An Empirical Comparison of Ideal and Empirical ROC-Based Reject Rules

Claudio Marrocco, Mario Molinara, and Francesco Tortorella

DAEIMI, Università degli Studi di Cassino
Via G. Di Biasio 43, 03043 Cassino (FR), Italia
{c.marrocco,m.molinara,tortorella}@unicas.it

Abstract. Two class classifiers are used in many complex problems in which the classification results could have serious consequences. In such situations the cost for a wrong classification can be so high that can be convenient to avoid a decision and reject the sample. This paper presents a comparison between two different reject rules (the Chow's and the ROC rule). In particular, the experiments show that the Chow's rule is inappropriate when the estimates of the a posteriori probabilities are not reliable.

Keywords: ROC curve, reject option, two-class classification, cost-sensitive classification, decision theory.

1 Introduction

Frequently, in two class classification problems, the cost for a wrong classification could be so high that it should be convenient to introduce a reject option [1]. This topic has been addressed by Chow in [2]. The rationale of the Chow's approach relies on the exact knowledge of the a posteriori probabilities for each sample to be recognized. Under this hypothesis, the Chow's rule is optimal because minimizes the error rate for a given reject rate (or vice versa). For the two-class classification cases in which the ideal setting assumed by the Chow's rule is not guaranteed and a real classifier must be used, an alternative method has been proposed in [3] where the information provided about the classifier performances by the empirical ROC curve is used to draw a reject rule which minimizes the expected cost for the application at hand. In [4] a review of the reject rule based on the empirical ROC and a comparison with the Chow's rule is presented; in the paper the authors claim to demonstrate the theoretical equivalence between the two rules and suggest that the Chow's reject rule should produce lower classification costs than those obtained by means of the reject rule based on the empirical ROC, even when real classifiers are employed.

A first comparison between the two approaches has been already proposed in [5] with reference to Fisher LDA. The experiments presented show that the empirical ROC reject rule works better than the Chow's rule in the majority of the cases considered.

P. Perner (Ed.): MLDM 2007, LNAI 4571, pp. 47–60, 2007.

In this paper we aim to analyze more extensively the two rules and to compare, by means of thorough experiments, their behavior in order to demonstrate that the Chow's rule is inappropriate when the the distributions of the two classes are not perfectly known. In the next sections we resume the main features of the two reject rules while in the last section the experiments performed on both artificial and real data sets are reported.

2 Two Class Classification and the ROC Curve

2.1 The Ideal Case

In two class classification problems, the goal is to assign a pattern $\mathbf{x}$ coming from an instance space X to one of two mutually exclusive classes that can be generically called Positive (P) class and Negative (N) class; in other words, $X = P \cup N$ and $P \cap N = \emptyset$. Let us firstly consider a typical Decision Theory scenario and suppose to have a complete knowledge of the distributions of the samples within X, i.e. we know the a priori probabilities of the two classes (π_P, π_N) and the class conditional densities $f_P(\mathbf{x}) = p(\mathbf{x}\,|\mathbf{x} \in P)$ and $f_N(\mathbf{x}) = p(\mathbf{x}\,|\mathbf{x} \in N)$. If $\begin{bmatrix} \lambda_{NN} & \lambda_{NP} \\ \lambda_{PN} & \lambda_{PP} \end{bmatrix}$ is the cost matrix defined for the problem at hand (where λ_{AB} is the cost of assigning a pattern to the class B when it actually belongs to the class A), the conditional risk associated to the classification of a given sample $\mathbf{x}$ is minimized by a decision rule which assigns the sample $\mathbf{x}$ to the class P if

$$lr(\mathbf{x}) = \frac{f_P(\mathbf{x})}{f_N(\mathbf{x})} > \frac{(\lambda_{NP} - \lambda_{NN})\,\pi_N}{(\lambda_{PN} - \lambda_{PP})\,\pi_P}$$

where $lr(\mathbf{x})$ is the likelihood ratio evaluated for the sample $\mathbf{x}$. A way to assess the quality of such rule as the costs and the a priori probabilities vary, is to evaluate the performance obtained on each class by the rule $lr(\mathbf{x}) > t$ as the threshold t is varied. For a given threshold value t, two appropriate performance figures are given by the *True Positive Rate* $TPR(t)$, i.e. the fraction of actually-positive cases correctly classified and by the *False Positive Rate* $FPR(t)$, given by the fraction of actually-negative cases incorrectly classified as "positive". If we consider the class-conditional densities of the likelihood ratio $\varphi_P(\tau) = p(lr(\mathbf{x}) = \tau\,|\mathbf{x} \in P)$ and $\varphi_N(\tau) = p(lr(\mathbf{x}) = \tau\,|\mathbf{x} \in N)$, $TPR(t)$ and $FPR(t)$ are given by:

$$TPR(t) = \int_t^{+\infty} \varphi_P(\tau)\,d\tau \qquad FPR(t) = \int_t^{+\infty} \varphi_N(\tau)\,d\tau \tag{1}$$

