The balanced accuracy and its posterior distribution

Kay H. Brodersen*†, Cheng Soon Ong*, Klaas E. Stephan† and Joachim M. Buhmann*

*Department of Computer Science, ETH Zurich, Switzerland; kay.brodersen@inf.ethz.ch
†Institute for Empirical Research in Economics, University of Zurich, Switzerland

***Abstract*—Evaluating the performance of a classification algorithm critically requires a measure of the degree to which unseen examples have been identified with their correct class labels. In practice, generalizability is frequently estimated by averaging the accuracies obtained on individual cross-validation folds. This procedure, however, is problematic in two ways. First, it does not allow for the derivation of meaningful confidence intervals. Second, it leads to an optimistic estimate when a biased classifier is tested on an imbalanced dataset. We show that both problems can be overcome by replacing the conventional point estimate of accuracy by an estimate of the posterior distribution of the balanced accuracy.**

***Keywords*-classification performance; generalizability; bias; class imbalance**

I. Introduction

Using a meaningful measure of generalizability is a key requirement for evaluating the performance that a classification algorithm has achieved on a given dataset. Since the true ability of an algorithm to correctly predict class labels of unseen data could only be determined if an infinite amount of test data was available, generalizability has to be approximated by an estimate instead (see [1] for an overview). Repeatedly splitting up the available data into a training set and a test set by means of cross-validation is a popular procedure for this, though it leaves an important question unanswered: based on a set of fold-wise cross-validation results, which measure of generalizability should be reported? In most classification settings, there is no specific need to impose different costs on different types of misclassification, and so the overall accuracy is of primary interest. In these cases, the most commonly adopted approach for summarizing cross-validation results is to report the average accuracy (or average error) across all folds. However, measuring performance in this way has two critical shortcomings. First, because the approach is non-parametric, it does not make it possible to compute meaningful confidence intervals of a true underlying quantity. In particular, computing the standard error of the mean across all folds is intrinsically flawed as it enforces symmetric limits and may lead to confidence intervals of accuracy including values above 100%. The second flaw in considering the average accuracy is that it may give a misleading idea about generalization performance in situations where a biased classifier is tested on an imbalanced dataset. Under these conditions, the average accuracy may lead to false conclusions about the significance with which an algorithm has performed better than chance.

In this paper, we argue that both shortcomings can be overcome by replacing the average accuracy by the posterior distribution of the balanced accuracy. In order to keep our treatment self-contained, we begin by briefly reviewing the current state of the art. Specifically, instead of giving a point estimate of the accuracy, we describe the well-known approach of estimating the *posterior distribution* of the accuracy (Section II). Based on this idea, our contribution is the following: we propose to replace the accuracy by the *balanced accuracy*, and we show how to estimate its posterior probability distribution under parametric assumptions (Section III). We illustrate the utility of our approach (Section IV) and briefly discuss our findings (Section V).

II. The posterior accuracy

In a binary classification setting, let n be the number of examples underlying a leave-m-out cross-validation scheme with k folds. We assume $k|m$, which implies that each fold contains $n-m$ training instances and m test cases.

A common way of computing an estimate of generalizability begins by summing the number of correctly labelled test cases, C, across all cross-validation folds, $C = \sum_{i=1}^{k} r_i$, where $r_i \in \{0, \dots, m\}$. The average accuracy can then be reported as the fraction $\frac{C}{n}$. It is worth noting that, since *classification error* = 1 − *classification accuracy*, the mean error could be reported instead, and this equivalence pertains to all other accuracy-related quantities discussed throughout this paper. In any case, however, point estimates by themselves are not sufficient to assess statistical significance.

There are two types of hypothesis that we often wish to test. First, is a classification algorithm operating at the level of guessing, or is its generalization accuracy significantly above chance? Second, more generally, does a classification algorithm significantly outperform an alternative algorithm? Both questions require statistical inference on a measure of generalizability.

In order to determine, for instance, whether a given classification outcome is the result of an algorithm that operates significantly above chance, one well-known possibility is to regard each test case as an independent Bernoulli experiment and compare $\frac{C}{n}$ to the level that must be reached by an

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.764

above-chance learning algorithm. The significance threshold is $F^{-1}(1-\alpha)$, where, for example, $\alpha = 0.05$, and where F^{-1} is the inverse cumulative density function of the Binomial distribution with parameters n and $p = \frac{1}{2}$. While this approach provides an estimate of significance, it does not associate the accuracy with a measure of precision. As a result, it does not readily support a principled way of directly comparing two algorithms both of which have been found to operate above chance.

