IDIAP RESEARCH REPORT

The Expected Performance Curve

Samy Bengio [1] Mikaela Keller [2]
Johnny Mariéthoz [3]

IDIAP–RR 03-85

January 8, 2004

submitted for publication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

1 IDIAP, CP 592, 1920 Martigny, Switzerland, bengio@idiap.ch
2 IDIAP, CP 592, 1920 Martigny, Switzerland, mkeller@idiap.ch
3 IDIAP, CP 592, 1920 Martigny, Switzerland, marietho@idiap.ch

IDIAP Research Report 03-85

The Expected Performance Curve

Samy Bengio Mikaela Keller Johnny Mariéthoz

January 8, 2004

submitted for publication

Abstract. In several research domains concerned with classification tasks, including person authentication and text categorization, ROC curves are often used to assess the quality of a particular model or to compare two or more models. Researchers also often publish some statistics coming from the ROC curve, such as so-called *break-even point* or *equal error rate*. The purpose of this paper is to argue that these measures can be misleading and should be avoided. Furthermore, we propose a replacement for them, called *Expected Performance Curves* (EPC). Empirical examples from several domains will be used throughout the paper to illustrate the problem and the novel measures.

Contents

1 Introduction

Two-class classification problems are common in machine learning. In several domains, on top of selecting the appropriate discriminant function, practitioners also modify the corresponding threshold in order to better suit an independent cost function. Moreover, they compare models with respect to the whole range of possible values this threshold could take, in the form of so-called ROC curves. In order to also provide quantitative comparisons, they often select one particular point of this curve (generally called *break-even point* or *equal error rate*).

The main purpose of this paper is to argue that such curves, as well as particular points on it like *break-even point* or *equal error rate* can not be used reliably to either compare two or more models, nor to obtain a realistic estimate of the performance of a given model.

We thus propose instead the use of a new set of curves, called *Expected Performance Curves* (EPC), which really reflect the expected (and reachable) performance of systems.

In Section 2, we review the various performance measures used in several research domains when confronted to 2-class classification tasks, such as person authentication, text categorization and medical studies. In Section 3, we explain, using two real case studies, why some of these measures can be misleading. In Section 4, we present the family of EPC curves, that really reflects the expected performance of a given model, hence enabling a fair comparison between models. Finally, Section 5 concludes the paper.

2 Performance Measures for 2-Class Classification Tasks

Let us consider two-class classification problems defined as follows: given a training set of examples (x_i, y_i) where x_i represents the input and y_i is the target class $\in \{0, 1\}$, we are searching for a function $f(\cdot)$ and a threshold θ such that

$$f(x_i) > \theta \text{ when } y_i = 1 \text{ and } f(x_i) <= \theta \text{ when } y_i = 0, \quad \forall i \; . \tag{1}$$

		Desired Class	
		1	0
Obtained Class	1	TP	FP
	0	FN	TN

Table 1: Types of errors in a 2-class classification problem.

The obtained function $f(\cdot)$ (and associated threshold θ) can then be tested on a separate test data set and one can count the number of utterances of each possible outcome: either the obtained class corresponds to the desired class, or not. In fact, one can decompose these outcomes further, as exposed in Table 1, in 4 different categories: *true positives* (where both the desired and the obtained class is 1), *true negatives* (where both the desired and the obtained class is 0), *false positives* (where the desired class is 0 and the obtained class is 1), and *false negatives* (where the desired class is 1 and the obtained class is 0). Let TP, TN, FP and FN represent respectively the *number of utterances* of each of the corresponding outcome in the test set.

Note once again that TP, TN, FP, FN and all other measures derived from them are in fact dependent both on the obtained function $f(\cdot)$ and the threshold θ. In the following, we will sometimes refer to, say, FP by FP(θ) in order to specifically show the dependency with the associated threshold.

Several tasks are in fact specific incarnation of 2-class classification problems. However, often for historical reasons, researchers specialized in these tasks have chosen different methods to measure the quality of their systems. In the following, we present three different tasks, namely Person Authentication, Text Categorization, and the general domain of medical studies, where different measures have been proposed to estimate the quality of their respective 2-class classification tasks.

2.1 Person Authentication

The general field of person authentication comprises several well-established research domains such as verification of voice, face, signature, fingerprints, etc [6]. Such a verification system has to deal with two kinds of events: either the person claiming a given identity is the one who he claims to be (in which case, he is called a *client*), or he is not (in which case, he is called an *impostor*). Moreover, the system may generally take two decisions: either *accept* the *client* or *reject* him and decide he is an *impostor*.

Thus, the system may make two types of errors: *false acceptances* (which we call here *false positives*, FP), when the system accepts an *impostor*, and *false rejections* (which we call here *false negatives*, FN), when the system rejects a *client*.