Taking into account the samples with likelihood ratio less than t, it is possible to evaluate the *True Negative Rate* $TNR(t)$ and the *False Negative Rate* $FNR(t)$, defined as:

$$\begin{aligned} TNR(t) &= \int_{-\infty}^{t} \varphi_N(\tau)\,d\tau = 1 - FPR(t) \\ FNR(t) &= \int_{-\infty}^{t} \varphi_P(\tau)\,d\tau = 1 - TPR(t) \end{aligned} \tag{2}$$

As it is possible to note from eq.(2), the four indices are not independent and the pair $(FPR(t), TPR(t))$ is sufficient to completely characterize the performance of the decision rule for a given threshold t. Most importantly, they are independent of the a priori probability of the classes because they are separately evaluated on the different classes. The *Receiver Operating Characteristic* (ROC) curve plots $TPR(t)$ vs. $FPR(t)$ by sweeping the threshold t into the whole real axis, thus providing a description of the performance of the decision rule at different operating points. An important feature of the ROC curve is that the slope of the curve at any point $(FPR(t), TPR(t))$ is equal to the threshold required to achieve the FPR and TPR of that point [6]. Therefore, the corresponding operating point on the ROC curve is the one where the curve has gradient $\frac{(\lambda_{NP}-\lambda_{NN})\pi_N}{(\lambda_{PN}-\lambda_{PP})\pi_P}$; such point can be easily found moving down from above in the ROC plane a line with slope $\frac{(\lambda_{NP}-\lambda_{NN})\pi_N}{(\lambda_{PN}-\lambda_{PP})\pi_P}$ and selecting the point in which the line touches the ROC curve [1]. The ROC generated by the decision rule based on the likelihood ratio is the optimal ROC curve, i.e. the curve which, for each $FPR \in [0, 1]$, has the highest TPR among all possible decision criteria employed for the classification problem at hand. This can be proved if we recall the *Neyman Pearson* lemma which can be stated in this way: if we consider the decision rule $lr(x) > \beta$ with β chosen to give $FPR = \varepsilon$, there is no other decision rule providing a TPR higher than $TPR(\beta)$ with a $FPR \leqslant \varepsilon$. The demonstration of the lemma can be found in [8,9]. The shape of the optimal ROC curve depends on how the class-conditional densities are separated: two perfectly distinguished densities produce an ROC curve that passes through the upper left corner (where $TPR = 1.0$ and $FPR = 0.0$), while the ROC curve generated by two overlapped densities is represented by a 45° diagonal line from the lower left to the upper right corner. Qualitatively, the closer the curve to the upper left corner, the more discriminable the two classes.

2.2 The Empirical Approach

The ideal scenario of the Bayesian Decision Theory considered so far unfortunately cannot be applied to the most part of real cases where we rarely have this kind of complete knowledge about the probabilistic structure of the problem. As a consequence, in real problems the optimal ROC is unknown since the actual class conditional densities are not known. In this case, the decision is performed by means of a trained classifier. Without loss of generality, let us assume that the classifier provides, for each sample $\mathbf{x}$, a value $\omega(\mathbf{x})$ in the range $(-\infty, +\infty)$ which can be assumed as a confidence degree that the sample belongs to one of the two classes, e.g. the class P. The sample should be consequently assigned to the class N if $\omega(\mathbf{x}) \to -\infty$ and to the class P if $\omega(\mathbf{x}) \to +\infty$. Also in this case it is possible to plot an ROC curve by considering the outputs provided by the trained classifier on a validation set V containing n_+ positive samples and n_- negative samples $V = \{p_i \in P, i = 1 \dots n_+\} \cup \{n_j \in N, j = 1 \dots n_-\}$. In this

way, we obtain an empirical estimator of the optimal ROC curve by evaluating, for each possible value of a threshold t in the range $(-\infty, +\infty)$, the empirical true and false positive rates as follows:

$$TPR(t) = \frac{1}{n_+} \sum_{i=1}^{n_+} S(\omega(p_i) > t) \qquad FPR(t) = \frac{1}{n_-} \sum_{j=1}^{n_-} S(\omega(n_j) > t)$$

where $S(.)$ is a predicate which is 1 when the argument is true and 0 otherwise. Let us call the obtained curve empirical ROC curve (see fig. 1) in order to distinguish it from the ideal ROC.

There are some differences to be highlighted between the empirical and the optimal ROC:

