An approach to optimizing abstaining area for small sample data classification

Blaise Hanczar, Jean-Daniel Zucker

▶ **To cite this version:**

Blaise Hanczar, Jean-Daniel Zucker. An approach to optimizing abstaining area for small sample data classification. Expert Systems with Applications, 2018, 95, pp.153–161. 10.1016/j.eswa.2017.11.013 . hal-01658150

HAL Id: hal-01658150
https://hal.science/hal-01658150

Submitted on 29 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ACCEPTED MANUSCRIPT

An approach to optimizing abstaining area for small sample data classification

Blaise Hanczar [1], Jean-Daniel Zucker [2,3]

[1] *IBISC, University Evry, IBGBI Building, 23 Boulevard de France, 91037 Evry, France*
blaise.hanczar@ibisc.univ-evry.fr
[2] *Université Pierre et Marie Curie Paris 6, Centre de Recherche des Cordeliers, UMR S 872, Paris, F-75006 France.*
[3] *IRD, UMI 209, UMMISCO, Centre IRD de l'Île de France, Bondy, F-93143 France*
jean-daniel.zucker@ird.fr
Corresponding author: blaise.hanczar@ibisc.univ-evry.fr (+33 1 64 85 34 61)

Abstract

Given a classification task, an approach to improve accuracy relies on the use of abstaining classifiers. These classifiers are trained to reject observations for which predicted values are not reliable enough: these rejected observations belong to an abstaining area in the feature space. Two equivalent methods exist to theoretically compute the optimal abstaining area for a given classification problem. The first one is based on the posterior probability computed by the model and the other is based on the derivative of the ROC function of the model. Although the second method has proved to give the best results, in small-sample settings such as the one found in omics data, the estimation of posterior probabilities and derivative of ROC curve are both lacking of precision leading to far from optimal abstaining areas. As a consequence none of the two methods bring the expected improvements in accuracy. We propose five alternative algorithms to compute the abstaining area adapted

to small-sample problems. The idea of these algorithms is to compute an accurate and robust estimation of the ROC curve and its derivatives. These estimation are mainly based on the assumption that the distribution of the output of the classfier for each class is normal or mixture of normal distributions. These distributions are estimated by a kernel density estimator or Bayesian semiparametric estimator. Another method works on the approximation of the convex hull of the ROC curve. Once the derivative of the ROC curve are estimated, the optimal abstaining area can be directly computed. The performance of our algorithms are directly related to their capacity to compute an accurate estimation of the ROC curve. A sensitivity analysis of our methods to the dataset size and rejection cost has been done on a set of experiments. We show that our methods improve the performances of the abstaining classifiers on several real datasets and for different learning algorithms.

Keywords: Supervised learning, reject option, small-sample setting, Abstaining classifier, ROC curve estimation

1. Introduction

In many domains, more and more data are produced because of recent technological advances and the increasing capabilities of modern computers to analyze and mine these data. One of the most interesting exploitations of these data is the construction of predictive classifiers [14]. For example, in genetic and molecular medicine, gene expression profiles are used to dif-

ferentiate different types of tumors with different outcomes and thus assist the physician in the selection of more suitable therapeutic treatment [26]. A huge number of different methods from pattern recognition or machine learning have been developed and applied on various domains. Even when these methods produce classifiers with a good accuracy, they are often still insufficiently accurate to be used routinely. For example, a diagnostic or a choice of therapeutic strategy must be based on a very high confidence classifier; an error of the predictive model may lead to tragic consequences. An avenue for improving this confidence is to use *abstaining classifiers* [21] also called *reject classifiers* [25] or *selective classifiers* [6]. Unlike *classical* classifiers that provide a predicted class for each test example, only a subset of the examples is assigned to a class. The abstaining classifiers define an abstaining area regrouping the examples whose confidence in the predicted class is low, these examples are rejected, i.e. no class is assigned to them [3, 24, 22, 4, 10, 19]. This type of classifier has thus a higher accuracy than the classic classifier at the expense of a positive rejection rate. As a consequence, there is a trade-off between accuracy and rejection rate to control [13]. In other words, the higher the classifier accuracy, the higher the rejection rate.

Chow has introduced the notion of abstaining classifier and his definition of the abstaining area is based on the exact posterior probabilities of each example [3]. For a given cost of rejection of an example, one can compute the optimal abstaining area. In practical cases, the exact posterior probabilities are not available since the class distribution is unknown. Chow's

rule must thus be used with an *estimation* of the posterior probabilities. To drop the necessity to rely on the exact posterior probabilities, Tortorella has proposed a method where the abstaining area is computed in selecting two points on the Receiver operating characteristic ROC curve [8] describing the performance of the classification model [24]. The two points are identified by their tangent on the ROC curve computed from the cost of rejection and type of error. As with Chow's rule, if one knows precisely the exact ROC curve, the resulting abstaining area is optimal. The problem is that on real data, one does not have the exact ROC curve either. The computation of the derivative is therefore done on the convex hull of the *empirical* ROC curve. Santos-Pereira proved that both the Chow's rule and ROC rule are equivalent in theory [23]. In practice, both rules do not lead to the optimal abstaining area since the exact posterior probabilities and the ROC curve are both unknown. As they thus rely on estimations, they compute therefore an *approximation* of the optimal abstaining area. Several studies have shown that, on real data, the ROC rule gives better performance than Chow's rules [17, 28]. The important point is that all of these studies have been done on datasets containing a large set of examples (several hundred or thousands). However in many domains the datasets contain few examples, for example in genomics studies where the acquisition cost of the data is expensive, the datasets contain generally less than 100 examples. In small sample data, the empirical ROC curve has to be constructed with only some tens of examples. As a consequence, the ROC convex hull curve is defined by a very limited

number of points and different tangent values. Figure 1 illustrates this fact on a concrete task, the bold line represents the exact ROC curve, the gray line is the empirical ROC curve computed from 40 examples and its convex hull curve is the black line. The ROC convex hull is defined by only four points and takes only three different values of the tangent. This means that only five different non-trivial abstaining areas are possible from this convex hull curve: $\{(P1, P2), (P1, P3), (P2, P3), (P2, P4), (P3, P4)\}$. From the exact ROC curve, we can define an infinite number of different abstaining areas. This example illustrates the limitation of the ROC rule in small sample data. In this paper, we propose some adaptations of the ROC rule for the small sample problem. We present five methods to approximate the ROC curve and its derivative in order to compute a reject area and improve the performance of the classifier. These methods are easy to implement and usable with any classification method. Our experiments show that we improve significantly the performance of the classifier on real data compared to the ROC and Chow's rules.

