# Performance visualization spaces for classification with rejection option

Blaise Hanczar

Version of Record: https://www.sciencedirect.com/science/article/pii/S0031320319302870
Manuscript_f4d412a0ab4ab1a3b7644a6cdb01bde1

# Performance visualization spaces for classification with rejection option

Blaise Hanczar [1]

[1] *IBISC, University Paris-Saclay, Univ. Evry, IBGBI Building, 23 Boulevard de France, 91037 Evry, France*
*Corresponding author: blaise.hanczar@ibisc.univ-evry.fr*

**Abstract**

The classification with reject option consists to train a classifier that rejects the examples when the confidence in its prediction is low. The objective is to improve the accuracy of the non-rejected examples and the reliability of the prediction. The performances of the reject classifiers depend on both the error rate and rejection rate. Since these two values are in opposition, we have to make a trade-off between them. This paper is focused on the visualization spaces the performances of the classifiers with rejection option. We analyze two common spaces, the ROC space and the error-rejection (ER) space, then we propose a new space: the cost-reject (CR) space. We show that the ROC space is the less convenient space to represent the performances of the reject classifier. However, it can be recommended for classification problems where the importance of the two classes is different. For the ER space, we point out that the linear interpolation that is commonly used to draw the error-reject curve is not correct and leads to an overestimation of the classifier performances. From the definition of the condition error and rejection rate, we propose a new interpolation of the error-rejection curve that is unbiased. We introduce a new visualization space called the cost reject space. The CR space plots the normalized classification cost in function on the normalized rejection cost. The performance of a classifier is represented in this space by a line. The three visualization spaces are compared on problems of classification algorithms comparison. The advantages and drawbacks of each spaces are discussed and some recommendations are provided in the conclusion.

*Keywords:* Classification with reject option, Classifier performances.

## 1. Introduction

The classification with reject option consists to train a model, called reject classifier, that is able to answer "I do not know" when the confidence in its prediction is low [2] [19]. Also called abstaining classifier [22], selective classification [8] or reject classification [25], this type of model receives an increasing interest in the machine learning community because of its application in many real world classification tasks. The classification with reject option is especially useful when a high level of confidence in the predictions are required. For example in medicine, classification based on genomics data is used to differentiate the types of tumors with different outcomes and thus assist the physician in the selection of more suitable therapeutic treatment [27]. In this application, an error of prediction may lead to a tragic consequence. The classifier should be able to reject a prediction, i.e. answer "I do not know" when the confidence in the prediction is low [14]. Notes that this mimics the behavior of the physician that does not take a decision when he has not enough information about a patient. Another application of the reject option is the problem of classification under budget constraint. Models based on a cascade of classifiers with reject option are used to both maximize the accuracy and minimize the "use cost" of the model [26] [15].

There are two different approaches to add a reject option in a classifier: the plug-in methods and the embedded methods. The principle of the plug-in methods is to add a rejection rule once the classifier has been fitted on the training set. This consists to compute a rejection area on the output of the classifier. There are many methods to compute this rejection area, some are based on the estimation of posterior probabilities [6] , ROC curve analysis [23][24] [7] or intensive empirical testing [20]. Jiang *et al.* proposed a trust score to measure the reliability of a prediction that we can use to reject some predictions [18]. In the embedded methods the rejection option is not added but included in the classifier. The rejection area is computed during the learning procedure. For example Cortes *et al.* proposes new loss function including the cost of a rejection [5]. Many works have been published on support vector machine with rejection in including the notion of reject in the Hinge loss function [13] [1]. Another approach is based on multi-classifier systems where the reject option is defined by the interaction of each classifier [16]. Recent works have proposed to include the rejection option in the boosting [4] or deep neural networks [12]. In both approaches, the objective of the classification with reject option is the minimization of both the error

rate and rejection rate. However these two values are in opposition i.e. the lower the error rate, the higher the rejection rate. It is therefore necessary to do a trade-off between the error rate and the rejection rate [17]. It is possible to set penalty values for rejection and an error in order to find the optimal trade-off between the two measures. However, in real world applications, we do not know which penalties to use. It is generally helpful to explore all trade-offs between the error rate and rejection rate to find a reject classifier with performances adapted to the problem. To perform this exploration we use different methods to visualize the performance of reject classifier in function the trade-off. The most used methods are the ROC space and the error-reject space also called ARC for accuracy rejection curve [21]. Condessa *et al.* have recently proposed a set of three performance measures for classifier with rejection [3].

In this paper, we present several tools to visualize and interpret the performances of the reject classifiers. In section two we formalize the classification with reject option and its different performance measures. In section three, we analyze the reject error space that is the most common tools to visualize the performances of reject classifier. We show how to interpret this space and point out the very common error of curve interpolation. In section four we introduce a new performance visualization space for the classification with rejection option, called the cost reject space. In section five, we show how to use the ROC space to visualize the error rate, the rejection rate, and their trade-offs. We conclude by a discussion on the advantages and drawbacks of these three different spaces and we provide some recommendations.

## 2. Classification with reject option

Let's consider a classification problem with $d$ classes: $\{c_1, ..., c_d\}$. We denote a training set of $N$ examples $Tr = \{(x_1, y_1), ..., (x_N, y_N)\}$ and a classifier $\Psi$. The classifier computes a set of continuous values, one for each class $\omega_i(x)$, representing the probabilities to assign $x$ to the corresponding class. In non-reject classification, the classifier is a function $\Psi : \mathbb{R}^m \rightarrow \{c_1, ..., c_d\}$ returning the class with the maximum $\omega_i(x)$. In classification with reject option, the classifier is a function $\Psi : \mathbb{R}^m \rightarrow \{c_1, ..., c_d, r\}$ where $r$ represents the rejection decision. A rejection threshold $t$ is fixed, if all of the $\omega_i(x)$ values are less than $t$ then the example is rejected, otherwise a prediction is

returned.

$$\Psi(x) = \begin{cases} r \quad \text{if} \quad \omega_i(x) < t \quad \forall i \\ c_i = argmax_i\{\omega_i(x)\} \quad \text{otherwise} \end{cases} \tag{1}$$

Different rejection thresholds can be fixed for each class if the seriousness of the different types of error is not equal [11]. This case has no impact on the two first visualization spaces presented in the section three and four. For the moment, we simplify the notations in considering that the rejection threshold $t$ is the same for each class. Note that this formulation describes the plug-in methods and all following work is applied to the plug-in methods. However, this work can also be applied to embedding methods. The embedding methods generally include a hyper-parameter controlling the trade-off between the error rate and rejection rate. This hype-parameter is equivalent to the rejection threshold of the plug-in methods. The error-reject, cost-reject and ROC space, presented in the following sections, can also be used for the embedding methods.

The performance measures for the classification with reject option are more complex than the measures for the classification without reject option. For a given example $(x, y)$, three cases are possible: the example can be well-classified, miss-classified or rejected. The probability of these three cases are represented respectively by the accuracy $A$, the error rate $E$ and the rejection rate $R$:

$$\begin{aligned} A &= p(\Psi(x) = y, \Psi(x) \neq r) \\ E &= p(\Psi(x) \neq y, \Psi(x) \neq r) \\ R &= p(\Psi(x) = r) \end{aligned} \tag{2}$$

We are also interested by the probability of good classifications and miss-classifications for the non-rejected examples. We define the conditional accuracy $A'$ and conditional error rate $E'$ as:

$$\begin{aligned} A &= p(\Psi(x) = y | \Psi(x) \neq r) = (1 - R)A' \\ E &= p(\Psi(x) \neq y | \Psi(x) \neq r) = (1 - R)E' \end{aligned} \tag{3}$$

We have the following equalities:

$$A + R + E = 1 \qquad A' + E' = 1 \tag{4}$$

Note that in practice all of these probabilities cannot be computed directly, they are estimated from a set of test examples.