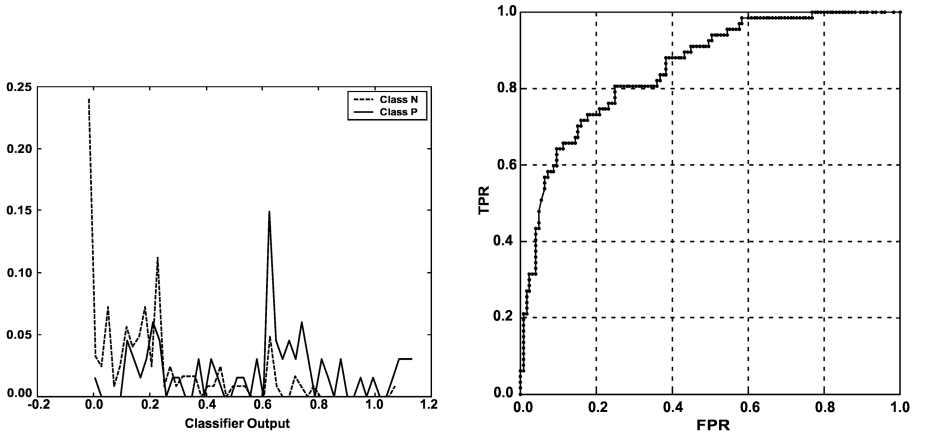
- once the two classes N and P have been specified through their conditional densities $f_P(\mathbf{x})$ and $f_N(\mathbf{x})$, the ideal ROC is unique, while different classifiers trained on the same problem have different empirical ROCs;
- for a continuous likelihood ratio, the ideal ROC is continuous and its slope in a particular point equals the value of the threshold required to achieve TPR and FPR of that point [1]; the empirical ROC is instead a discrete function and the relation between slope and threshold does not hold. However it is still possible to find the optimal operating point also on the empirical ROC by moving down from above in the ROC plane a line with slope and selecting the point in which the line touches the ROC curve. Provost and Fawcett [7] have shown that the point is one of the vertices of the convex hull which contains the empirical ROC curve (see fig. 2);
- the ideal ROC is the optimal ROC curve, i.e. the curve which, for each FPR [0,1], has the highest TPR among all possible decision criteria employed for the classification problem at hand. In other words, the ideal ROC curve dominates every empirical ROC and consequently has the highest *area under the ROC curve* (*AUC*) attainable.

3 Two-Class Classification with Reject

When dealing with cost sensitive applications which involve a reject option, the possible outcomes of the decision rule include the reject and thus the cost matrix changes accordingly (see tab. 1).

Table 1. Cost matrix for a two-class problem with reject

		Predicted Class		
		N	P	R
True Class	N	λ_{NN}	λ_{NP}	λ_R
	P	λ_{PN}	λ_{PP}	λ_R



(a) (b)

Fig. 1. (a) The densities of the confidence degree obtained by the classifier output on real data and (b) the corresponding ROC curve

It is worth noting that λ_{NN} (cost for a True Negative) and λ_{PP} (cost for a True Positive) are negative costs since related to benefits, while λ_{NN} (cost for a False Negative) and λ_{NP} (cost for a False Positive) are positive costs. λ_R weights the burden of managing the reject (e.g. by calling another, more proficient classifier) and thus it is positive but smaller than the error costs.

3.1 The Ideal Case

Let us firstly suppose that we are working within a Bayesian scenario, i.e. we know the a priori probabilities of the two classes (π_P, π_N) and the class conditional densities $f_P(\mathbf{x})$ and $f_N(\mathbf{x})$. In this ideal setting, the classification cost is minimized by the Chow's rule [2] which can be expressed in terms of the likelihood ratio $lr(\mathbf{x})$ as follows:

$$\begin{aligned} \mathbf{x} \rightarrow N \quad & \text{if} \;\; lr(\mathbf{x}) < \frac{\pi_N}{\pi_P} \frac{\lambda_R - \lambda_{NN}}{\lambda_{PN} - \lambda_R} = u_1 \\ \mathbf{x} \rightarrow P \quad & \text{if} \;\; lr(\mathbf{x}) > \frac{\pi_N}{\pi_P} \frac{\lambda_{NP} - \lambda_R}{\lambda_R - \lambda_{PP}} = u_2 \\ reject \quad & \text{if} \;\; u_1 \leq lr(\mathbf{x}) \leq u_2 \end{aligned} \tag{3}$$

The rule can be also defined in terms of the *a posteriori* probability $\Pr(P|\mathbf{x})$. If we recall that

$$\Pr(P|\mathbf{x}) = \frac{\pi_P f_P(\mathbf{x})}{\pi_N f_N(\mathbf{x}) + \pi_P f_P(\mathbf{x})} = \frac{\pi_P lr(\mathbf{x})}{\pi_N + \pi_P lr(\mathbf{x})}$$

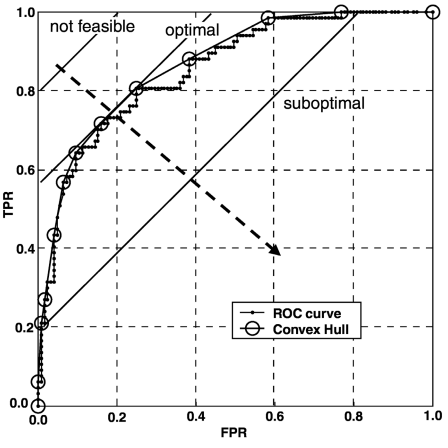


Fig. 2. The ROC curve shown in fig. 1 and its convex hull. Three level lines with the same slope are also shown: the line touching the ROC convex hull determines the optimal operating point since it involves the minimum risk. The line above the optimal line does not determine any feasible point, while the line below identifies only suboptimal points.

