IDIAP RESEARCH REPORT

Significance Tests for *Bizarre* Measures in 2-Class Classification Tasks

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

IDIAP–RR 04-34

June 7, 2004

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, mkeller@idiap.ch

[2] IDIAP, CP 592, 1920 Martigny, Switzerland, marietho@idiap.ch

[3] IDIAP, CP 592, 1920 Martigny, Switzerland, bengio@idiap.ch

IDIAP Research Report 04-34

Significance Tests for *Bizarre* Measures in 2-Class Classification Tasks

Mikaela Keller Johnny Mariethoz Samy Bengio

June 7, 2004

Abstract. Statistical significance tests are often used in machine learning to compare the performance of two learning algorithms or two models. However, in most cases, one of the underlying assumptions behind these tests is that the error measure used to assess the performance of one model/algorithm is computed as the sum of errors obtained on each example of the test set. This is however not the case for several well-known measures such as F_1, used in text categorization, or DCF, used in person authentication. We propose here a practical methodology to either adapt the existing tests or develop non-parametric solutions for such *bizarre* measures. We furthermore assess the quality of these tests on a real-life large dataset.

Contents

abstract

1 Introduction

In Machine Learning like in Physics, Biology or Psychology, when a result is presented, in order to make it reliable, the question of its statistical significance should be addressed. Indeed, the level of confidence we have in a result should be part of the report. In Machine Learning, for example, when comparing two trained models on a common test set, we would like to know if the better performance we obtained for model A (as compared to model B) is only an artifact of the test set or if it is a more general result.

Several researchers (see for instance [1] and [4]) have proposed statistical tests suited for 2-class classification tasks where the performance is measured in terms of the classification error (ratio of the number of errors and the number of examples), which enables the use of assumptions based on the fact that the error can be seen as a sum of random variables over the examples.

However in a wide range of domains, 2-class classification tasks are solved within a domain specific framework which differs from the general framework in their choice of performance measure; instead of using the *classification error*, some find it more convenient to first consider separately the error (or performance) of each class and combine them in some way. These *bizarre* measures can therefore not be considered as sums of random variables over the examples, thus the usual assumptions are no longer correct, and most of the currently proposed tests cannot be applied directly.

Following the taxonomy of questions of interest defined by Dietterich in [1], we can differentiate between statistical tests that analyze learning algorithms and statistical tests that analyze classifiers. In the first case, one intends to be robust to possible variations of the train and test sets, while in the latter, one intends to be only robust to variations of the test set. While the methods discussed in this paper can be applied alternatively to both approaches, we will concentrate here on the second one, as it is more tractable (for the empirical section) while still corresponding to real life situations where the training set is fixed and one wants to compare two solutions (such as during a competition).

We thus propose in this paper a practical methodology to apply statistical tests to bizarre performance measures such as the well-known F_1 measure used in text categorization or the so-called DCF measure used in person authentication. Moreover, using a very large dataset (the extended Reuters dataset [5]), we propose to verify how reliable these tests are on real data.

In the following section we quickly recall some of the currently used statistical significance tests for 2-class classification tasks and introduce the problem of bizarre measures. In section 3, we present one kind of measure for which the usual tests can in fact be adapted quite easily using some simple algebra, while section 4 gives an example of a performance measure for which such adaptation is simply not possible and a non-parametric approach is thus proposed. Finally, in section 5 we present the results of an experiment done with real data in order to assess the quality of the non-parametric test proposed in section 4.

2 Statistical Significance Tests for 2-Class Classification Tasks

Let us first remind the basic classification framework in which statistical significance tests are used in machine learning. We consider comparing two models A and B on a two-class classification task where the goal is to classify input examples x_i into the corresponding class $y_i \in \{-1, 1\}$, using already trained models $f_A(x_i)$ or $f_B(x_i)$. One can estimate their respective performance on some test data by counting the number of utterances of each possible outcome: either the obtained class corresponds to the desired class, or not. Let N_e be the number of errors of a model and N the total number of test examples; the classification error C can thus be seen as a proportion of errors:

$$C = \frac{N_e}{N} \,. \tag{1}$$

The difference between two models A and B can then be written as

$$D = C_A - C_B = \frac{N_{e,A} - N_{e,B}}{N} \qquad (2)$$

where $N_{e,A}$ is the number of errors of model A while $N_{e,B}$ is the number of errors of model B.

