# Probabilistic Performance Evaluation for Multiclass Classification Using the Posterior Balanced Accuracy

Henry Carrillo[1], Kay H. Brodersen[2], and José A. Castellanos[1]

[1] Instituto de Investigación en Ingeniería de Aragón, Universidad de Zaragoza, C/ María de Luna 1, 50018, Zaragoza, Spain
{hcarri,jacaste}@unizar.es

[2] Translational Neuromodeling Unit, Department of Information Technology and Electrical Engineering, Swiss Federal Institute of Technology (ETH Zurich), 8032 Zurich, Switzerland
brodersen@biomed.ee.ethz.ch

**Abstract.** An important problem in robotics is the empirical evaluation of classification algorithms that allow a robotic system to make accurate categorical predictions about its environment. Current algorithms are often assessed using sample statistics that can be difficult to interpret correctly and do not always provide a principled way of comparing competing algorithms. In this paper, we present a probabilistic alternative based on a Bayesian framework for inferring on balanced accuracies. Using the proposed probabilistic evaluation, it is possible to assess the balanced accuracy's posterior distribution of binary and multiclass classifiers. In addition, competing classifiers can be compared based on their respective posterior distributions. We illustrate the practical utility of our scheme and its properties by reanalyzing the performance of a recently published algorithm in the domain of visual action detection and on synthetic data. To facilitate its use, we provide an open-source MATLAB implementation.

**Keywords:** multiclass classifiers, accuracy, balanced accuracy, probabilistic performance.

## 1 Introduction

A central theme in the development of intelligent, autonomous robots has been the challenge of decision problems. Typical examples include the critical tasks of object detection [3,14], scene recognition [8,27], active SLAM [10,11] or loop closing [15,17]. All of these domains have seen significant progress in the development of increasingly accurate classification algorithms.

By contrast, there has been less focus on the evaluation of the *performance* of such algorithms. Assessing the performance of a given classifier is crucial as it allows us to (*i*) obtain an interpretable estimate of the degree to which its results generalize to unseen examples from the same distribution from which the existing data were drawn, (*ii*) compare competing approaches, and (*iii*) tune the (hyper)parameters of a classifier in light of the estimated performance in a given domain.

A common basis for evaluating the performance of a classifier is the confusion matrix. It provides a summary of classification outcomes and permits the inspection of

M.A. Armada et al. (eds.), *ROBOT2013: First Iberian Robotics Conference*,
Advances in Intelligent Systems and Computing 252,
DOI: 10.1007/978-3-319-03413-3_25, © Springer International Publishing Switzerland 2014

the number of correct and incorrect predictions in each class. However, in the absence of an appropriate summary statistic, reporting a confusion matrix by itself is generally insufficient and easily leads to highly subjective interpretations of performance.

Commonly used summary statistics that are based on confusion matrices include the overall sample accuracy; the per-class, or balanced, sample accuracy; the Kappa coefficient; and the $F_\mu$-score. Other statistics include the area under the receiver-operating characteristic (ROC) curve and the area under the precision-recall (PR) curve, although these are typically limited to two-class (binary) classification problems (see [21] for a generalization).

While all of the above performance metrics can be helpful in understanding the behaviour of a classifier, the key quantity of interest in most practical domains of application is the generalization ability, i.e., the probability of the classifier to make a correct prediction on an unseen example. It is tempting to try and answer this question by considering the *accuracy* of a classifier alone. However, classification accuracy is a misleading performance indicator when the data are not perfectly balanced [1,6,12,18].

A straightforward way of resolving the above limitation is to replace the accuracy by the *balanced accuracy*, defined as the arithmetic mean of class-specific accuracies. Critically, however, it is not sufficient to report the mean of class-specific *sample* accuracies. Rather, we must infer on the latent class-specfic accuracies of which the observed sample accuracies are an instantiation. Inferring on the balanced accuracy then means, for example, to report a point estimate as well as a measure of uncertainty about this estimate.