2. Abstaining Classifiers

2.1. Definition

We consider a classification problem with two classes: positive P and negative N. Let's a training set of m examples $T = \{(x_1, y_1), ..., (x_m, y_m)\}$ ($x_i \in \mathbb{R}^d$, $y_i \in \{P, N\}$) and a classifier rule Ψ. The classifier output $\omega(x)$ is a continuous value of an example x. In fixing a threshold α on this output

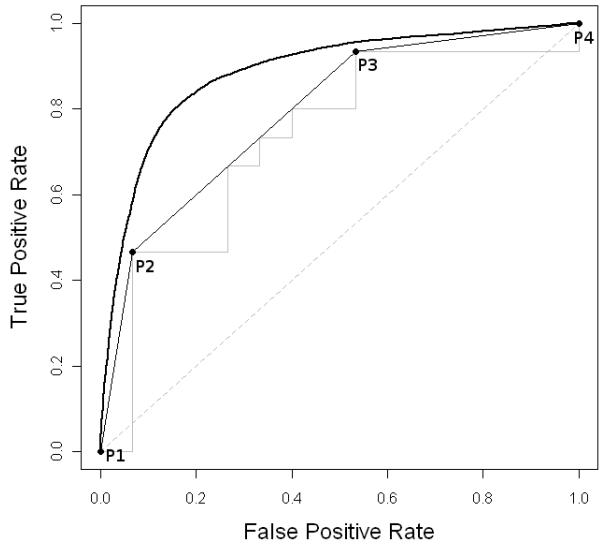

Figure 1: Exact ROC curve (bold line) and empirical ROC curve (gray line) and empirical ROC curve convex hull (black line).

Table 1: Cost matrix of the classification problem.

		actual	
		P	N
Predicted	P	λ_{PP}	λ_{NP}
class	N	λ_{PN}	λ_{NN}
	R	λ_R	λ_R

we define a classic classifier Ψ_α that assigns one of the two classes to each example. In fixing two thresholds $\{\alpha_N, \alpha_P\}$, we define an abstaining classifier that rejects some examples and assigns the others to one of the two classes.

$$\Psi_{\alpha_n,\alpha_p}(x) = \begin{cases} N & \text{if} \quad \omega(x) \leq \alpha_N \\ R & \text{if} \quad \alpha_N < \omega(x) < \alpha_P \\ P & \text{if} \quad \omega(x) \geq \alpha_P \end{cases}$$

with the constraint $\alpha_N \leq \alpha_P$. R represents the rejection of the example x. Figure 2 shows the distribution of the two classes on the classifier output. The two thresholds α_N and α_P divide the classifier output into three decision regions ($\{N, P, R\}$). The performance of the classifier depends on the following values: the rate of true negative (TNR), true positive (TPR), false negative (FNR), false positive (FPR), positive rejection (RPR), negative rejection (RNR) and the prior probabilities of the two classes π_P and π_N. For each of the classification type a cost is defined, they are represented by the cost matrix as in table 2.1. The performance of a classifier is measured by its expected loss:

$$\begin{aligned} L(\Psi_{\alpha_N,\alpha_P}) = \pi_P \left[\lambda_{PP} TPR + \lambda_{NP} FNR + \lambda_R RPR\right] \\ + \pi_N \left[\lambda_{NN} TNR + \lambda_{PN} FPR + \lambda_R RNR\right] \end{aligned}$$

The objective is to find the thresholds α_N and α_P minimizing the expected loss of the classifier.

2.2. Chow's rule

With Chow's rule [3], we considered that the posterior probability of the positive class is given $\omega(x) = p(P|x)$. We define the three loss functions L_N, L_R and L_P that represent the expected loss that is obtained in assigning an example x to respectively the class N, R or P.

From this formulation we solve the following equations :

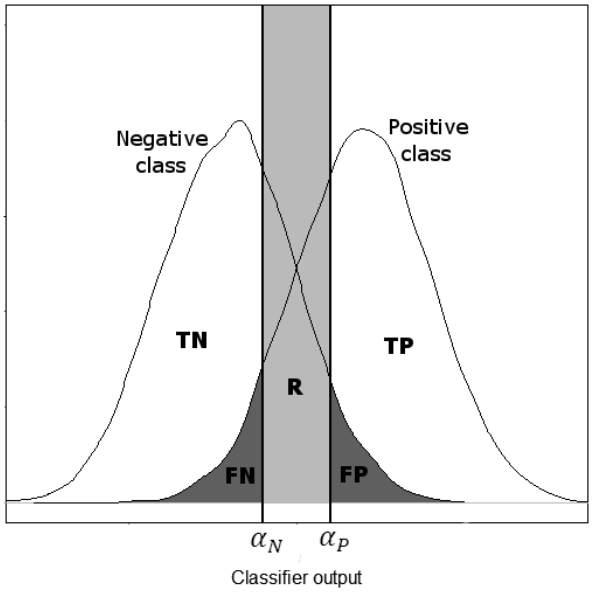

Figure 2: Distribution of the classes on the classifier output:

$$Ł_P(x) = L_R(x) \quad \Leftrightarrow \quad \lambda_{PP}p(P|x) + \lambda_{NP}p(N|x) = \lambda_R$$
$$\Leftrightarrow \quad \lambda_{PP}p(P|x) + \lambda_{NP}(1 - p(P|x)) = \lambda_R \quad \Leftrightarrow \quad p(P|x) = \frac{\lambda_R - \lambda_{NP}}{\lambda_{PP} - \lambda_{NP}}$$

$$Ł_N(x) = L_R(x) \quad \Leftrightarrow \quad \lambda_{PN}p(P|x) + \lambda_{NN}p(N|x) = \lambda_R$$
$$\Leftrightarrow \quad \lambda_{PN}p(P|x) + \lambda_{NN}(1 - p(P|x)) = \lambda_R \quad \Leftrightarrow \quad p(P|x) = \frac{\lambda_R - \lambda_{NN}}{\lambda_{PN} - \lambda_{NN}}$$

We obtain the optimal decision thresholds :

$$\alpha_P^* = \frac{\lambda_R - \lambda_{NP}}{\lambda_{PP} - \lambda_{NP}} \qquad \alpha_N^* = \frac{\lambda_R - \lambda_{NN}}{\lambda_{PN} - \lambda_{NN}} \tag{1}$$

It is interesting to note that if the classifier output is the likelihood ratio $lr(x) = \frac{p(x|P)}{p(x|N)}$ then the optimal decision thresholds are the following:

$$lr(x)_P^* = \frac{\lambda_R - \lambda_{NP}}{\lambda_{PP} - \lambda_R} \frac{\pi_N}{\pi_P} \qquad lr(x)_N^* = \frac{\lambda_R - \lambda_{NN}}{\lambda_{PN} - \lambda_R} \frac{\pi_N}{\pi_P}$$

These decisions threshold are equal to the optimal ROC curve derivative found in the ROC rule as we will see in the next section. This illustrates the theoretical results stating that Chow's rule and ROC rule are equivalent.

In theory the abstaining area defined by these two thresholds is optimal. In practice since the exact posterior probabilities are unknown, we have to rely on their estimations, we therefore obtain an approximation of the abstaining rejection area.

