1

Learning multiclass rules with class-selective rejection and performance constraints

Nisrine Jrad, Pierre Beauseroy and Edith Grall-Maës
Université de Technologie de Troyes ICD (FRE CNRS 2848), LM2S
France

1. Introduction

The task of classification occurs in a wide range of human activity. The problem concerns learning a decision rule that allows to assign a pattern to a decision option on the basis of observed attributes or features. Contexts in which a classification task is fundamental include, sorting letters on the basis of machine-read postcodes, the preliminary diagnosis of a patient's disease or the fraud currency and documents detection. In the classical framework, decision options are given by the pre-defined classes and a decision rule is designed by optimizing a given loss function, for instance the misclassification rate.

In some cases, the loss function should be more general.

First, for some applications, like face identification or cancer diagnosis, one may favor withholding decision instead of taking a wrong decision. In such cases, the introduction of rejection options should be considered in order to ensure a higher reliability Ha (1997); Horiuchi (1998); Jrad, Grall-Maës & Beauseroy (2008); Jrad et al. (2009d). Basic rejection consists of assigning a pattern to all classes which means that no decision is taken. More advanced rejection methods enable to assign a pattern ambiguously to a subset of classes. In this class-selective rejection scheme, decision options are given by the pre-defined classes as well as by defined subsets of different combinations among these classes. In order to define a decision rule, a general loss function can be defined by costs that penalize differently the wrong decisions and the ambiguous ones.

Some applications may require to control the performance of the decision rule or more specifically, the performance measured by indicators related to the decision rule. These latter could be formulated as the performance constraints. Hence, the decision problem should also take into account these constraints. A general formulation of this problem was proposed in Grall-Maës & Beauseroy (2009). The decision problem is formulated as an optimization problem with constraints. It was shown that the optimal rule can be obtained by optimizing its Lagrangian dual function which consists of finding the saddle point of this Lagrangian function. This optimal theoretical rule is applicable when the probability distributions are known. However, in many applications, only amounts of training set is available. Therefore, one should infer a classifier from a more or less limited set of training examples. In the classical decision framework, referred as the classical framework, many historical strands of research can be identified: statistical, Support Vector Machines, Neural Network Bishop (2006); Guobin & Lu (2007); Hao & Lin (2007); Husband & Lin (2002); Vapnik (1998); Yang et al. (2007)... In the class-selective rejection scheme, fewer works have been done Ha (1997); Horiuchi (1998).

One approach based on ν-1-SVM was proposed in Jrad, Grall-Maës & Beauseroy (2008) and tested on five cancer genes datasets in Jrad et al. (2009d). A cascade of classifiers with class-selective rejection learned on different feature sets was used as a good way to provide improved supervised diagnosis. In this chapter, multiclass problem is studied in the general framework of class-selective rejection subject to constraints. Two approaches are presented and discussed; a class-modeling approach and a boundary based approach.

The class-modeling approach is defined within the statistical community. It exploits flexible classes of models to provide an estimate of the joint distribution within each class, which in turn provides a classification rule. Estimators may be either parametric or not. In the parametric case, an additional hypothesis about the underlying probability density function should be made. To illustrate that approach, a parametric estimator, Gaussian Mixture Models (GMM) Titterington et al. (1985), and a non-parametric estimator, Parzen Windows estimator (PW) Emanuel (1962), are explored. The proposed approach Jrad et al. (2009c) consists of optimizing the class-conditional density estimates on the basis of a goodness of fit criterion. The GMM and PW densities are plugged into the hypothesis tests framework to get the decision rule associated to the estimates. The decision rule is selected by optimizing the Lagrangian function.

The boundary based approach is defined in the SVM community. It avoids the estimation of the complete density functions which is unnecessary since only densities in the neighborhood of borders need to be precisely known. A multiclass support vector machine algorithm (MSVM), based on ν-1-SVM, is used. The proposed method divides the multiple class problem into several unary classification problems and train one ν-1-SVM for each class Scholkopf et al. (2001); Scholkopf & Smola (2001); Tax (2001) coupled with its regularization path Hastie et al. (2004); Rakotomamonjy & Davy (2007). The winning class or subset of classes is determined using a prediction function that takes into consideration the different costs. The parameters of all the ν-1-SVMs are optimized jointly in order to minimize the Lagrangian function. Taking advantage of the regularization path method, the entire parameters searching space is considered. Compared to similar approaches Bottou et al. (1994); Hao & Lin (2007); Yang et al. (2007), since the searching space is widely extended, the selected decision rule is more likely to be the optimal one. Note that standard multiclass learning strategy is a particular case of the proposed approaches where the different decision options are given by the pre-defined classes, the loss function is given by the error rate and no constraint is considered. We will refer to this case as the classical framework.

The class-conditional approach and the boundary based approach were applied to several artificial datasets or toy problems. The datasets were constructed such that they differ in modal complexity and sample size. By using toy problems, the characteristics of the datasets can be exactly set and the performances of the supervised rule can be deduced by a simple comparison to the theoretical ones. General comments about the behavior of the different approaches can be made by analyzing these results. As a final example, the boundary approach is tested on five well-known cancer genes data sets, LEUKEMIA72 Golub et al. (1999), OVARIAN Welsh et al. (2001), NCI Ross et al. (2000); Scherf et al. (2000), LUNG CANCER Garber et al. (2001) and LYMPHOMA Alizadeh et al. (2000) in order to study the performance of this approach on real world datasets.

This chapter is outlined as follows. Paragraph 2 introduces the general framework of multiclass problems with class-selective rejection and performance constraints. It presents the optimal solution in the statistical theory framework. After describing the problem, we turn to an exploration of a supervised solution in the class-modeling framework by exploiting the non-parametric Parzen Windows estimator and the parametric Gaussian Mixture Models in paragraph 3. The boundary strategy based on ν-1-SVM is presented in paragraph 4. The efficiency of the latter approach is illustrated through a supervised cancer diagnosis. Results on the five genes datasets are reported in paragraph 5. The last paragraph discusses and compares the class-conditional and boundary based approaches.

2. Classification with class selective rejection and performance constraints

This section addresses the problem of multiclass decision with class-selective rejection and performance constraints. It gives a general framework for specifying such a problem. The optimal solution is presented in the statistical hypothesis testing framework.

2.1 Multiclass problem

Let us consider a multiclass decision problem with N classes. A given pattern $x \in \Re^d$ belongs to the class j noted w_j, for $j = 1, \ldots, N$, with the class-conditional probability density function $P(x/w_j)$. Each class is characterized by its a priori probability $P_j = P(w_j)$. The unconditional probability function (mixture density) $P(x)$ and the posterior probabilities $P(w_j/x)$ are provided through the total probability theorem and Bayes' formula.

The proposed general framework, introduced in Grall-Maës et al. (2006a) and developed in Grall-Maës & Beauseroy (2009), allows to define a multiclass decision problem subject to performance constraints using three kinds of criteria:

- the decision options: they correspond to the assignment subsets of classes that are deemed as admissible for the problem. In the class-selective rejection scheme, there are $2^N - 1$ assignment subsets of classes. They correspond to the possible subsets in a set of N elements excluding the empty set. They can be referred to ψ_i with $i = 1, \ldots, 2^N - 1$. For example, assigning a pattern to $\psi_i = \{1; 3\}$ means that it is assigned to both classes w_1 and w_3 with ambiguity.

  Thus, the decision options are defined by the set Ψ composed of only the admissible subsets of classes ψ_i:

  $$\Psi = \{\psi_1, \psi_2, \ldots, \psi_I\},$$

  where $I \leq 2^N - 1$ is the number of decision options. Any decision rule $Z : \Re^d \to [1, 2, \ldots, I]$ is defined such that $Z(x) = i$ when x is assigned to the set ψ_i.

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

  $$P(D_i/w_j) = \int_{\{x|Z(x)=i\}} P(x/w_j)dx.$$

- the performance constraints to be satisfied. They are defined by inequalities, each of them defining a threshold on a linear combination of class conditional decision probabilities. Any performance constraint $C^{(k)}$ where k is a integer between 1 and the number of constraints K is defined by its expression $e^{(k)}(Z)$ and its threshold $\gamma^{(k)}$:

  $$C^{(k)} : e^{(k)}(Z) = \sum_{i=1}^{I} \sum_{j=1}^{N} \alpha_{i,j}^{(k)} P_j P(D_i/w_j) \leq \gamma^{(k)} \quad (1)$$

  with $e^{(k)}(Z)$ a linear combination of class conditional decision probabilities, $\alpha_{i,j}(k)$, for $i = 1, \ldots, I$ and $j = 1, \ldots, N$, is the cost of deciding that an element x belongs to the set ψ_i when it is assigned to the class j, in the expression of the kth constraint.

- the average expected loss: it corresponds to the cost function to be minimized. It is also expressed as a linear combination of class-conditional decision probabilities:

$$\overline{c}(Z) = \sum_{i=1}^{I} \sum_{j=1}^{N} c_{i,j} P_j P(D_i/w_j), \tag{2}$$

  where $c_{i,j}$, for $i = 1, \ldots, I$ and $j = 1, \ldots, N$, is the cost of deciding to assign an element x to the set ψ_i when it belongs to the class j. The values of $c_{i,j}$ are relative since the aim is to minimize $\overline{c}(Z)$, thus, without loss of generality, the values are defined in the interval $[0; 1]$.