In order to be independent on the specific dataset distribution, the performance of the system is often measured in terms of *false acceptance rate* (FAR) and *false rejection rate* (FRR), defined as follows:

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \ , \tag{2}$$

$$\mathrm{FRR} = \frac{\mathrm{FN}}{\mathrm{FN+TP}} \ . \tag{3}$$

A unique measure often used combines these two ratios into the so-called *detection cost function* (DCF) proposed by [3] and defined as follows:

$$\mathrm{DCF} = \mathrm{Cost(FN)} \cdot P(Y=1) \cdot \mathrm{FRR} + \mathrm{Cost(FP)} \cdot P(Y=0) \cdot \mathrm{FAR} \tag{4}$$

where $P(Y=1)$ is the prior probability that a client will use the system, $P(Y=0)$ is the prior probability that an impostor will use the system, Cost(FN) is the cost of a false rejection, and Cost(FP) is the cost of a false acceptance. Finally, a particular case of the DCF is known as the *half total error rate* (HTER) where the costs are equal to 1 and the probabilities are 0.5 each:

$$\mathrm{HTER} = \frac{\mathrm{FAR} + \mathrm{FRR}}{2} \ . \tag{5}$$

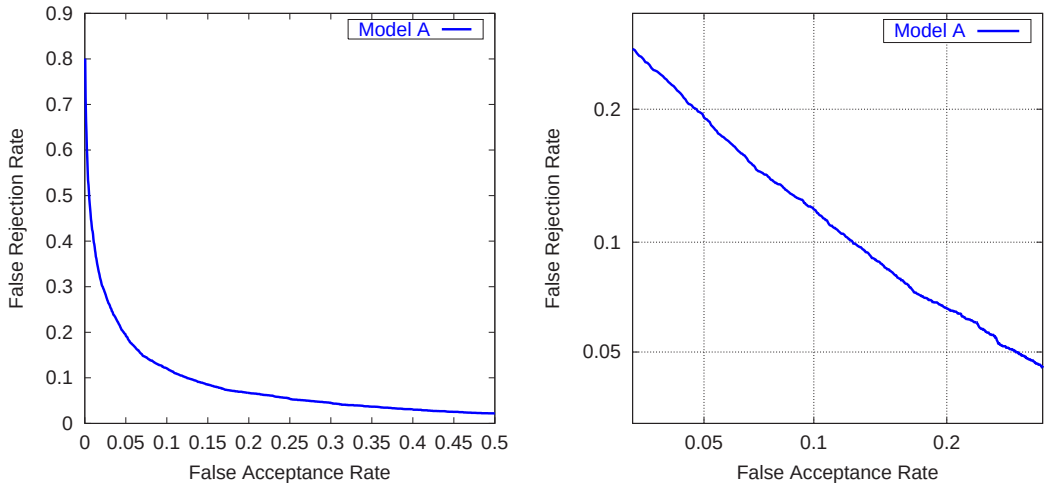
Most person authentication systems can be tuned to either favor FARs or FRRs, by changing the value of the threshold θ which controls whether the decision will be a client or an impostor. Hence, in fact FAR and FRR depend on the particular value of θ. One particular θ often chosen in the literature is when $\mathrm{FAR}(\theta)$ equals $\mathrm{FRR}(\theta)$, and the corresponding HTER measure is called *Equal Error Rate* (ERR):

$$\mathrm{EER} = \mathrm{HTER}(\theta^\star) \tag{6}$$

such that

$$\theta^\star = \arg\min_\theta |\mathrm{FAR}(\theta) - \mathrm{FRR}(\theta)| \ . \tag{7}$$

Several researchers, instead of selecting a threshold θ corresponding to an operating point such as the EER, prefer to present their results through the use of the so-called *Receiver Operating Characteristic* (ROC) curve [5], which, in the case of person authentication, shows values of $\mathrm{FRR}(\theta)$ with respect to corresponding $\mathrm{FAR}(\theta)$ for all possible values of θ. An example of such a graph can be seen on Figure 1(a). Note that in this version of the ROC curve, being nearer the $(0, 0)$ coordinate yields the best ROC curve. [2] have also proposed the use of the DET curve, which is a non-linear transformation of the ROC curve in order to make results easier to compare in the case of person authentication. The non-linearity is in fact a normal deviate, coming from the hypothesis that the scores of client accesses and impostor accesses follow a Gaussian distribution. If this hypothesis is true, the DET curve should be a line. Figure 1(b) shows the DET curve corresponding to the ROC curve of Figure 1(a).



(a) Typical ROC Curve

(b) Corresponding DET Curve

Figure 1: Typical ROC and DET curves for Person Authentication

2.2 Text Categorization

Text Categorization (TC) is a particular issue in the field of Information Retrieval. It can be defined as the task of assigning one or several predefined categories to documents. Common applications of TC are Text Filtering (for instance SPAM/non-SPAM email), Document Organization, Automated Meta-data Generation, etc. Several approaches to solving this problem are present in the related literature, of which [4] gives an exhaustive review. What is of interest to us is that, most of the time, Text Categorization is considered as a set of 2-class classification tasks. Indeed, the problem of multi-label classification under $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ (*i.e.* assigning one or several categories from $\mathcal{C}$ to a document d) can be transformed into $|\mathcal{C}|$ independent 2-class classification problems, *i.e.* to classify d under c or $\bar{c}$ (non-c) for each $c \in \mathcal{C}$.

In order to measure the performance of a classifier for a class c, the following measures are used:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}, \tag{8}$$

that is, the proportion of documents assigned to c that really are class members, and

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}, \tag{9}$$

that is, the proportion of class members assigned to the class c. These measures represent the effectiveness of the model, so the closer to 1 the better.