This paper describes a simple Bayesian framework for assessing the performance of classifiers. The proposed model makes it possible to compute the full posterior distribution of the balanced accuracy given the available classification outcomes. This approach extends previous work [6, 7] by providing a generalization to multiclass classifiers. In addition, we suggest a concrete method for comparing two balanced accuracies based on the posterior distribution of their difference. This method allows one to rank competing classifiers in a probabilistically interpretable fashion.

Using a Bayesian model for multiclass balanced accuracies offers three strengths over previous schemes: (*i*) the useful properties of the balanced accuracy are generalized to a multiclass setting; (*ii*) a Bayesian perspective allows us to explicitly incorporate prior knowledge (e.g., domain-specific information or a cost function that assigns a measure of importance to each class), account correctly for posterior uncertainty, and easily derive other posterior inferences; (*iii*) the model enables cross-algorithm comparisons that correctly account for the posterior uncertainty about each algorithm's performance.

The paper is structured as follows: Section 2 briefly reviews the merits of a Bayesian approach to performance evaluation. Section 3 provides a concrete example of a decision problem from the domain of robotics, followed by a brief overview of previous approaches to performance evaluation. Section 4 develops the proposed Bayesian model for performance evaluation in multiclass classification and a method for comparing competing classifiers. Section 5 presents a set of experiments in order to characterize the properties of the approach and illustrate its application. We conclude the material in Section 6 with a brief discussion.

## 2 Bayesian Inference on Classification Performance

In most situations, evaluating the performance of a classifier aims at characterizing the classifier's ability to predict the correct class of data that has not yet been seen. Abstracting away from specific implementations, we can denote the performance of a classifier by the variable $\lambda \in [0, 1]$. The two limits $0$ and $1$ refer to the ability of making incorrect or correct predictions on all future instances, while $0.5$ refers to classifications at random. Despite the advantages of a Bayesian framework, $\lambda$ has mostly been evaluated by adopting a classical, or frequentist, approach to inference.

Classical inference considers distributions over data but does not permit distributions over parameters such as $\lambda$. As a result, it is restricted to point estimates, $\hat{\lambda}$, and, most commonly, $95\%$ confidence intervals, representing the interval in which the true value would be in $95\%$ of cases if the experiment was repeatedly carried out an infinite number of times. A 'test' is then carried out by asking whether the value of a summary statistic (i.e., a $t$-score), or a more extreme value, could be observed under a 'null' hypothesis. The main advantage of point estimates, confidence intervals, and hypothesis tests is their computational simplicity. However, their correct interpretation is prone to errors [4].

In a Bayesian framework, inference proceeds by passing from a prior distribution, $p(\lambda)$, to a posterior distribution, $p(\lambda|\mathcal{D})$, that is informed by the data $y$. Depending on the given cost function, the posterior mean, mode, or median then replaces classical point estimates. Posterior intervals replace confidence intervals. And Bayesian model comparison replaces hypothesis tests. The main advantage of Bayesian inference is its conceptual simplicity (providing a probabilistic statement about the quantity of interest rather than providing a sampling statistic about a summary statistic) and the flexibility with which posterior inferences can be summarized [16]. A downside is that Bayesian inference is often computationally more complex than classical inference. In the application considered in this paper, however, this is not an issue, since the classifier evaluation is usually small-scale and carried out offline.

## 3 A Motivating Example

To illustrate the importance of assessing overall classification performance, we consider the example of a service robot, designed to clean the dishes in a kitchen. A core component of such a robot is the capability of visually classifying objects as a 'mug', a 'sink', or a 'bottle of wash-up liquid'. Clearly, any classification algorithm designed for this task must be proficient at detecting all three types of object, since all are required to complete the task. Thus, we are interested in the overall performance of the classifier (i.e., the variable $\lambda$) rather than its performance on individual classes.