One way of estimating the variance of the accuracy is to consider the standard error of the mean, $\hat{\sigma}/\sqrt{n}$, where $\hat{\sigma}$ is the empirical standard deviation of $\frac{C}{n}$, observed across all cross-validation folds. However, this quantity is dependent on arbitrary design choices such as m, the number of test cases in each cross-validation fold, and, worse still, may easily lead to error bars including values above 100%.

An alternative to the schemes described above is to adopt a probabilistic view on generalizability [2, pp. 68–74]. Rather than averaging the outcomes obtained on different cross-validation folds, we can use Bayesian statistics to express our uncertainty about the underlying generalizability [1]. Specifically, we can treat cross-validation results as outcomes of a Binomial experiment and view each test case as drawing with replacement from a bucket with an unknown mixture of 'correct' and 'incorrect' balls. Given C and I draws of such 'correct' and 'incorrect' balls, respectively, and assuming a uniform (flat) prior on the interval $[0, 1]$, the posterior distribution of the fraction A of 'correct' balls is given by the the conjugate prior of the Binomial distribution, i.e., the Beta distribution

$$A \sim Beta(a, b) \tag{1}$$

with $a = C + 1$ and $b = I + 1$. Hence, when parameterized in this way, it is the posterior probability density of the probability x of classifying correctly unseen examples drawn from the same source as the training data,

$$p_A(x; C, I) = \frac{1}{B(C + 1, I + 1)} x^C (1 - x)^I, \tag{2}$$

where $B(\cdot)$ is the Beta function and C and I denote the number of correct and incorrect predictions, respectively [2, pp. 68–74]. Thus, the generalizability of a classification algorithm can be described in terms of its posterior accuracy distribution. For example, we could report:

$$\begin{aligned}
\text{the mean} \quad & \frac{C+1}{C+I+2} && (3)\\
\text{the median} \quad & F_B^{-1}\left(\tfrac{1}{2}; C+1, I+1\right) && (4)\\
\text{the mode} \quad & \frac{C}{C+I} && (5)\\
\text{a posterior} \quad & [F_B^{-1}\left(\tfrac{\alpha}{2}; C{+}1, I{+}1\right); && \\
\text{probability interval} \quad & F_B^{-1}\left(1{-}\tfrac{\alpha}{2}; C{+}1, I{+}1\right)] && (6)
\end{aligned}$$

where F_B^{-1} is the inverse cumulative density function of the Beta distribution. Note that the conventional average accuracy discussed at the beginning can now be interpreted as the mode of the posterior of the true accuracy under the assumption of a flat Beta prior over accuracies (see Section IV for an illustration). In the next section, we will proceed to the main contribution of this paper, by showing how a key limitation of the above approach can be overcome by considering a different performance measure.

III. The posterior balanced accuracy

Given a confusion matrix of classification results, the accuracy can be a misleading performance measure. Specifically, it may falsely suggest above-chance generalizability. How may this situation arise?

It is a well-known phenomenon in binary classification that a training set consisting of different numbers of representatives from either class may result in a classifier that is *biased* towards the more frequent class. When applied to a test set that is imbalanced in the same direction, this classifier may yield an optimistic accuracy estimate. In an extreme case, the classifier might assign every single test case to the large class, thereby achieving an accuracy equal to the fraction of the more frequent labels in the test set.

Previous studies have examined different ways of addressing this problem, see [3], [4], [5]. One strategy, for example, is to restore balance on the training set by undersampling the large class or by oversampling the small class. Another strategy is to modify the costs of misclassification in such a way that no bias is acquired. However, while these methods may under some circumstances prevent a classifier from becoming biased, they do not provide generic safeguards against reporting an optimistic accuracy estimate. This observation motivates the use of a different generalizability measure: the *balanced accuracy*, which can be defined as the average accuracy obtained on either class. Based on a confusion matrix

	actual	
	+	−
predicted+	TP	FP
predicted−	FN	TN

the balanced accuracy (before we adopt a probabilistic viewpoint) is given by $\frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right)$. If the classifier performs equally well on either class, this term reduces to the conventional accuracy (number of correct predictions divided by number of predictions). In contrast, if the conventional accuracy is high only because the classifier takes advantage of an imbalanced test set, then the balanced accuracy, as desired, will drop to chance.

Unlike the measure described in [6], the balanced accuracy used here is symmetric about the type of class. If desired, this symmetry assumption can be dropped, yielding $c \times \frac{TP}{P} + (1-c) \times \frac{TN}{N}$, where $c \in [0, 1]$ is the cost associated with the misclassification of a positive example.