the Chow's rule can be written as:

$$\begin{aligned} \mathbf{x} \rightarrow N \quad & \text{if} \quad \Pr(P|\mathbf{x}) < \frac{\lambda_R - \lambda_{NN}}{\lambda_{PN} - \lambda_{NN}} = t_1 = \frac{\pi_P u_1}{\pi_N + \pi_P u_1} \\ \mathbf{x} \rightarrow P \quad & \text{if} \quad \Pr(P|\mathbf{x}) > \frac{\lambda_{NP} - \lambda_R}{\lambda_{NP} - \lambda_{PP}} = t_2 = \frac{\pi_P u_2}{\pi_N + \pi_P u_2} \\ reject \quad & \text{if} \quad t_1 \leq \Pr(P|\mathbf{x}) \leq t_2 \end{aligned} \tag{4}$$

It is worth noting that in the ideal scenario, the slope of the ROC curve at any point is equal to the threshold on the likelihood ratio which has generated that point [6], and thus the points corresponding to the two thresholds u_1 and u_2 can be easily identified on the ideal ROC.

3.2 The Empirical Approach

In real problems, however, the class conditional densities are not available and thus the optimal decision rule in 3 or in 4 cannot be applied. In such cases, the typical approach is to train a classifier $\omega(\mathbf{x})$ on a set of samples representative of the classes to be discriminated and to use it to estimate the class of new samples. Even though the Chow's rule cannot be directly applied, a reject option can be still defined on the empirical ROC, as it has been shown in [3]. The decision rule is still based on two thresholds τ_1 and τ_2 applied on the output of the classifier $\omega(\mathbf{x})$:

$$\begin{aligned} \mathbf{x} \rightarrow N \quad & \text{if} \quad \omega(\mathbf{x}) < \tau_1 \\ \mathbf{x} \rightarrow P \quad & \text{if} \quad \omega(\mathbf{x}) > \tau_2 \\ reject \quad & \text{if} \quad \tau_1 \leq \omega(\mathbf{x}) \leq \tau_2 \end{aligned} \tag{5}$$

As a consequence, the values of TPR and FPR change as:

$$TPR(\tau_2) = \frac{1}{n_+} \sum_{i=1}^{n_+} S(\omega(p_i) > \tau_2) \qquad FPR(\tau_2) = \frac{1}{n_-} \sum_{j=1}^{n_-} S(\omega(n_j) > \tau_2) \tag{6}$$

It is worth noting that the condition described in eq.(2) is no more satisfied since now there are two thresholds. In fact, the values of TNR and FNR are given by:

$$\begin{aligned} FNR(\tau_1) &= \frac{1}{n_+} \sum_{i=1}^{n_+} S(\omega(p_i) < \tau_1) \\ TNR(\tau_1) &= \frac{1}{n_-} \sum_{j=1}^{n_-} S(\omega(n_j) < \tau_1) \end{aligned} \tag{7}$$

Moreover, we have now a portion of samples rejected given by:

$$\begin{aligned} RP(\tau_1, \tau_2) &= \frac{1}{n_+} \sum_{i=1}^{n_+} S(\tau_1 \leqslant \omega(p_i) \leqslant \tau_2) = 1 - TPR(\tau_2) - FNR(\tau_1) \\ RN(\tau_1, \tau_2) &= \frac{1}{n_-} \sum_{j=1}^{n_-} S(\tau_1 \leqslant \omega(n_j) \leqslant \tau_2) = 1 - TNR(\tau_1) - FPR(\tau_2) \end{aligned} \tag{8}$$

As a consequence, the classification cost obtained when imposing the threshold τ_1 and τ_2 is given by:

$$\begin{aligned} C(\tau_1, \tau_2) = \pi_P \cdot \lambda_{PN} \cdot FNR(\tau_1) + \pi_N \cdot \lambda_{NN} \cdot TNR(\tau_1) + \\ \pi_P \cdot \lambda_{PP} \cdot TPR(\tau_2) + \pi_N \cdot \lambda_{NP} \cdot FPR(\tau_2) + \\ \pi_P \cdot \lambda_R \cdot RP(\tau_1, \tau_2) + \pi_N \cdot \lambda_R \cdot RN(\tau_1, \tau_2) \end{aligned} \tag{9}$$

The values of the thresholds should be chosen in order to minimize $C(\tau_1, \tau_2)$; to this aim the classification cost can be written as:

$$C(\tau_1, \tau_2) = k_2(\tau_2) - k_1(\tau_1) + \pi_P \cdot \lambda_{PN} + \pi_N \cdot \lambda_{NN} \tag{10}$$

where:

$$\begin{aligned} k_1(\tau_1) &= \pi_P \cdot \lambda'_{PN} \cdot TPR(\tau_1) + \pi_N \cdot \lambda'_{NN} \cdot FPR(\tau_1) \\ k_2(\tau_2) &= \pi_P \cdot \lambda'_{PP} \cdot TPR(\tau_2) + \pi_N \cdot \lambda'_{NP} \cdot FPR(\tau_2) \end{aligned}$$

and

$$\begin{aligned} \lambda'_{PP} &= \lambda_{PP} - \lambda_R \quad & \lambda'_{PN} &= \lambda_{PN} - \lambda_R \\ \lambda'_{NN} &= \lambda_{NN} - \lambda_R \quad & \lambda'_{NP} &= \lambda_{NP} - \lambda_R \end{aligned}$$