Neither Precision nor Recall, alone, is enough to evaluate the system. For example the *trivial acceptor* (*i.e.* the classifier that accepts all examples as belonging to c), obtains a Recall = 1. In fact, once again, some threshold θ can often be tuned in such a way that decreasing θ makes Precision decrease and Recall increase. This behavior is often presented in the literature through the use of ROC curves similar to those used in Person Authentication, but where the axes represent Precision(θ) with respect to Recall(θ) (hence in that case, good models tend to have their corresponding curve toward the top right end of the graph). In general, a unique measure is chosen which combines both values. The common choices present in the related literature are:

- the F_1 measure:

$$F_1 = \frac{2 * \text{Precision}(\theta) * \text{Recall}(\theta)}{\text{Precision}(\theta) + \text{Recall}(\theta)}, \tag{10}$$

which is the harmonic mean[1] of Precision and Recall;

- the *break-even point*:

$$\text{BEP} = \frac{\text{Precision}(\theta^\star) + \text{Recall}(\theta^\star)}{2} \tag{11}$$

such that

$$\theta^\star = \arg\min_\theta |\text{Precision}(\theta) - \text{Recall}(\theta)|, \tag{12}$$

which is the point where the ROC curve intercepts the line Precision = Recall and is what we expect, *i.e* both Precision and Recall high;

- the *eleven-point average precision*:

$$\text{11pt-avg} = \frac{1}{11} \sum_{\alpha=0,0.1,\cdot,0.9,1} \text{Precision}(\theta_\alpha) \tag{13}$$

such that

$$\theta_\alpha = \arg\min_\theta |\alpha - \text{Recall}(\theta)|, \tag{14}$$

which is proportional to an approximation of the area under the ROC curve.

2.3 Medical Studies

The Diagnostic Test Interpretation is an important step in Medical Studies. It addresses the critical problem of deciding whether a patient has a disease or not, given the result of a medical test. If this result can be presented as a value (such as rate of iron in blood, etc), the threshold θ represents a critical value that will determine whether the test is positive or negative. In this field ROC curves are also used to evaluate the ability of a test to discriminate diseased cases from normal cases and to compare the diagnostic performances of different tests [7]. In general, in this domain, the ROC curves represent, the Sensitivity(θ) with respect to (1 − Specificity(θ)), where

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{15}$$

is the proportion of patients with disease who test positive, and

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{16}$$

is the proportion of patients without disease who test negative. These two functions of θ measure the effectiveness of the medical test. They have values in the $[0, 1]$ interval, and the nearer to 1 the better.

Like in the previously presented domains, a unique measure summarizing the ROC curve or combining the values of Sensitivity and Specificity, is desired. Several choices are proposed, among them:

- the *area under the ROC curve (AUROCC)*, which is computed by summing the rectangles under the curve or by using maximum likelihood to fit a curve to the data (note that this measure is similar to the *11pt-avg* defined by equation (13));

[1] The harmonic mean $\left(\frac{1}{2}(a^{-1} + b^{-1})\right)^{-1}$ tends to be closer than the arithmetic mean $\frac{1}{2}(a+b)$ to the smallest element:

harmonic mean < geometric mean < arithmetic mean .

- the *break-even point*

$$\text{BEP} = \frac{\text{Sensitivity}(\theta^\star) + \text{Specificity}(\theta^\star)}{2} \tag{17}$$

such that

$$\theta^\star = \arg\min_\theta |\text{Sensitivity}(\theta) - \text{Specificity}(\theta)| \tag{18}$$

which is the intersection of the ROC curve with the line of equation $y = 1 - x$.

3 Mismatch Between *A Posteriori* and *A Priori* Measures

As presented in the previous section, several measures are used to assess the quality of 2-class classification problems in different domains. Most of these measures depend, either explicitly or implicitly, on the critical value of the threshold θ.

However, while some researchers select θ on a separate validation set, most of them unfortunately select it directly on the test set. For instance, when published results show EERs (in person authentication) or BEPs (in text categorization or medical studies), that automatically means they have chosen the corresponding θ on the test set (there is no other way to obtain such a measure). As it is well known, these results should thus be regarded with caution, since they are optimistically biased, having chosen at least one parameter on the test set. Of course, it might be argued that we are talking about only one tiny parameter, which should thus not have a significant impact on the overall quality of the models, but as it will be seen, it does, and this in general reflects some sort of mismatch between validation data and test data, either at the level of the raw data itself, or at the level of the scores obtained by the models on this data.

To make things clearer, we will call a measure *a priori* when it has been computed without looking at the test set for any purpose, hence for instance using a threshold θ chosen on a separate validation set, and *a posteriori* when some information from the test set was used, such as θ.