2.3. Optimal ROC based abstaining area

An alternative to Chow's rule is to compute the optimal rejection area from the ROC curve. The expected loss of an abstaining classifier can be

expressed by the following form:

$$
\begin{aligned}
&L(\Psi_{\alpha_N,\alpha_P}) = \\
&\pi_P \left[\lambda_{PP} TPR_{\alpha_P} + \lambda_{PN} FNR_{\alpha_N} + \lambda_R (FNR_{\alpha_P} - FNR_{\alpha_N})\right] \\
&+ \pi_N \left[\lambda_{NP} FPR_{\alpha_P} + \lambda_{NN} TNR_{\alpha_N} + \lambda_R (FPR_{\alpha_N} - FPR_{\alpha_P})\right]
\end{aligned}
$$

where TPR_{α_P} is the true positive rate obtained by the no-abstaining classifier Ψ_{α_P}. The values $(FNR_{\alpha_P} - FNR_{\alpha_N})$ and $(FPR_{\alpha_N} - FPR_{\alpha_P})$ represent the rejection rate of respectively positive and negative class. On real data, we do not have the true values of $TPR, TNR, FPR, FNR, \pi_P, \pi_N$ they have to be estimated empirically from an independent finite validation set of examples: $\widehat{TPR} = \frac{TP}{N_P}, \widehat{TNR} = \frac{TN}{N_N}, \widehat{FPR} = \frac{FP}{N_N}, \widehat{FNR} = \frac{FN}{N_P}, \widehat{\pi_P} = \frac{N_P}{N_P+N_N}, \widehat{\pi_N} = \frac{N_N}{N_P+N_N}$ where TP, TN, FP, FN are the number of examples in the different types of classification computed on the validation set. We compute an approximation of the expected loss function:

$$
\begin{aligned}
&\widehat{L}(\Psi_{\alpha_N,\alpha_P}) = \\
&\widehat{\pi_P} \left[\lambda_{PP} \widehat{TPR_{\alpha_P}} + \lambda_{PN} \widehat{FNR_{\alpha_N}} + \lambda_R (\widehat{FNR_{\alpha_P}} - \widehat{FNR_{\alpha_N}})\right] \\
&+ \widehat{\pi_N} \left[\lambda_{NP} \widehat{FPR_{\alpha_P}} + \lambda_{NN} \widehat{TNR_{\alpha_N}} + \lambda_R (\widehat{FPR_{\alpha_N}} - \widehat{FPR_{\alpha_P}})\right]
\end{aligned}
\tag{2}
$$

The ROC curve is a function that links the true positive rate to the false positive rate $\frac{TP}{N_P} = f_{ROC}(\frac{FP}{N_N})$. In including the definition to the previous

formulas, we can express $\hat{L}(\Psi_{\alpha_N,\alpha_P})$ only in function of FP_{α_P} and FP_{α_N}.

$$\hat{L}(\Psi_{\alpha_N,\alpha_P}) = \widehat{\pi_P}[(\lambda_{PP} - \lambda_R) f_{ROC}(\widehat{FPR_{\alpha_P}}) + (\lambda_R - \lambda_{PN}) f_{ROC}(\widehat{FPR_{\alpha_N}}) + \lambda_{PN}]$$
$$+ \widehat{\pi_N}[(\lambda_{NP} - \lambda_R)\widehat{FPR_{\alpha_P}} + (\lambda_R - \lambda_{NN})\widehat{FPR_{\alpha_N}} + \lambda_{NN}]$$

In solving the equations $\frac{\partial \hat{L}(\Psi_{\alpha_N,\alpha_P})}{\partial \widehat{FPR_{\alpha_N}}} = 0$ and $\frac{\partial \hat{L}(\Psi_{\alpha_N,\alpha_P})}{\partial \widehat{FPR_{\alpha_P}}} = 0$ the conditions to reach the minimum of the loss function are given by the derivatives of the ROC function :

$$\begin{aligned} f'_{ROC}\left(\widehat{FPR_{\alpha_N}}\right) &= \frac{N_N(\lambda_R - \lambda_{NN})}{N_P(\lambda_{PN} - \lambda_R)} \\ f'_{ROC}\left(\widehat{FPR_{\alpha_P}}\right) &= \frac{N_N(\lambda_R - \lambda_{NP})}{N_P(\lambda_{PP} - \lambda_R)} \end{aligned} \tag{3}$$

f_{ROC} is a function strictly increasing and concave, a given value of derivative corresponds to an unique point on the ROC curve. Since each point on the ROC curve (FPR_α, TPR_α) is associated to a decision threshold, we can obtain the rejection area from the two previous derivatives.

3. Computing the abstaining area

We present here six methods to compute the abstaining area in a real condition. The first one is the implementation of the ROC rule given by Tortorella [25], the others are the new methods that we proposed. The methods presented in section 3.2, 3.3, 3.4 and 3.5 are based on an approximation of the ROC curve. A lot of ROC approximation methods has been proposed, a taxonomy of the existing methods can be found in [15, 9]. Our methods use

respectively the Gaussian kernel approximation (3.2), Bayesian semiparametric estimator (3.3), binormal model (3.4) and Bezier curve from the convex hull of the ROC points. The last method is an empirical and exhaustive search of the best rejection area.

3.1. The ROC rule

In practice, generally only one dataset T is available to construct the classifier. We, therefore, have to split this dataset into a training set Tr and a validation set Tv. The training set is used to learn the model that returns the value of the classifier output $\omega(x)$ for any example x. The validation set is used to compute the ROC curve of the model and define the abstaining area. The ROC curve is defined by a set of v points $U_{ROC} = \left\{ \binom{FPR_i}{TPR_i}; 1 \leq i \leq v \right\}$ and a decision threshold is associated to each of these points. The ROC convex hull is the minimum convex set dominating all points of the ROC curves, the set of points defining the ROC convex hull $U_{ROCCH} = \left\{ \binom{FPR_j}{TPR_j}; 1 \leq j \leq w \right\}$ is a subset of U_{ROC}. Computing the abstaining area consists in identifying the two points of the ROC convex hull whose tangents are the closest from the target derivative defined by equation (2). These two points are associated to the two decision thresholds that will define the abstaining area. As mentioned in the introduction, the ROC convex hull is a piecewise linear curve and takes a finite number of different tangents. It is not a unique tangent but a range of tangents that is associated to each point of the ROC convex hull. The point $\binom{FPR_j}{TPR_j}$ will be selected for

any target derivative contained in the interval $\left[\frac{TPR_j - TPR_{j-1}}{FPR_j - FPR_{j-1}}; \frac{TPR_{j+1} - TPR_j}{FPR_{j+1} - FPR_j}\right]$. Recall that this method is possible only if the tangent ranges for each point are disjoint. This is the case since it has been proved that the convex hull of the ROC curve is concave (i.e. the area below the curve is convex) [27].

3.2. ROC approximation by Gaussian kernel density estimator (roc.kernel)

To avoid the problem of finite number tangents, we propose to approximate the ROC curve by a kernel density estimator from R_{ROC}. Kernel estimators are known to be simple with good theoretical and practical properties for the ROC curve estimation [16].

The ROC curve can be formulated as :

$$\widehat{f_{ROC}(t)} = 1 - F_P(F_N^{-1}(1-t)) \tag{4}$$

with $t \in [0,1]$ where $F_N^{-1}(1-t) = inf x \in \mathbb{R} | F(x) \geq 1-t$. F_P and F_N represent the distributions of respectively the positive and negative class on the classifier outout ω. We have therefore $TPR_\alpha = 1 - F_P(\alpha)$ and $FPR_\alpha = 1 - F_N(\alpha)$. Since the true distributions F_P and F_N are unknown, they are estimated by a Gaussian kernel estimator:

$$\begin{aligned} \widehat{F_P(\alpha)} &= \frac{1}{N_P} \sum_{i=1}^{N_P} \Phi\left(\frac{\alpha - w_i^{(P)}}{h_P}\right) = 1 - \widehat{TPR_\alpha} \\ \widehat{F_N(\alpha)} &= \frac{1}{N_N} \sum_{i=1}^{N_N} \Phi\left(\frac{\alpha - w_i^{(N)}}{h_N}\right) = 1 - \widehat{FPR_\alpha} \end{aligned} \tag{5}$$

where $(\omega_1^{(P)}, ..., \omega_{N_P}^{(P)})$ and $(\omega_1^{(N)}, ..., \omega_{N_N}^{(N)})$ are the classifier output of respectively the positive and negative examples. h_P and h_N are the bandwidths of the Gaussian kernel for each class. The bandwidths are computed using the Altman ans Leger's approach [1]. From the formulas (5), the TPR_α and FPR_α are computed for a large value range of $\alpha : \{\alpha_1, ..., \alpha_K\}$. This set of ROC points $\{(TPR_{\alpha_i}, FPR_{\alpha_i})\}$ gives an estimation of the ROC curve. Each value of α_i is associated to a derivative of the ROC curve by :

$$\widehat{f'_{ROC}(FPR_{\alpha_i})} = \frac{\widehat{TPR_{\alpha_{i+1}}} - \widehat{TPR_{\alpha_{i-1}}}}{\widehat{FPR_{\alpha_{i+1}}} - \widehat{FPR_{\alpha_{i-1}}}}$$

The rejection area is formed by the two values of α with the corresponding derivative value closest to the target ROC curve derivative defined in (3).