In this way, the optimization problem can be simplified and the optimal values for thresholds τ_{1opt} and τ_{2opt} can be separately obtained by maximizing $k_1(\tau_1)$ and minimizing $k_2(\tau_2)$:

$$\begin{aligned} \tau_{1opt} &= \arg\max_{\tau} \left[\pi_P \cdot \lambda'_{PN} \cdot TPR(\tau_1) + \pi_N \cdot \lambda'_{NN} \cdot FPR(\tau_1)\right] \\ \tau_{2opt} &= \arg\min_{\tau} \left[\pi_P \cdot \lambda'_{PP} \cdot TPR(\tau_2) + \pi_N \cdot \lambda'_{NP} \cdot FPR(\tau_2)\right] \end{aligned} \tag{11}$$

As described in [3], the optimal thresholds can be found by considering the empirical ROC evaluated on a validation set; in particular, they correspond to the points T_1 and T_2 of the empirical ROC touched by two lines with slopes:

$$m_1 = -\frac{\pi_N}{\pi_P}\frac{\lambda_{NN}-\lambda_R}{\lambda_{PN}-\lambda_R} \qquad m_2 = -\frac{\pi_N}{\pi_P}\frac{\lambda_{NP}-\lambda_R}{\lambda_{PP}-\lambda_R} \tag{12}$$

and the thresholds are the values of the confidence degree which have generated those points (see fig. 3).

It is worth noting that τ_{1opt} must be less than τ_{2opt} to achieve the reject option, i.e. the slopes must be such that $m_1 < m_2$. If $m_1 \geqslant m_2$, the reject option is not practicable and thus the best choice is to work at 0 reject. Taking into account eq.(12), the condition for the reject option to be applicable is $\frac{\lambda_{NN}-\lambda_R}{\lambda_{PN}-\lambda_R} > \frac{\lambda_{NP}-\lambda_R}{\lambda_{PP}-\lambda_R}$. This condition depends only on the cost values; however, there could be situations in which the condition is verified but the geometry of the ROC curve of the classifier at hand is such that the level curves corresponding to m_1 and m_2 touch the same point [3]. In other words, in spite of the costs which could allow the reject, the characteristics of the classifier could make not applicable the reject option.

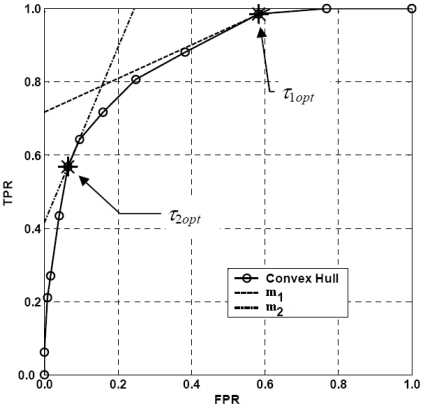


Fig. 3. The ROC curve, the level curves and the optimal thresholds for a given cost combination

3.3 Ideal and Empirical ROC Reject Rules. Are They Equivalent?

It is worth noting that, in the empirical ROC, the values of the thresholds are not an immediate function of the slopes (like in the ideal case) but the relation between the slopes (and the costs) and the threshold values is provided by the geometry of the ROC curve and after all by the output of the classifier. As a consequence, such values change when considering a different classifier.

The empirical rule reduces to the Chow's rule when dealing with the ideal ROC instead of an empirical ROC. In fact, in the ideal case, the lines with slopes in eq. (12) identify two points in which the likelihood ratio has the same value of the two slopes. In other words, $u_2 = m_2$, $u_1 = m_1$ and the reject rule in

eq. (5) reduces to the reject rule in eq. (3). This means that the empirical rule is certainly suboptimal with respect to the Chow's rule, but this latter does not work when the ideal setting assumptions (i.e. that the distributions of the two classes are completely known) are not verified and a real dichotomizer should be used instead of the optimal decision rule based on the likelihood ratio or on the *aposteriori* probabilities.

However, on the basis of this observation, authors in [4] claim to demonstrate the theoretical equivalence between the two rules and suggest that the adoption of the ideal thresholds in eq.(3) (or thresholds derived from those in eq. (4) if post probabilities are adopted in the decision rule instead of the likelihood ratio) should produce lower classification costs than those obtained by means of the reject rule based on the empirical ROC. This would mean that the reject thresholds are independent of the classifier chosen or, in other words, that every real classifier can be considered an effective estimator of the true likelihood ratio or of the true post probabilities.