### 3.1 Confusion Matrices

A common way of reporting classification results, especially in a multiclass setting, is to compile a confusion matrix, also referred to as a contingency table or accumulation matrix. Let each element $x_{r,c}$ of a confusion matrix $C \in \mathbb{N}^{l \times l}$ represent the number of

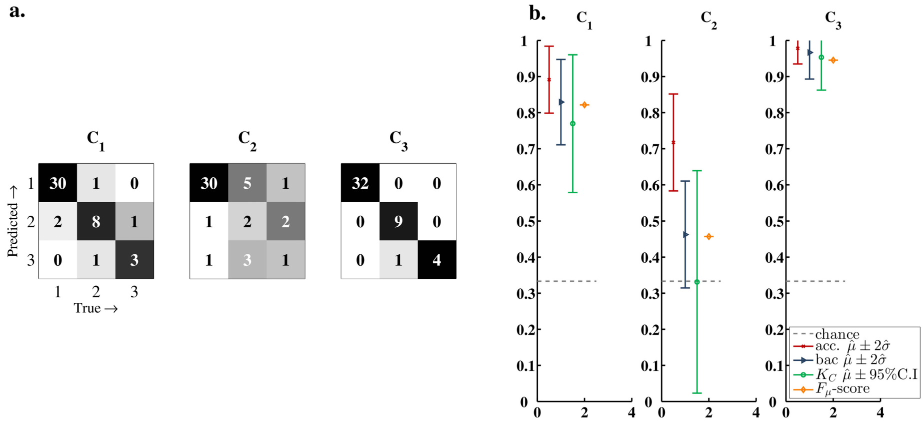

**Fig. 1.** (a) Three example confusion matrices, summarizing the classification outcomes described in Section 3. (b) Conventional performance metrics with classical error bars stemming from the standard error of the mean. To reproduce these results, see `MotivatingExample.m`, available online [9].

times a classifier predicted class $r$ when the true class label was $c$. Thus, diagonal and off-diagonal elements indicate the number of correct and incorrect predictions, respectively.

Let us suppose that three competing classifiers were tested on a given dataset, resulting in three matrices $C_1$, $C_2$, and $C_3$ (see Fig. 1). Using a graphical representation of the matrices, one can easily obtain an intuitive sense about which classifier performed best ($C_3$). However, such an assessment remains vague and does not obviate the need for a quantitative evaluation. For instance, how confident are we that classifier 2 performed better than chance? How certain is it that classifier 3 outperformed classifier 1? Several metrics have been proposed to address such questions. In the next part, we will review some of these, focusing on their properties regarding the assessment of prediction accuracy of a classifier.

### 3.2 Performance Metrics

The literature on performance metrics on the basis of confusion matrices is large and diverse, comprising both frequent propositions of new statistics and the development of statistical models for their estimation. Here, we briefly consider some of the most common metrics and outline where balanced accuracies fit in.

The most common statistic for reporting the performance of a multiclass classifier is its *sample accuracy* (acc), defined as the number of correct predictions across all classes, $k$, divided by the number of examples, $n$. While conceptually simple, assessing performance using the sample accuracy alone has long been known to be prone to erroneous interpretation. This is because the accuracy does not account for the degree of class imbalance that may be present in a given dataset [1, 6, 12, 18], which means it can only be correctly interpreted in relation to a dataset-dependent baseline accuracy.

An example of the misleading nature of inferences based on accuracies can be seen in the results of the example in Fig. 1b. The plot shows that the sample accuracy of $C_2$ is close to that of $C_1$, despite $C_2$ misclassifying classes 2 and 3 most of the time.

In robotics, a common way of overcoming the above limitation is to resort to the $F_\mu$-*score* [24, p. 183], defined as the across-classes average of the $F_\beta$*-score* [26]. The $F_\beta$-*score* itself, frequently used in binary classification, is given by the $\beta$-weighted average between precision and recall. Setting $\beta = 2$, as is commonly done, yields the harmonic mean of precision and recall. Thus, as depicted in Fig. 1b, the $F_\mu$-score accounts for the degree of class imbalance in the test examples. At present, however, there has been no established convention of computing its corresponding confidence or credible intervals.

Another approach to overcoming the limitation of sample accuracies is based on the *Kappa coefficient* [13], which has been one of the dominating metrics in the domain of remote sensing. It quantifies the degree of overall agreement within a given confusion matrix $C \in \mathbb{N}^{l \times l}$,

$$K_c = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

with

$$p_0 = \frac{1}{l} \sum_{i=1}^{l} k_i, \quad p_e = \frac{1}{l^2} \sum_{i=1}^{l} C_{i+} \times C_{+i}, \tag{2}$$

where $k_i$ is the number of correct predictions in class $i$ and $l$ is the number of classes. $C_{i+}$ and $C_{+i}$, respectively, are the row-wise and column-wise sums of row and column $i$ in the confusion matrix.