3.3. ROC approximation by Bayesian semiparametric estimator (roc.bayesian)

Bayesian modeling has been applied successfully to ROC curve approximation. The Bayesian semi-parametric ROC analysis, proposed by Erkanli [7], is one of the most popular approximation methods. The classifier output of positive examples is represented by a mixture of normal distributions.

$$\omega_{K,\theta_K}^{(P)} \sim N(\mu_K, \sigma_K^2)$$

where K is the number of components and $\theta_K = (\mu_K, \sigma_K^2)$ are the parameters of the distributions. Let a set of classifier output values from positive examples $\{w_1^{(P)}, ..., w_{N_P}^{(P)}\}$, the posterior predictive density of the future output of

a positive example given the past outputs can be approximated by :

$$f_N(\omega^{(P)}|\omega_1^{(P)},...,\omega_{N_P}^{(P)} = \frac{1}{N_P}\sum_{i=1}^{N_P}\sum_{k=1}^{K}\eta_k^{(i)}f(\omega|\theta_k^{(i)})$$

where $f(\omega|\theta_k^{(i)})$ is the normal density. $\eta_1,...,\eta_K$ are the probabilities of the components, they are computed from Beta distributions. All details about the computation of these densities and the choice of the number of components can be found in the original paper [7]. Having approximated the posterior predictive density $f(\omega^{(P)}|\omega_1^{(P)},...,\omega_{N_P}^{(P)}$, it is then straightforward to obtain the cumulative distribution function $F_P(\omega|\omega_1^{(P)},...,\omega_{N_P}^{(P)}) = \int_{-\inf}^{\omega} f_N(\omega|\omega_1^{(P)},...,\omega_{N_P}^{(P)}d\omega$. The true positive rate is therefore

$$\widehat{TPR_\alpha} = 1 - F_P(\omega|\omega_1^{(P)},...,\omega_{N_P}^{(P)}) \tag{6}$$

With the same procedure, we compute the false positive rate in estimating the posterior predictive density of negative examples $f_N(\omega|\omega_1^{(N)},...,\omega_{N_N}^{(N)}$. From the formulas (6), the TPR_α and FPR_α are computed for a large value range of α : $\{\alpha_1,...,\alpha_K\}$. This set of ROC points $\{(TPR_{\alpha_i},FPR_{\alpha_i})\}$ gives an estimation of the ROC curve. Each value of α_i is associated to a derivative of the ROC curve by :

$$f'_{ROC}\widehat{(FPR_{\alpha_i})} = \frac{\widehat{TPR_{\alpha_{i+1}}} - \widehat{TPR_{\alpha_{i-1}}}}{\widehat{FPR_{\alpha_{i+1}}} - \widehat{FPR_{\alpha_{i-1}}}}$$

The rejection area is formed by the two values of α with the corresponding derivative value closest to the target ROC curve derivative defined in (3).

3.4. ROC approximation by binormal models (roc.binorm)

In ROC curve approximation, a usual approach is to use a binormal model [5, 12] considering that the probability distribution of the two classes on the classifier output is Gaussian. This model implies that the positive and negative class follow respectively the normal distribution $N(\mu_P, {\sigma_P}^2)$ and $N(\mu_N, {\sigma_N}^2)$. The estimation of the ROC curve is defined by:

$$\widehat{f_{ROC}(t)} = \Phi(a + b\Phi^{-1}(t))$$

where Φ is the cumulative distribution function of the normal distribution, a and b are the parameters of the model estimated from the data : $a = \frac{\mu_P - \mu_N}{\sigma_P}$ and $b = \frac{\sigma_N}{\sigma_P}$. The true and false positive rate can be estimated by $\widehat{TPR_t} = \Phi(a - bt)$ and $\widehat{FPR_t} = \Phi(-t)$. This classic binormal model gives good results in most of the cases, but if it is used on data with only a few examples or poorly distributed, we may obtain a 'hooked' ROC curve that has non-monotonic slope. It may even lead to "degenerate" ROC curve with a zigzag shape. This may be very problematic for the derivative ROC estimation and rejection area computation. To deal with this problem, Metz et al. have proposed a proper version of the binormal model which is based on a monotonic transformation of the likelihood ratio. The proper binormal model is very similar to the classic binormal model when no "hook" or "degeneracy"

is present. All details of the ROC estimation by the proper binormal model can be found in [18]. From the ROC curve estimation we obtain a set of ROC points $\{(TPR_{\alpha_i},FPR_{\alpha_i})\}$ that will be used to construct the rejection area with he same procedure explained in the end of section 3.2.

3.5. ROC approximation by Bezier curves (rocch.approx)

In this method, the ROC curve is approximated from its convex hull. The ROC convex hull curve is formed by the set of points $U_{ROCCH} = \left\{\binom{FPR_j}{TPR_j}; 1 \leq j \leq w\right\}$. We construct a Bezier curve of degree $w-1$ from this set of points to approximate the ROC curve. The bezier curves have no parameters and are adapted to small set of points. The TPR and FPR are estimated by parametric curves for $t \in [0,1]$ as:

$$\widehat{TPR(t)} = \sum_{i=0}^{w} \binom{w}{i} (1-t)^{w-1} t^i TPR_i$$
$$\widehat{FPR(t)} = \sum_{i=0}^{w} \binom{w}{i} (1-t)^{w-1} t^i FPR_i$$

We point out that the points of the ROC convex hull U_{ROCCH} have the following characteristics: $0 \leq FPR_i < FPR_j \leq 1$ and $0 \leq TPR_i < TPR_j \leq 1$ for all $i < j$. The Bezier curve constructed from this set of points will have

the following properties:

$$
\begin{aligned}
0 \leq \widehat{FPR(t_1)} < \widehat{FPR(t_2)} \leq 1 \qquad & \forall \quad t_1 < t_2 \\
0 \leq \widehat{TPR(t_1)} < \widehat{TPR(t_2)} \leq 1 \qquad & \forall \quad t_1 < t_2 \\
\frac{\widehat{TPR(t_1)}}{\widehat{FPR(t_1)}} > \frac{\widehat{TPR(t_2)}}{\widehat{FPR(t_2)}} \qquad & \forall \quad t_1 < t_2
\end{aligned}
$$

The two first properties show that the estimated ROC curve is defined in $[0,1] \rightarrow [0,1]$ and is increasing. The third property says that it is concave. Once the approximation is performed, we identify the false positive rate FPR^* corresponding to the target derivative FPR^* given in the equation (3). Since it is likely that FPR^* does not correspond to a point of U_{ROCCH}, we find the two points containing FPR^* i.e. we find i^* such that $FPR_{i^*} \leq FPR^* \leq FPR_{i^*+1}$. The decision thresholds is given by $\alpha^* = \kappa FPR_{i^*} + (1-\kappa) FPR_{i^*+1}$ with $\kappa = \frac{TPR_{i^*+1} - TPR^*}{TPR_{i^*+1} - TPR_{i^*}}$.