Each of these measures gives a view on the performances of the reject classifier. They can be combined in an unique value, the classification cost $L$, that represents the final performance of the classifier:

$$L = \lambda_A A + \lambda_E E + \lambda_R R = (1 - R)(\lambda_A A' + \lambda_E E') + \lambda_R R \tag{5}$$

$L$ is a value to minimize that represents the cost to pay in making a prediction with the classifier. The value of L for a given classifier is depending on the rejection threshold $t$ defined in equation (1). $\lambda_A$, $\lambda_E$ and $\lambda_R$ represent respectively the cost of a good classification, a miss-classification, and a rejection. There is generally no reason to penalize a good classification so we set $\lambda_A = 0$ and to simplify the notation we set $\lambda_E = 1$. The classification cost becomes:

$$L = E + \lambda_R R = (1 - R)E' + \lambda_R R \tag{6}$$

The trade-off between the error and rejection rate is controlled by the $\lambda_R$. We can define an interval of $\lambda_R$. A rejection must always be penalized, $\lambda_R$ is therefore strictly positive. The cost of rejection must be lower than the cost of a random prediction, so we have $0 < \lambda_R < 1 - \frac{1}{d}$. In real world applications, it is very difficult to set a precise value of $\lambda_R$ for a given classification problem. It is generally helpful to test different values of $\lambda_R$ and explore the performances of the reject classifier through different trade-offs between the error rate and rejection rate. In the three next sections, we present different tools in order to visualize these performances.

## 3. The error reject (ER) space

The most popular tool to visualize the performances of a reject classifier is the error reject (ER) space [21]. This space represents in x-axis the rejection rate $R$ and in the y-axis the conditional error rate $E'$. The advantage of the ER space is to show explicitly the tradeoff between the rejection rate and the error rate. The figure 1 illustrates some interesting properties of the error reject space. The point (0,0) represents the performances of the best classifier i.e. all examples are well-classified with no rejection. The worst performances are produced by the random classifiers i.e. classifiers making predictions at random. The random classifiers have their conditional error rate equal to $1 - \frac{1}{d}$ whatever the rejection rate, they are represented by the horizontal line ($E' = 1 - \frac{1}{d}$), in the figure 1 we set $d = 2$ the horizontal line is therefore at $E' =$

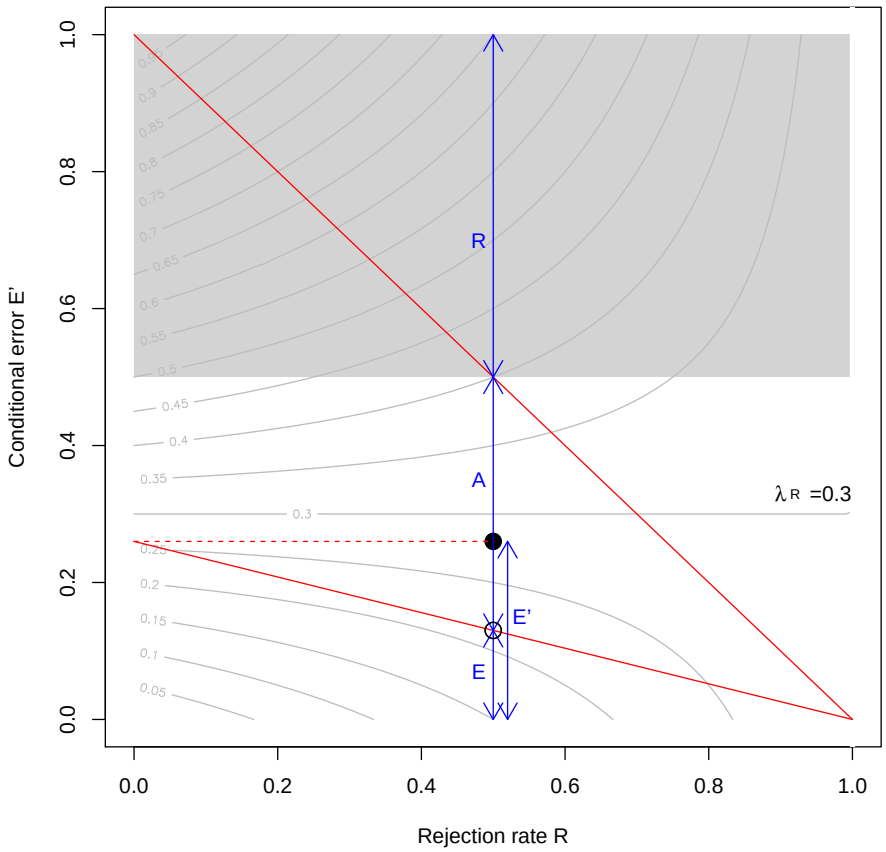


Figure 1: Visualisation of a reject classifier in the error reject space with $\lambda_R = 0.3$.

0.5. All points above this line (gray area) represents classifiers worst than a random classifier. This part of the space is therefore not interesting. We focus only on the white area on this graph that we call the area of interest of the ER space. The performances of a reject classifier are represented by a point (black dot) at the coordinate $(R, E')$. The unconditional accuracy $A$ and error rate $E$ of the classifier can be obtained by a simple construction. The red line from $(0, E')$ to $(1, 0)$ is crossing the vertical line that is passing through the black dot $(R, E)$ and this crossing is denoted with the white dot at the coordinate $(R, E)$. This vertical line gives the value of the unconditional performances, $E$ is represented by the distance between the x-axis and the white dot, $A$ by the distance between the white dot and the red diagonal ($E' = 1 - R$) and

$R$ by the distance between the red diagonal ($E' = 1 - R$) and the horizontal line $E' = 1$. The gray lines represent the iso-cost lines. The iso-cost lines represent all points in the ER space with the same classification cost. The iso-cost lines depends only on the value of $\lambda_R$ (in the figure 1 we set $\lambda_R = 0.3$). These iso-cost lines show the evolution of the classification cost in function on $E'$ and $R$.

Generally, we plot the performances in the ER space of classifiers with different trade-offs between the error and rejection rate. To explore different trade-offs, several rejection thresholds $t$ are tested. When $t$ changes, the rejection rate $R$ and conditional error rate $E'$ also change. The performances of the reject classifier without fixed rejection threshold $t$, can be therefore represented by a curve on the ER space. In theory, the number of possible values for the rejection threshold $t$ is infinite. However, in practice, the number of test examples is finite. The number of relevant values for $t$ to be tested is also finite. We therefore do not have a curve but only a set of points in the ER space. The question is how to interpolate a curve from this set of point. The widely used solution is to do a linear interpolation. We show that the linear interpolation is not appropriate to the ER curve and may lead to wrong conclusions in classifier comparison or error-rejection trade-off selection problems. Let's $\Psi_0$ and $\Psi_X$ two reject classifiers based on the same classifier output but with different rejection thresholds. The performances of $\Psi_0$ and $\Psi_x$ are represented on the ER space respectively by the point $(R_0, E'_0)$ and $(R_X, E'_X)$. We want to compute the interpolation between these two points. Let's call $r_0$ and $e_0$ the number of rejected examples and miss-classifications of $\Psi_0$. Assume that we increase the reject area of $\Psi_0$ in rejecting one more example, the performances of this new classifier are $(R_{0+1}, E'_{0+1})$. The rejection rate increases :

$$R_{0+1} = \frac{r_0 + 1}{N} = R_0 + \frac{1}{N}$$

For the error rate, there are two cases. In the first one, the newly rejected example was a good classification for $\Psi_0$, the conditional error rate increases :

$$E'_{0+1} = \frac{e_0}{N - r_0 - 1} = E'_0 \frac{N - r_0}{(N) - r_0 - 1}$$

In the second case, the newly rejected example was a miss-classification for

$\Psi_0$, the conditional error rate decreases :

$$E'_{0+1} = \frac{e_0 - 1}{N - r_0 - 1} = E'_0 \frac{N - r_0}{N - r_0 - 1} - \frac{1}{N - r_0 - 1}$$