Such assumption does not hold at all for a large class of classification systems such as margin based classifiers which do not provide an estimate of the post probabilities. In these cases the Chow's rule does not work, while the reject rule based on the empirical ROC is still applicable (see, e.g., [10]). However, such assumption seems to be excessive even for classification systems which provide estimates of the likelihood ratio or of the post probabilities (e.g. Multi Layer Perceptrons), since there are many elements (e.g. the limited size of the training set, the learning parameters which are not univocal, etc.) which make the estimate not very accurate.

In particular, the differences between the two rules should be higher and higher in favor of the empirical ROC-based rule as the ideal setting assumption becomes less verified. To experimentally prove such hypothesis, we have designed a set of experiments using both synthetic data sets and real data sets. The synthetic data sets are built by adding noise to the post probabilities generated from some chosen distributions. The aim is to simulate in a controlled way a realistic situation, i.e. a classification problem in which the post probabilities cannot be exactly obtained since the distributions of the two classes are not completely known and must be estimated by means of some method which inevitably provides a certain amount of error. For the real data sets, we train some well known classifiers and we use the outputs of the classifiers as estimates of the post probabilities. For each sample to be classified, the output is compared with the thresholds provided by the two rules thus obtaining two decisions which can be compared.

In the next section we present the methodology adopted in the experiments and the results obtained.

4 Experiments

4.1 Synthetic Data Set

To create an artificial problem, a gaussian model for the distribution of the samples of the two classes has been adopted. In particular, we simulate the

output of a classifier ω as $\omega(\mathbf{x}) = Pr(P|\mathbf{x}) + \varepsilon(\mathbf{x})$ where $Pr(P|\mathbf{x})$ is the a posteriori probability of the class P given the input vector $\mathbf{x}$ and $\varepsilon(\mathbf{x})$ is the error associated to that sample. In our framework the distribution of the two classes is supposed to be gaussian with known mean and covariance matrix $\Sigma = I$.

In particular, we generate the likelihood probabilities for the two classes P and N:

$$f_N(\mathbf{x}) = (2\pi)^{-K/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_N)^T(\mathbf{x}-\mu_N)\right)$$
$$f_P(\mathbf{x}) = (2\pi)^{-K/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_P)^T(\mathbf{x}-\mu_P)\right) \quad (13)$$

and from the Bayes theorem we find the a posteriori probability $Pr(P|\mathbf{x})$. Knowing the distribution of the samples it is possible to vary the vector of the mean μ, so as to create different data sets according to a value M that measures the distance between the means of the distributions of the two classes. In this paper, we considered three cases of interest: $M = 4.5$, i.e. the two classes are completely separated; $M = 3$, i.e. the two classes are partially overlapped and $M = 1.5$, i.e. the two classes are quite completely overlapped. Then, the term $\varepsilon(\mathbf{x})$ that simulates the error committed by a classifier is modeled according to two distributions: a gaussian distribution with zero mean and variance varying among 0 and 1 with step 0.1 and an uniform distribution varying among 0 and 1 with step 0.1. For each value of M 1000 samples have been generated and the distributions have been truncated so that each output ω is in the interval [0,1].

4.2 Real Data Set

Four data sets publicly available at the UCI Machine Learning Repository [11] have been used in the following experiments; all of them have two output classes and numerical input features. All the features were previously rescaled so as to have zero mean and unit standard deviation. More details for each data sets are reported in table 2. The employed classifiers are neural networks and Support Vector Machines (SVM). In particular, four Multi Layer Perceptron (MLP) with a variable number of units in the hidden layer between two and five have been trained for 10000 epochs using the back propagation algorithm with a learning rate of 0.01 and four Radial Basis Function (RBF) have been built with a variable number of units in the hidden layer between two and five. Then, four Support Vector Machine (SVM) with different kernels have been used;in particular, the kernels used were linear, polynomial of degree 2, RBF with $\sigma = 1$ and sigmoidal with $\sigma = 0.01$.

Table 2. Data sets used in the experiments

Data Sets	Features	Samples	% Major Class	Train Set	Valid. Set	Test Set
Pima	8	768	65.1	538	115	115
German	24	1000	70.0	700	150	150
CMC	9	1473	57.3	1031	221	221
Heart	13	303	54.1	213	45	45

4.3 Results

In the comparison of the two reject rules 12 runs of a multiple hold out procedure have been performed to reduce the bias in the data. In each run, each data set has been divided into three subsets: a training set used to train the classifiers, a validation set to evaluate the thresholds of the empirical reject rule and a test set to compare the two methods. In the experiments with artificial data we had only the validation and the test set containing respectively the 23% and the 77% of the whole data set. In the experiments with real data, the three subsets contain respectively the 70%, the 15% and the 15% of the samples of the whole data set. In this way, 12 different values of the required costs have been obtained.

Another possible cause of bias is given by the employed cost matrix. To achieve a result independent of the particular cost values, we have used a matrix (called cost model) in which each cell contains a distribution instead of a fixed value. In this way, 1000 different cost matrices have been randomly generated on the basis of the cost model adopted. In our experiments, an uniform distribution over the interval $[-10, 0]$ for λ_{PP} and λ_{NN}, over the interval $[0, 50]$ for λ_{NP} and λ_{PN} and over the interval $[0, 30]$ for the reject cost λ_R.