The usual starting point of most statistical tests is to define the so-called *null hypothesis* which considers the two models as equivalent, and then verifies how probable this hypothesis is. Hence, assuming that D is an instance of some random variable $\mathbf{D}$ which follows some distribution, we are interested in

$$p(|\mathbf{D}| < |C_A - C_B|) = \delta \qquad (3)$$

where δ represents the risk of selecting the *alternate hypothesis* (the two models are different) while the *null hypothesis* is in fact true. This can in general be estimated easily when the distribution of $\mathbf{D}$ is known. In the simplest case, known as the *proportion test*, one assumes (reasonably) that the decision taken by each model on each example can be modeled by a Binomial, and further assumes (wrongly) that C_A and C_B are independent. When N is large, this leads to estimate $\mathbf{D}$ as a Normal distribution with zero mean and standard deviation σ_D

$$\sigma_D = \sqrt{\frac{2\bar{C}(1-\bar{C})}{N}} \qquad (4)$$

where $\bar{C} = \frac{C_A - C_B}{2}$ is the average classification error.

In order to get rid of the wrong independence assumption between C_A and C_B, [7] proposes to only take into account the examples for which the two models A and B disagree. Let N_{AB} be the number of test examples for which models A and B gave different decisions. The difference between the two models can now be written as

$$D = C_A - C_B = \frac{N_{AB}}{N} \qquad (5)$$

and one can show that in that case $\mathbf{D}$ follows a Normal distribution with zero mean and standard deviation σ_D

$$\sigma_D = \frac{\sqrt{N_{AB}}}{N} \,. \qquad (6)$$

This is also very similar to the well-known McNemar test, which instead considers $\mathbf{D}$ as a χ^2 distribution but yields basically the same result for large N, which is the case in Machine Learning [1].

More recently, several other statistical tests have been proposed, such as the 5x2cv method [1] or the variance estimate proposed in [4], which both claim to better estimate the distribution of the errors (and hence the confidence on the statistical significance of the results).

However, all these solutions assume that the error of one model (C_A or C_B) is the average of some random variable (the error) estimated on each example. Intuitively, this average will thus tend to be Normally distributed as N grows, following the central limit theorem. On the other hand, often for historical reasons, several machine learning tasks are not measured in this way. For instance, in text categorization or information retrieval, researchers use the well-known F_1 measure [6], and in person authentication, researchers use the DCF measure [3]. Both are in fact aggregate measures of the whole test set which cannot be seen directly as proper proportions which denominator would be the number of examples. In the following two sections, we propose two solutions that can be used to measure statistical significance in such cases.

3 Adapted Proportion Test for Person Authentication Tasks

The Detection Cost Function (DCF) is a general performance measure used in person authentication, where the task is to verify the claimed identity of a person using his or her biometric information;

voice, face, form of the hand, etc. If the claimed identity is right, the person is considered as a client (which we will call hereafter the positive class), otherwise he or she is considered as an impostor (the negative class).

Let us consider the following decomposition of the various outcomes of a classification system: let N_{tp} be the number of true positives (clients that were recognized as such), N_{tn} the number of true negatives (impostors recognized as such), N_{fp} the number of false positives (impostors recognized as clients) and N_{fn} the number of false negatives (clients recognized as impostors). The DCF can be written as:

$$\mathrm{DCF} = \underbrace{\mathrm{Cost}(fn) \cdot P(p) \cdot \frac{N_{fn}}{N_{fn} + N_{tp}}} + \underbrace{\mathrm{Cost}(fp) \cdot P(n) \cdot \frac{N_{fp}}{N_{fp} + N_{tn}}} \tag{7}$$

where $\mathrm{Cost}(fn)$ is the cost of rejecting a true client, $\mathrm{Cost}(fp)$ is the cost of accepting an impostor, $P(p)$ is the prior probability of a client (positive) access, while $P(n)$ is the prior probability of an impostor (negative) access. The DCF thus enables to fix the relative costs of each class while being robust to the actual train distribution, which may not reflect the real one.