In this framework, finding the optimal decision rule consists in determining the decision rule Z^* so that the cost $\overline{c}$ is minimum and the constraints given by equation (1) are satisfied. The decision problem to be solved is expressed by the following optimization problem:

$$\begin{gathered} \min_Z \overline{c}(Z) \\ \text{s.t. } e^{(k)}(Z) \leq \gamma^{(k)} \quad \forall k = 1, \ldots, K. \end{gathered}$$

Given this primal problem, the Lagrangian dual problem is defined by:

$$\max_{\boldsymbol{\mu} \in \Re^{K+}} \{\min_Z \{L(Z, \boldsymbol{\mu})\}\} \tag{3}$$

in which $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_K]^T$ is the vector of Lagrangian multipliers associated with the constraints and $\boldsymbol{\gamma} = [\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(K)}]^T$ is the vector of the constraint thresholds and

$$\begin{aligned} L(Z, \boldsymbol{\mu}) =& \overline{c}(Z) + \sum_{k=1}^{K} \mu_k (e^{(k)}(Z) - \gamma^{(k)}) \\ =& \sum_{i=1}^{I} \sum_{j=1}^{N} \left(c_{i,j} + \sum_{k=1}^{K} \mu_k \alpha_{i,j}^{(k)} \right) P_j P(D_i/w_j) - \sum_{k=1}^{K} \mu_k \gamma^{(k)} \end{aligned} \tag{4}$$

2.2 Theoretical optimal decision rule

According to Grall-Maës & Beauseroy (2009) the optimal objective values of the primal and dual problems are equal. Thus, solving the dual problem provides the optimal decision rule. The Lagrangian $L(Z, \boldsymbol{\mu})$ can be rewritten as:

$$L(Z, \boldsymbol{\mu}) = \sum_{i=1}^{I} \int_{\{x|Z(x)=i\}} \lambda_i(x, \boldsymbol{\mu}) dx - \boldsymbol{\mu}^T \boldsymbol{\gamma} \tag{5}$$

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

$$\lambda_i(x, \boldsymbol{\mu}) = \sum_{j=1}^{N} P_j P(x/w_j) \left(c_{i,j} + \boldsymbol{\mu}^T \boldsymbol{\alpha}_{i,j} \right)$$

with $\boldsymbol{\alpha}_{i,j} = [\alpha_{i,j}^{(1)}, \alpha_{i,j}^{(2)}, \ldots, \alpha_{i,j}^{(K)}]^T$. For a given $\boldsymbol{\mu}$ the minimum value of $L(Z, \boldsymbol{\mu})$ is obtained when the integrated expression is minimum, that is by choosing the decision rule $\widetilde{Z}_{\boldsymbol{\mu}}$ so that:

$$\widetilde{Z}_{\boldsymbol{\mu}}(x) = i \quad \text{if} \quad \lambda_i(x, \boldsymbol{\mu}) < \lambda_l(x, \boldsymbol{\mu}), \quad \forall i = 1, \ldots, I, \quad l = 1, \ldots, I, \quad l \neq i. \tag{6}$$

The solution of the dual problem (3), which defines the optimal decision rule Z^* is obtained with $\boldsymbol{\mu}^*$ that maximizes $L(\widetilde{Z}_{\boldsymbol{\mu}}, \boldsymbol{\mu})$ as follows:

$$Z^* = \widetilde{Z}_{\boldsymbol{\mu}^*} \quad \text{with} \quad \boldsymbol{\mu}^* \quad \text{given by} \quad \boldsymbol{\mu}^* = arg \max_{\boldsymbol{\mu} \in \Re^{K+}} L(\widetilde{Z}_{\boldsymbol{\mu}}, \boldsymbol{\mu})$$

Note that the same decision rule could be obtained by minimizing a modified loss function:

$$c_{modif}(Z) = \overline{c}(Z) + \sum_{k=1}^{K} \mu_k^*(e^{(k)}(Z) - \gamma^{(k)}) = L(Z, \boldsymbol{\mu}^*) \tag{7}$$

If $\boldsymbol{\mu}^*$ is known, the optimal rule Z^* is obtained by minimizing $\overline{c}(Z) + \sum_{k=1}^{K} \mu_k^* e^{(k)}(Z)$.
A particular case of the stated rule can be given by the non constrained rule for which decision options are given by the admissible classes and the loss function is defined by the probability of error. We will refer to this case as the classical case.
In the supervised learning framework, P_j and $P(D_i/w_j)$ are unknown, the process is described by a training sample set. To tackle the decision problem in that case, two approaches are developed and tested in the following: one is based on density estimation and the other based on boundary methods. For each of the two approaches, different models can be designed. In the following, we will present and discuss some of these methods and their characteristics.

3. Class-modeling approach

The most straightforward method to obtain a multiclass rule is to estimate the density of the training data and exploit them within the decision theory framework to provide a classification rule. We will refer to this approach as class-modeling approach. It exploits flexible classes of models which attempt to provide an estimate of the joint distribution of the features within each class. The decision rule is determined by plugging these estimations in the statistical hypothesis framework and solving the latter optimization problem (3). Many estimators may be used Bishop (2006).
In this section we consider that the data is described by a set of observations $\{x_1, \ldots, x_n\}$ drawn from a given class w. Two estimators, the non-parametric Parzen Windows estimator PW Emanuel (1962) and the parametric Gaussian Mixture Models GMM Titterington et al. (1985) are investigated. Simulations on artificial datasets were carried out to study the efficiency of the proposed method and discuss its sensibility to estimator choice. In the coming subsections PW and GMM estimators will be presented, followed by the description of the supervised learning method. Then two different data sets are introduced. They show very specific features in order to illustrate the advantages and disadvantages of the proposed methods.

3.1 Parzen Windows estimator

Let us suppose that n observations are being drawn from some unknown probability density $P(x)$ in some d-dimensional space, which we shall take to be Euclidean, and we wish to estimate the value of $P(x)$. Suppose that, for some particular applications, no specific hypothesis about the probability law can be made. In this case, a non-parametric density estimator should be used. Kernels or Parzen Windows estimator (PW) Emanuel (1962) is exploited in this study. It contains parameters that control the model complexity rather than the form of the distribution. Thus the density model is obtained by placing kernels over each data point and adding up the contributions over the whole dataset. Most often, Gaussian kernels are chosen with a mean m centered on the individual training objects ($m = x_p$) and diagonal covariance matrices $S = hI$, where h is the smoothness parameter of the windows width. This choice gives rise to the following density model:

$$\widehat{P}(x;h) = \frac{1}{n}\sum_{p=1}^{n} G(x;x_p,hI) \tag{8}$$

where G is a gaussian component given by:

$$G(x;m,S) = \frac{1}{(2\pi)^{\frac{d}{2}} \mid S \mid^{\frac{1}{2}}} \exp[-0.5(x-m)^T S^{-1}(x-m)] \tag{9}$$

in which x is a given pattern, m is the mean and S is the covariance matrix of the kernel. When the covariance matrix S is set equal to hI, the Parzen density estimator assumes equally weighted features. The estimation of the density only depends on one parameter h and on the sample set. Thus, training a Parzen density consists of the determination of a single parameter. The smoothness parameter gives a trade-off between sensitivity to noise at small h and oversmoothing at large h. The optimal width of the kernel h can be obtained by maximizing the likelihood function. Because only one parameter is estimated the data model is easy to learn.

3.2 Gaussian Mixture Models estimator

A mixture of Gaussians is a linear combination of normal distributions Titterington et al. (1985). The GMM are given by:

$$\widehat{P}_T(x;\theta_T) = \sum_{t=1}^{T} \pi_t G(x;m_t,S_t) \tag{10}$$

where $\theta_T = \{m_1,\ldots,m_T,S_1,\ldots,S_t,\pi_1,\ldots,\pi_T\}$ is the vector parameter of the Gaussian components of a given class, T is the number of components per class, $G(x;m_t,S_t)$ is the d-dimensional gaussian density given by equation (9) and π_t is the mixing weight of the t-th component satisfying:

$$\sum_{t=1}^{T} \pi_t = 1 \quad \text{and} \quad \pi_t \geq 0.$$

When the number of Gaussians T is defined beforehand by the user, the means m_t and covariances S_t of the individual Gaussian components can efficiently be estimated by an Expectation Maximization routine Dempster et al. (1977). The total number of free parameters in the mixture of Gaussians is $T(d + \frac{d(d+1)}{2} + 1)$.

When T is not defined a priori, the task is to estimate the parameters π_t, m_t, S_t and the number T of components that maximize the log-likelihood:

$$\mathcal{L}_T(x_1,\ldots,x_n;\theta_T) = \sum_{p=1}^{n} \log \widehat{P}_T(x_p;\theta_T).$$

The log-likelihood maximization can be carried out by the greedy EM algorithm based on the theoretical results of Li & Barron (1999). In this latter, Li and Barron show that the difference in Kullback-Leibler divergence achievable by T-component mixtures and the Kullback-Leibler distance achievable by any (possibly non-finite) mixture from the same family of components tends to zero with the rate c/T with c a constant dependant from the component family. Furthermore, this bound is reachable by employing the greedy procedure. Therefore, the maximum likelihood of the mixture can be determined by adding iteratively a new component to the mixture. In this chapter, the greedy EM Verbeek et al. (2003); Vlassis & Likas (2002) algorithm for learning GMM is used since it is able to find the global likelihood maxima and to estimate the unknown number of the mixture components. This algorithm can be summarized as follows.

- Starting from a 1-component mixture ($T = 1$), the optimal parameters are obtained by an EM procedure until convergence($| \mathcal{L}^{iteration}(x_1, \ldots, x_n; \theta_T) - \mathcal{L}^{iteration-1}(x_1, \ldots, x_n; \theta_T) | \leq \varepsilon$). Then, a search for a new component $G(x; m^*_{T+1}, S^*_{T+1})$ location and a corresponding weight π^*_{T+1} is performed in order to maximize the new log-likelihood:

$$\begin{aligned}\mathcal{L}_{T+1}(x_1, \ldots, x_n; \theta_{T+1}) &= \sum_{p=1}^{n} \log \widehat{P}_{T+1}(x_p; \theta_{T+1}) \\ &= \sum_{p=1}^{n} \log[(1 - \pi^*_{T+1})\widehat{P}_T(x_p; \theta_T) + \pi^*_{T+1} G(x_p; m^*_{T+1}, S^*_{T+1})]\end{aligned} \tag{11}$$

  with $\widehat{P}_T(x; \theta_T)$ remaining unchanged. It is obvious that the crucial step of this algorithm is the search of a new component location. It can be shown that $\mathcal{L}_{T+1}(x_1, \ldots, x_n; \theta_{T+1})$ is concave as function of π_{T+1} but can have multiple maxima as function of m^*_{T+1} and S^*_{T+1}. Hence, a global search is required.