3.6. ROC-curve exhaustive search (exhaustive)

The exhaustive search is not based on the theoretical consideration presented in the previous section. It is an ad hoc method where a large number of different abstaining areas is evaluated and the best one is returned. The following algorithm describes the method:

The performance of all abstaining areas are estimated by $loss(\Psi_{(\alpha_N, \alpha_P)})$ which represents the expected loss (formulas (2)) of the classifier estimated on the validation test. δ is a parameter to fix that controls the number and

size of the tested abstaining areas. The number of tested abstaining areas is $\sum_{i=0}^{1/\delta}(\frac{1}{\delta} - i)$ and their sizes are in $\{k\delta | k \in [1, \frac{1}{\delta}]\}$. If δ is too high, few different abstaining areas will be tested and it is unlikely that we obtain an abstaining area close to the optimal one. The computing time depends on the number of tester abstaining area determined by δ. There is, therefore, a trade-off between the performance and the computing time in choosing δ. In our experiments we set $\delta = 0.001$.

4. Results and Discussion

We have performed a set of experiments on both artificial and real datasets in order to empirically investigate the behavior of the five approaches we have introduced above and compared their performances to the state-of-the-art methods.

4.1. Accuracy of the ROC derivative estimation

The artificial datasets are based on Gaussian distributions in dimension 10. We change the notation for more clarity. The positive class follows the distribution $N(\mathbf{1}, I)$ and negative class $N(-\mathbf{1}, I)$ where $\mathbf{1}$ is a vector of size 10 containing only 1s, μ is drawn from a uniform distribution $U[0.5, 1.5]$ and I is the identity matrix. In each experiment, 100 iterations have been done.

The computation of the abstaining area is based on the approximation of the ROC derivative. Figure 3 shows the precision of this approximation. In using a Gaussian artificial dataset, we can compute the exact ROC

curve and its derivative. In using the ROC derivative methods described int he section 3, we approximate the derivative for each value of false positive rate. The derivative approximation has been computed from a validation set of 60 examples with the ROC curve estimation used in the methods roc.kernel, roc.bayesian, roc.binormal and rooch. The full gray curve shows the logarithm of the derivative the real ROC curve. The derivative of the ROC convex hull is a piecewise constant curve. Derivative approximation by roc.kernel, roc.bayesian, roc.binorm or rocch.approx is much closer from the exact derivative than the ROC convex hull. The ROCCH approximation curve is very smooth because it is based on Bezier curve. It is inaccurate on the extremities of the graph. Note that it is rare that the decision values of the abstaining area are defined by extreme values of FPR, so the precision of derivative approximation is less important at the extreme of figure 3. Other simulations with different artificial data have been made and lead to the same conclusions. These results show that our approximation methods produce more accurate derivative that the ROC convex hull curve.

4.2. Impact of the validation set size

In another experiment based on artificial Gaussian dataset, the impact of the validation set on the performance of the final classifier is investigated. A training set, a validation set and a test set of size respectively 100, N_{valid}, 10000 are generated. N_{valid} varies from 20 to 500. The training set is used to learn the model, the validation set to compute the abstaining area and the

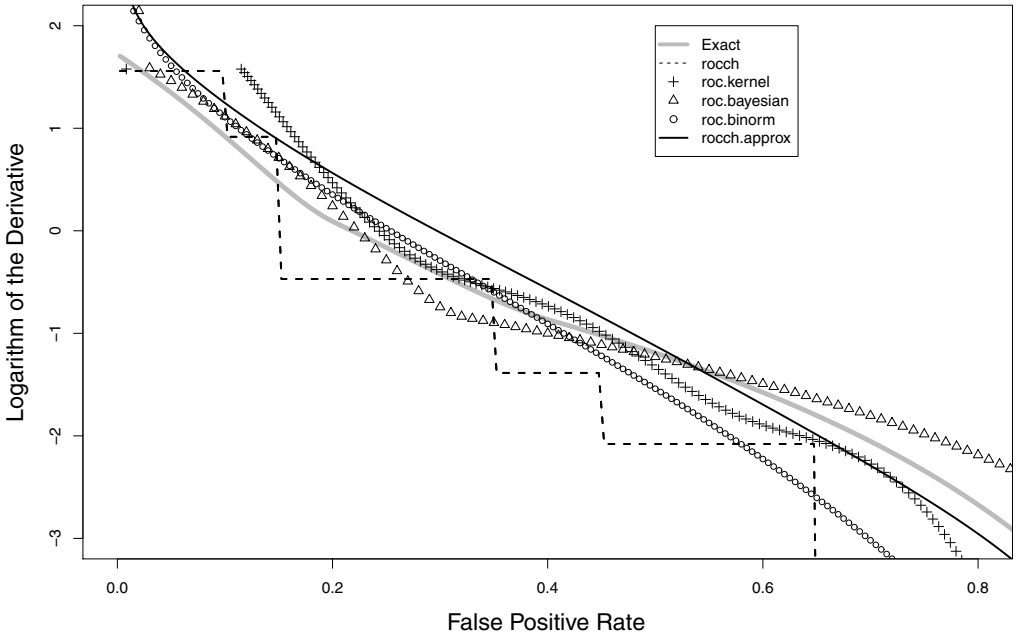

Figure 3: Precision of the ROC derivative approximation. The gray line gives the true derivative of the ROC curve in function on the false positive rate. The other lines represent the approximations of the derivative given by the different methods.

test set to compute the performance of the classifier. Figure 4 gives the loss of the classifier in function of the size of the validation set. For all methods, the loss is decreasing with the number of validation examples. The ROC rule has a much higher loss than other methods for small validation set (i.e. $N_{valid} < 100$). For small validation set the ROC convex hull curve contains few points, so the number possible abstaining area is very small, it is unlikely that one of this abstaining area is close to the optimal one. When the number of validation examples increases, the approximation of the ROC derivative becomes accurate for all methods and their performances are similar. The

performance of the exhaustive search does not depend on the ROC derivative approximation but on the estimation of the expected loss on the validation set. The larger the validation set, the more accurate the loss estimation and the abstaining area. If we plot the loss in function on the training set size, we will obtain the same kind of graphic than in figure 4. The loss is decreasing with the number of training examples. The performance of an abstaining classifier thus depends on both the training and the validation set. On real data, we do not have a training and a validation set, but only one dataset. We have to split this dataset into a training and a validation set. For the next experiments, we split the original data into a training and a validation set of the same size.