Let's denote $X$ as the number of examples rejected by $\Psi_X$ and classified by $\Psi_0$. Within these $X$ examples, $M$ are miss-classified and $G$ well-classified by $\Psi_0$:

$$X = N(R_X - R_0) \qquad M = N(E'_X - E'_0) \qquad G = X - M$$

The curve between $\Psi_0$ and $\Psi_x$ on the ER space depends only on the sequence of good-classifications and miss-classifications, it can be computed directly from the previous formulas. We denote $R_{0+x}$ the rejection rate when we improve the number of rejected examples of $\Psi_0$ by $x$. We denote $E'_{0+(g,m)}$ the conditional error rate when we improve the number of rejected examples of $\Psi_0$ by $g$ well-classified and $m$ miss-classified examples.

$$\begin{cases} R_{0+x} = \frac{r_0+1}{N} = R_0 + \frac{x}{N} \\ E'_{0+(g,m)} = E'_0 \frac{N-r_0}{N-r_0-g-m} - \frac{m}{N-r_0-g-m} \end{cases}$$

We call the pessimistic interpolation the highest conditional error rate for a given rejection rate. It is defined when the $G$ well-classified examples are rejected before any miss-classified example is rejected. In the first step, we reject the well-classified examples, the conditional error rate is therefore $E'_{0+(g,0)}$ for $0 \leq g \leq G$. Once all $G$ well-classified examples have been rejected, we reject the $M$ miss-classified examples. The conditional error rate becomes $E'_{0+(G,m)}$ for $0 \leq m \leq M$. The curve representing the pessimistic interpolation is defined in the ER space by the following set of points :

$$\begin{cases} (R_g, E'_{0+(g,0)}) & \forall g \quad 0 \leq g \leq G \\ (R_{G+m}, E'_{0+(G,m)}) & \forall m \quad 0 \leq m \leq M \end{cases}$$

In the same way, we define the optimistic interpolation representing the lowest conditional error rate for a given rejection rate. It happens when the $M$ miss-classified examples are rejected before any good-classified example is rejected. The curve representing the optimistic interpolation is defined in the ER space by the following set of points :

$$\begin{cases} (R_m, E'_{0+(0,m)}) & \forall m \quad 0 \leq m \leq M \\ (R_{g+M}, E'_{0+(g,M)}) & \forall g \quad 0 \leq g \leq G \end{cases}$$

We note $E'_{0+x}$ the conditional error when we improve the number of rejected examples of $x$ whatever the proportion of well-classified and miss-classified examples. It is a random variable depending on the number of rejected good-classifications and miss-classifications. The probability that these $x$ examples are composed of $g$ good-classifications and $m$ miss-classifications is binomial :

$$p(g,m) = \binom{x}{g} \left(\frac{G}{X}\right)^{g} \left(\frac{M}{X}\right)^{(x-g)}$$

The expected error $\mathbb{E}[E'_{0+x}]$ is therefore :

$$\mathbb{E}[E'_{0+x}] = \sum_{g=0}^{x} p(g, x-g) E'_{0+(g,x-g)}$$

We recommend using this interpolation to draw the error-reject curve of a given classifier. The method is to vary the rejection threshold $t$, and for each value $t$ computing the corresponding conditional error rate and rejection rate. We obtain a set of points $\{(E'_i, R_i)\}$ in the ER space. This set of points is used to interpolate the error-reject curve in using algorithm 1.

**Algorithm 1** Compute the interpolation of the ER curve

```
InterpolationER {(E'_i, R_i), i = 1..p}
// {(E'_i, R_i)} are the set of conditional error rates and rejection rates representing the points to interpolate in the ER space.
Inter ← { }
for i = 0 to (p-1) do
  Inter ← Inter + (R_i, E'_i)
  compute g and m the number of respectively well-classified and miss-classified examples that are not rejected in R_i and rejected in R_{i+1}
  for j = 1 to (g+m) do
    Inter ← Inter + (R_{i+j}, E[E'_{0+x}])
  end for
end for
Inter ← Inter + (R_p, E'_p)
return Inter
```

To check that our interpolation is correct, we test it on simulations based on artificial datasets with two classes. Each class is defined in a 10-dimension

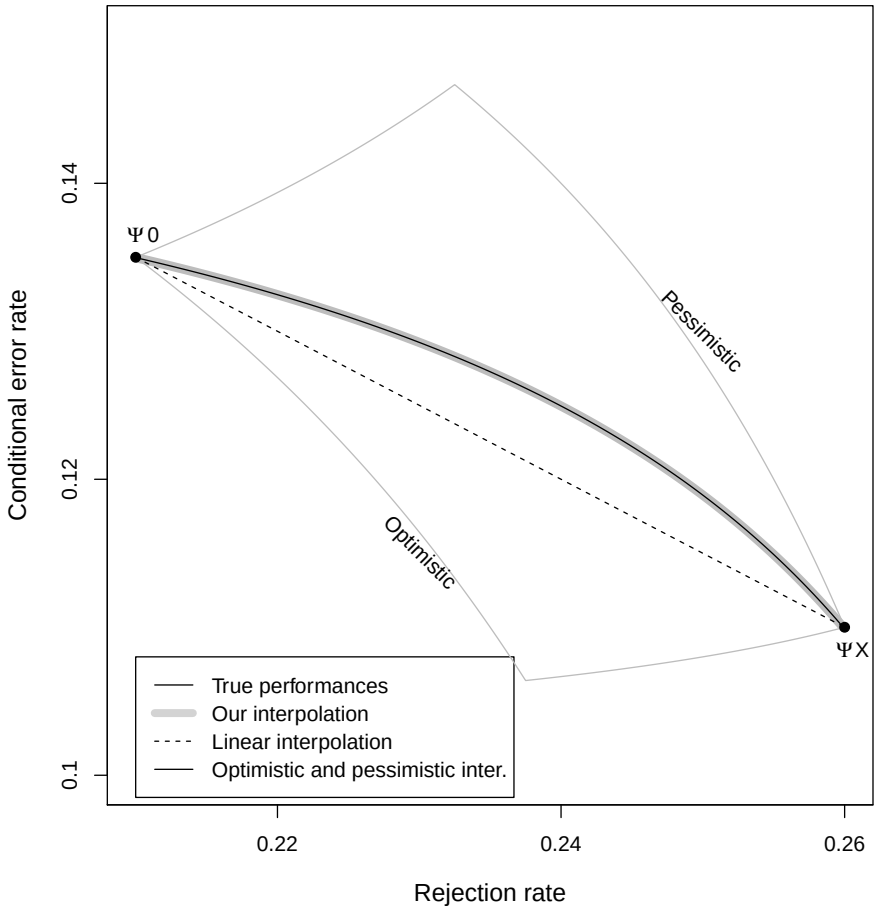


Figure 2: Interpolation in the error reject space.

space by a Gaussian distribution with a diagonal covariance matrix. We generate 100 training examples and 100 test examples for each class. A LDA classifier is fitted on the training set and the rejection option is defined by fixing a rejection threshold on the posterior probabilities estimated by the classifier. We test the performances with different rejection thresholds from 0 to 1 and plot the corresponding points on the ER space. The figure 2 shows two of these points ($\Psi_0$ and $\Psi_X$) and the different interpolations between them. The pessimistic and optimistic interpolations are represented by the gray lines, linear by the dotted line and our interpolation by the bold gray

line. We generate another very large set of test examples and used them to compute the true performance of the classifier between $\Psi_0$ and $\Psi_X$. This true performance is represented by the black curve. We see that our interpolation and true performance curve are identical. This validates our interpolation method.