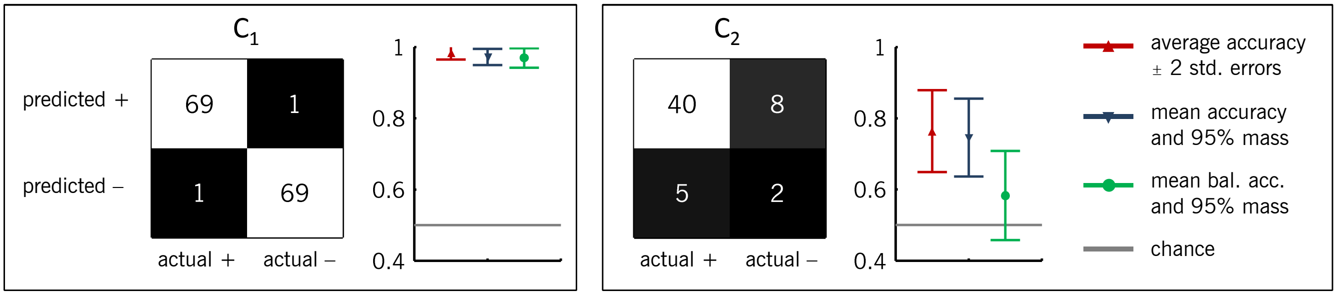

Figure 1. Comparison of accuracy measures for two illustrative confusion matrices C_1 and C_2. The first example shows how the conventional average accuracy (red) may imply a confidence interval that includes values above 100%. The second example shows how accuracies, unlike balanced accuracies (green), falsely suggest above-chance generalizability in the case of a biased classifier that has taken advantage of an imbalanced test set.

In analogy with our discussion of accuracies in Section II, a probabilistic view allows us to treat the balanced accuracy as a random variable and reason about its posterior distribution. Depending on whether the true label of a given test case is positive or negative, let us regard a prediction as a draw either (i) from a bucket of 'true positive' or 'false negative' balls, or (ii) from a bucket of 'true negative' or 'false positive' balls. We are interested in the probability density of $\frac{1}{2}(A_P + A_N)$, where A_P and A_N are random variables specifying the accuracy on positive and negative examples, respectively. This density can be derived from the convolution of two Beta distributions,

$$p_B(x;\ TP, FP, FN, TN) = \int_0^1 p_A\left(2(x-z); TP+1, FN+1\right) \cdot p_A\left(2z; TN+1, FP+1\right) \quad dz, \qquad (7)$$

where $p_A(x)$ is the density of the accuracy as defined in (2) and $p_B(x)$ is the density of the balanced accuracy. Thus, assuming a flat prior on the true balanced accuracy, we can report cross-validation results by describing the posterior distribution of the balanced accuracy. Note that the mean (mode) of the distribution of the balanced accuracy does not necessarily equal the mean of the means (modes) of the separate accuracy distributions for positive and negative examples. There are no analytical forms for the moments considered in eqns. (3) through (6). However, we can compute numerical approximations (see Section IV). A complete set of MATLAB routines for this is available online.[1]

IV. Illustrative Examples

The utility of the balanced accuracy and its posterior distribution could be illustrated using real-world data, but the key properties can be demonstrated best on the basis of a small set of hand-crafted examples.

As a result of training and testing two independent classifiers on different datasets, let C_1 and C_2 be the confusion matrices of the respective results, summed across all cross-validation folds. We wish to compare the average accuracy (along with standard errors) to the posterior accuracy and the posterior balanced accuracy (see Figure 1).

In the first example, the test set is perfectly balanced (70 positive vs. 70 negative examples). As a result, the differences between the three accuracy measures are not substantial. However, the simulation does illustrate how 2 standard errors around the average accuracy yield an interval that includes values above 100% (Figure 1, left box, red interval). In contrast, the probability intervals of the posterior accuracy and balanced accuracy show the desired asymmetry (blue and green intervals).

In the second example, both the average accuracy and the mean of the posterior accuracy seem to indicate strong classification performance (Figure 1, right box, red and blue intervals). The balanced accuracy, by contrast (green interval), reveals that in this simulation the test set was imbalanced (45 positive vs. 10 negative examples) and, in addition, the classifier had acquired a bias towards the large class (48 positive vs. 7 negative predictions). Accuracy measures on their own fail to detect this situation and give the false impression of above-chance generalizability.

The difference between accuracies and balanced accuracies is further illustrated in Figure 2. Based on the confusion matrix C_2, the two plots show all of the statistics mentioned in Sections II and III superimposed on the central 95% probability interval of the respective posterior distributions. The figure also contains the 'average balanced accuracy' as originally defined in the context of (7), computed as the mean of the modes of the accuracies on positive and negative examples. The simulation shows how a biased classifier applied to an imbalanced test set leads to a hugely optimistic estimate of generalizability when measured in terms of the accuracy rather than the balanced accuracy.

V. Discussion

In binary classification, confusion matrices form the basis of a multitude of informative measures of generalizability.