4.3. Impact of the cost of rejection

The abstaining area computation depends on the cost of rejecting an example λ_R. In this experiment, we check if the comparison of the different methods is robust to the value of λ_R. We compute the loss of all methods in function of the value of λ_R on the different datasets and different error costs for each datasets. We show in this section a representative example of the results obtained on this set of experiments. This example is on the Alon dataset, the cost of good classification is 0 ($\lambda_{PP} = \lambda_{NN} = 0$), the cost of error is 1 ($\lambda_{PN} = \lambda_{NP} = 1$). We remind that the truly important magnitude is the relative value of the cost of a given type of classification with respect to the cost of the remaining classifications. Here we fix the cost of

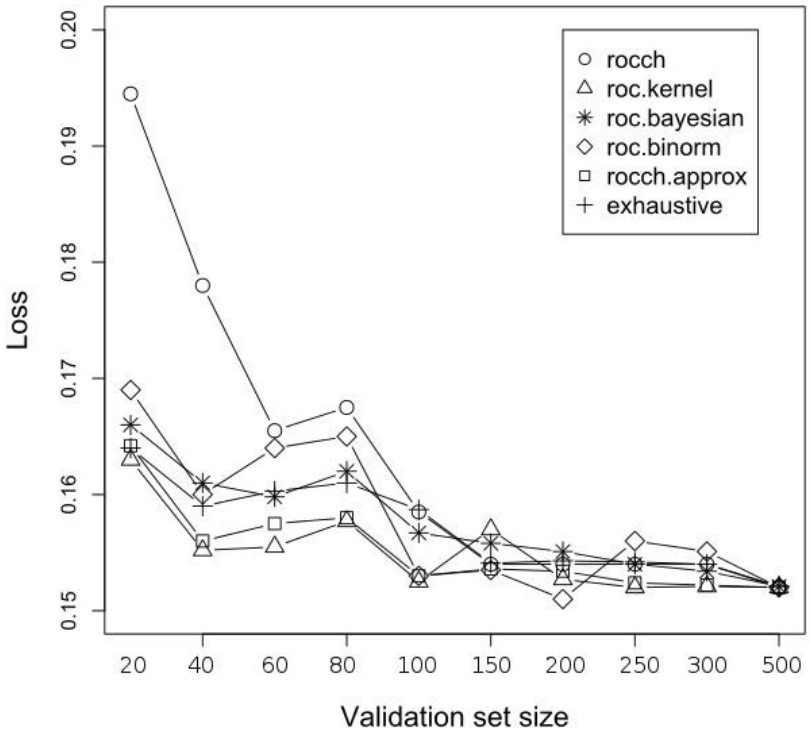

Figure 4: The loss of an abstaining classifier in function of the size of the validation set.

$\lambda_{PN}, \lambda_{NP}, \lambda_{PP}, \lambda_{NN}$ and we vary the rejection cost λ_R in order to vary the ration of the rejection cost on the error cost. The expected loss is estimated by a 10-times 10-fold cross-validation. Although the cross-validation presents some problems in small sample data, it is still one of the best available estimation method [11]. At each cross-validation iteration, 1/10 of the examples are selected to form the test set. The subset of examples not selected for the test, is split into a training and a validation set of equal size. A t-test based

feature selection is applied on the training set in order to reduce the data to the 100 most discriminant genes. The model is fitted on the training set and the abstaining area is computed on the validation set. Note that all these steps are included in the cross-validation procedure in order to avoid the problem of estimation bias [2]. The used classification rules are the linear discriminant analysis (LDA), random forest (given 500 trees) and support vector machine with linear kernel (given $C = 1$).

Figure 5 gives the loss in function on the value of the cost of rejection λ_R. For all methods, the loss is increasing with the value of λ_R, that makes sense because the cost are not normalized, when the cost of rejection increase, the cost of error is still constant, so the overall cost increases. Note that if we want to compare classifiers with different costs, we have to normalize the costs such that $\sum_i \lambda_i = 1$. At $\lambda_R = 0.7$ all methods, except the Chow's rule, give the same loss because the rejection cost is so high that no example is rejected, this corresponds to the performance of classic classifier with no abstaining area. The Chow's rule gives different results because it does not need a validation set, all examples are used for the learning of the model. At $\lambda_R = 0$ (not shown on figure) all examples are rejected the expected loss of all classifiers is therefore 0. These two points are trivial; the range of interest is between them. We see that the ranking of the methods is rather stable and does not depend on the λ_R. The choice of λ_R depends totally on the context of the classification problem, it should be defined in interaction with the biologists and physicians. In the next experiments, we have fixed the

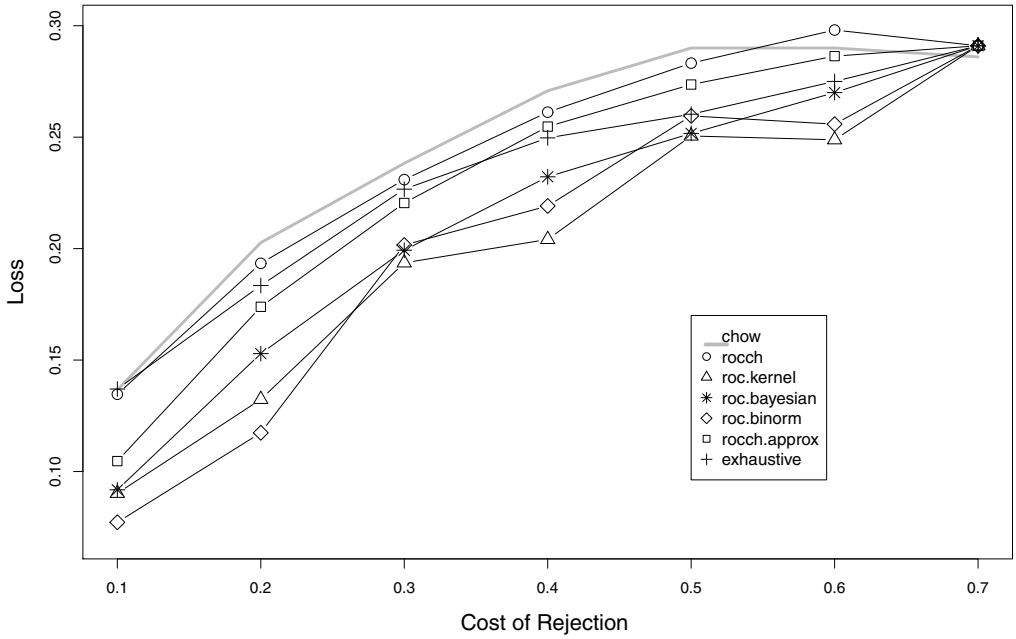

Figure 5: Performance of the abstaining classifier depending on the cost of rejection and the method used.

rejection cost at $\lambda_R = 0.3$.

4.4. Results on real data

We made an empirical comparison of the performance of the different methods on real small sample datasets. Medical decision based on genomics data is one of the main application of reject classification in small-sample settings. For the experiments, we choose eight microarray datasets covering a wide range of size and dimensionality. The choice of the cost of false positive and negative depends on the context. Alon (Colon cancer), Singh (prostate cancer) and Gordon (lung cancer) datasets are related to the problem of

cancer detection. The consequence of the non-detection of an actual cancer is more dramatic than the false cancer detection. Thus the cost of false positive is considered higher than that of false negative. In Golub (leukemia), Alizadeh (leukemia), Batthacharjee (lung cancer) and West (breast cancer) the objective is to identify the type of cancer. All these types of cancer have a priori the same seriousness. Thus for these tasks, the cost of false positive and negative have been considered equal. In VandeVijer (breast cancer) the task is to predict the outcome of cancer (be it good or poor), in both cases an error of prognostic (FP or FN) may lead to wrong therapeutic decisions whose consequences are difficult to a priori order like in the task of cancer detection. The cost of false positive and negative are therefore equal. The cost of good classification is fixed to $\lambda_{PP} = \lambda_{NN} = 0$ and cost of rejection to $\lambda_R = 0.3$. The value assigned to $\lambda_R = 0.3$ is a relative value that must be considered in comparison with the value $\lambda_{PN} = \lambda_{NP} = 1$ assigned to the cost of the errors and to the value $\lambda_{PP} = \lambda_{NN} = 0$ assigned to the cost of the correct classifications. The characteristics of all datasets are presented in Table 2. The loss of each abstaining classifier is estimated by a 10-times 10-fold cross-validation procedure as explained in section 4.3.