  One way pointed in Vlassis & Likas (2002) proposes to use all the points as initial candidates of the sought component. Every point is the mean of a corresponding candidate ($m_{T+1} = x_p$) with the same covariance matrix $\sigma^2 I$, where σ is set according to Weston & Watkins (1999). For each candidate component, π_{T+1} is set to the mixing weight maximizing the second order Taylor approximation of $\mathcal{L}_{T+1}(x_1, \ldots, x_n; \theta_{T+1})$ around $\pi_{T+1} = 0.5$. The candidate yielding to the highest log-likelihood when added to $\widehat{P}_T(x; \theta_T)$ in (11) is selected and updated using EM until convergence. The new component is added to $\widehat{P}_T(x; \theta_T)$ and the research is repeated until reaching the maximum likelihood on a validation set.

- An improved version of this global search Verbeek et al. (2003) is used in this work. For each insertion problem, the proposed method constructs a fixed number of candidates per existing mixture component. Based on the posterior distributions, we partition the data set in T disjoint subsets A_t. For each set, C candidate components are constructed (for the following experiences $C = 10$ candidates). To generate new candidates from A_t, we pick uniformly random two data points x_l and x_r in A_t. Then, we partition A_t into two disjoint subsets A_{tl} and A_{tr}. For elements of A_{tl} the point x_l is closer than x_r and vice versa for A_{tr}. The mean and covariance of the sets A_{tl} and A_{tr} are used as parameters for two candidate components. The initial mixing weights for candidates generated from A_t are set to $\pi_t/2$. To obtain the next two candidates we draw new x_l and x_r, until the desired number of candidates is reached. After computing the log-likelihood of each of the CT candidates, we set the new component as the candidate that maximizes the log-likelihood $\widehat{P}_{T+1}(x; \theta_{T+1})$ when added to the existing mixture with its corresponding mixing weight.

3.3 Supervised decision rule

In the statistical decision theory framework, the determination of a multiclass rule that satisfies performance constraints consists in finding the optimal Z^* and the optimal Lagrange multipliers $\boldsymbol{\mu}^{\boldsymbol{*}}$ by solving the optimization problem defined in (3). However, in the supervised learning framework, P_j and $P(D_i/w_j)$ are unknown. One strategy to learn a supervised classifier is to estimate these probabilities and determine the corresponding optimal supervised rule $\widehat{Z}^*$ and Lagrange multipliers $\widehat{\boldsymbol{\mu}}^*$. In these experiments, we study the repercussion due to the estimation of $P(x/w_j)$ and we consider that P_j is known. The two estimators introduced above are used. The probability estimates depend on the labeled set and on the density estimators parameters, h of Parzen and T, $(m_t, S_t, \pi_t), t = 1, \ldots, T$ of the GMM. These parameters are determined by maximizing the log-likelihood of a validation set using 10-Cross Validation.

Each estimator produces its own optimal solution $\widehat{Z}^*$ and $\widehat{\boldsymbol{\mu}}^*$ which minimizes the corresponding loss function $\widehat{L}(\widehat{Z}^*, \widehat{\boldsymbol{\mu}}^*) = \widehat{\bar{c}}(\widehat{Z}^*) + \sum_{k=1}^{K} \widehat{\mu}_k^* \left(\widehat{e}^{(k)}(\widehat{Z}^*) - \gamma^{(k)}\right)$. To assess the quality of the supervised rules a criterion, proposed in Grall-Maës et al. (2006b) is used. It is given by $L(\widehat{Z}^*, \boldsymbol{\mu}^*) = \bar{c}(\widehat{Z}^*) + \sum_{k=1}^{K} \mu_k^* \left(e^{(k)}(\widehat{Z}^*) - \gamma^{(k)}\right)$. It has to be estimated on the true optimal lagrange multipliers $\boldsymbol{\mu}^*$ and an infinite test set (theoretical densities) in order to get the theoretical performance of the rule. The learning-testing procedures of the GMM and Parzen Windows estimator algorithm are as follow:

1. For each class w_j, estimating the GMM or the Parzen Windows distributions $\widehat{P}(x/w_j)$ using a training set and a validation set.
2. Learning the decision rule by solving the optimization problem (3) with the estimated $\widehat{P}(x/w_j)$, namely, finding the optimal supervised rule $\widehat{Z}^*$ and the optimal $\widehat{\boldsymbol{\mu}}^{\boldsymbol{*}}$.
3. Assessing the quality of the rule by computing the Lagrangian $\widehat{L}(\widehat{Z}^*, \boldsymbol{\mu})^*$ of the supervised rule $\widehat{Z}^*$ on an infinite test set using theoretical $\boldsymbol{\mu}^*$.

In a supervised framework, $\boldsymbol{\mu}^*$ and the infinite test set are unknown. A supervised procedure to compute this criterion was proposed in Jrad, Grall & Beauseroy (2008) and tested experimentally to show the validity and the relevance of this criterion.

To sum up, we frame the problem of classification with rejection option subject to constraints as a Bayesian inference problem. We formulate a model of how patterns are generated and then derive an algorithm for making optimal inferences under this model.

3.4 Toy problem

To evaluate the performances of the proposed supervised learning approaches, two 2-D problems with three equiprobable classes and performance constraints were considered. For both problems, GMM and PW estimators were used and compared. Synthetic data were constructed and used to investigate different characteristics of the methods. By using artificial data instead of real world data, we avoid focusing on unknown and unsuspected distributions. It gives the opportunity to just focus on some important aspects of data distributions. Experimental results are presented and discussed below.

The first problem is defined by three classes, each one is a 3-gaussian component distribution with unbalanced weights, leading to a trimodal distribution. The aim of this experiment is to study the case where the distributions correspond to the hypothesis of GMM. The second problem is given by three bivariate gamma distributions in order to study the case where

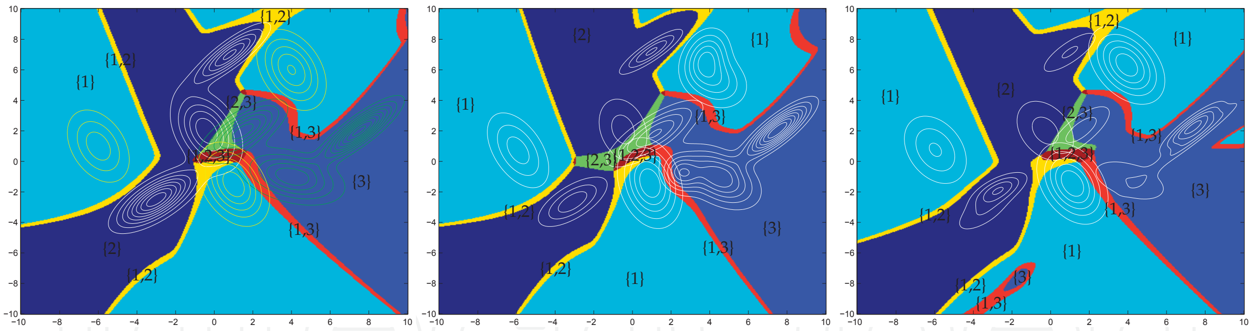


Fig. 1. 3-gaussian component problem: density probabilities and the corresponding partition. From left to right: Theoretical case (left) and an example of estimators in the case of Parzen (middle) and GMM (right)

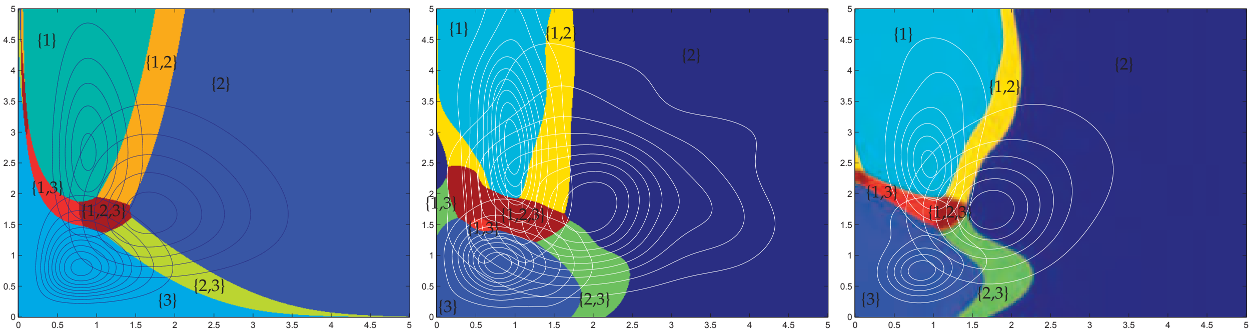


Fig. 2. Bivariate gamma problem: density probabilities and the corresponding partition. From left to right: Theoretical case (left) and an example of estimators in the case of Parzen (middle) and GMM (right)

the hypothesis of GMM is not fulfilled by data distributions. The corresponding theoretical densities are represented using isodensity curves in figures 1 (left) and 2 (left).

For both problems, the decision options are given by: $\psi_1 = \{1\}$, $\psi_2 = \{2\}$, $\psi_3 = \{3\}$, $\psi_4 = \{1,2\}$, $\psi_5 = \{1;3\}$, $\psi_6 = \{2;3\}$ and $\psi_7 = \{1;2;3\}$. The constraints are defined by $P_E \leq 0.05$ and $P_I \leq 0.1$ for the first problem and $P_E \leq 0.1$ and $P_I \leq 0.15$ for the second one. P_E is the probability of error and P_I is the probability of indistinctness, namely, the probability to assign a pattern x of w_j to a subset of classes ψ_i that contains more than one class of which one is w_j. The average loss is defined by $\overline{c}(Z) = P_E + 0.5P_I + P(D_7)$ where $P(D_7)$ corresponds to no decision. Theoretical decision rules are given by the partitions reported in figures 1 (left) and 2 (left).

For both experiments, the influence of the sample size is investigated by proceeding the learning-testing algorithm described above for three sets with different sizes (50, 100 and 200 observations per class). These sets were randomly drawn from the theoretical distributions. Experiments were repeated 40 times for the first set, 20 times for the second and 10 times for the third in order to study the bias and the variance of the results.