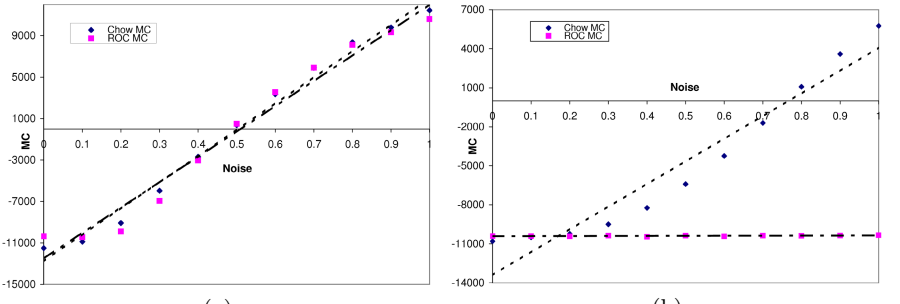


Fig. 4. Results obtained on artificial data sets: (a) $M = 4.5$ and additive Gaussian noise, (b) $M = 4.5$ and additive uniform noise

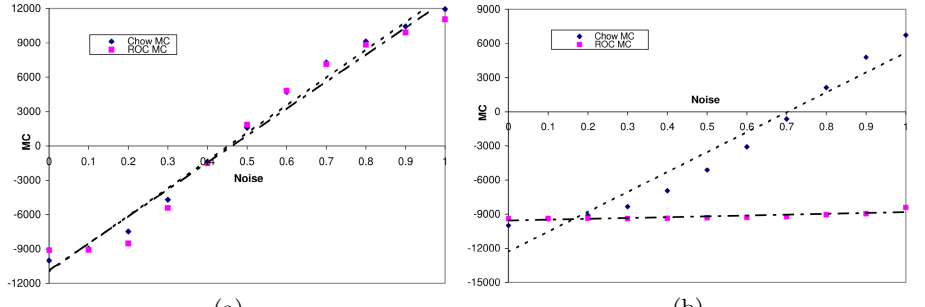


Fig. 5. Results obtained on artificial data sets: (a) $M = 3$ and additive Gaussian noise, (b) $M = 3$ and additive uniform noise

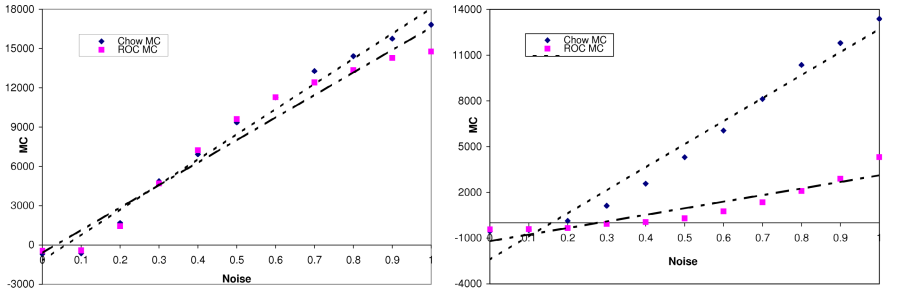


Fig. 6. Results obtained on artificial data sets: (a) $M = 1.5$ and additive Gaussian noise, (b) $M = 1.5$ and additive uniform noise

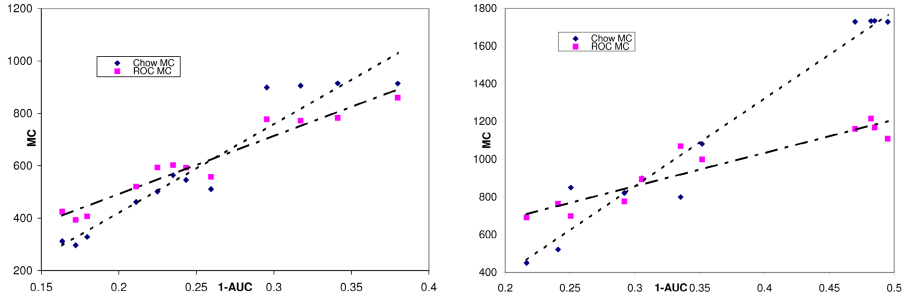


Fig. 7. Results obtained on real data sets: (a) Pima, (b) German

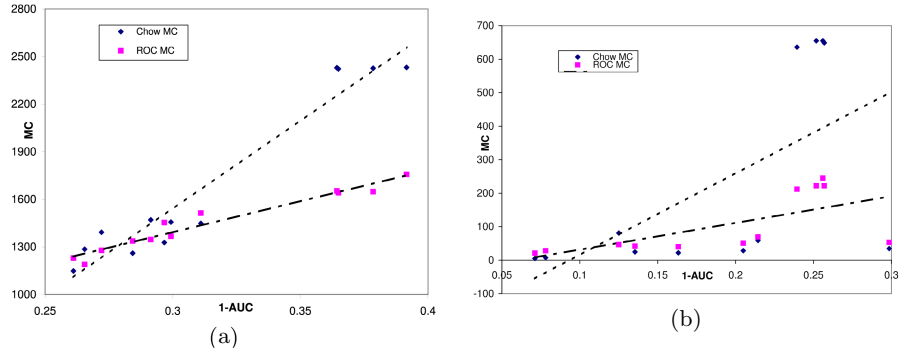


Fig. 8. Results obtained on real data sets: (a) CMC, (b) Heart

The obtained results are shown in figs. 4-6 for the synthetic data and in figs. 7-8 for the real data sets. In both cases we report the comparison in terms