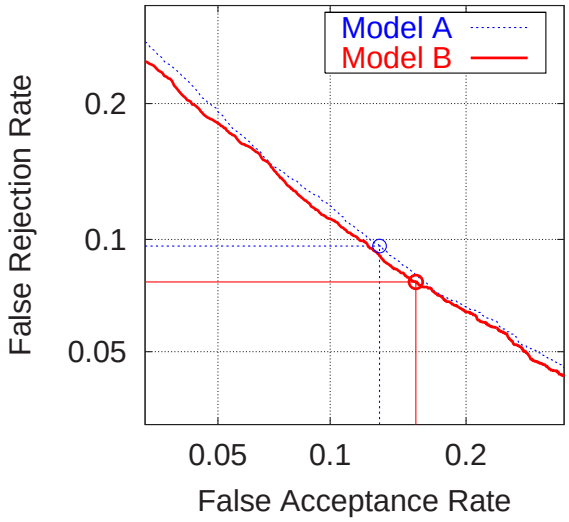
As we will now see, when results are published in terms or the various representation of ROC curves, they implicitly let the reader select his own threshold directly on the test set, hence biasing his interpretation of the measure. As explained in Section 2, ROC curves show the performance of the system on the test set for different thresholds (also called operating points) without making a choice among them. However, in a real-life application, one would normally have to select the threshold before looking at the test set.

While ROC curves certainly represent all the possible outcome of a system with respect to θ, the choice of θ may have a significant impact on the underlying performance. And since θ is probably different for two different systems, two curves representing two different systems cannot be compared so easily. In the following two subsections, we will show on real data, for both person authentication and text categorization, how such comparison can be misleading.

3.1 Illustration with Person Authentication

We present here a real case study in the field of person authentication, on the well-known text-independent speaker verification benchmark database NIST'2001, consisting of up to 63573 test accesses. We compared the performance of 2 models, hereafter called *model A* and *model B*, which correspond to two different techniques (one based on a generative approach and the other on a more discriminant approach[2]). Their respective DET curves and HTER performances yielded incompatible results, suggesting that one of the measures (or both) does not fully represent the expected performance of the system. Figure 2(a) shows the *a posteriori* DET curves of the two models, while Table 2(b) shows the HTER performances of the two models on three different operating points: one that minimizes the HTER on a separate validation set, one such that FAR = FRR on the validation set, and one such that FAR = FRR on the test set itself.

[2]While this is not the topic of this paper (since it should apply to any data/model), people interested in knowing more about the problem tackled in this case study are referred to [1].



(a) DET curves for models A and B

Criterion	Model	FAR	FRR	HTER
HTER (valid)	A	0.114	0.108	**0.111**
	B	0.138	0.086	0.112
EER (valid)	A	0.131	0.096	**0.114**
	B	0.158	0.078	0.118
EER (test)	A	0.110	0.110	0.110
	B	0.107	0.107	**0.107**

(b) Performance comparison between models A and B using three different criteria: minimum HTER and EER on a separate validation set, and EER on the test set.

Figure 2: Illustration of the problem for person authentication

Looking at the DET curves of Figure 2(a), we see that model B's performance is always below (better) model A's performance, letting think that for any threshold, model B should always be better. However, looking at Table 2(b), we see that for the two operating points computed *a priori* (on a separate validation set), model A is indeed better that model B, while on the operating point computed *a posteriori* (on the test set), model B is better than model A. Moreover, results obtained with either the *a priori* EER criterion or the *a posteriori* EER criterion are both statistically significant[3] with a confidence level of 95%, although showing opposite behaviors. So, what is going on? obviously, at least one of these measures is wrong. Which one? How can we correct this?

In order to explain why the DET curves misled us, consider the two circles on these curves (Figure 2(a)). They represent the performance of the model when the threshold was selected using the same criterion (EER) on a separate validation set. The selected thresholds are quite different from each other and from the test data EERs, thus the circles are far from each other. Moreover, it might happen, and it is the case here, that the HTER of a given point of model A becomes less than the HTER of another point of model B.

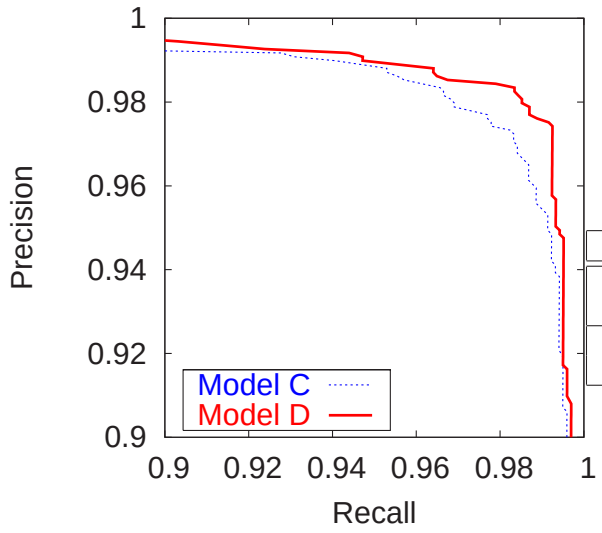
3.2 Illustration with Text Categorization

In the following, we show a real example, in the field of text categorization, which also illustrates the mismatch between *a priori* and *a posteriori* measures. Two models, *model C* and *model D*, based on two different classification algorithms, have been trained on the Reuters-21578 news stories database, with the ModApte split, for the EARN category against the 114 others. This database is widely used, and many classifiers have been compared on it. It has 9603 documents in the training set and 3299 in the test set. The EARN category is the most represented in the database. As in the case of person authentication the performances we obtained show contradictory results.

[3]with a standard proportion test on the real classification error, assuming a binomial distribution for the decisions, and using a normal approximation.