3.5 Experimental results and discussions

For both problems, P_E, P_I, $\overline{c}(Z^*)$ and $L(Z^*, \boldsymbol{\mu}^*)$ were computed for the optimal rule. Besides, their estimated values $\widehat{P}_E$, $\widehat{P}_I$, $\widehat{\overline{c}}(\widehat{Z}^*)$ and $\widehat{L}(\widehat{Z}^*, \boldsymbol{\mu}^*)$ were computed on supervised partitions using an infinite test set and the theoretical $\boldsymbol{\mu}^*$ as mentioned previously. Their mean and

	GMM			Parzen			Theo.
	50 obs	100 obs	200 obs	50 obs	100 obs	200 obs	
$\widehat{P}_E$	0.084±0.023	0.060±0.011	0.053±0.006	0.033±0.015	0.034±0.010	0.030±0.008	0.050
$\widehat{P}_I$	0.082±0.027	0.094±0.016	0.098±0.006	0.090±0.011	0.087±0.008	0.085±0.006	0.100
$\widehat{\overline{c}}$	0.161±0.040	0.133±0.017	0.136±0.018	0.248±0.039	0.212±0.028	0.207±0.029	0.131
$\widehat{L}$	0.205±0.033	0.147±0.012	0.140±0.011	0.224±0.024	0.189±0.015	0.177±0.017	0.132

Table 1. Values of the theoretical and estimated $\widehat{P}_E$, $\widehat{P}_I$, $\widehat{\overline{c}}(\widehat{Z}^*)$ and $\widehat{L}(\widehat{Z}^*, \boldsymbol{\mu}^*)$ using GMM and Parzen estimators for the 3-gaussian component problem

	GMM			Parzen			Theo.
	50 obs	100 obs	200 obs	50 obs	100 obs	200 obs	
$\widehat{P}_E$	0.114±0.021	0.110±0.010	0.109±0.008	0.083±0.028	0.079±0.018	0.086±0.010	0.100
$\widehat{P}_I$	0.143±0.023	0.143±0.011	0.143±0.012	0.147±0.019	0.148±0.012	0.147±0.010	0.150
$\widehat{\overline{c}}$	0.235±0.022	0.239±0.012	0.234±0.009	0.304±0.044	0.290±0.030	0.268±0.016	0.229
$\widehat{L}$	0.251±0.021	0.249±0.008	0.247±0.006	0.280±0.025	0.260±0.010	0.248±0.003	0.229

Table 2. Values of the theoretical and estimated $\widehat{P}_E$, $\widehat{P}_I$, $\widehat{\overline{c}}(\widehat{Z}^*)$ and $\widehat{L}(\widehat{Z}^*, \boldsymbol{\mu}^*)$ using GMM and Parzen estimators for the gamma distributions problem

their standard deviations are reported in tables 1 and 2. An example of optimal rules built with GMM and Parzen densities using 200 observations per class set are shown in figures 1 (middle) and 1 (right) (for the 3-gaussian component distributions problem) and 2 (middle) and 2 (right) (for the gamma distributions problem).

Results show that decision rules built with GMM and Parzen estimates are relevant. Their accuracy increases as long as the learning set size increases. Furthermore, GMM can be considered as a good family of non-symmetrical density estimators. They achieve results superior to Parzen estimators in term of losses, especially when the learning set size decreases. These results can be explained by several reasons:

i Parzen estimators converge asymptotically to the real densities.

ii Parzen density estimates have not a compact form; they are sums of as many local windows as the size of learning set, while the GMM estimates are compact functions parameterized according to a global search over all the learning set. Thus, the local nature of the Parzen estimator can lead to overfitting.

Moreover, for the first problem, the errors, as the function of the number of observations, of the GMM decrease faster than those of Parzen. It is an expected result since the distributions corresponds to the GMM hypothesis and GMM are more accurate fitting a multimodal distribution. Thus the choice of the estimator is a crucial point for class-modeling approaches. The estimator must fit the data pretty well to get good results. The performances of the classification rule are strictly related to those of the density estimator.

To sum up, we can note that when the sample size is sufficiently large and a flexible density model is used, this approach should work very well if the probability density estimator in use is convergent. Unfortunately, as the dimension of the representation space increases, it requires an exponentially increasing number of training samples to overcome the curse of dimensionality Duda & Hart (1973). Finding the right estimator to describe the dataset distribution and the given sample size is a typical incarnation of the bias-variance dilemma.

When a good probability model is assumed (one for which the bias is small) and the sample size is sufficient, this approach has some advantages. Since probability estimates of the patterns are computed explicitly, integration with other potential classes or constraints, not necessarily considered at design time, is facilitated. Besides, progress towards more complicated applications like solving classification problems with time-evolutionary constraints Jrad et al. (2009a) can be also facilitated.

4. Boundary approach

Vapnik argued in Vapnik (1998) that when just a limited amount of data is available, one should avoid solving a more general problem as an intermediate step to solve the original problem. To solve this more general problem more data might be required than for the original problem. In a bayesian framework minimizing the error rate loss, estimating a complete data density for each of the N classes might be too demanding when only the data boundary is required. Therefore, only a boundary between or around the dataset is determined. In the general framework of classe-selective rejection with constraints, a complete data density estimation is also not required. However, one should take into consideration the conditional probabilities estimation in order to evaluate the performance.

In this work, ν-1-SVM will be used. The proposed method is based on the "decomposition-reconstruction" approach. It decomposes the initial problem into N problems and trains one ν-1-SVM Scholkopf et al. (2001); Scholkopf & Smola (2001); Tax (2001) coupled with the regularization path of each class Hastie et al. (2004); Rakotomamonjy & Davy (2007). A reconstruction step is required to decide the winning decision and consequently to derive a decision rule that satisfy the constraints. It uses weighted distance measure between objects and classes.

In the coming subsections we will present the ν-1-SVM concept and explains briefly the derivation of the entire regularization path which enables to get rapidly all ν-1-SVM models for a wide range of values of ν. The proposed method is described and validated on two artificial datasets.

4.1 ν-1-SVM

Considering a set of n vectors $X = \{x_1, x_2, \ldots, x_n\}$ drawn from an input space $\mathcal{X}$, ν-1-SVM computes a function $f_X^\lambda(.)$ and a real number b^λ in order to determine the region $\mathcal{R}^\lambda$ in $\mathcal{X}$ such that:

$$\begin{cases} f_X^\lambda(x) - b^\lambda \geq 0 & \text{if } x \in \mathcal{R}^\lambda \\ f_X^\lambda(x) - b^\lambda < 0 & \text{otherwise} \end{cases}$$

The function $f_X^\lambda(.)$ is designed by minimizing the volume of $\mathcal{R}^\lambda$ under the constraint that all the vectors of X, except a portion λ, must lie in $\mathcal{R}^\lambda$. This portion corresponds to the outliers and parameterizes the function $f_X^\lambda(.)$. An alternative parameter can be refereed to $\nu = \frac{\lambda}{n}$. It corresponds to the fraction of outliers with $0 \leq \nu \leq 1$.

In order to determine $\mathcal{R}^\lambda$, the space of possible functions $f_X^\lambda(.)$ is reduced to a Reproducing Kernel Hilbert Space (RKHS) with kernel function $K(.,.)$. Let $\phi : \mathcal{X} \rightarrow \mathcal{H}$ be the mapping defined over the input space $\mathcal{X}$. Let $< .,. >_{\mathcal{H}}$ be a dot product defined in $\mathcal{H}$. The kernel $K(.,.)$ over $\mathcal{X} \times \mathcal{X}$ is defined by:

$$\forall (x_p, x_q) \in \mathcal{X} \times \mathcal{X} \quad K(x_p, x_q) =< \phi(x_p), \phi(x_q) >_{\mathcal{H}}$$

Without loss of generality, $K(.,.)$ is supposed normalized such that for any $x \in \mathcal{X}$, $K(x,x) = 1$. Thus, all the mapped vectors $\phi(x_p)$, $p = 1, \ldots, n$ are in a subset of a hypersphere with radius

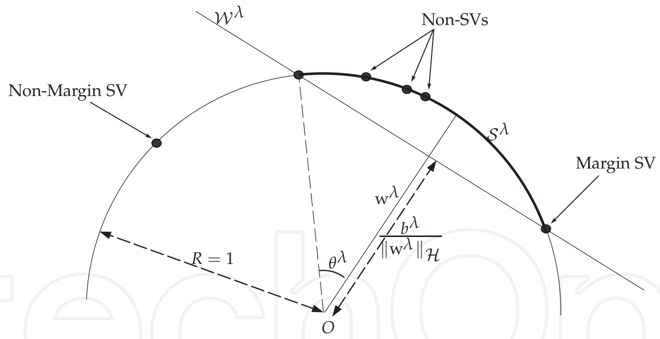


Fig. 3. Training data mapped into the feature space on a portion $\mathcal{S}^\lambda$ of a hypersphere.