We see that the true performance curve is very different from the widely used linear interpolation. The figure 3 illustrates the consequences of this problem. On the same artificial dataset defined previously, we construct two classifiers and plot their performances (black and gray points) on the ER space. We see that with the linear interpolation the black curve is always below than the gray curve. This means that the "black" classifier is absolutely better than the "gray" classifier. However, this conclusion is wrong. The true performance curves ( or our interpolation) tell us that the gray classifier is the best for some ranges of rejection rate. The use of linear interpolation may, therefore, lead to wrong conclusions in comparative studies or error-rejection trade-off selection.

## 4. The cost reject (CR) space

We propose a new tool, called the Cost Reject (CR) space, in order to visualize the performances of the reject classifiers in function of the rejection cost. The x-axis and y-axis represent respectively the normalized rejection cost $\tilde{\lambda}$ and the normalized classification cost $\tilde{L}$. $\tilde{L}$ is a normalization of the classification cost such that it becomes a convex combination of the error rate and the rejection rate:

$$\tilde{L} = \frac{L}{1+\lambda_R} = \frac{1}{1+\lambda_R}E + \frac{\lambda_R}{1+\lambda_R}R = (1-\tilde{\lambda}_R)E + \tilde{\lambda}_R R \tag{7}$$

The normalized rejection cost $\tilde{\lambda}$ is the coefficient of this convex combination. $\tilde{\lambda}$ is in the interval $[0,1]$ and is defined by : $\tilde{\lambda}_R = \frac{\lambda_R}{1+\lambda_R}$. Note that in the normalized classification cost, the trade-off between the error and rejection is equivalent to the trade-off in the classification cost since the rates of the error cost on the rejection cost is the same $\frac{1}{\lambda_R} = \frac{1-\tilde{\lambda_R}}{\tilde{\lambda_R}}$.

The performances of a classifier are represented on this space by a segment defined by the points $[(0,E),(1,R)]$. The figure 4 shows the performances of a classifier with $E = 0.23$ and $R = 0.44$ in the CR space. When $\tilde{\lambda}_R = 0$ (resp. $\tilde{\lambda}_R = 1$) we have $\tilde{L} = E$ (resp. $\tilde{L} = R$). We can read the error rate

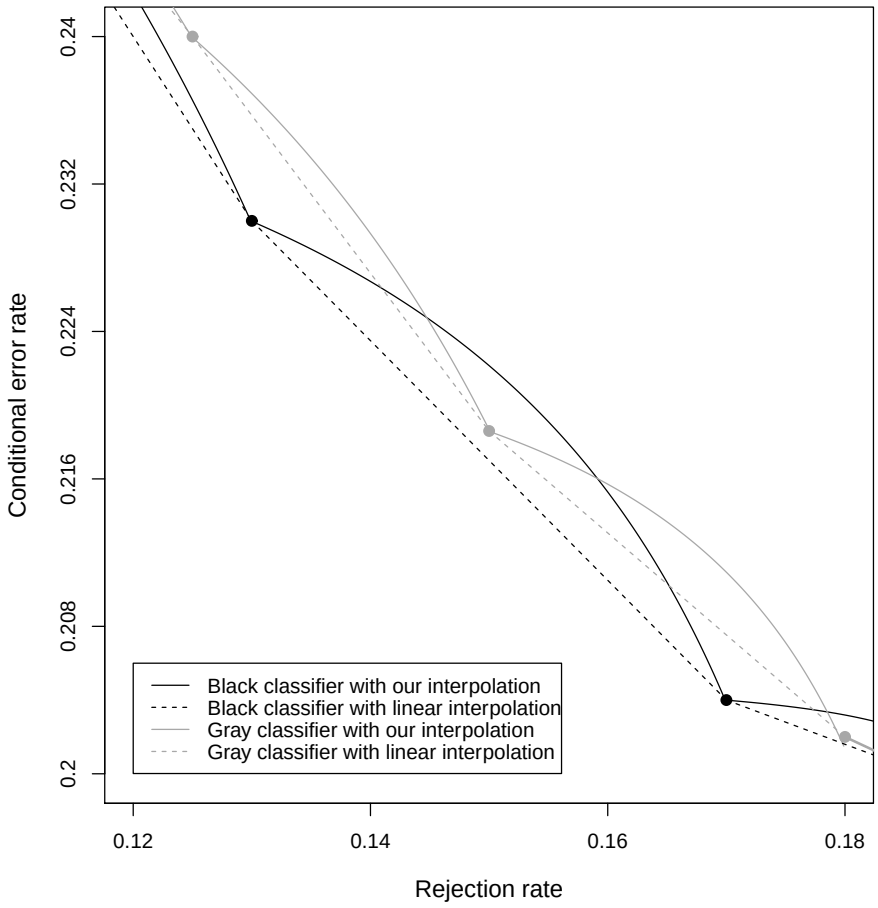


Figure 3: Comparison of two reject classifier in the ER space with different interpolation methods.

and rejection rate of the classifier at the extremities of the segment. The red segment $[(0,0),(1,0)]$ represents the best classifier making no errors and no rejections. We point out some relations between the ER and CR space. There is a bijection between the points in the ER space and the segments on the CR space. The is another bijection between the point in the CR space and the iso-cost curve in the ER space. These two bijections are defined by

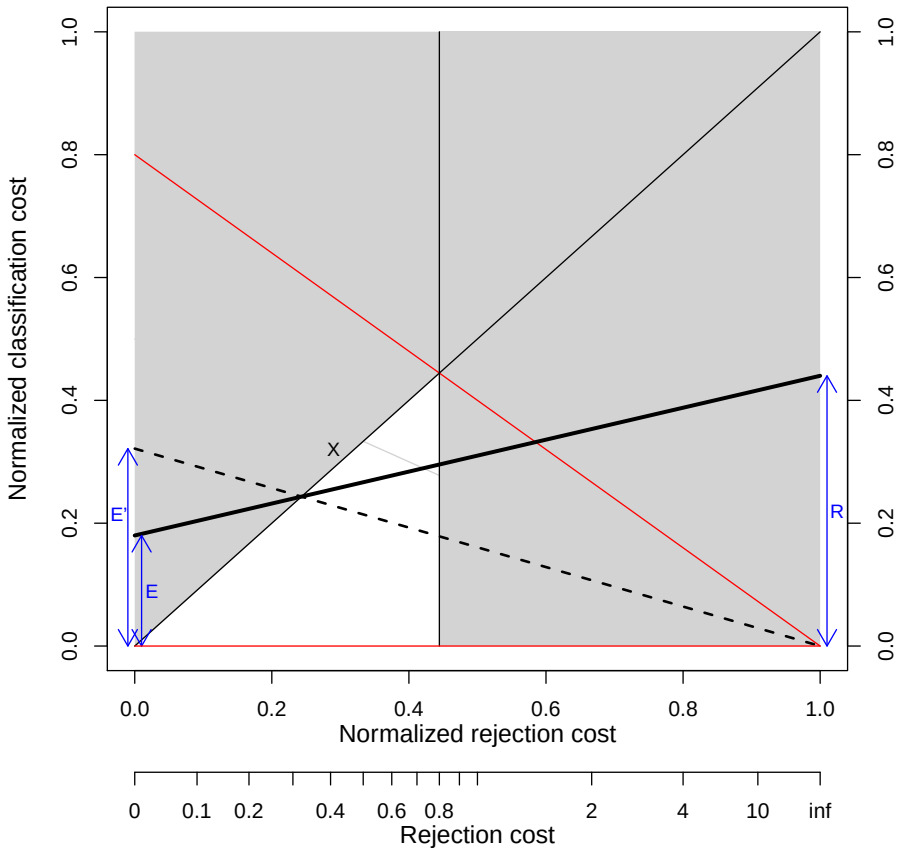


Figure 4: Visualisation of a reject classifier in the CR space.

