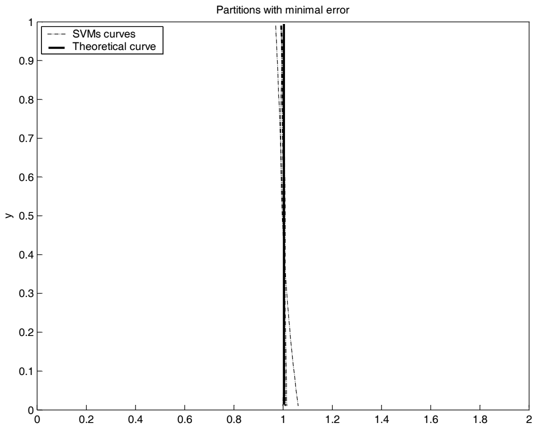

Figure 2: Partitions corresponding to minimal error.

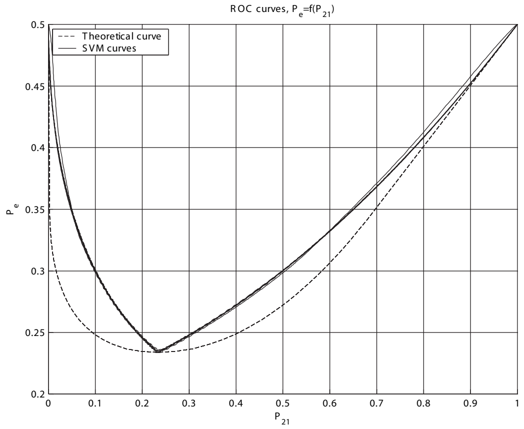

Figure 4: ROC curves $P_e = f(P_{21})$.

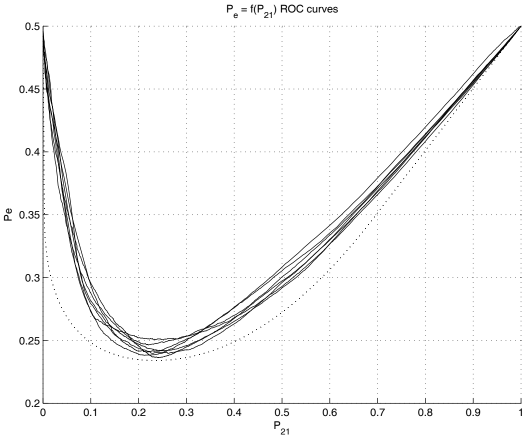

Figure 5: ROC curves $P_e = f(P_{21})$ for both theoretical (dashed line) and SVM (solid lines) results.

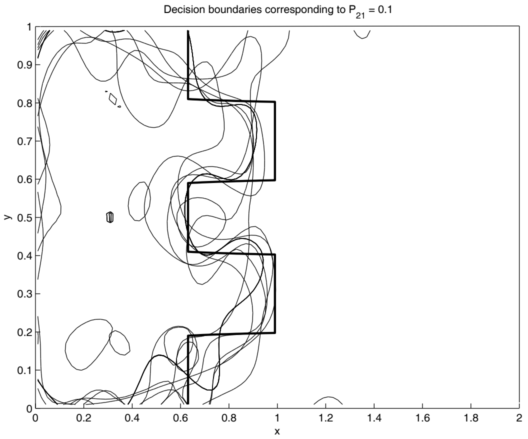

Figure 6: Decision boundaries corresponding to $P_{21} = 0.1$: theoretical (bold solid line) and SVM (sild lines).

vious 7 training sets using this new strategy. It appears that the values of P_e corresponding to $P_{21} = 0.1$, are all smaller than the ones obtained by varying only the intercept of optimal SVMs (figure 4). The decision boundaries obtained with these SVMs are all reported on Figure 6, they are undoubtedly closer to the theoretical one.

5. DISCUSSION

These experiments represent a first approach to deal with constrained binary classification problems, precisely first type error constrained classification, on sampled data using SVM. We showed with tests that varying only the intercept of SVMs, by analogy with shifting thresholds of theoretical rules, leads to poor performances. Particularly for $P_{21} = 0.1$, we obtained nearly the same value of P_e (0.3 instead of the optimal value 0,249) for all the training sets. This is due to the fact that the discriminant functions $w^T x$ (characterized by w) of optimal SVMs are not pertinent for the actual problem. In order to remedy and obtain better performances, an alternative approach to choose SVM's parameters leading to more pertinent discriminant functions (or equivalently the slope w) and that gives better performances for all the training sets, is introduced. Resulting values of P_e vary between 0.295 and 0.272 with a mean of 0.283, this corresponds to an average improvement of $\frac{0.3-0.283}{0.3-0.249} = 34\%$ upon P_e, which is an interesting result.

6. CONCLUSIONS

The study of two particular problems of constrained binary classification in a Bayesian framework has been driven in this paper. The obtained decision rules consist in comparing likelihood ratio with thresholds depending on the constraints. In order to infer such decision rules from sampled data using SVMs, one may try to construct the decision rule by tuning the intercept of the optimal hyperplane of SVM. The experimental analysis described in section 4 showed that this leads to poor performances for first type error constrained classification. In order to improve classification performances for such problems, a new approach has been proposed to tune SVM parameters. It consists in determining the optimal values of these parameters minimizing total error P_e for a given value of first type error P_{21}. This approach allows to improve performances, but remains lower than theoretical results. To go further, an other solution taking into account the classification constraint when optimizing the SVM parameters, has to be found. That will be the subject of future works.

REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Patern Rcognition*, Academic Press, New York, 2 edition, 1990.

[2] C.K. Chow, "On Optimum Recognition Error and Reject Tradeoff", *IEEE Transactions on Information Theory*, Vol. IT-16, N.1, pp. 41-46, January 1970.

[3] B. Dubuisson and M. Masson, "A statistical decision rule with incomplete knowledge about classes", *Pattern Recognition*, 26(1):155-165, 1993.

[4] G. Fumera, F. Roli , and G. Giacinto, "Reject Option with Multiple Thresholds", *Pattern Recognition*, vol. 33, no. 12, pp. 2099-2101, 2000.

[5] A. Bounsiar, P. Beauseroy and E. Grall, "Etude de la classification binaire avec different sénarios de contraintes", internal report, University of Technology of Troyes, Troyes, 2004.

[6] V. Vapnik. *The Nature of Statistical Learning Theory*. Spring Verlag, 1995.

[7] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[8] N. Cristianini, and J. Shawe-Taylor. *Support vector machines and other kernel-based learning methods*. Combridge University Press, 2000.

[9] C. J. C. Burges. "A tutorial on support vector machines for pattern recognition". *Data Mining and Knowledge Discovery* , 2(2):121-167, 1998.

[10] C. Cortes and V. Vapnik. "Support Vector Networks". *Machine Learning*, 20:273-279, 1995.

[11] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukhrjee. "Choosing kernel parameters for support vector machines". *Machine Learning*, 2000.