Like the $F_\mu$-score, $K_c$ accounts for the degree of class imbalance in the data. However, it can be invariant to the number of misclassifications and does not necessarily reflect what one would intuitively consider prediction strength [25].

An alternative is the *balanced accuracy* (bac), defined as the average accuracy obtained on all classes. In the case of a multiclass classification problem, its sample statistic is given by

$$\hat{\lambda} = \frac{1}{l} \sum_{i=1}^{l} \frac{k_i}{n_i}, \tag{3}$$

where $k_i$ is the number of correct predictions in class $i$, $l$ is the number of classes and $n_i$ is the number of examples in class $i$. The balanced accuracy is frequently used in practice and has several conceptual strengths over the conventional accuracy while maintaining its simplicity. However, a probabilistic approach is not always being adopted when interpreting it, despite the fact that the limits of a frequentist confidence interval, as can be seen in classifier 3 in Fig. 1b, can easily lie outside of its $[0, 1]$ domain. (One possible remedy is to apply a $z$-transform prior to computing the interval.)

In summary, current approaches to multiclass performance evaluation face multiple challenges: (*i*) assessing performance on the basis of sample statistics does not replace principled probabilistic inference; (*ii*) error bars are often based on ill-justified distributional assumptions, such as in the case of classical confidence intervals without a $z$-transform; (*iii*) classification accuracy remains a popular metric even in those cases where an imbalanced dataset leads to misleading conclusions; (*iv*) alternatives to the above, such as the use of a Bayesian framework [6, 7] have not yet been generalized to multiclass classification problems.

# 4 Theory

To help address the challenges outlined above, we describe a Bayesian approach to estimating the accuracy and balanced accuracy of a classifier in a multiclass setting. The adoption of a Bayesian perspective has long been considered helpful in this context (cf. [5, p. 68-74]; [24, p. 72-78]) and has in particular been described previously for binary classifiers [6, 7], where $C \in \mathbb{N}^{2\times 2}$. Here, we develop a generalization to the multiclass case where $C \in \mathbb{N}^{l\times l}$ with $l \geq 2$. In order to keep our treatment self-contained, we will begin with the binary case and then demonstrate its extension to multiclass classification.

It is worth pointing out that the approach in this paper differs from the implementation in [6, 7] in that we suggest a feasible strategy for computing the posterior distribution of $\lambda$ in a multiclass setting. Our strategy is based on the characteristic function of the per-class posterior distribution and its Fourier transforms. This strategy is computationally more efficient than a direct extension of the previous implementation, which would otherwise require an $l$-fold convolution, where $l$ is the number of classes in the classifier.

Another focus of the present paper is the comparison of competing algorithms based on the posterior distribution of the difference of their respective balanced accuracies.

## 4.1 Problem Statement

We consider the solution to a decision problem in which each one of $n$ i.i.d. examples (or trials) is associated with a class label from a finite set of categories $\{1, \ldots, l\}$. We wish to assess the generalizability of the classifier. In others words, we wish to characterize the classifier's ability of predicting the correct class on future, unseen data, i.e., estimate the variable $\lambda$.

## 4.2 Solution Sketch

In a Bayesian framework, the performance of a classifier is considered a latent (unobservable) variable, and we use probabilities to express our uncertainty about classification performance before and after observing actual classification outcomes. Under this view, evaluating the performance $\lambda$ of a classifier means passing from a prior distribution $p(\lambda)$ to a posterior conditioned on observed data $p(\lambda|\mathcal{D})$. The posterior encodes the plausibility of all possible true performance values in light of the observed data, and there are many ways in which it can be summarized, for example, in terms of the posterior mean, mode, or a posterior interval.