Figure 3(a) shows the *a posteriori* ROC curves of the two models. According to this graph, *model D* seems to be better than *model C* for every threshold, since its ROC curve is always above the *model C* one. In Table 3(b), *a posteriori* and *a priori* values of Precision, Recall and F_1 at the BEP, are shown. The second section of the Table presents the value which is most often used in the related literature, that is the BEP computed *a posteriori*. It corresponds to the results reflected by Figure 3(a) curves, and states that *model D* is statistically significantly better with 95% confidence. But if we choose the threshold value θ of the BEP on the training set and then compute the Precision and Recall values at θ on the test set, then, as shown in the first section of the Table, *model C* is statistically significantly[4] better with 95% confidence.

Hence, we can see that using *a posteriori* BEP value or ROC curve information can be misleading. In the following section, some replacement measures are thus proposed.



(a) ROC curves for models C and D

Criterion	Model	Prec.	Recall	F1
BEP (valid)	C	0.978	0.975	**0.977**
	D	0.995	0.923	0.958
BEP (test)	C	0.977	0.977	0.977
	D	0.983	0.983	**0.983**

(b) Performance comparison between models C and D using two different criteria: BEP on a separate validation set and BEP on the test set.

Figure 3: Illustration of the problem for text categorization

4 The Expected Performance Curve

In the context of a real application, one has in general some criterion to optimize which reflects the relative costs of each type of error (FPs and FNs), or a relative gain for each type of correct decision (TPs and TNs). The choice of the threshold θ should thus reflect this knowledge. Hence we would like to propose the use of new curves which would let the user select a threshold according to some criterion, in an unbiased way. In this Section, we propose three such series of curves, each reflecting a particular way to express this criterion. We shall call these curves Expected Performance Curves (EPC).

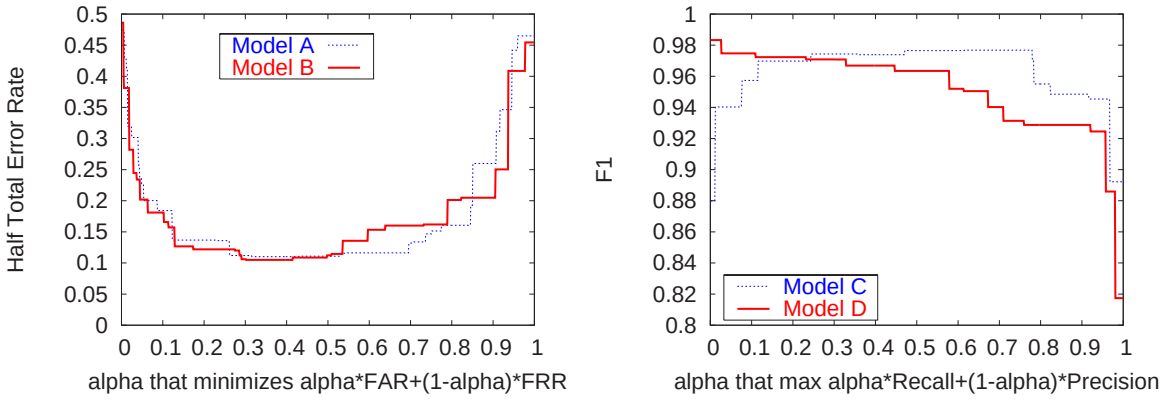
[4]For this value and the previous one we used a standard proportion test on the real classification error, assuming a binomial distribution for the decisions, and using a normal approximation.

4.1 General Framework

The general framework of EPC curves is to present the obtained performance on a test set (for instance using the HTER for person authentication tasks, or the F1 measure for text categorization) with respect to some measure of expected performance of the system when the threshold is set on a separate validation set. This threshold could be chosen in several ways. Note that all the curves that are presented in this section have been computed using the freely available EPC software[5].

4.2 Some Examples of EPCs

The first solution is to select it in order to minimize or maximize a global criterion such as the DCF criterion (4) for person authentication tasks, or the F1 criterion (10) for text categorization, *on a separate validation set*. Algorithm 1 presents the general method to obtain such a curve for DCF, but the modification in order to obtain the same algorithm for F1 in text categorization should be obvious. In fact, we show in Figure 4 the result of the algorithm for both of our case studies.