one and center O. Provided $K(.,.)$ is always positive, $\phi(X)$ is a subset of the positive orthant of the hypersphere. A common choice of $K(.,.)$ is the Gaussian Radial Basis Function (RBF) kernel $K(x_p, x_q) = \exp[\frac{-1}{2\sigma^2} \parallel x_p - x_q \parallel^2_{\mathcal{X}}]$ with σ the parameter of the Gaussian RBF kernel.
ν-1-SVM consists of separating the training vectors in $\mathcal{H}$ from the center O with a hyperplane $\mathcal{W}^\lambda$ while maximizing the margin $\frac{b^\lambda}{\|\mathrm{w}^\lambda\|}_{\mathcal{H}}$ with w^λ the normal vector of $\mathcal{W}^\lambda$. The solution will be given by the function $f^\lambda_{\mathrm{X}}(.)$ such that $f^\lambda_{\mathrm{X}}(x) - b^\lambda =< \mathrm{w}^\lambda, \phi(x) >_{\mathcal{H}} -b^\lambda \geq 0$ for the $(1-\nu)n$ mapped training vectors.
This yields $f^\lambda_{\mathrm{X}}(.)$ as the solution of the following convex quadratic optimization problem:

$$\begin{aligned} &\min_{\mathrm{w}^\lambda, b^\lambda, \xi_p} \sum_{p=1}^{n} \xi_p - \lambda b^\lambda + \frac{\lambda}{2} \parallel \mathrm{w}^\lambda \parallel^2_{\mathcal{H}} \\ &\text{subject to} \quad < \mathrm{w}^\lambda, \phi(x_p) >_{\mathcal{H}} \geq b^\lambda - \xi_p \\ &\text{and} \quad \xi_p \geq 0 \quad \forall p = 1, \ldots, n \end{aligned} \tag{12}$$

where ξ_p are the slack variables. This optimization problem is solved by introducing Lagrange multipliers α_p. As a consequence to Kuhn-Tucker conditions, w^λ is given by:

$$\mathrm{w}^\lambda = \frac{1}{\lambda} \sum_{p=1}^{n} \alpha_p \phi(x_p)$$

which results in:

$$f^\lambda_{\mathrm{X}}(.) - b^\lambda = \frac{1}{\lambda} \sum_{p=1}^{n} \alpha_p K(x_p, .) - b^\lambda.$$

The dual formulation of (12) is obtained by introducing Lagrange multipliers as:

$$\min_{\alpha_1, \ldots, \alpha_n} \frac{1}{2\lambda} \sum_{p=1}^{n} \sum_{q=1}^{n} \alpha^\lambda_p \alpha^\lambda_q K(x_p, x_q) \tag{13}$$

$$\text{with} \quad \sum_{p=1}^{n} \alpha^\lambda_p = \lambda \quad \text{and} \quad 0 \leq \alpha^\lambda_p \leq 1 \quad \forall p = 1, \ldots, n$$

A geometrical interpretation of the solution in the RKHS is given by figure 3. The function $f^\lambda_{\mathrm{X}}(.)$ and the number b^λ define a hyperplane $\mathcal{W}^\lambda$ orthogonal to w^λ. The hyperplane $\mathcal{W}^\lambda$

separates the $\phi(x_p)$s from the sphere center, while having $\frac{b^\lambda}{\|\mathrm{w}^\lambda\|_{\mathcal{H}}}$ maximum which is equivalent to minimize the portion $\mathcal{S}^\lambda$ of the hypersphere bounded by $\mathcal{W}^\lambda$ that contains the set $\{\phi(x) \quad \text{s.t.} \quad x \in \mathcal{R}^\lambda\}$.

4.2 Regularization Path

Regularization path was first introduced by Hastie et al. (2004) for a binary SVM. Later, Rakotomamonjy & Davy (2007) developed the entire regularization path for a ν-1-SVM. The basic idea of the ν-1-SVM regularization path is that the Lagrange multipliers of a ν-1-SVM is a piecewise linear function of λ. Thus the principle of the method is to start with large λ $(ie. \quad \lambda = n - \epsilon)$ and decrease it towards zero, keeping track of breaks that occur as λ varies.

As λ decreases $\| \mathrm{w}^\lambda \|_{\mathcal{H}}$ increases and hence the distance between the sphere center and $\mathcal{W}^\lambda$ decreases. Points move from being outside (Non-Margin SVs with $\alpha_p^\lambda = 1$ in figure 3) to inside the portion $\mathcal{S}^\lambda$ (Non-SVs with $\alpha_p^\lambda = 0$). By continuity, points must linger on the hyperplane $\mathcal{W}^\lambda$ (Margin SVs with $0 < \alpha_p^\lambda < 1$) while their α_p^λs decrease from 1 to 0. Break points occur when a point moves from a position to another one. Since α_p^λ is piecewise-linear in λ, $f^\lambda(.)$ and b^λ are also piecewise-linear in λ. Thus, after initializing the regularization path (computing α_p^λ by solving (13) for $\lambda = n - \epsilon$), almost all the α_p^λs are computed by solving linear systems. Only for some few integer values of λ smaller than n, α_p^λs are computed by solving (13) according to Rakotomamonjy & Davy (2007).

Using simple linear interpolation, this algorithm enables to determine very rapidly the ν-1-SVM corresponding to any value of λ.

4.3 Supervised decision rule

Given N classes and N trained ν-1-SVMs, one should design a supervised decision rule Z that minimizes the loss function (2) and satisfies the constraints (1). The reconstruction step consists of moving from unary to multiclass classifier by assigning samples to a decision option. The assignment condition (6) can not be used since the distributions $P(x/w_j)$ are unknown. The reconstruction step relies on a distance of an unlabelled pattern x to each of the training class set w_j $(j = 1, \ldots, N)$, using the ν-1-SVM parameterized by λ_j, is defined as follows:

$$d^{\lambda_j}(x) = \frac{\cos(\widehat{\mathrm{w}^{\lambda_j}, \phi(x)})}{\cos(\theta^{\lambda_j})} = \frac{\| \mathrm{w}^{\lambda_j} \|_{\mathcal{H}}}{b^{\lambda_j}} \cos(\widehat{\mathrm{w}^{\lambda_j}, \phi(x)}) \tag{14}$$

where θ^{λ_j} is the angle delimited by w^{λ_j} and the support vector as shown in figure 3, $\cos(\theta^{\lambda_j})$ is a normalizing factor which is used to normalize all the $d_j^\lambda(x)$.

Using $\| \phi(x) \| = 1$ in (14) leads to the following:

$$d^{\lambda_j}(x) = \frac{< \mathrm{w}^{\lambda_j}, \phi(x) >_{\mathcal{H}}}{b^{\lambda_j}} = \frac{\frac{1}{\lambda_j} \sum_{p=1}^{n_j} \alpha_p^{\lambda_j} K(x_p, x)}{b^{\lambda_j}} \tag{15}$$

Since the lagrange multipliers $\alpha_p^{\lambda_j}$ are obtained by the regularization path for any value of λ_j, computing d^{λ_j} is considered as an easy-fast task. The distance measure $d^{\lambda_j}(x)$ is inspired from Davy et al. (2006). When data are distributed in a unimodal form, the $d^{\lambda_j}(x)$ is a decreasing function with respect to the distance between a sample x and the data mean. The probability density function is also a decreasing function with respect to the distance from the mean. Thus, $d^{\lambda_j}(x)$ preserves distribution order relations. In such case, and under optimality of the ν-1-SVM classifier, the use of $d^{\lambda_j}(x)$ should reach the same performances as the one obtained using the distribution.

In the simplest case of multiclass problems where the loss function is defined as the error probability, a sample x is assigned to the class maximizing $d^{\lambda_j}(x)$ as follows:

$$x \in \arg \max_{j=1\ldots N} \quad d^{\lambda_j}(x).$$

To extend the multiclass prediction process to the class-selective scheme, a weighted form of the distance measure is proposed. A weight β_j is associated to the distance d^{λ_j} to pull the location of a pattern toward the class for which a wrong decision costs the most. Thus, introducing weights performs a distance adjustment and helps solving problems with different costs c_{ij} on the classification decisions. The decision rule is defined as:

$$\widehat{Z}(x) = i \quad \text{if} \quad \sum_{j=1}^{N} c_{ij}\widehat{P}_j\beta_j d^{\lambda_j}(x) \leq \sum_{j=1}^{N} c_{lj}\widehat{P}_j\beta_j d^{\lambda_j}(x), \forall i, l = 1, \ldots, I, l \neq i.$$

where $\widehat{P}_j$ is the empirical estimators of P_j.

The decision rule depends on the RBF vector parameter $\boldsymbol{\sigma}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ vectors of σ_j, λ_j and β_j for $j = 1, \ldots, N$. Tuning λ_j is the most time expensive task since changing λ_j leads to solve the optimization problem formulated in (13). Moreover, tuning λ_j is a crucial point, it enables to control the boundary around data. In fact, it was shown in Scholkopf et al. (2001) that this regularization parameter is an upper bound on the fraction of outliers and a lower bound on the fraction of the SVs. In Husband & Lin (2002); Yang et al. (2007) a smooth grid search was supplied in order to choose the optimal values of $\boldsymbol{\lambda}$. The N values λ_js were chosen equal to reduce the computational costs. However, this assumption reduces the search space of parameters too. To avoid this restriction, the proposed approach optimizes all the λ_j with $j = 1, \ldots, N$ corresponding to the N ν-1-SVMs using regularization path and consequently explores the entire parameters space. Thus the tuned λ_j are most likely to be the optimal ones. The parameter $\boldsymbol{\sigma}$ are set equals $\sigma_1 = \sigma_2 = \ldots = \sigma_N$.

In the general framework of class-selective rejection subject to constraints, the decision rule for a given $\boldsymbol{\mu}$ is given by:

$$\widehat{\widehat{Z}}_{\boldsymbol{\mu}}(x) = i \quad \text{if} \tag{16}$$
$$\sum_{j=1}^{N}\left(c_{i,j} + \sum_{k=1}^{K}\mu_k\alpha_{i,j}^{(k)}\right)\widehat{P}_j\beta_j d^{\lambda_j}(x) \leq \sum_{j=1}^{N}\left(c_{l,j} + \sum_{k=1}^{K}\mu_k\alpha_{l,j}^{(k)}\right)\widehat{P}_j\beta_j d^{\lambda_j}(x), \forall i, l = 1, \ldots, I, l \neq i.$$

Since the problem is described by a sample set, an estimate $\widehat{L}(\widehat{Z}, \widehat{\boldsymbol{\mu}})$ of $L(Z, \boldsymbol{\mu})$ given by (1) is used:

$$\widehat{L}(\widehat{Z}, \widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{I}\sum_{j=1}^{N}\left(c_{i,j} + \sum_{k=1}^{K}\widehat{\mu}_k\alpha_{i,j}^{(k)}\right)\widehat{P}_j\widehat{P}(D_i/w_j) - \widehat{\boldsymbol{\mu}}^T\boldsymbol{\gamma} \tag{17}$$

where $\widehat{P}(D_i/w_j)$ is the empirical estimators of $P(D_i/w_j)$.