It is easy to see that the two under-braces consider different categories of examples: the first one computes the weighted error on the positive class examples (the client accesses), while the second one computes the weighted error on the negative class examples (the impostor accesses). Hence, the error made on each category of examples can still be seen as a (weighted) proportion and both proportions can be considered independent as they are computed on two different populations of examples (clients and impostors). We can thus use the fact that the sum of two independent Normal distributions is a Normal which variance is the sum of the underlying variances. Let us denote $N_n = N_{fp} + N_{tn}$ the number of negative examples and $N_p = N_{fn} + N_{tp}$ the number of positive examples. One can derive the proportion test when assuming (wrongly) independence between DCF_A and DCF_B, which, after some simple algebra, yields a Normal distribution with standard deviation

$$\sigma_{dcf} = \sqrt{2\left[\mathrm{Cost}(fn)^2 P(p)^2 \frac{\bar{N}_{fn}}{N_p^2} \cdot \left(1 - \frac{\bar{N}_{fn}}{N_p}\right) + \mathrm{Cost}(fp)^2 P(n)^2 \frac{\bar{N}_{fp}}{N_n^2} \cdot \left(1 - \frac{\bar{N}_{fp}}{N_n}\right)\right]} \tag{8}$$

where $\bar{N}_{fn} = \frac{N_{fn,A}+N_{fn,B}}{2}$ is the average of the number of false negatives of models A and B, and $\bar{N}_{fp} = \frac{N_{fp,A}+N_{fp,B}}{2}$ is the average of the number of false positives of models A and B.

Taking into account the fact that the errors of models A and B are computed on the same test set (hence the two DCFs are not independent), we can again concentrate on only those examples which were differently classified by both models A and B, as suggested in [7]. Let $N_{n,AB}$ be the number of negative examples for which models A and B gave different decisions, and similarly for $N_{p,AB}$, we then obtain

$$\sigma_{dcf} = \sqrt{\frac{\mathrm{Cost}(fn)^2 P(p)^2 \cdot N_{p,AB}}{N_p^2} + \frac{\mathrm{Cost}(fp)^2 P(n)^2 \cdot N_{n,AB}}{N_n^2}}\ . \tag{9}$$

Hence, for measures such as DCF, it is still possible to modify classical statistical tests (note that using the same strategy, it is also possible to use the statistical tests proposed in [1] or [4]). On the other hand, we will see in the next section that there are still other performance measures for which no such simple adaptation can hold.

4 Bootstrap Percentile Test for F_1

Text categorization is the task of assigning one or several categories, among the predefined set $\mathcal{C} = \{c_1, \ldots, c_K\}$, to textual documents. As explained in [6], text categorization is usually solved as K 2-class classification problems, in a one-against-the-others approach. In this field two measures are considered of importance:

$$\text{Precision} = \frac{N_{tp}}{N_{tp} + N_{fp}}, \quad \text{and} \quad \text{Recall} = \frac{N_{tp}}{N_{tp} + N_{fn}}.$$

These are effectiveness measures, *i.e.* inside $[0, 1]$ interval, the closer to 1 the better. For each category k, Precision$_k$ measures the proportion of documents of the class among the ones considered as such by the classifier and Recall$_k$ the proportion of documents of the class correctly classified.

To summarize these two values, it is common to consider the so-called F_1 measure [8], which is the inverse of the harmonic mean of Precision and Recall:

$$F_1 = \left(\frac{1}{2}\left[\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}\right]\right)^{-1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

$$= \frac{2N_{tp}}{2N_{tp} + N_{fn} + N_{fp}}. \tag{11}$$

Let us consider two models A and B, which achieve a performance measured by F_{1_A} and F_{1_B} respectively. The difference $dF_1 = F_{1_B} - F_{1_A}$ does not fit the assumptions of the tests presented in Section 2. Indeed, it cannot be decomposed into a sum over the documents of independent random variables, since the numerator and the denominator of dF_1 are non constant sums over documents of independent random variables For the same reason F_1, while being a proportion, cannot be considered as a random variable following a Normal distribution for which we could easily estimate the variance.

Hence, we would like to present here an alternative solution to measure the statistical significance of dF_1, based on the Bootstrap Percentile Test proposed in [2]. The idea of this test is to approximate the unknown distribution of dF_1 by an estimate based on bootstrap replicates of the data.

Notice that for each document the intersection of the 2 classifiers response can be transcribed into eight possible events, described in Table 1. For example, the event e_1 represents a positive response of both classifiers when the document belongs to the tested class.

class 1	TP_A	FN_A	class -1	FP_A	TN_A
TP_B	e_1	e_2	FP_B	e_5	e_6
FN_B	e_3	e_4	TN_B	e_7	e_8

Table 1: Eight possible joint outcomes between models A and B.