To model classification performance, we code correct predictions as $y = 1$ and incorrect predictions as $y = 0$. A classification result can then be viewed as a sequence of outcomes $y_1, \ldots, y_n$. Modelling each outcome as the i.i.d. result of a Bernoulli experiment, we obtain

$$p(y_i|\lambda) = \mathrm{Bern}(y_i|\lambda) = \lambda^{y_i}(1-\lambda)^{1-y_i}, \tag{4}$$

where $\lambda$ is the probability of any one trial being classified correctly. This implies that the number of correct predictions $k$ in a sequence of trials $y_1, \ldots, y_n$ follows a Binomial distribution:

$$p(k|\lambda, n) = \text{Bin}(k|\lambda, n) = \binom{n}{k} \lambda^k (1-\lambda)^{n-k} \tag{5}$$

Finally, we express any available prior knowledge about classification performance by placing a prior on $\lambda$. A natural choice for this is to use the conjugate prior of the Binomial distribution, i.e., the Beta density. In the absence of any preceding classification results, we express maximal prior uncertainty (i.e., all values in the domain $[0 \ldots 1]$ of performance $\lambda$ are considered equally plausible *a priori*) using a uniform distribution

$$p(\lambda) = \text{Beta}(\lambda|a, b) = \text{Beta}(\lambda|1, 1). \tag{6}$$

Thus, given the observed data $k$ and $n$, we obtain the posterior performance as:

$$\begin{aligned} p(\lambda|k, n) &= \frac{\text{Bin}(k|\lambda, n) \times \text{Beta}(\lambda|1, 1)}{p(k)} \\ &= \text{Beta}(\lambda|k+1, n-k+1) \end{aligned} \tag{7}$$

This posterior encodes our knowledge about $\lambda$ in light of the observed classification result. Critically, it goes beyond point estimates of performance (such as, e.g., the sample accuracy $k/n$), since it reflects how uncertain we are about our estimate. For instance, observing only very few classification outcomes will be correctly accounted for by a wide posterior distribution; whereas the observation of a large additional number of outcomes would cause the posterior to shrink to a more precise distribution.

Having described the model in general terms, we will now turn to the special cases of posterior accuracies and balanced accuracies in a multiclass setting, as described in the following two sections.

### 4.3 The Posterior Multiclass Accuracy

In what follows, we describe how to obtain the posterior distribution of the accuracy for a multiclass setting. It should be noted that this section merely serves as a preparation of the next section; using the accuracy to describe the performance of a classifier is often misleading and is discouraged [1, 6, 12, 18].

In order to infer on the posterior classification accuracy of a multiclass classifier, we can use the model described above as is. In the multiclass setting, the variable[1] $\theta$ then simply represents the probability with which an individual trial is classified correctly, i.e., classification accuracy. In other words, the posterior distribution of the overall accuracy is given by

$$p(\theta|k, n, a, b) = \text{Beta}(\theta|a+k, b+n-k), \tag{8}$$

where $a = 1$ and $b = 1$ encode our prior ignorance about the classifier's performance (see [19] for a comparison of alternative priors). The key point to note here is that the

[1] The variable $\theta$ represents the same as the variable $\lambda$, i.e., the performance of a classifier. We changed its notation to prevent abuse of notation in the next section.

availability of a full posterior distribution yields a plethora of useful ways of forming summary statistics. For example, we could report a central 95% credible interval,

$$\left[F^{-1}_{\mathrm{B}(a+k,b+n-k)}(\tfrac{0.05}{2});\; F^{-1}_{\mathrm{B}(a+k,b+n-k)}(1-\tfrac{0.05}{2})\right], \tag{9}$$

where $F^{-1}_{\mathrm{B}(\cdot)}(\cdot)$ is the inverse cumulative density function of the Beta distribution, evaluated at the desired quantile.

Alternatively, we could derive a point estimate of classification accuracy that minimizes the expectation of a given loss function. For example, the optimal point estimate under an $\ell_2$-loss function is the posterior mean:

$$\langle\theta\rangle_{p(\theta|k,n,a,b)} = \frac{a+k}{a+b+n} \tag{10}$$

In contrast, the expected loss of a $(0,1)$-loss function is minimized by the posterior mode:

$$\arg\max_{\theta} p(\theta|k,n,a,b) = \frac{k+a-1}{a+b+n-2} \tag{11}$$

This shows that we can reinterpret the conventional sample accuracy $k/n$ as the optimal estimate under a flat prior and a loss function that is $L(\theta,\hat{\theta}) = 0$ if $\theta = \hat{\theta}$ and 1 otherwise.