The parameters λ_j, β_j, σ_j and μ_k are optimized so that the estimated Lagrange function $\widehat{L}(\widehat{Z}, \widehat{\boldsymbol{\mu}})$ is optimum on a validation set. This is accomplished by employing an iterative search over

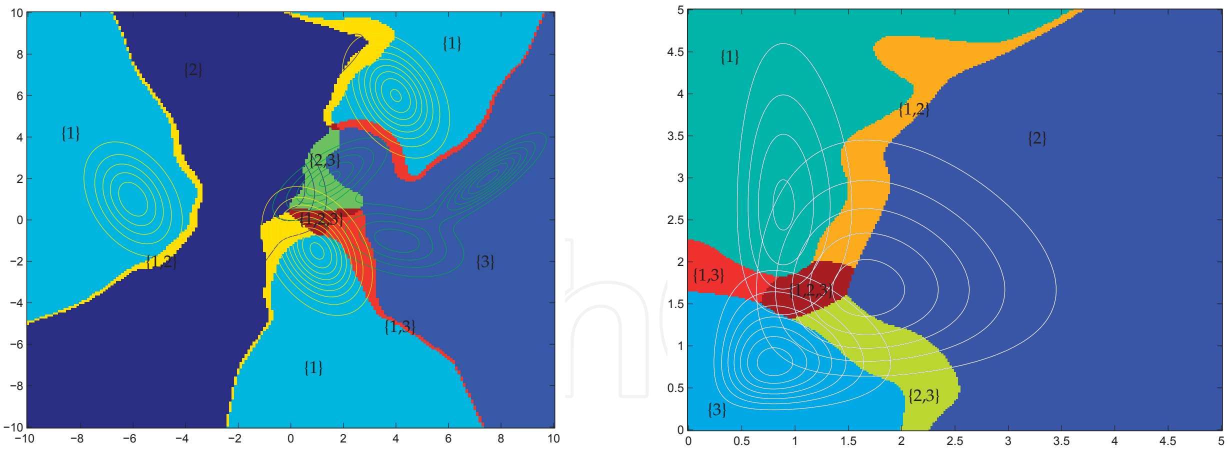


Fig. 4. 3-gaussian component (left) and Bivariate gamma (right) problems: supervised partitions and theoretical density probabilities.

	3-gaussian			Gamma		
	50 obs	100 obs	200 obs	50 obs	100 obs	200 obs
$\widehat{P}_E$	0.076±0.013	0.061±0.008	0.050±0.005	0.101±0.012	0.094±0.014	0.088±0.005
$\widehat{P}_I$	0.138±0.005	0.121±0.002	0.107±0.001	0.176±0.027	0.179±0.033	0.173±0.013
$\widehat{c}$	0.163±0.017	0.153±0.008	0.149±0.004	0.267±0.020	0.262±0.018	0.258±0.011
$\widehat{L}$	0.206±0.028	0.165±0.010	0.151±0.006	0.275±0.015	0.271±0.032	0.247±0.006

Table 3. Values of the estimated $\widehat{P}_E$, $\widehat{P}_I$, $\widehat{c}(\widehat{Z}^*)$ and $\widehat{L}(\widehat{Z}^*, \boldsymbol{\mu}^*)$ using the boundary method for the 3-gaussian component and the gamma distributions problems

the kernel parameter and a global search over the other ones. More explicitly, the kernel parameters are chosen from a previously defined set of real numbers $[\sigma_0, \ldots, \sigma_s]$ with $s \in \aleph$. For each given value of σ_j, a decision rule is sought by solving an alternate optimization problem over λ_j, β_j and μ_k. The optimal rule is given by the set of parameters minimizing the Lagrange estimate on a validation set.

4.4 Toy problem

To evaluate the efficiency of the proposed boundary approach and compare it with the class-modeling one, the same bidimensional toy problems considered in section 3 were studied under the same constraints. The same synthetic data with the same number of repetitions were considered. Thus theoretical densities and theoretical decision rules are those illustrated in figures 1 (left) and 2 (left). The theoretical performances are those reported on tables 1 and 2.

Supervised decision rule is optimized according to the supervised learning algorithm defined in the boundary scheme. The estimated values $\widehat{P}_E$, $\widehat{P}_I$, $\widehat{c}(\widehat{Z}^*)$ and $\widehat{L}(\widehat{Z}^*, \boldsymbol{\mu}^*)$ were computed on supervised partitions using an infinite test set and the theoretical $\boldsymbol{\mu}^*$ as mentioned previously. Their mean and their standard deviations are reported in tables 3. An example of optimal rules learned with 200 observations per class set are shown in figures 4 (left) for the 3-gaussian distribution problem and 4 (right) for the gamma distribution problem.

Experimental results show that the proposed boundary method or MSVM is relevant. It achieves good results on some complex distributions like multimodal and non-symmetrical

Dataset	LEUKEMIA72	OVARIAN	NCI	LUNG CANCER	LYMPHOMA
# Gene	6817	7129	9703	918	4026
# Sample	72	39	60	73	96
# Class	3	3	9	7	9

Table 4. Multiclass gene expression datasets

distributions. For both problems under study, standard deviations are relatively small. The boundary approach is less sensitive to the variation of the dataset representativity. For the 3-gaussian component problem, MSVM performs as good as the GMM based method and better than PW based method. For the gamma distribution problem, MSVM, PW and GMM based methods show similar accuracy for sufficiency large data sets, while GMM based method outperforms PW and MSVM in term of losses for moderated or small amount of data.

5. Cancer diagnosis

To illustrate the proposed approach a biomedical application dealing with cancer tumors is presented using the boundary method. Recently, cancer diagnosis based on gene profiles has received more attention. Since cancer diagnosis problems are usually described by a small set of samples with a large number of genes, feature or gene selection was considered as an important issue in analyzing multiclass microarray data.

In this section, five well-known gene expression datasets are considered. Two experiments based on the boundary approach are presented. The first considers the five cancer diagnosis problems in the classical framework to make results comparable with those of Chen et al. (2005). The second experiment considers the LUNG CANCER problem in the general framework of class-selective rejection and performance constraints.

5.1 Problem description

In this chapter, five multiclass gene expression datasets are studied: LEUKEMIA72 Golub et al. (1999), OVARIAN Welsh et al. (2001), NCI Ross et al. (2000); Scherf et al. (2000), LUNG CANCER Garber et al. (2001) and LYMPHOMA Alizadeh et al. (2000). Table 4 describes the five genes datasets.

Given these microarray datas with N tumor classes, a small amount n of tumor samples and a large number g of genes per sample, one should identify a small subset of d informative genes that contribute most to the prediction task before solving this task. Various feature selection methods exist in literature. One way pointed in Chen et al. (2005) is to use test statistics.

For each dataset, six test statistics are evoked as a first process in a gene-based cancer diagnosis: ANOVA F or F Kutner et al. (2005), Brown-Forsythe test or B Brown & Forsythe (1974), Welch test WELCH (1951) or W, Adjusted Welch test or W^* Hartung & Makambi (2002), Cochran test Cochran (1937) or C and Kruskal-Wallis test or H Daniel (1999). For each test statistics, 50 and 100 informative genes were selected.

The second step is a classification step which is performed, in the classical framework, according to the proposed boundary approach and five existing ones: Naive Bayes, Nearest Neighbor, Linear Perceptron, Multilayer Perceptron Neural Network. The classification step is also studied in the general framework of class-selective rejection and performance constraints according to the proposed boundary approach.

		F	B	W	W*	C	H
LEUKEMIA	Proposed Algorithm	4	3	5	5	3	2
	Mean	3.4	2.4	2.8	2.8	3.2	3.0
	Median	3	2	3	3	3	3
OVARIAN	Proposed Algorithm	0	0	0	0	0	0
	Mean	0.2	0.0	0.0	0.0	0.0	0.0
	Median	0	0	0	0	0	0
NCI	Proposed Algorithm	31	26	27	27	27	33
	Mean	36.0	32.0	27.4	26.0	27.0	35.4
	Median	35	29	27	27	27	35
LUNG CANCER	Proposed Algorithm	14	16	16	16	16	15
	Mean	17.6	17.0	17.6	17.6	18.0	18.0
	Median	17	17	18	18	18	18
LYMPHOMA	Proposed Algorithm	18	16	9	10	9	15
	Mean	23.8	19.8	14.0	14.0	12.8	22.0
	Median	23	19	12	12	13	20

Table 5. Prediction errors of the proposed classifier, mean and median values of the 5 classifiers prediction errors according to Chen et al. (2005) with 50 informative selected genes

5.2 Experimental settings

The cancer diagnosis is accomplished using the classification algorithm introduced in section 4 in both classical and class-selective rejection subject to constraints frameworks. Results are reported in the following sections as a prediction error. Mean and median values of the prediction errors of the five classifiers mentioned above are also reported from Chen et al. (2005). The generalization accuracy of all the classifiers was computed using Leave One Out (LOO) resampling method. LOO divides a gene dataset of n patients into two sets, a set of $n-1$ patients and a test set of 1 blinded patient. This method involves n separate runs. For each run, the first set of $n-1$ are used to learn the rule and the test set of 1 blinded sample is used to assess the performance of the rule. The overall prediction error is the sum of the patients misclassified on all n runs. In the following, results are explored and discussed.

5.2.1 Classical framework

First, the cancer diagnosis problem is considered in the traditional Bayesian framework with no constraints. The decisions are given by the possible set of tumor classes and the loss function is defined as the probability of error to make results comparable with those of Chen et al. (2005). In this case, the costs of misclassification $c_{i,j}$ are known, equal and there is no penalty for a correct classification. The decision rule becomes the solution of a minimization problem without constraints (Lagrange multipliers are null). The performance of the proposed method was measured by evaluating its accuracy rate and it was compared to results obtained by the five predictors evoked in Chen et al. (2005): Naive Bayes, Nearest Neighbor, Linear Perceptron, Multilayer Perceptron Neural Network with five nodes in the middle layer, and Support Vector Machines with second order polynomial kernel.