We can then rewrite dF_1 as follows:

$$dF_1 = 2\left[\frac{N_{e_1} + N_{e_3}}{N_p + N_{e_1} + N_{e_3} + N_{e_4} + N_{e_6}} - \frac{N_{e_1} + N_{e_2}}{N_p + N_{e_1} + N_{e_2} + N_{e_4} + N_{e_5}}\right] \tag{12}$$

where N_{e_j} is the number of times the event e_j occurs in the sample. Let S_0 be the sample of test set document's events, *i.e.* the i^{th} element of S_0 corresponds to the intersection of the classifiers response for the i^{th} document in the test set. Let us assume that we have N such documents ($card(S_0) = N$). We draw *with replacement*, as illustrated in Table 2, among the elements of the set S_0, N elements to form the first bootstrap sample S_1. We then compute $dF^*_{1_1}$ from S_1 using eq. (12), the first bootstrap replicate of dF_1. We repeat this process M times and thus obtain M replicates of dF_1, which altogether represent a non-parametric estimate of $p(dF_1)$. Using this estimate, we can finally compute the probability that dF_1 is positive by simply counting the fraction of these M values that were positives:

$$p(dF_1 > 0) = \frac{N_{dF^*_{1_l} > 0}}{M} \tag{13}$$

where $N_{dF^*_{1_l} > 0}$ is the number of bootstrap replicates which yielded a positive dF^*_1.

	doc_1	doc_2	doc_3	doc_4	$\dots$	doc_N	
model A	TP_A	FP_A	FN_A	TP_A	$\dots$	TN_A	
model B	TP_B	TN_B	TP_B	FN_B	$\dots$	FP_B	
S_0	e_1	e_7	e_2	e_3	$\dots$	e_6	$\rightarrow dF_{1_0}$
S_1	e_1	e_6	$\dots$	$\dots$	$\dots$	e_7	$\rightarrow dF^*_{1_1}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
S_M	e_8	e_6	e_5	$\dots$	$\dots$	e_3	$\rightarrow dF^*_{1_M}$

Table 2: Illustration of the bootstrap process

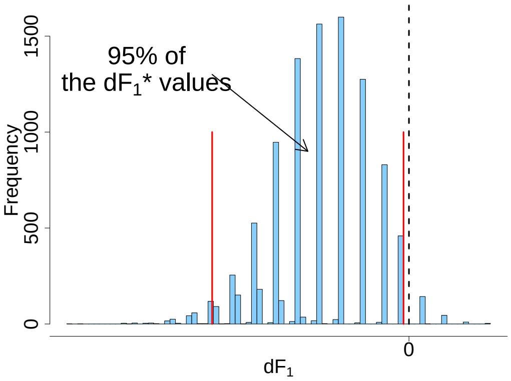

Figure 1: Histogram of dF^*_1 on the category set X

Similarly, we can estimate the probability that dF_1 is inside an interval $[a, b]$ as:

$$p(a < dF_1 < b) = \frac{N_{a<dF^*_{1_l}<b}}{M} \,. \qquad (14)$$

Let us consider selecting a and b such that $p(a < dF_1 < b) = 95\%$, centered around the mean of $p(dF_1)$, as illustrated in Figure 1. If 0 is outside the interval, we can say that $dF_1 = 0$ is not among the most frequent results, thus we can say that model A and model B are different with 95% confidence. On the contrary, if 0 is inside the interval, we notice that among the most frequent dF_1 values some are positives and others are negatives, that is, sometimes model A is better than model B and sometimes it is the converse. In this case we conclude that we do not have enough confidence to say that the two models are really different.

Other non-parametric tests such as the sign test are sometimes advocated in the related literature (see for instance [8] or [9]). They may assess whether model A is better than model B, but the assessment will not be based on the difference of F_1, and thus will not give any answer about the significance of the difference of such measure.

5 Experiments

In order to assess the quality of the bootstrap estimate of the significance of the difference of F_1, we have run several experiments on the very large RCV1 Reuters dataset [5], which contains up to 806,791 documents. We divided it as follows: 7982 documents were used as a training set D_{tr} (to train models A and B), 7909 documents were used as a test set D_{te} (to estimate which of model A or B is the best one, and with which confidence), and the rest, called D_{true} and containing 790900 documents, was used to verify if our estimate of the confidence we have on dF_1 was reasonable. There was a total of 100 categories and we defined 3 sub-problems with 3 disjoints subsets of 21 categories, which we will refer to as X, Y and Z.