of mean cost (MC) intended as the average of the classification costs obtained on the 12 runs of the hold out procedure and on the 1000 cost matrices employed on the considered problem. To order the classifier performance in the artificial case we refer to the added noise since we have a complete knowledge of the post probabilities and in particular we refer to the variance for the Gaussian distribution and to the width of the interval in the uniform distribution. However, while in the synthetic model we know the effective distribution of the classes and so it is possible to relate the obtained results to the noise added to the true post probabilities, when dealing with the real data sets we cannot do the same unless we have a measure of the accuracy with which the real classifier estimates the true post probabilities. To this aim, the AUC can be seen as a reliable estimate of the discriminating quality of the classifier [12]. Moreover, since the ideal ROC curve represents the "upper bound" of any empirical ROC curves (i.e. it is dominant with respect to any empirical ROC curve), we can reasonably assume that the greater is the AUC, the closer is the empirical to the ideal ROC curve and the better is the estimate of the true post probabilities. For this reason, in the graphs the value 1-AUC is reported on x-axis to be consistent with the previous figures. In each graph, beyond the scatter plot of the mean costs values also the regression lines are reported to emphasize the trend of the mean costs obtained by the two analyzed rules.

If we look at the behavior of the two rules on the synthetic data sets it is possible to note that the Chow reject rule outperforms the ROC rule only when the added noise is low. On the contrary, when the noise becomes higher the empirical ROC rule becomes better since the estimate of the post probabilities becomes worse and worse. This behavior is more visible when the added noise follows an uniform distribution (figs. 4-(b), 5-(b), 6-(b)) while a similar behavior is shown if the added noise is gaussian (figs. 4-(a), 5-(a), 6-(a)) because it produces less bias in the data.

The same behavior obtained on the artificial data is shown on the real data sets (figs. 7, 8) where the improvement obtained with the ROC rule is very evident when AUC decreases, i.e. when the classifier is not able to estimate a reliable post probability for the two classes and the ideal conditions are less verified.

5 Conclusions and Future Work

In this paper we have experimentally compared the Chow's reject rule and the ROC based reject rule presented in [3]. Despite what claimed in [4] we have found that the Chow's rule is inappropriate when the estimates of the a posteriori probabilities are not sufficiently accurate, while the ROC based reject rule gives good results. One could argue that such result could be not surprising, but we believe that the strong assertion about the robustness of the Chow's rule made in [4] is worth a critical analysis based on the evidence of specific experiments besides theoretical arguments. Finally, the analysis begun in this paper has pointed out the need of a further investigation to characterize the type of situations when one rule has advantage over another.

References

1. Webb, A.R.: Statistical Pattern Recognition. John Wiley and Sons Ltd, West Sussex (2002)
2. Chow, C.K.: On optimum recognition error and reject tradeoff. IEEE Trans. Information Theory, IT10, 41–46 (1970)
3. Tortorella, F.: A ROC-based reject rule for dichotomizers. Pattern Recognition Letters 26, 167–180 (2005)
4. Santos-Pereira, C.M., Pires, A.M.: On optimal reject rules and ROC curves, Pattern Recognition Letters. Pattern Recognition Letters 26, 943–952 (2005)
5. Xie, J., Qiu, Z., Wu, J.: Bootstrap methods for reject rules of Fisher LDA. In: Proc. 18th Int. Conf. on Pattern Recognition, 3rd edn. pp. 425–428. IEEE Press, NJ (2006)
6. van Trees, H.L.: Detection, Estimation, and Modulation Theory. Wiley, New York (1968)
7. Provost, F., Fawcett, T.: Robust classification for imprecise environments. Machine Learning 42, 203–231 (2001)
8. Mukhopadhyay, N.: Probability and Statistical Inference. Marcel Dekker Inc. New York (2000)
9. Garthwaite, P.H., Jolliffe, I.T., Jones, B.: Statistical Inference, 2nd edn. Oxford University Press, Oxford (2002)
10. Tortorella, F.: Reducing the classification cost of support vector classifiers through an ROC-based reject rule. Pattern Analysis and Applications 7, 128–143 (2004)
11. Blake, C., Keogh, E., Merz, C.J.: UCI Repository of Machine Learning Databases, Irvine, University of California, Department of Information and Computer Science (1998), `http://www.ics.uci.edu/mlearn/MLRepository.html`
12. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms, IEEE Trans. Knowledgde and Data Engineering 17, 299–310 (2005)