The learning step of the proposed approach consists of finding the minimal value of the loss function estimate. The $n-1$ samples are divided, using 5-Cross Validation (5-CV), into a

		F	B	W	W*	C	H
LEUKEMIA	Proposed Algorithm	5	2	3	3	4	6
	Mean	3.4	3.0	3.0	3.0	3.2	3.0
	Median	3	3	4	3	3	3
OVARIAN	Proposed Algorithm	0	0	0	0	0	0
	Mean	0.2	0.0	0.0	0.0	0.0	0.0
	Median	0	0	0	0	0	0
NCI	Proposed Algorithm	33	21	26	25	26	36
	Mean	33.0	22.6	23.8	25.2	25.2	31.6
	Median	33	22	25	26	26	31
LUNG CANCER	Proposed Algorithm	11	10	11	11	11	13
	Mean	12.2	12.2	11.4	12.2	12.2	15.8
	Median	12	12	11	11	11	14
LYMPHOMA	Proposed Algorithm	16	16	11	10	11	17
	Mean	21.8	19.2	13.0	13.8	14.4	18.2
	Median	17	16	12	12	12	18

Table 6. Prediction errors of the proposed classifier, mean and median values of the 5 classifiers prediction errors according to Chen et al. (2005) with 100 informative selected genes

training set and a validation set. N ν-1-SVMs are trained using the training set for all values of ν_j. The decision is obtained by tuning the parameters β_j and λ_j for $j = 1, \ldots, N$ for a given kernel parameter $\boldsymbol{\sigma}$ and by testing different values of $\boldsymbol{\sigma}$ in the set $[2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}]$. Finally, the decision rule which minimizes the loss function estimate is selected and used to classify the blinded patient.

Table 5 reports the errors of the proposed algorithm, the average value and the median value of the 5 classifiers prediction errors reported in Chen et al. (2005) when 50 informative genes are used. Table 6 reports values when 100 informative genes are used. F, B, W, W^*, C and H represent the six test statistics.

Experimental results show that, for OVARIAN, NCI, LUNG CANCER and LYMPHOMA multiclass genes problems, the prediction error is data dependent. The proposed boundary approach achieves competitive performances compared to the 5 classifiers reported in Chen et al. (2005). For these datasets, prediction errors of the proposed approach are less than the mean and median values of the 5 classifiers prediction errors reported in Chen et al. (2005). However, for LEUKEMIA72, the proposed algorithm performances are almost in the same range of those provided by the 5 classifiers reported in Chen et al. (2005). The proposed approach prediction error is equal, or in the worst case, slightly higher than the mean and median errors. Focusing on the data nature, the five data under study, described by table 4 show that even though data are all described by a large number of genes and small number of samples they have different nature. Genes number varies from 918 for LUNG CANCER problem to 9703 for NCI. NCI has ten times genes more than LUNG CANCER with less number of patients and more classes. According to the table 4, NCI is the hardest problem since it is described by the largest number of genes with the smallest number of patients per class (60 patients for

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	6	0	0	0	0	0	0
	SCLC	0	4	0	0	0	1	0
	LCLC	0	0	3	0	0	4	1
	SCC	0	0	0	16	0	3	0
	AC2	0	0	0	0	4	0	0
	AC3	0	1	1	0	1	4	0
	AC1	0	0	1	0	2	1	20

Table 7. Confusion Matrix of $50W^*$ LUNG CANCER dataset. Total of misclassified is equal to 16.

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	5	0	0	0	0	0	0
	SCLC	0	4	0	0	0	0	0
	LCLC	0	0	1	1	0	2	2
	SCC	0	0	2	14	0	1	0
	AC2	0	0	0	0	7	0	0
	AC3	0	0	2	1	0	8	0
	AC1	1	1	0	0	0	2	19

Table 8. Confusion Matrix of $50H$ LUNG CANCER dataset. Total of misclassified is equal to 15.

9 classes). Thus, it is expected that the gene selection task is difficult and consequently the prediction accuracy is not high.

Moreover, it is worthy to note that these data are more or less imbalanced which makes the discrimination step harder. For example, the ratio of the large to the small classes reaches 23 for the LYMPHOMA problem. For this problem, the proposed boundary method results are considerably more accurate than the 5 existing ones. Thus, this approach can be considered as an adapted method solution to solve imbalanced problems: because both minority and majority classes are learned separately, descriptions are not dominated by the majority classes. Finally, we can note that focusing on the test statistics comparison, experimental results confirm those of Chen et al. (2005). B, W and W^* can be the most performing tests under variances heterogeneity assumptions. For the LUNG CANCER dataset where the gene-patient ratio is the smallest, the prediction accuracy is almost the same for all the test statistics.

5.2.2 Class-selective rejection framework

In order to illustrate the interest of considering the multiclass cancer diagnosis in class-selective rejection scheme subject to constraints, one gene dataset is considered and studied.

In the following, we present the study of LUNG CANCER problem in the class selective-rejection scheme subject to two constraints. Let's start by defining the decision options. In fact, LUNG CANCER diagnosis problem is determined by the gene expression profiles of 67 lung tumors and 6 normal lung specimens from patients whose clinical course was followed for up

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	6	0	0	0	0	0	0
	SCLC	0	3	0	0	0	0	0
	LCLC	0	0	3	0	0	4	0
	SCC	0	0	0	16	0	2	0
	AC2	0	0	0	0	4	0	0
	AC3	0	0	0	0	1	3	0
	AC1	0	0	1	0	1	1	20
	$\{LCLC, SCC, AC3\}$	0	0	1	0	0	2	0
	All tumors	0	2	0	0	1	1	1
	All classes	0	0	0	0	0	0	0

Table 9. Confusion matrix of the $50W^*$ LUNG CANCER problem with class selective rejection. Total of misclassified is equal to 10, total of partially and totally rejected samples is equal to 8.

to 5 years. The tumors comprised 41 Adenocarcinomas (ACs), 16 squamous cell carcinomas (SCCs); 5 cell lung cancers (LCLCs) and 5 small cell lung cancers (SCLCs). ACs are subdivided into three subgroups 21 AC of group 1 tumors, 7 AC of group 2 tumors and 13 AC of group 3 tumors. Thus, the multiclass diagnosis cancer consists of 7 classes.

Authors in Garber et al. (2001) observed that AC of group 3 tumors shared strong expression of genes with LCLC and SCC tumors. Thus, poorly differentiated AC is difficult to distinguish from LCLC or SCC. Confusion matrices (tables 7 and 8) computed in the classical framework, with $50W^*$ and $50H$ prove well these claims. It can be noticed that 8 of the 16 misclassified $50W^*$ patients and 8 of the 15 misclassified $50H$ patients correspond to confusion between these three subcategories. Therefore, one may define a new decision option as a subset of these three classes to group these errors and set up an additional test to differentiate them.

Moreover, researches affirm that distinction between patients with nonsmall cell lung tumors (SCC, AC and LCLC) and those with small cell tumors or SCLC is extremely important, since they are treated very differently. Thus, a confusion or wrong decision among patients of nonsmall cell lung tumors should cost less than a confusion within nonsmall cells tumors classes or within small cells tumors classes. Besides, one may provide an extra decision option that includes all the subcategories of tumors to avoid this kind of confusion. Finally, another natural decision option can be the set of all classes, which means that the classifier has totally withhold taking a decision.

Given all these information, the classification problem can be defined as follows:

- Ten decision options can be defined. The possible decision options are given by: $\{Normal\}$, $\{SCLC\}$, $\{LCLC\}$, $\{SCC\}$, $\{AC2\}$, $\{AC3\}$, $\{AC1\}$, $\{LCLC, SCC, AC3\}$, $\{SCLC, LCLC, SCC, AC2, AC3, AC1\}$ and $\{Normal, SCLC, LCLC, SCC, AC2, AC3, AC1\}$.
- The chosen comprimised is given by $P_E \leq 0.15$ and $P_I \leq 0.1$ where P_E is the probability of error and P_I is the probability of indistinctness.

- The loss function defined by (2) with the costs $c_{i,j}$ given by:

$$c_{i,j} = \begin{cases} 1, & w_j \notin \psi_i; \\ \frac{|\psi_i|-1}{N-1}, & w_j \in \psi_i \text{ et } |\psi_i| > 1; \\ 0, & \psi_i = \{w_j\}. \end{cases}$$

Solving this problem with $50W^*$ LUNG CANCER problem leads to the confusion matrix presented in table 9. As a comparison with table 7, one may mainly note that the number of misclassified patients decreases from 16 to 10 and 8 withhold decisions or rejected patients. The probability of error has decreased from 0.219 to 0.136 with a probability of ambiguity equal to 0.109. This partial rejection contributes to avoid confusion between nonsmall and small lung cells tumors and reduces errors due to indistinctness among LCLC, SCC and AC3. Besides, according to the example under study, no patient is totally rejected. It is an expected result since initially (table 7) there exists no confusion between normal and tumor samples.
To take a decision concerning the rejected patients, we may refer to clinical analysis. It is worth to note that for partially rejected patients, clinical analysis is less expensive in terms of time and money than those on completely blinded patients. Moreover, a supervised solution can be also proposed. It aims to use genes selected from another test statistic in order to assign rejected patients to one of the possible classes Jrad et al. (2009b;d). Many factors play an important role in the cascade classifiers system such as the choice of test statistics, the number of classifiers in a cascade system,... Such concerns are under study.