We first extracted the dictionary from the training set, removed stopwords and applied stemming to it, as normally done in text categorization. Each document was then represented as a bag-of-words using the usual *tfidf* coding. We then trained two different models (model A was a set of one-against-the-others Support Vector Machines (SVMs) with linear kernel, while model B was also a set of SVMs but with a Gaussian kernel, properly tuned using cross-validation on the training set).

After training, models A and B were used to compute dF_1 on the test set D_{te}. We then applied the bootstrap technique described in section 4 on the test set D_{te} in order to compute the estimator dF_1^* (using 10000 bootstrap replicates) of the distribution of dF_1. Finally, we used the remaining dataset D_{true} in order to obtain a more precise estimate of the distribution of dF_1 by bootstraping a large number (1000) of subsets of size 7909 from D_{true} and estimated dF_1 on each of them. It can be argued that doing so, the subsets may not be totally independent. While this is true, it can be shown that the average overlap between 2 subsets is less than 1%, in the case of sub-sampling sets 100 times smaller than the actual set[1]. This is of course not the case when bootstraping sets of the same size as the actual set (as done on D_{te}), where the average overlap between 2 sets is almost 40%. We would thus like to verify whether, despite this overlap, the results obtained with the Bootstrap Percentile Test are reliable.

Results of the experiments over the category sets X, Y and Z are displayed in Table 3.

	Bootstrap percentile Test confidence	Percentage on D_{true} sub-sampling
Experiment over X	95.6% ($F_{1_B}^* > F_{1_A}^*$)	92.4%
Experiment over Y	54.2% ($F_{1_B}^* > F_{1_A}^*$)	69.1%
Experiment over Z	96.5% ($F_{1_A}^* > F_{1_B}^*$)	82.5%

Table 3: Confidence on $|F_{1_B} - F_{1_A}| > 0$

As can be seen, in the experiment over X, both the bootstrap percentile test, plotted in Figure 1, and the estimation over D_{true}, are quite confident on F_{1_B} being higher than F_{1_A}. However, the first seems to over-estimate this confidence. We can make the same observation in the experiment over Z, where the bootstrap on D_{te} is over-optimistic as compared to the estimate on D_{true}. On the contrary, on the experiment over Y, the bootstrap test gives a smaller level of confidence than the estimate over D_{true}, to any model being better than the other, the latter not being very confident either. A possible explanation for these untidy results is that the percentile estimation is based on the tails of the distribution and thus the smaller the dataset the less data it has to rely on. While it is clear that the proposed estimate on the small test set does not seem to be very accurate, it is important to verify whether it includes an intrinsic bias. This will be verified in future experiments.

6 Conclusion

In this paper we have proposed two different methods to estimate the statistical significance of the difference between models when this difference is computed in terms of *bizarre* measures, for which the assumptions underlying the usual statistical tests do no hold anymore.

We have first shown that in some cases, simple arithmetics could be used to derive a correct test (as for the DCF measure). For other cases, such as for the F_1 measure, we proposed a non-parametric approach based on bootstrap replicates.

Finally, using a very large (non-artificial) dataset, we proposed an empirical framework to evaluate the quality of the proposed tests.

[1] On the other hand, dividing the data into 100 non-overlaping subsets would also result in some dependencies between the subsets, since one example drawn in one subset could never appear in any others (another way to see the dependency is to note that the last subset is not really drawn, as it is composed of what was left from the other subsets).

Acknowledgments

This research has been partially carried out in the framework of the Swiss NCCR project (IM)2. It was also supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss OFES. This publication only reflects the authors' views.

References

[1] T.G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924, 1998.

[2] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap.* Chapman and Hall, 1993.

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

[4] Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003.

[5] T.G. Rose, M. Stevenson, and M. Whitehead. The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In *In Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria*, 2002.

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

[7] G. W. Snedecor and W. G. Cochran. *Statistical Methods.* Iowa State University Press, 1989.

[8] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, UK, 1975.

[9] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In Marti A. Hearst, Fredric Gey, and Richard Tong, editors, *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, pages 42–49, Berkeley, US, 1999. ACM Press, New York, US.