Table 4.4 gives the expected loss for all methods. We see that, in almost all cases, the performance of reject classifiers is much better than the non-reject classifiers. Unlike the conclusion of several papers [17, 28], the Chow's rule gives better performance than ROC rule in small sample data. Chow's rule outperforms ROC rule in 18 experiments out of 24. In some experiments,

Table 2: Characteristics of the eight real datasets.

Dataset	#examples	#features	λ_{PN}	λ_{NP}	class P	class N
Gordon	181	1626	2	1	healthy	tumor
Alon	62	2000	2	1	healthy	tumor
Singh	77	798	2	1	healthy	tumor
Golub	72	7129	1	1	AML	ALL
Alizadeh	42	1095	1	1	DLBCL1	DLBCL2
Batthacharjee	203	12600	1	1	adenocarcinoma	other types
West	49	7129	1	1	ER+	ER-
Van de Vijver	295	7129	1	1	good	poor

Table 3: Expected loss of the abstaining classifiers depending on the different methods on the eight real datasets. The best performances are in bold. A double line separates our five proposed methods roc kernel, roc bayesian, roc binorm, rocch approach and exhaustive

Method	Golub	Alizadeh	Gordon	Alon	Singh	Van de Vijver	Bhattacharjee	West
Support Vector Machine								
No rejection	0.210	0.233	0.080	0.321	0.224	0.365	0.141	0.297
Chow's rule	0.161	0.145	0.047	0.265	0.154	0.298	0.095	0.165
roc rule	0.152	0.196	0.071	0.274	0.174	0.301	0.111	0.232
roc kernel	0.133	0.117	**0.011**	0.255	0.140	0.312	0.089	**0.156**
roc bayesian	0.142	0.117	0.015	**0.249**	0.152	0.300	0.084	0.161
roc binorm	0.140	0.152	0.017	0.269	0.167	0.307	0.096	0.168
rocch approx	0.142	0.158	0.022	0.263	0.163	**0.296**	0.094	0.161
exhaustive	**0.117**	**0.112**	0.017	0.275	**0.137**	0.297	**0.091**	0.167
Linear Discriminant Analysis								
No rejection	0.201	0.211	0.092	0.367	0.311	0.362	0.154	0.264
Chow's rule	0.164	0.193	0.051	0.300	0.227	0.351	0.097	0.183
roc rule	0.150	0.260	0.078	0.256	0.224	0.317	0.117	0.255
roc kernel	0.144	0.162	**0.035**	0.251	0.220	0.301	0.084	**0.162**
roc bayesian	0.139	0.178	**0.035**	0.267	0.205	**0.294**	**0.081**	0.178
roc binorm	0.144	**0.150**	0.040	0.283	0.199	0.305	0.089	0.170
rocch approx	**0.131**	0.209	0.041	0.257	0.203	0.301	0.091	0.200
exhaustive	0.135	0.178	0.047	**0.245**	**0.196**	0.306	0.093	0.163
Random Forest								
No rejection	0.250	0.245	0.029	0.314	0.250	0.303	0.132	0.192
Chow's rule	0.169	0.207	0.016	0.239	0.152	0.262	0.076	0.162
roc rule	0.261	0.221	0.015	0.261	0.154	0.261	0.091	0.201
roc kernel	0.148	**0.133**	**0.011**	0.238	0.153	0.263	**0.061**	**0.128**
roc bayesian	**0.141**	0.148	0.012	**0.234**	0.147	0.274	0.070	0.145
roc.binorm	0.159	0.155	0.014	0.260	0.144	0.271	0.071	0.161
rocch approx	0.179	0.179	0.015	0.245	0.148	**0.259**	0.069	0.150
exhaustive	0.151	0.143	0.014	0.258	**0.132**	0.261	0.066	0.149

the ROC rule gives an expected loss much higher than the other methods, for example in West dataset with LDA, Alizadeh dataset with SVM or Golub

Table 4: P-values of the Wilcoxon signed-rank test where the alternative hypothesis is the methods gives better results than the Chow's rule and ROC rule.

	Chow rule			ROC rule		
Method	SVM	LDA	RF	SVM	LDA	RF
roc kernel	0.078	0.008	0.06	0.039	0.008	0.023
roc bayesian	0.025	0.008	0.08	0.008	0.023	0.039
roc binorm	0.679	0.004	0.230	0.011	0.019	0.025
rocch approx	0.181	0.070	0.156	0.004	0.008	0.011
exhaustive	0.055	0.004	0.055	0.008	0.007	0.011

dataset with random forest. this kind of failure of the ROC rule occurs when both very few points from the ROC convex hull (3 or 4 points) and the shape of ROC convex hull is very different from the shape of the true ROC curve, as illustrated in figure 1. An important result coming out of this table is that our five methods give much better performances than the two state-of-the-art methods whatever the dataset or the classification rules. Roc.kernel, roc.bayesian, roc.binorm, rocch.approx and exhaustive outperforms Chow's rule and ROC rule in all experiments. Roc.kernel gives the best results in 6 experiments, roc.bayesian in 5 experiments, roc.binorm in 1 experiment, rocch.approx in 4 experiments and exhaustive in 8 experiments. According to the table 4.4 our methods improve the performance of the state of the art by 30% in average. For the VandeVijver dataset the difference of performance between the state of the art and our methods is very small, the reason comes from the number of examples. This dataset contains 295 examples, the state of the art methods becomes good with this number of examples, as shown the figure 4. A Wilcoxon signed-rank test is performed to check the significance

of the improvement of our methods compared to the Chow's rule and ROC rule for each classifier. The alternative hypothesis is that our method gives better performance than state-of-the-art methods on all datasets. Table 4.4 gives the p-value of these statistical tests for each classification method. The p-value of the tests against Chow's rule are higher than against ROC rule, however in all case our methods are significantly better than the state-of-the-art. The only exception is *Roc binorm* against Chow's rule for SVM. There is no criterion to choose the best method among these three for a given dataset or classification rule. It only depends on the precision of the distribution of the classes on the validation set. However, we can give the following practical recommendation: once a model has been fitted on the training set, plot the ROC curve and its convex hull curve from the validation set. If the ROC curve and its convex hull are very different, it is likely that the approximation of the derivative will be inaccurate. In this case, it is safer to use an exhaustive search. In the opposite case, it is better to use the approximation of ROC or ROCCH derivative. A noticeable drawback of the exhaustive search is its computing time, which is much higher than the other ones.