### 4.4 The Posterior Multiclass Balanced Accuracy

Classification accuracy, as defined above, is a misleading measure of performance when the data are not perfectly balanced. This is because a classifier may take advantage of an imbalanced dataset and trivially achieve a classification accuracy equal to the fraction of the majority class, and thus potentially much higher than the $1/l$ baseline. Put differently, the baseline accuracy that can always be achieved by a classifier, even in the case of zero mutual information between data features and class labels, depends on the degree of class imbalance; it is not always $1/l$ (i.e., 0.5 in the case of binary classification).

This issue can be resolved by replacing the accuracy by the balanced accuracy, i.e., by the arithmetic mean of class-specific accuracies,

$$\lambda := \frac{1}{l}\sum_{i=1}^{l}\theta_i, \tag{12}$$

where $\theta_i$ is the (latent) accuracy of the classifier on class $i$. When the data are perfectly balanced (i.e., the data contain the same number of examples from each class), the balanced accuracy reduces to the conventional accuracy. Critically, however, its baseline performance is always $1/l$, regardless of the degree of class imbalance. Thus, if a classifier has achieved class-specific accuracies above $1/l$ only by exploiting the class imbalance, its balanced accuracy will drop to $1/l$, as desired [7].

Under a Bayesian perspective, we wish to pass from a prior distribution over the balanced accuracy to a posterior distribution in light of the observed classification outcomes $\mathcal{D} = \{(k_1,n_1),\ldots,(k_l,n_l)\}$. We have seen in (8) how we can obtain the posterior distribution of the overall accuracy. Thus, to obtain the posterior balanced accuracy, we first apply (8) to each class in turn; we then find the conditional distribution over $\lambda$.

**Computing $\boldsymbol{\lambda}$.** In [6, 7], a convolution is used to compute $\lambda$ for the binary case. The direct extension of its convolution approach would require an $l$-fold convolution, where $l$ is the number of classes in the classifier. This would in turn require the numerical computation of an $l$-dimensional integral, which can be both computationally complex and numerically unstable.

An alternative is to consider the inverse Fourier transform of the products of the characteristic functions of the individual distributions, as described next.

The probability distribution of a given random variable $\theta$ is fully specified by its characteristic function $\Phi_\theta$ which is given by the Fourier transform $\mathcal{F}\{\theta\}$ [22, p. 145]. Thus, owing to the product property of the Fourier transform, the convolution of two functions is identical to the product of the functions' Fourier transforms. In the context of classification, we can exploit this fact to obtain the posterior distribution of the balanced accuracy as

$$p(\lambda|\mathcal{D}) = \mathcal{F}^{-1}\{\Phi_{\check{\theta}_1} \times \cdots \times \Phi_{\check{\theta}_l}\}, \tag{13}$$

where $\Phi_{\check{\theta}_i}$ is the characteristic function of $\check{\theta}_i := \frac{1}{l}\theta_i, \forall i = 1 \ldots l$, and where $\mathcal{F}^{-1}\{\cdot\}$ is the inverse Fourier transform [22, p. 272]. This step is facilitated by the fact that the posterior distributions of all individual class-specific accuracies, $p(\theta_i|k_i, n_i)$, are Beta densities and can be obtained using (8).