(a) DCF EPC for models A and B

(b) F1 EPC for models C and D

Figure 4: DCF and F1 Expected Performance Curves

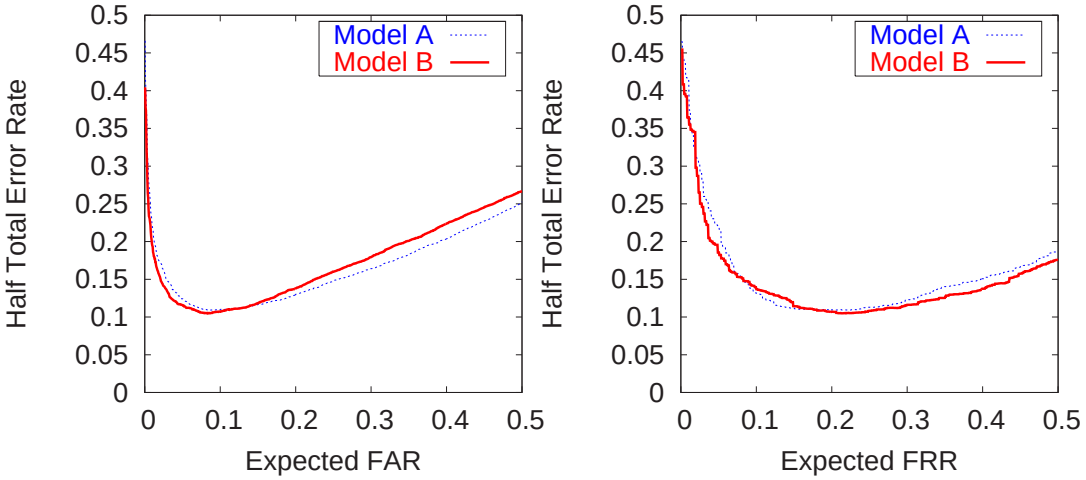
Algorithm 1 Method to generate the DCF Expected Performance Curve

Let *valid* be the validation set
Let *test* be the test set
Let $\text{FAR}(\theta, valid)$ be the FAR obtained on the validation set for threshold θ
for values $\alpha \in [0, 1]$ **do**
  $\theta^\star = \arg\min_\theta(\alpha \cdot \text{FAR}(\theta, valid) + (1 - \alpha) \cdot \text{FRR}(\theta, valid))$
  compute $\text{FAR}(\theta^\star, test)$, $\text{FRR}(\theta^\star, test)$ and $\text{HTER}(\theta^\star, test)$
  plot $\text{HTER}(\theta^\star, test)$ with respect to α
end for

We can see for instance in Figure 4(a) that if one selects the threshold such that it minimizes the HTER on a separate validation set (which corresponds to the performances obtained when $x = 0.5$ on this Figure), model A is slightly better than model B (as confirmed in Table 2(b)), while if the threshold is chosen to minimize, say, (0.2 FAR + 0.8 FRR) on a separate validation set, then model B is better than model A. More generally, this Figure shows that neither of the two models is better

[5]EPC is available at `http://www.Torch.ch/extras/epc` as a package of the Torch machine learning library.

for a wide range of α values. On the other hand, looking at Figure 4(b), we see that selecting the threshold in order to maximize, for instance, (0.5*Recall + 0.5*Precision), we are in a position where model C becomes better than model D, and this is true for a wide range of values for α, which controls the trade-off between expected Precision and expected Recall.



(a) FAR EPC for models A and B

(b) FRR EPC for models A and B

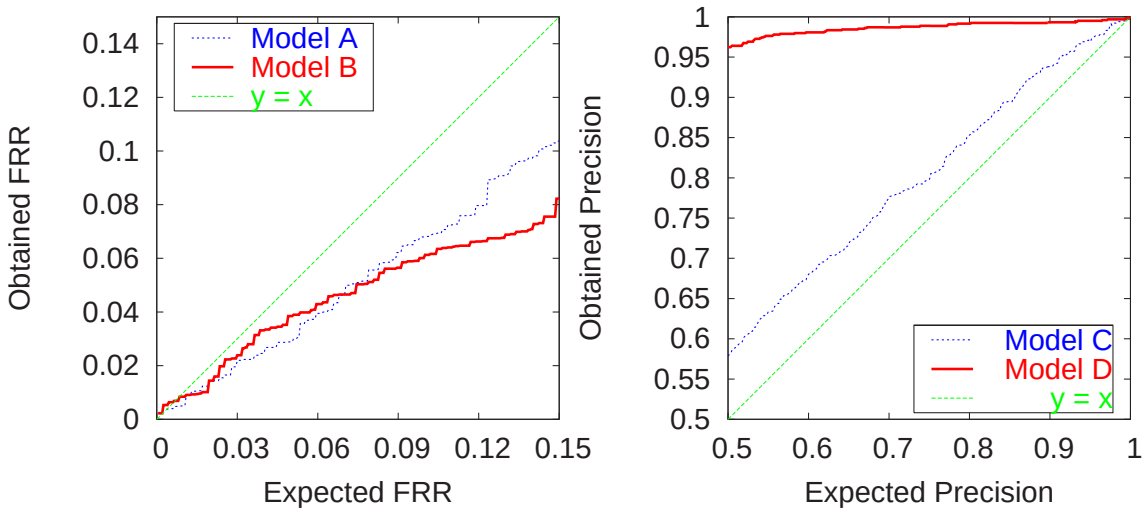
Figure 5: FAR and FRR Expected performance Curves

Algorithm 2 Method to generate the FAR Expected Performance Curve

Let *valid* be the validation set
Let *test* be the test set
Let FAR($\theta, valid$) be the FAR obtained on the validation set for threshold θ
for values v of expected FAR: between 0.0 and 0.5 **do**
  $\theta^\star = \arg\min_\theta |v - \text{FAR}(\theta, valid)|$
  compute FAR($\theta^\star, test$), FRR($\theta^\star, test$) and HTER($\theta^\star, test$)
  plot HTER($\theta^\star, test$) with respect to v
end for

When the criterion is to control the particular value of one of FAR, FRR, Precision, Recall, Sensitivity or Specificity, another EPC curve can then be used. For instance, in person authentication, some applications in the banking domain often expect to control the expected FAR. In this case, Algorithm 2 shows how to prepare such a curve, and Figure 5(a) shows the result for our case study, which compares models for several values of the expected FAR (using again a separate validation set to select the corresponding thresholds). Using this algorithm and the resulting graph, it is clear that model B is better than model A for small values of expected FAR. Figure 5(b) shows the same graph when the criterion is to control the expected FRR instead of FAR. Here, depending on the expected FRR, there is no clear winner between models A and B in this case. Again these two Expected Performance Curves can also be defined for Precision, Recall, Sensitivity and Specificity in the same way.