the following formulas :

$$\tilde{L} = (1-R)E' + (R-(1-R)E')\tilde{\lambda_R} \qquad E' = \frac{\tilde{L} - \tilde{\lambda}_R}{(1-R)(1-\tilde{\lambda_R})} \tag{8}$$

The segment $[(0,0),(1,1)]$ represents the classifier rejecting all examples. All points above this line represent classifier performances worse than the trivial classifier that reject all examples. The red segment $\left[(0,1-\frac{1}{d}),(1,0)\right]$ corresponds to the classifier assigning a random class to all examples, in this example we set $d = 5$, so $E = 0.8$ for a random classifier. All points above this line represent classifier performances worse than a trivial classifier that makes

random predictions. In the section two, we showed that the cost of rejection has to be lower than the cost of random guessing, we have therefore the constraint $\lambda_R < 1-1/d$, this is represented by the vertical line at $\tilde{\lambda}_R^{max} = 0.44$. The gray area defined by these lines represents cases where the classification with reject option is trivial or not possible. The area of interest in the CR space is therefore defined by the triangle $(0,0)(\tilde{\lambda}_R^{max},0)(\tilde{\lambda}_R^{max},\tilde{\lambda}_R^{max})$.

The conditional error rate can be easily constructed. Let $X$ the point at the intersection of the segment representing the performances of the classifier and the segment $[(0,0),(1,1)]$. The line passing through $(1,0)$ and $X$ cut the y-axis at $E'$. This can be proven in applying the Thales's theorem twice and showing that $\frac{E}{1-R} = \frac{|X(0,0)|}{|X(1,1)|} = \frac{E'}{1}$.

We use the normalized rejection cost to define the CR space whose scale increases linearly from 0 to 1. However, it is more practical for performance visualization to plot the rejection cost on the x-axis whose scale increases exponentially from 0 to $+\inf$. In the figure 4 the two scales are plotted on the x-axis.

We showed in the previous section that the performances of a classifier can be represented by a curve in the ER space if the rejection threshold is not fixed. The same principle can be applied in the CR space. The different performances produced by the different values of the rejection threshold are represented by a set of segments as illustrated in the figure 5. We clearly see on the extremities of the figure that the rejection rate goes from 0 to 1 and the error rate goes from $E$ to 0. For each value of the rejection cost, we want to keep the best classifier i.e. the one minimizing the classification cost. The lower envelope of this set of segments (bold curve), that we call the CR curve, represents the best performances for any rejection cost. For $\lambda_R \leq 0.12$ the CR curve is confused with the segment $[(0,0),(1,1))]$. This means that below this rejection cost the best solution is to reject all examples. For $\lambda_R \geq 0.44$ the CR curve is confused with the segment representing the classifier with no rejection. These two values give the interval of rejection cost where the rejection is not trivial. In the ER space, it is difficult to identify a similar interval. We can not know which part of the ER curve is relevant because it depends on the rejection cost.

The CR space is particularly useful for the classifier comparison. We illustrate this in an example where we compare the performances of two classifiers in a two-classes problem. We use the same type of artificial dataset described in section 4. We compare the performances of a LDA classifier and SVM with a radial kernel. The posterior probabilities of each class are

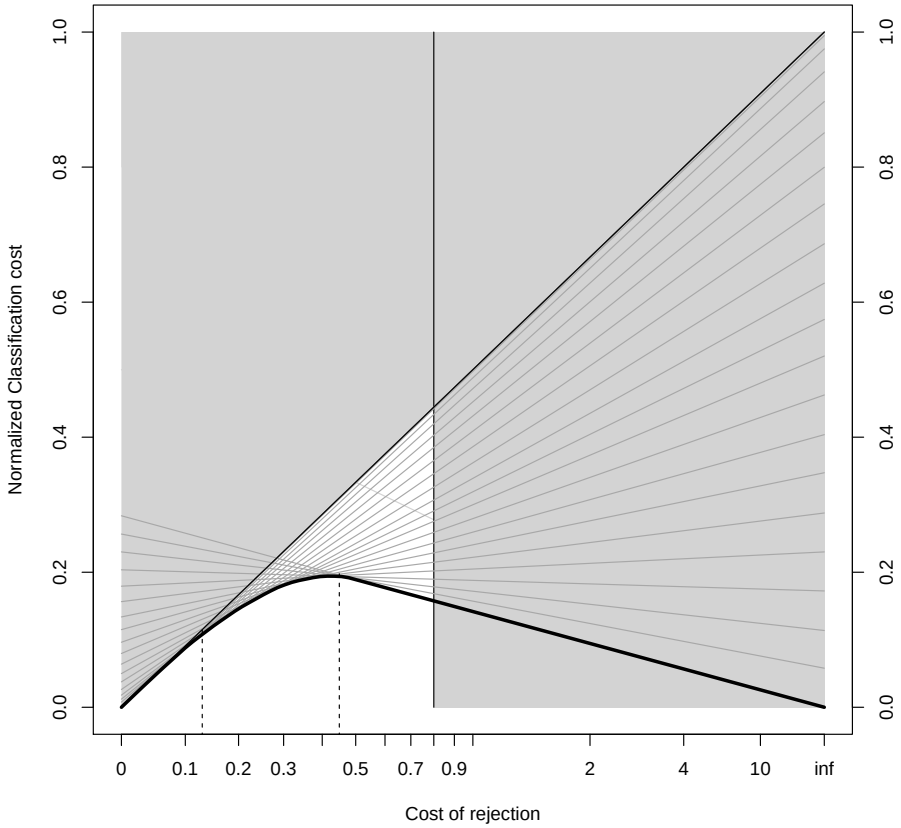


Figure 5: The CR curve of the classifier is defined by the lower envelope of the set of segment representing the performances with different values for the rejection threshold.

estimated from the classifiers and a rejection threshold is applied. Different trade-offs between rejection and error are tested in varying the value of threshold from 0 to 1. The figure 6 shows the ER (left) and CR (right) curves of both classifiers (SVM in black and LDA in gray). A natural interpretation of the ER space on the figure 6 is that no classifier is absolutely better than the other. The gray classifier is better with low rejection rate and the black classifier is better for high rejection rate. The choice of the best classifier depends on the trade-off between error and rejection rate that we want. The CR space tells something very different. The black curve is always below the gray curve on the area of interest this lead to the conclusion that the black