[1] http://people.inf.ethz.ch/bkay/downloads

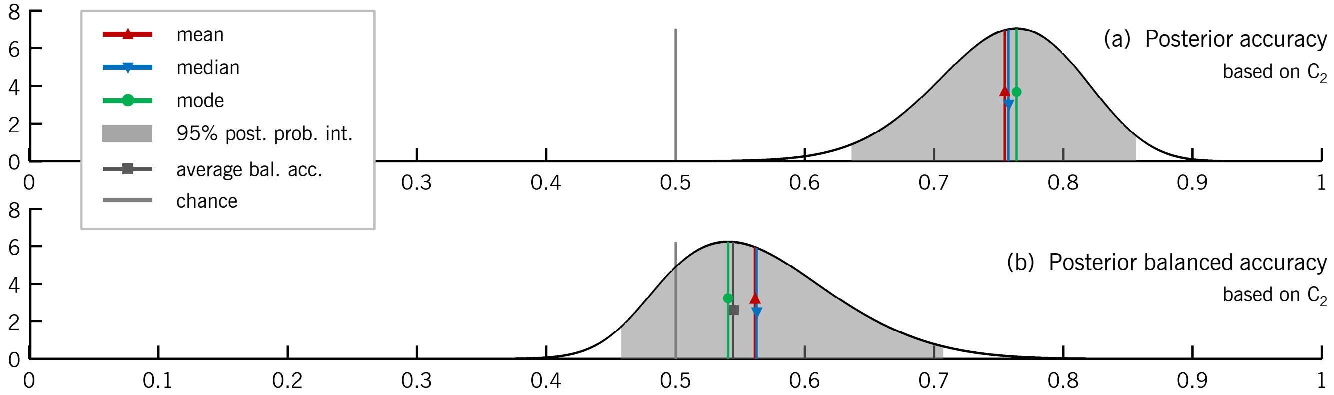

Figure 2. Comparison between the posterior distribution of the accuracy and the balanced accuracy, based on the confusion matrix C_2 depicted in Fig. 1.

Yet, it is still common to average accuracies across cross-validation folds. This approach neither supports meaningful confidence intervals; nor does it provide safeguards against a biased classifier that has taken advantage of an imbalanced test set. The first limitation can be overcome by the well-known approach of considering the full posterior distribution of the accuracy instead of a point estimate [2, pp. 68–74]; and the second one by the idea of replacing conventional accuracies by balanced accuracies.

Throughout this paper, we have made no distinction between individual classifiers on the one hand and classification algorithms on the other, since the idea of considering the posterior distribution of the balanced accuracy can be applied to either. The only way in which the two cases differ is whether we look at the confusion matrix that results from a single train/test cycle (yielding the posterior of the balanced accuracy of an individual classifier) or whether we sum the confusion matrices across all cross-validation folds (leading to the posterior of the algorithm as a whole). In most practical applications, it is the generalizability of the algorithm that will be of primary interest. The approach can therefore be used for any number of underlying cross-validation folds: it solely requires the overall confusion matrix, as obtained by summing individual confusion matrices across all folds.

The notions of the posterior distribution of both the accuracy and the balanced accuracy can be generalized in a natural way to a multiclass setting. Specifically, the posterior of the relative fractions of class frequencies in the test data can be estimated by replacing the Beta distribution by the Dirichlet distribution.

Another important generalization will be the notion of balancing not only class labels themselves but also other variables that correlate with class labels. This is important, for instance, in the case of a test set with balanced class labels in which another binary variable, closely correlated with class labels, is imbalanced. A biased classifier could then falsely suggest high generalizability while, in fact, it has learnt to separate examples according to the additional variable rather than according to the original class labels.

The relationship between posterior probability intervals and other measures, e.g., based on the binomial tail inversion or a ROC analysis [7], will be investigated in future studies.

Acknowledgment

This work was funded by the NEUROCHOICE project of SystemsX.ch (KES) and the University Research Priority Program 'Foundations of Human Social Behaviour' at the University of Zurich (KHB, KES).

References

[1] J. Langford, "Tutorial on practical prediction theory for classification," *Journal of Machine Learning Research*, vol. 6, pp. 273–306, 2005.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.

[3] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Machine Learning: ECML 2004*, 2004, pp. 39–50.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 3, p. 321357, 2002.

[5] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent Data Analysis*, vol. 6, pp. 429–449, 2002.

[6] D. R. Velez, B. C. White, A. A. Motsinger, W. S. Bush, M. D. Ritchie, S. M. Williams, and J. H. Moore, "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genetic Epidemiology*, vol. 31, no. 4, pp. 306–315, May 2007.

[7] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The binormal assumption on precision-recall curves," in *Proc. ICPR*, 2010.