6. Discussions and conclusion

This chapter presents the multiclass decision problem in a new framework where the performances of the decision rule must satisfy some constraints. A general formulation of the problem with class-selective rejection subject to performance constraints was expounded. The definition of the problem takes into account three kinds of criteria: the label sets, the performance constraints, and the average expected loss. The solution of the stated problem was given within the statistical decision theory framework. Some supervised learning strategies were presented. Two approaches are proposed; a class-modeling approach and a boundary based approach. The first named class-modeling approach is defined within the statistical community. Class-modeling approaches are generally characterized by having an explicit underlying probability model, which provides a probability of being in each class rather than simply a classification. The second is defined in the Support Vector Machines community. It focuses on the boundary of the data. It avoids the estimation of the complete density of the data, which might be difficult using small sample sizes.
Experimental results on artificial datasets show that, on the first hand, class-modeling approaches require big amounts of data because it is based on a complete density estimate. Furthermore, the performances of the classifier is conditioned by the choice of a good convergent estimator. As a comparison between GMM and PW algorithms, it is worthy to note that even though PW is widely used, for some complex distributions like multimodal distributions, GMM fitting can be a better model yielding to an accurate decision rule. GMM produce not only memory and computational advantages, but also superior results in terms of solving the under vs. overfitting compromise. When a large sample of typical data is available, the density method is expected to work well.

On the second hand, MSVM methods based on ν-1-SVM methods is a relatively new approach that avoids the estimation of the complete probability density. This not only gives an advantage when just a limited sample is available, it is even possible to learn from data when the density distribution is difficult to estimate (representation space of high dimension).

Experimental results on real datasets show that, in the particular case where decisions are given by the possible classes and the loss function is set equal to the error rate, the proposed boundary approach, compared with the state of art multiclass algorithms, can be considered as a competitive one. Moreover, this method seems to be an interesting solution to solve imbalanced problems. Because both minority and majority classes are learned separately, descriptions are not dominated by the majority classes. Consequently, the performance of the learning procedure is supposed to outperform multiclass rules learned from imbalanced data sets. In the class-selective rejection scheme with constraints, the proposed classifier ensures higher reliability and reduces time and expense costs by introducing partial and total rejection and restricting the misclassified the ambiguously classified samples.

Finally, we can say that the expounded approaches are a new way to learn accurate multiclass decision rules satisfying users requirements on the global performances of the classification system. To avoid too demanding constrains like $P_E < \epsilon$ and $P_I < \epsilon'$ which lead most of the time to very large total rejection (total rejection should not be subject to constraint otherwise the problem may have no solution), we advocate to choose reasonable constraints as an initial target and then to tune the value of the obtained Lagrange multipliers to select a satisfactory compromise. The main interest of the proposed method is to provide a nice initial starting point for the decision rule design and also to avoid costs adjustments which may be difficult to achieve when the number of decision options and consequently the number of costs is large. In the proposed approach only Lagrangian multipliers need final adjustments. Since the number of the Lagrangian multipliers is equal to the number of monitored performance constraints which is generally limited, the choice of a good trade-off among the performances constraints is not too difficult to achieve.

7. References

Alizadeh, A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Hudson, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P. & Staudt, L. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, *Nature* **403**(6769): 503–511.

Bishop, C. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer.

Bottou, L., Cortes, C., Denker, J., Drucker, H., Guyon, I., Jackel, L., LeCun, Y., Muller, U., Sackinger, E., Simard, P. & Vapnik, V. (1994). Comparison of classifier methods: a case study in handwriting digit recognition, *International Conference on Pattern Recognition*, pp. 77–87.

Brown, M. B. & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means, *Technometrics* **16**(1): 129–132.

Chen, D., Liu, Z., Ma, X. & Hua, D. (2005). Selecting genes by test statistics, *Journal of Biomedicine and Biotechnology* **2005**(2): 132–138.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments, *Journal of the Royal Statistical Society* **4**: 102–118.

Daniel, W. W. (1999). *Biostatistics:A Foundation for Analysis in the Health Sciences*, NY: Wiley.

Davy, M., Desobry, F., Gretton, A. & Doncarli, C. (2006). An online support vector machine for abnormal events detection, *Signal Process.* **86**(8): 2009–2025.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm, *J. Roy. Statist* **39**: 1–38.

Duda, R. & Hart, P. (1973). *Pattern Classification and Scene Analysis*, John Wiley and Sons, pp. 98–105.

Emanuel, P. (1962). On estimation of a probability density function and mode, *The Annals of Mathematical Statistics* **33**(3): 1065–1076.

Garber, M., Troyanskaya, O., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G., Perou, C., Whyte, R., Altman, R., Brown, P., Botstein, D. & Petersen, I. (2001). Diversity of gene expression in adenocarcinoma of the lung, *Proc Natl Acad Sci*, Vol. 98, USA, pp. 13784–13789.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**(5439): 531–537.

Grall-Maës, E. & Beauseroy, P. (2009). Optimal decision rule with class-selective rejection and performance constraints, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **99**(1).

Grall-Maës, E., Beauseroy, P. & Bounsiar, A. (2006a). Multilabel classification rule with performance constraints, *Proceedings of IEEE conference ICASSP'06*, France.

Grall-Maës, E., Beauseroy, P. & Bounsiar, A. (2006b). Quality assessment of a supervised multilabel classification rule with performance constraints, *EUSIPCO'06*, Italy.

Guobin, O. & Lu, M. Y. (2007). Multi-class pattern classification using neural networks, *Pattern Recogn.* **40**(1): 4–18.

Ha, T. M. (1997). The optimum class-selective rejection rule, *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(6): 608–615.

Hao, P. & Lin, Y. (2007). A new multiclass support vector machine with multisphere in the feature space, *IEA/AIE*, pp. 756–765.

Hartung, J. & Makambi, K. H. (2002). Small sample properties of tests on homogeneity in one-way anova and meta-analysis, *Statistical Papers* **43**: 197–235.

Hastie, T., Rosset, S., Tibshirani, R. & Zhu, J. (2004). The entire regularization path for the support vector machine, *J. Mach. Learn. Res.* **5**: 1391–1415.

Horiuchi, T. (1998). Class selective rejection rule to minimize the maximum distance between selected classes, *PR* **31**(10): 1579–1588.

Husband, C. & Lin, C. (2002). A comparison of methods for multiclass support vector machines, *IEEE Transactions on Neural Networks* **13**: 415–425.

Jrad, N., Grall, E. & Beauseroy, P. (2008). Supervised learning rule selection for multiclass decision with performance constraints, *IEEE conference ICPR*, USA.

Jrad, N., Grall-Maës, E. & Beauseroy, P. (2008). A supervised decision rule for multiclass problems minimizing a loss function, *Seventh International Conference on Machine Learning and Applications* pp. 48–53.

Jrad, N., Grall-Maës, E. & Beauseroy, P. (2009a). Apprentissage supervisé de règles de décision multiclasses avec contraintes de performances évolutives, *XXII Colloque Gretsi*.

Jrad, N., Grall-Maës, E. & Beauseroy, P. (2009b). Classification supervisée de tumeurs cancéreuses avec rejet sélectif, *XXII Colloque Gretsi*.

Jrad, N., Grall-Maës, E. & Beauseroy, P. (2009c). Gaussian mixture models for multiclass problems with performance constraints, *ESANN*.

Jrad, N., Grall-Maës, E. & Beauseroy, P. (2009d). Gene-based multiclass cancer diagnosis with class-selective rejections, *Journal of Biomedicine and Biotechnology* .

Kutner, M. H., Nachtsheim, C. J., Neter, J. & Li, W. (2005). *Applied Linear Statistical Models*, 5th edn, McGraw Hill.

Li, J. Q. & Barron, A. R. (1999). Mixture density estimation, *In Advances in Neural Information Processing Systems 12*, MIT Press, pp. 279–285.

Rakotomamonjy, A. & Davy, M. (2007). One-class svm regularization path and comparison with alpha seeding, *ESANN 2007*, Bruge, Belgium, pp. 221–224.

Ross, D., Scherf, U., Eisen, M., Perou, C., Rees, C., Spellman, P., V, V. I., Jeffrey, S., de Rijn, M. V., Waltham, M., Pergamenschikov, A., Lee, J., Lashkari, D., Shalon, D., Myers, T., Weinstein, J., Botstein, D. & Brown, P. (2000). Systematic variation in gene expression patterns in human cancer cell lines, *Nat Genet.*, Vol. 24, USA, pp. 227–235.

Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., Brown, P. O. & Weinstein, J. N. (2000). A gene expression database for the molecular pharmacology of cancer, *Nat Genet* **24**(3): 236–244.

Scholkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution, *Neural Computation* **13**(7): 1443–1471.

Scholkopf, B. & Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, USA.

Tax, D. (2001). *One-class classification: concept learning in the absence of counter-examples*, PhD thesis, Technische Universiteit Delft.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, John Wiley, New York.

Vapnik, V. N. (1998). *Statistical Learning Theory*, Wiley-Interscience.

Verbeek, J., Vlassis, N. & Krose, B. (2003). Efficient greedy learning of gaussian mixture models, *Neural Computation* **15**(2): 469–485.

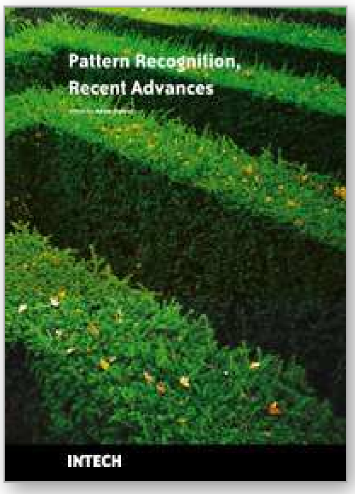
Vlassis, N. & Likas, A. (2002). A greedy em algorithm for gaussian mixture learning, *Neural Processing Letters* **15**(1): 77–87.

WELCH, B. L. (1951). On the comparison of several mean values: an alternative approach, *Biometrika* **38**(3-4): 330–336.

Welsh, J., Zarrinkar, P., Sapinoso, L., Kern, S., Behling, C., Monk, B., Lockhart, D., Burger, R. & Hampton, G. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer, *Proc Natl Acad Sci*, Vol. 98, USA, pp. 1176–1181.

Weston, J. & Watkins, C. (1999). Multiclass support vector machines, *ESANN*.

Yang, X., Liu, J., Zhang, M. & Niu, K. (2007). A new multiclass svm algorithm based on one-class svm, *ICCS 2007*, Beijing, China, pp. 677–684.



Pattern Recognition Recent Advances
Edited by Adam Herout

ISBN 978-953-7619-90-9
Hard cover, 524 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821