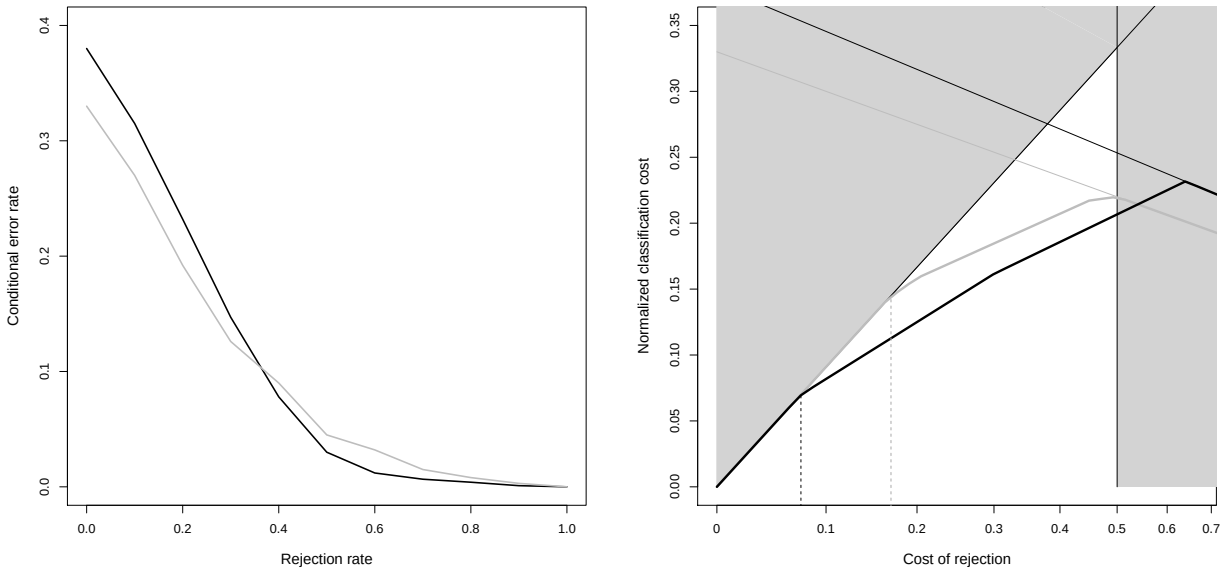


Figure 6: Comparison of two classifiers in the ER (left) and CR (right) space.

classifier is absolutely better than the gray classifier. For $\lambda_R \leq 0.08$ the best solution is to reject all examples for both classifiers, so their performances are equal. For $0.08 < \lambda_R \leq 0.17$ it becomes useful for the black classifier to reject some examples, the cost of the black classifier becomes lower than gray classifier. For $0.17 < \lambda_R < 0.5$ both classifiers improve their performances in rejecting examples but the black classifier is still the better. For $\lambda_R \geq 0.5$ the rejection becomes too costly, it is better to classify randomly the examples than to reject them. The gray classifier becomes better than the black classifier only when the best solution is to use a random classifier. Actually, the ER space is misleading, it gives the impression that the gray classifier can be interesting in some cases. The problem that the area of interest does not take into account all constraints on $\lambda_R$ and especially the constraints induced by the assumption that the cost of rejection should be lower than the cost of random guessing. Some points in the area of interest in the ER space are irrelevant for classifier comparison or error-rejection trade-off selection.

## 5. The ROC space

Now we consider the specific problem of classification with two classes ($d = 2$): positive ($p$) and negative ($n$). For two classes problems, a very useful tool to visualize the performances of non-reject classifiers is the ROC space [9]. The ROC space represents the true positive rate ($TPR$) in function of the false positive rate ($FPR$). In varying the decision threshold, the performances of the classifier are represented by the ROC curve. In this section, we show how to use the ROC space to visualize the performances of reject classifiers.

In two classes problem, the classifier returns only one continuous value of $\omega(x)$ representing the estimated probability that the example belongs to the positive class. The rejection area is depending on two decision thresholds $\{t_N, t_P\}$ (with the constraint $t_N \leq t_P$) and the reject classifier is defined by :

$$\Psi(x) = \begin{cases} n & \text{if} \quad \omega(x) \leq t_N \\ p & \text{if} \quad \omega(x) \geq t_P \\ r & \text{if} \quad t_N < \omega(x) < t_P \end{cases} \tag{9}$$

The performances of the reject classifier are measured by the following values: the rate of true negative ($TNR$), true positive ($TPR$), false negative ($FNR$), false positive ($FPR$), positive rejection ($RPR$) and negative rejection ($RNR$).

$$\begin{array}{ll} TPR = p(\Psi(x) = p, \Psi(x) \neq r | y = p) & TNR = p(\Psi(x) = n, \Psi(x) \neq r | y = n) \\ FNR = p(\Psi(x) = n, \Psi(x) \neq r | y = p) & FPR = p(\Psi(x) = p, \Psi(x) \neq r | y = n) \\ RPR = p(\Psi(x) = r, | y = p) & RNR = p(\Psi(x) = r, | y = n) \end{array} \tag{10}$$

The rate of true negative ($TNR'$), true positive ($TPR'$), false negative ($FNR'$), false positive ($FPR'$) represent the performances of classification for the non-rejected examples :

$$\begin{array}{ll} TPR' = p(\Psi(x) = p | \Psi(x) \neq r, y = p) & TNR' = p(\Psi(x) = n | \Psi(x) \neq r, y = n) \\ FNR' = p(\Psi(x) = n | \Psi(x) \neq r, y = p) & FPR' = p(\Psi(x) = p | \Psi(x) \neq r, y = n) \end{array} \tag{11}$$

We have the following properties on these values:

$$\begin{array}{ll} TPR = (1 - RPR)TPR' & TNR = (1 - RNR)TNR' \\ FNR = (1 - RPR)FNR' & FPR = (1 - RNR)FPR' \\ TPR + FNR + RPR = 1 & TNR + FPR + RNR = 1 \\ TPR' + FNR' = 1 & TNR' + FPR' = 1 \end{array} \tag{12}$$

A reject classifier should be viewed as the combination of two non-reject classifiers. Let's called $\Psi_P$ and $\Psi_N$ the two non-reject classifiers based on the classifier output $\omega(x)$ and using respectively the decision threshold $t_P$ and $t_N$. The vote of these two classifiers defines a reject classifier. If $\Psi_P$ and$\Psi_N$ agree then we assign the class returned by the classifiers if $\Psi_P$ and $\Psi_N$ are disagree the example is rejected:

$$\Psi(x) = \begin{cases} n & \text{if} \quad \Psi_P(x) = \Psi_N(x) = n \\ p & \text{if} \quad \Psi_P(x) = \Psi_N(x) = p \\ r & \text{if} \quad \Psi_P(x) \neq \Psi_N(x) \end{cases} \tag{13}$$

The performances of a reject classifier can be decomposed into the combination of the performances of the two corresponding non-reject classifiers. Let's $TPR_P, TNR_P, FPR_P, FNR_P$ (resp. $TPR_N, TNR_N, FPR_N, FNR_N$) the performances of the classifier $\Psi_P$ (resp. $\Psi_N$), we can express the performances of the reject classifier by :

$$\begin{array}{lll} TPR = TPR_P & FNR = FNR_N & RPR = TPR_N - TPR_P \\ TNR = TNR_N & FPR = FPR_P & RNR = FPR_N - FPR_P \end{array} \tag{14}$$