5. Conclusion

Abstaining classifiers are important extensions of the classic classifiers supporting the significant improvement of accuracy and reliability of predictions. The optimal abstaining area can theoretically be computed by two

methods: the Chow's rule based on the computation of the posterior probabilities and the ROC rule based on the derivative of the ROC curve. The literature shows that these two methods are equivalent in theory but that the ROC rule seems to give better performance in practice. In this paper, we show that in small sample problems the ROC convex hull is formed from very few points leading to a very small number of possible abstaining areas. The constructed abstaining area is often far from the optimal one. We have proposed in this paper five new methods of abstaining area construction adapted to small sample problem. Our methods (excepted the exhaustive method) are theoretically equivalent to both Chow and Roc rule. The differences between our methods and the state of the art are related to the ROC curve estimation. Our results show that our methods give a better approximation that both the ROC rule and Chow's rule. Our methods are adapted to small sample problem since with small validation set we obtain much better performances that the ROC rule. Our results are robust with respect to the cost of rejection. On eight real datasets and three classification rules, we show that our methods give better performance that both Chow's rule and ROC rule. Finally, the proposed methods are easy to implement and can be used with any domain of application or classification rule. In summary, the strength of our methods is clearly to address the high-dimensional problem where N¡¡p and especially when the number of observations N is small. Theirs weakness is that whenever the sample is large enough its gain may be not significant. They should improve the reliability of the classifiers for real world small sam-

ple applications that are increasingly available in bioinformatics and medical applications.

Even when state of the art machine learning methods produce classifiers that have good generalization accuracy, they are often difficult to be used in routine by clinician. In effect, for clinician it is often difficult to decide how confident they can be in the classifier predictions. As claimed by Pepe [20] "new clinical classifiers lag far behind the well established standards that exist for evaluating new clinical treatments. Indeed, a diagnostic or a choice of therapeutic strategy must be based on a very high confidence classifier; an error of the predictive model may lead to tragic consequences. In other words if a classical classifier is used in a hospital cancer department to identify the lymphoblastic from the myelogenous leukemia of patients suffering from acute leukemia there is no consensus on the accuracy threshold that is considered to be reliable. If a classifier gives a probability 60% for lymphoblastic and 40% for myelogenous for a given patient; although, the probability of myelogenous is the highest, it is unlikely that a clinician will have confidence in the classifier assignment to the highest confidence prediction i.e. here lymphoblastic) to the patients cancer. This is especially true in medicine where observations available for learning are scarce and costly. On the contrary an abstaining classifier will reject the patient because in this case no reliable diagnosis can be done. This implies that clinician may have much higher confidence in predictions when the patient is not rejected.

In future work we will use classifiers with reject option for the construc-

tion of cascade of classifiers in order to reduce the acquisition cost. The acquisition cost is what you pay to obtain the variables of an example. It can be money, time, memory, or any other non-infinite resource. In most prediction problems, some examples are easier to predict than others. They can be predicted in using fewer variables i.e. with a lower acquisition cost. A set of abstaining classifiers with different acquisition cost could be combined into a cascade in increasing acquisition cost. To compute a prediction, the example is sent to the first abstaining classifier. If the first classifier rejects the example, then it is sent to the second classifier of the cascade that has a higher acquisition cost than the first classifier. This principle is repeated until the last classifier using all variables. The key problem of this model is the simultaneous computation of the abstaining area of all classifiers of the cascade.

References

[1] Naomi Altman and Christian Leger. Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46(2):195–214, 1995.

[2] C. Ambroise and G. McLachlan. Selection bias in gene extraction on the basis of microarray gene expression data. *Proc. Natl. Acad. Sci.*, 99(10):6562–6566, may 2002.

[3] C.K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

[4] U.R. Devarakota, B. Mirbach, and B. Ottersten. Reliability estimation of a statistical classifier. *Pattern recognition letters*, 29(3):243–253, 2008.

[5] D. D. Dorfman and E. Alf. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals: Rating method data. *Journal of Mathematical Psychology*, 6:487–496, 1969.

[6] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 99:1605–1641, 2010.

[7] Alaattin Erkanli, Minje Sung, E Jane Costello, and Adrian Angold. Bayesian semi-parametric roc analysis. *Statistics in medicine*, 25(22):3905–3928, 2006.

[8] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letter*, 27(8):861–874, 2006.

[9] Luzia Gonçalves, Ana Subtil, M Rosário Oliveira, and Patricia de Zea Bermudez. Roc curve estimation: an overview. *REVSTAT–Statistical Journal*, 12(1):1–20, 2014.

[10] B. Hanczar and E.R. Dougherty. Classification with reject option in gene expression data. *Bioinformatics*, 24(17):1889–1895, September 2008.

[11] B. Hanczar and E.R. Dougherty. On the comparison of classifiers for microarray data. *Current Bioinformatics*, 5(1):29–39, 2010.

[12] J. A. Hanley. The robustness of the binormal assumptions used in fitting roc curves. *Medical Decision Making*, 8:197–203, 1988.

[13] Lars Kai Hansen, Christian Liisberg, and Peter Salamon. The error-reject tradeoff. *Open Systems & Information Dynamics*, 4:159–184, 1997.

[14] Lawrence Kelley and Michael Scott. The evolution of biology. a shift towards the engineering of prediction-generating tools and away from traditional research practice. *EMBO reports*, 9(12):1163–1167, 2008.

[15] Wojtek J Krzanowski and David J Hand. *ROC curves for continuous data*. CRC Press, 2009.

[16] Chris J Lloyd and Zhou Yong. Kernel estimators of the roc curve are better than empirical. *Statistics & Probability Letters*, 44(3):221–228, 1999.

[17] Claudio Marrocco, Mario Molinara, and Francesco Tortorella. An empirical comparison of ideal and empirical roc-based reject rules. In *Proceedings of the 5th international conference on Machine Learning and Data Mining in Pattern Recognition*, MLDM '07, pages 47–60, 2007.

[18] Charles E. Metz and Xiaochuan Pan. Proper binormal roc curves: the-

ory and maximum-likelihood estimation. *J. Math. Psychol.*, 43(1):1–33, 1999.

[19] M.S.A. Nadeem, JD. Zucker, and B. Hanczar. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. *Journal of Machine Learning Research - Proceedings Track*, 8:65–81, 2010.

[20] MS Pepe. Evaluating technologies for classification and prediction in medicine. *Statistics in medicine*, 24(24):3687–3696, 2005.

[21] Tadeusz Pietraszek. Optimizing abstaining classifiers using roc analysis. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 665–672, 2005.

[22] Tadeusz Pietraszek. On the use of roc analysis for the optimization of abstaining classifiers. *Machine Learning*, 68(2):137–169, August 2007.

[23] Carla M. Santos-Pereira and Ana M. Pires. On optimal reject rules and roc curves. *Pattern Recognition Letters*, 26(7):943–952, 2005.

[24] F. Tortorella. A roc-based reject rule for dichotomizers. *Pattern Recognition Letters*, 26(2):167–180, 2005.

[25] Francesco Tortorella. An optimal reject rule for binary classifiers. In *Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition*, pages 611–620, 2000.

[26] M. van de Vijver. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med.*, 347:1999–2009, 2002.

[27] Stijn Vanderlooy, Ida G. Sprinkhuizen-Kuyper, Evgueni N. Smirnov, and H. Jaap van den Herik. The roc isometrics approach to construct reliable classifiers. *Intellence Data Anaysis.*, 13(1):3–37, 2009.

[28] Jigang Xie, Zhengding Qiu, and Jie Wu. Bootstrap methods for reject rules of fisher lda. In *Proceedings of the 18th International Conference on Pattern Recognition*, volume 3 of *ICPR '06*, pages 425–428, 2006.