In order to understand a little bit more the behavior of each model, we can also compare, for each measure of interest (such as FAR, FRR, Precision, Recall, Sensitivity, Specificity), its expected value (computed on a separate validation set) with the obtained one (on the test set). Figure 6(a)



(a) Obtained FRR with respect to expected FRR for models A and B

(b) Obtained Precision with respect to expected Precision for models C and D

Figure 6: Two other EPC curves

shows this curve for FRR applied on models A and B. We see that both models A and B are far from the correct answer (which is represented by the line $y = x$) and always overestimate the actual FRR. The same graph comparing expected and obtained Precision for models C and D can be seen in Figure 6(b). Here, clearly, model D is far from a good estimate! In fact, this bad estimation has a significant impact on the choice of the threshold, which then impacts on the obtained results, hence explaining why the original ROC cannot be used to compare models: they do not take into account the error made during this threshold estimation.

More generally, any measure that does not compute a threshold before applying it on the test set (hence, the ROC and DET curves, but also single measures such as EER and BEP), is subject to this problem. In fact, either implicitly or explicitly, all of these measures compute at least one parameter on the test set, which of course will have a tendency to optimistically bias the corresponding measure.

4.3 Areas Under the Expected Performance Curves

In general, people often prefer to compare their models according to a unique quantitative performance measure, rather than through the use of curves which can be difficult to interpret. The problem however is that one number only (such as the HTER or the F1 measures) could be misleading. One solution proposed by several researchers is to summarize the ROC curve by some approximation of the area under it (see for instance the 11pt-avg measure of equation (13)).

Knowing that the ROC curve is in fact a biased measure of the performance of a system, the corresponding area under it will also be biased. Would it be possible to obtain a corresponding unbiased measure? We propose here to compute the expected average of the two antagonist measures (*i.e.* FAR&FRR, Precision&Recall, etc) given a criterion (such as the expected FAR, Precision, DCF, etc.), and we will show that it is in fact related to the area under the ROC curve (AUROCC).

Let us present this new measure in the context of Text Categorization where the AUROCC is more often used. The average of Precision and Recall, noted $\frac{1}{2}$(P+R), reflects the behavior of the system at one particular threshold θ. Since the precise domain of θ is model dependent, the threshold is in fact inferred through the use of a chosen criterion as explained in section 4.2 (for example, our

model would associate a particular value of θ for an expected Precision of 0.1). Thus, computing the expectation of $\frac{1}{2}$(P+R) over all the possible values of a criterion should give a reasonable expected performance measure of the system at every *reachable* operating point given this particular criterion.

Let $\theta_{f=\alpha}$ be the threshold such that

$$\theta_{f=\alpha} = \arg\min_{\theta} |\alpha - f(\theta)|, \tag{19}$$

we can now write the expected value of $\frac{1}{2}$(P+R) using Precision as threshold selection criterion, as follows:

$$E\left[\frac{1}{2}(\text{P+R})|\text{Precision}\right] = \frac{1}{2}\int_{\alpha\in[0,1]} \left[\text{Precision}(\theta_{\text{Precision}=\alpha}) + \text{Recall}(\theta_{\text{Precision}=\alpha})\right] d\alpha, \tag{20}$$

and using Recall as criterion,

$$E\left[\frac{1}{2}(\text{P+R})|\text{Recall}\right] = \frac{1}{2}\int_{\beta\in[0,1]} \left[\text{Precision}(\theta_{\text{Recall}=\beta}) + \text{Recall}(\theta_{\text{Recall}=\beta})\right] d\beta. \tag{21}$$

Note that if we select the thresholds θ *a posteriori* (directly on the test set) then,

$$\begin{gathered}\text{Precision}(\theta_{\text{Precision}=\alpha}) = \alpha, \quad \text{Recall}(\theta_{\text{Recall}=\beta}) = \beta, \\ \int_{\alpha\in[0,1]} \text{Recall}(\theta_{\text{Precision}=\alpha}) d\alpha = AUROCC \quad \text{and} \\ \int_{\beta\in[0,1]} \text{Precision}(\theta_{\text{Recall}=\beta}) d\beta = AUROCC,\end{gathered} \tag{22}$$

with $AUROCC$ being the area under the ROC curve. Thus, using the fact that $\int_0^1 \gamma d\gamma = \frac{1}{2}$, we can obtain the relation between the expected *a posteriori* $\frac{1}{2}$(P+R) and the area under the ROC curve, by computing the average $G\left(\frac{1}{2}(\text{P+R})\right)$ of equation (20) and equation (21) *a posteriori*:

$$\begin{aligned} G\left(\frac{1}{2}(\text{P+R})\right)_{post} &= \frac{1}{2}\left\{E\left[\frac{1}{2}(\text{P+R})|\text{Precision}\right]_{post} + E\left[\frac{1}{2}(\text{P+R})|\text{Recall}\right]_{post}\right\} \\ &= \frac{1}{2}\left\{AUROCC + \frac{1}{2}\right\}. \end{aligned} \tag{23}$$