These performances can be represented in the ROC space as illustrated in the figure 7. The performances of $\Psi_P$ and $\Psi_N$ are represented by the two black dots. The white dot represents is the upper left corner of the rectangle defined by the black points, so its coordinates are $(FPR_P, TPR_N)$. This point is useful to visualize the performances of the reject classifier. On the vertical line passing through this point, we visualize the performances on the positive examples $TPR + RPR + FNR = 1$. On the horizontal line, we visualize the performances on the negative examples $FPR + RNR + TNR = 1$. The triangle represents the conditional $TPR'$ and $FPR'$ of the reject classifier. Note that if we have $RPR = RNR$ we can construct this point in plotting the line $((0,0), \Psi_P)$ and the line $((1,1), \Psi_N)$. The intersection of these two lines represents the $TPR'$ and $FPR'$. The reason is that the line

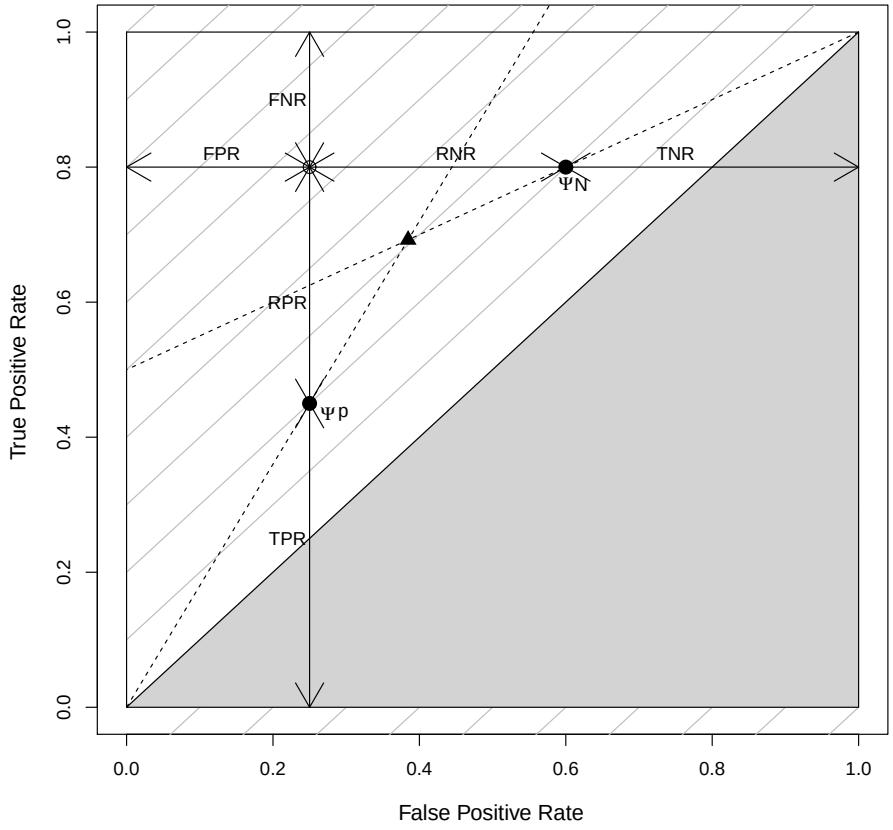


Figure 7: Visualisation of a reject classifier in the ROC space.

$((0,0), \Psi_P)$ is the iso-precision line and the conditional precision $pre'$ can be expressed in the function of the precision $pre_P$ of the classifier $\Psi_P$:

$$pre' = \frac{TPR'}{TPR' + FPR'} = pre_P \left( \frac{TPR_P + \frac{1-RPR}{1-RNR} FPR_P}{TPR_P + FPR_P} \right) \tag{15}$$

When $RPR = RNR$ we have $pre' = pre_P$. The same demonstration can be done on the negative example in order to define the second line.

Let the cost of the predictions defined in the table 5. If we consider that the good classifications are not penalized ($\lambda_{TP} = \lambda_{TN} = 0$), the classification cost of a reject classifier is defined by :

$$L = \pi_P \left( \lambda_{FN} FNR + \lambda_{RP} RPR \right) + \pi_N \left( \lambda_{FP} FPR + \lambda_{RN} RNR \right) \tag{16}$$

Table 1: Cost matrix of a two-class prediction problem.

| | | actual | |
|---|---|---|---|
| | | P | N |
| Predicted | P | $\lambda_{TP}$ | $\lambda_{FP}$ |
| class | N | $\lambda_{FN}$ | $\lambda_{TN}$ |
| | R | $\lambda_{RP}$ | $\lambda_{RN}$ |

where $\pi_P$ and $\pi_N$ are the prior probabilities of the two classes.

The ROC space can represent the classification cost only for non-reject classifier. The iso-cost lines are the parallel gray lines in the figure 7, their slope depends on the ratios $\frac{\pi_P}{\pi_N}$ and $\frac{\lambda_{FN}}{\lambda_{FP}}$ [10]. The ROC space can not represent the classification cost of reject classifier because the classification cost depends on the trade-off between error and rejection rates that is controlled by $\lambda_{RP}$ and $\lambda_{RN}$ and these values are not represented in the ROC space. A solution is to rewrite the classification cost (16) in introducing the formulas (14) as follow:

$$\begin{aligned} L &= \pi_P \lambda_{FN} \left( \frac{\lambda_{RP}}{\lambda_{FN}} FNR_N + (1 - \frac{\lambda_{RP}}{\lambda_{FN}}) FNR_P \right) + \pi_N \lambda_{FP} \left( \frac{\lambda_{RN}}{\lambda_{FP}} FPR_P + (1 - \frac{\lambda_{RN}}{\lambda_{FP}}) FPR_N \right) \\ &= \pi_P \lambda_{FN} FNR_\lambda + \pi_N \lambda_{FP} FPR_\lambda \end{aligned} \tag{17}$$

The classification cost of the reject classifier is equal to the classification cost of a non-reject classifier whose the performances, noted $(FNR_\lambda, FPR_\lambda)$, are convex combinations of the performances of $\Psi_P$ and $\Psi_N$. The point $(FPR_\lambda, TPR_\lambda)$ represents the performances of a non reject classifier that has the same classification cost than the reject classifier :

$$\begin{aligned} TPR_\lambda &= \frac{\lambda_{RP}}{\lambda_{FN}} TPR_P + \left(1 - \frac{\lambda_{RP}}{\lambda_{FN}}\right) TPR_N \\ FPR_\lambda &= \frac{\lambda_{RN}}{\lambda_{FP}} FPR_P + \left(1 - \frac{\lambda_{RN}}{\lambda_{FP}}\right) FPR_N \end{aligned} \tag{18}$$

The point$(FPR_\lambda, TPR_\lambda)$ is an indirect visualization of the performances of a reject classifier in the ROC space. This representation depends naturally on the rejection cost and more specially on the ratios $\frac{\lambda_{RP}}{\lambda_{FN}}$ and $\frac{\lambda_{RN}}{\lambda_{FP}}$.

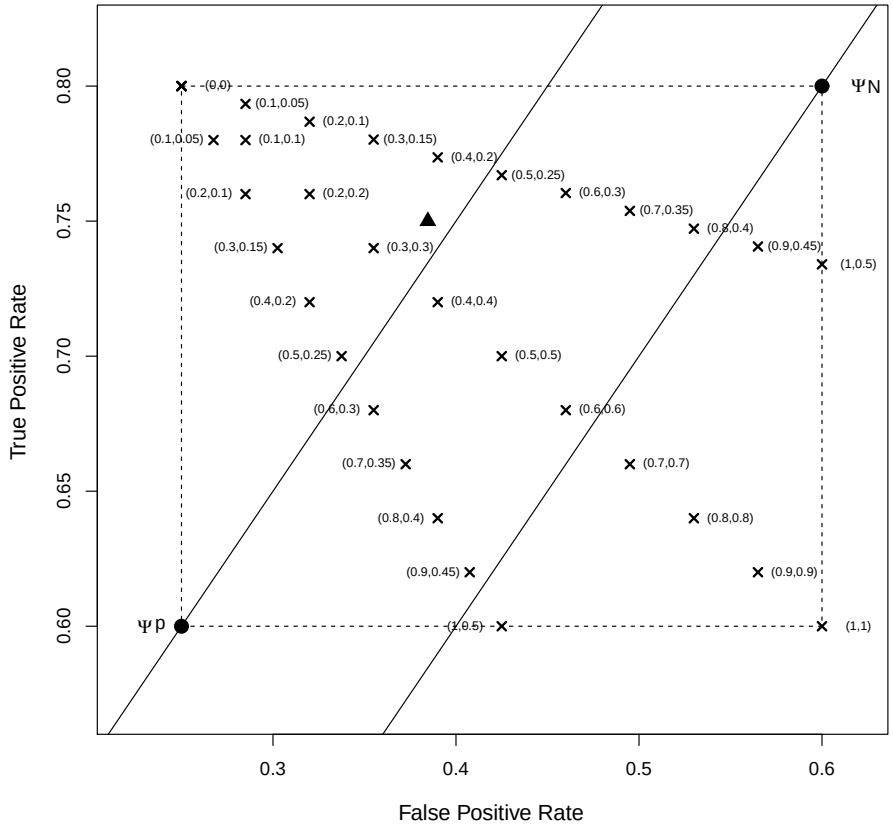


Figure 8: Visualization of the classification cost of a reject classifier in the ROC space with ratios ($\frac{\lambda_{RP}}{\lambda_{FN}}$ , $\frac{\lambda_{RN}}{\lambda_{FP}}$ ).

The figure 8 shows some examples of the point $(FPR_\lambda, TPR_\lambda)$ for a given reject classifier. Each cross corresponds to the classification cost of the reject classifier for given ratios $\frac{\lambda_{RP}}{\lambda_{FN}}$ and $\frac{\lambda_{RN}}{\lambda_{FP}}$. In varying these ratios from 0 to 1, the cross spans the rectangle defined by $\Psi_P$ and $\Psi_N$. The upper left corner corresponds to the case where the costs of rejection are null. The diagonal from the upper left to the bottom right corner corresponds to the cases where the ratios are equal. The two black lines are the iso-cost lines passing respectively through $\Psi_P$ and $\Psi_N$. These lines can be used to define the area of interest of the reject classifier. If a cross is above both iso-cost lines then the reject classifier has a lower classification cost than both non-

reject classifiers. The reject option will improve the performances and should therefore be used. All crosses below one of these two iso-cost lines represent cases where the rejection cost are too high such that the reject option can improve the performances of the classifier.

## 6. Discussion and conclusion

In this paper, we presented three spaces (ER, CR, ROC) to represent the performances of the reject classifiers. Each one has its advantages and drawbacks, the choice of the space depends on the context of the classification problem. Both ER and CR space plot the performances of the reject classifiers in exploring different trade-offs between the error and the rejection. However, they do not use the same tools to do this exploration. The ER space represents explicitly the relation between the conditional error and the rejection rate, but the comparison of the performances with different costs of rejection is difficult. The CR uses the cost of rejection and classification cost. It is easy to make a comparison with different rejection cost but the relation between error and rejection rate is less visible than in the ER space. We recommend using the ER space if a value of rejection cost can be fixed and the CR space when the rejection cost varies. Another major difference between the ER and CR space is the values that are used to represent the performance. The ER space represents the conditional performances $E'$, $A'$ and $R$ whereas the CR space represents the unconditional performances $E$, $R$, $A$ and $\tilde{L}$. In function of the performances measure that we use, we can prefer the ER and CR space. However, we point out that in the CR space we can easily construct all other values $E'$ and $A'$. In the ER space, $E$ and $A$ can also be easily constructed, however, the representation of $L$ is more complex since the iso-cost lines are not linear.

In both ER and CR space we defined an area of interest, representing the part of the graph where we have to focus our attention. A point out from this area of interest represents cases where the rejection option is not possible or trivial. In CR space the area of interest is defined by the three constraints: 1) a classifier must have better performances than random classifier 2) a reject classifier cannot have worse performances than the classifier rejecting all examples 3) $\lambda_R < 1 - \frac{1}{d}$. In the ER space, only the first condition can be represented. The second and third conditions cannot be used because when $R = 1$ then $E'$ is not defined and $\lambda_R$ is not represented in the ER space. There are some parts in the area of interest of the ER space that

are actually irrelevant for the performance analysis of a reject classifier. The consequences of this point, that have been illustrated in the experiments of on the figure 6, may lead to wrong conclusions in classifiers comparison or error-rejection trade-off selection. We recommend checking always the area of interest in the CR space.

The ROC space is clearly the less convenient space to represent the performances of the reject classifier. Only the unconditional values can easily be represented, the conditional $TPR'$ and $FPR'$ can be constructed, the classification cost can only be visualized indirectly in plotting the performances of a non-reject classifier with equivalent performances. The only case where we recommend the use of the ROC space is in the two-classes problem when the importance of the two classes is different $\lambda_{FP} \neq \lambda_{FN}$ or/and $\lambda_{RP} \neq \lambda_{RN}$. The ER and CR space deal with error and rejection rate computed over all classes. Only the ROC space can plot separately the performances of the two classes.

## References


[1] Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(Aug):1823–1840, 2008.

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

[3] Filipe Condessa, José Bioucas-Dias, and Jelena Kovačević. Performance measures for classification systems with rejection. *Pattern Recognition*, 63:437–450, 2017.

[4] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In *Advances in Neural Information Processing Systems*, pages 1660–1668, 2016.

[5] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82. Springer, 2016.

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

[7] Clément Dubos, Simon Bernard, Sébastien Adam, and Robert Sabourin. Roc-based cost-sensitive classification with a reject option. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3320–3325. IEEE, 2016.

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

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

[10] Peter A. Flach. The geometry of roc space: Understanding machine learning metrics through roc isometrics. In *ICML*, pages 194–201, 2003.

[11] Giorgio Fumera, Fabio Roli, and Giorgio Giacinto. Multiple reject thresholds for improving classification reliability. In Springer Berlin / Heidelberg, editor, *Advances in Pattern Recognition: Joint IAPR International Workshops, SSPR 2000 and SPR 2000, Alicante, Spain,*, page 863, August/September 2000.

[12] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in neural information processing systems*, pages 4878–4887, 2017.

[13] Yves Grandvalet, Alain Rakotomamonjy, Joseph Keshet, and Stéphane Canu. Support vector machines with a reject option. In *Advances in neural information processing systems*, pages 537–544, 2009.

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

[15] Blaise Hanczar and Avner Bar-Hen. Controlling the cost of prediction in using a cascade of reject classifiers for personalized medicine. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 42–50. SCITEPRESS-Science and Technology Publications, Lda, 2016.

[16] Blaise Hanczar and Michèle Sebag. Combination of one-class support vector machines for classification with reject option. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 547–562. Springer, 2014.

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

[18] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, pages 5546–5557, 2018.

[19] Hoel Le Capitaine. A unified view of class-selection with probabilistic classifiers. *Pattern Recognition*, 47(2):843–853, 2014.

[20] 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.

[21] 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.

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

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

[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] Kirill Trapeznikov and Venkatesh Saligrama. Supervised sequential classification under budget constraints. In *Artificial Intelligence and Statistics*, pages 581–589, 2013.

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