Of course, if we select the thresholds using a separate validation set, the result obtained in (23) is not true anymore. However, in this case the average $G\left(\frac{1}{2}(\text{P+R})\right)$ remains interesting since it can be interpreted as a measure summarizing two EPC curves. Indeed, the two components of the average, (20) and (21), are the area under an EPC curve similar to the curve of Figure 5(a) but with $\frac{1}{2}$(P+R) as vertical axis and the expected Precision or Recall as horizontal axis. Obviously, any other threshold selection criterion can also be considered instead. For instance we can consider the following new DCF criterion, similar to (4), that selects θ_γ such that:

$$\theta_\gamma = \arg\min_{\theta}(\gamma \cdot \text{Precision}(\theta) + (1-\gamma)\cdot \text{Recall}(\theta)), \tag{24}$$

and then compute $E\left[\frac{1}{2}(\text{P+R})|\text{DCF}\right]$.

Table 2 show $G\left(\frac{1}{2}(\text{P+R})\right)$ and $E\left[\frac{1}{2}(\text{P+R})|\text{DCF}\right]$ for methods C and D of our text categorization problem. As we can see, the *a posteriori* and *a priori* techniques give similar results here, although slightly contradicting each other.

Note that in equations (20) and (21), we integrate $\frac{1}{2}$(P+R) over values of Precision and Recall from 0 to 1. We can argue that a value of Precision around 0 may be reachable but of no interest. In

Model	$G(\frac{1}{2}(\text{P+R}))_{post}$	$G(\frac{1}{2}(\text{P+R}))$	$E[\frac{1}{2}(\text{P+R})\|\text{DCF}]$
C	0.747	0.778	**0.964**
D	**0.748**	**0.787**	0.954

Table 2: Areas under the EPC curves, for Precision and Recall (text categorization) using both *a priori* and *a posteriori* methods to compute the underlying thresholds.

the field of person authentication it is common to only pay attention to "reasonable" values of FAR and FRR. The values of "reasonable" bounds are task dependent, but their choice can be decisive. Table (3) shows the values of

$$E[\text{HTER}|\text{FAR}] = \frac{1}{2}\int_{\alpha\in[0,u_\alpha]} [\text{FAR}(\theta_{FAR=\alpha}) + \text{FRR}(\theta_{FAR=\alpha})]\, d\alpha, \tag{25}$$

computed for two different bounds u_α (0.1 and 0.5). This shows that depending on the tolerance to false acceptances, the best model can be interpreted as A or B.

Model	$E[\text{HTER}\|\text{FAR} < 0.1]$	$E[\text{HTER}\|\text{FAR} < 0.5]$
A	0.014	**0.082**
B	**0.013**	0.086

Table 3: Areas under the EPC curves, for Half Total Error Rate (person authentication), with different bounds and criteria.

5 Conclusion

In this paper, we have explained why the current use of ROC and DET curves, as well as *a posteriori* measures such as EER and BEP, used regularly in several publications related to domains such as person authentication, text categorization, or medical applications, can be misleading when used to compare performances between models or to assess the expected performance of a given model.

We have thus proposed the use of new curves called Expected Performance Curves (EPC), which reflect more precisely the criteria underlying the real application and hence enable a more realistic comparison between models as well as a better analysis of their respective expected performances. From these curves, several single measures can also be obtained, and all of them should reflect a realistic performance comparison for a particular (and reachable) operating point of the system. Moreover, a summary measure, reflecting the expected performance of the system under all reachable conditions, has also been proposed. Note that a free software is available to compute these curves and statistics (`http://www.torch.ch/extras/epc`).

It might be argued that one weakness of this new set of measures is the need for a separate validation set. While this is true and necessary in order to obtain realistic expected performances, one could always rely on cross-validation techniques to solve this problem of a lack of training data.

ROC and DET curves can certainly still be used when the goal is to understand the behavior of a model without taking into account the selection of the threshold, however this should be done with caution, since it does not correspond to a real application setting.

6 Acknowledgments

This research has been carried out in the framework of the Swiss NCCR project (IM)2.

References

[1] J. Mariéthoz and S. Bengio. An alternative to silence removal for text-independent speaker verification. Technical Report IDIAP-RR 03-51, IDIAP, Martigny, Switzerland, 2003.

[2] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proceedings of Eurospeech'97, Rhodes, Greece*, pages 1895–1898, 1997.

[3] A. Martin and M. Przybocki. The NIST 1999 speaker recognition evaluation - an overview. *Digital Signal Processing*, 10:1–18, 2000.

[4] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[5] H. L. Van Trees. *Detection, Estimation and Modulation Theory, vol. 1.* Wiley, New York, 1968.

[6] P. Verlinde, G. Chollet, and M. Acheroy. Multi-modal identity verification using expert fusion. *Information Fusion*, 1:17–33, 2000.

[7] Zweig and Campbell. ROC plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4):561–577, 1993.