Just as in the case of the posterior accuracy, it is useful and important to obtain summary statistics, such as the mean, mode, or a credible interval. In contrast to the closed-form expressions we saw in the previous section, **analytical solutions for balanced-accuracy statistics are not available**. To address this limitation and facilitate their use, we provide a numerical implementation in MATLAB.[2]

### 4.5 Comparing Competing Multiclass Classifiers Using Their Posterior Performance Distribution

Given the posterior distribution of the performance of one classifier $p(\lambda_1|\mathcal{D})$, a critical question is how this given classifier compares to a competing classifier with posterior performance $p(\lambda_2|\mathcal{D})$ or others, e.g., $p(\lambda_3|\mathcal{D})$. This question can be answered by considering the pairwise posterior differences between the respective competing classifiers,

$$p\,(\delta \mid \mathcal{D})\,, \tag{14}$$

where $\delta := \lambda_{(j)} - \lambda_{(i)}$ denotes the difference between the posterior balanced accuracies $\lambda_{(i)}$ and $\lambda_{(j)}$ of two competing classifiers (cf. the *difference between proportions* [20, p. 175-176]) and $\mathcal{D} = \{k_1^{(1)}, \ldots, k_l^{(1)}, k_1^{(2)}, \ldots, k_l^{(2)}\}$ is the classifiers outcomes.[3]

**Properties of $\boldsymbol{\delta}$.** The domain of the random variable $\delta$ is continuous in $[-1 \ldots 1]$, Its distribution can summarized, for example, by reporting the posterior expectation. It represents the expected algebraic distance in performance between the two classifiers.

[2] The software can be downloaded from: `http://mloss.org/software/view/447/`

[3] The dataset sizes $\{n_1^{(1)}, \ldots, n_l^{(1)}, n_1^{(2)}, \ldots, n_l^{(2)}\}$ have been omitted for brevity.

**Computing $\delta$.** Algorithmically, we can compute the posterior density of the difference of two balanced accuracies using a stochastic approximate inference approach. Specifically, under a Monte Carlo scheme [24, p. 154-155], we repeatedly draw samples from $p(\lambda_{(i)}|\mathcal{D})$ and $p(\lambda_{(j)}|\mathcal{D})$ and collect the differences between each pair. This approach will result in an approximation of $p\left(\delta \mid \mathcal{D}\right)$. In practice, a high number of samples (e.g., 5 000) is required for a suitable approximation.

A simple heuristic method for comparing competing classifiers can be done by ranking them by their performance's algebraic distance against each other. We suggest the following simple scheme:

1. Given $T$ competing classifiers, compute the pairwise posterior difference between them (using eq. 14) and its posterior mean. Decide on a winning classifier for each comparison based on the sign of the posterior mean.
2. Count the number of times each classifier wins.
3. Rank the competing classifiers according to its number of victories.

This simple scheme assumes that the classifiers are tested with the same dataset. It also assumed that the posterior mean of $\delta$ approximates the $p(\lambda_{(i)} > \lambda_{(j)}|\mathcal{D})$ for every pairwise comparison, which could be an optimistic approximation. Non-heuristic schemes of comparison are of interest and objective of future research.

# 5 Experiments

In this section, we present a set of experiments to compare the proposed probabilistic evaluation to previous frequentist approaches such as sample accuracy, sample balanced accuracy, Kappa coefficient and $F_\mu$-score. We begin by considering the synthetic data from the motivating example in Section 3 and conclude the section by reanalyzing an empirical dataset from the domain of action detection.

## 5.1 A Bayesian Look to the Motivating Example

Using the confusion matrix of Fig. 1a as input, the posterior distribution of the balanced accuracy (henceforth PDBAC) of each classifier can be computed as described in Section 4.4. The PDBAC of each classifier is depicted in Fig. 2.

As expected, the posterior mean of the classifiers show a clear difference in performance between ($C_2$) and the others. This behavior is also exhibited by the sample balance accuracy, the $F_\mu$-score and the Kappa coefficient, but not for the sample accuracy, as discussed in Section 3. Moreover, the uncertainty associated with the posterior mean is in the correct domain of the random variable $\lambda$ for all classifiers, unlike the (non-$z$-transformed) accuracies, which for $C_3$ exhibit non-sensical uncertainty bounds.

To illustrate the effect of sample size on posterior inferences, Table 1 details the performance metrics after evaluating $C_1, C_2$, and $C_3$, as well as two scaled versions of them. As expected, as the number of trials grows, the limits of the classical confidence intervals assume more sensical values. It is further worth noting that the classical point estimate itself is ignorant to the sample size (since the confusion matrices retain the same proportions). By contrast, the posterior mean reflects a shrinking influence from

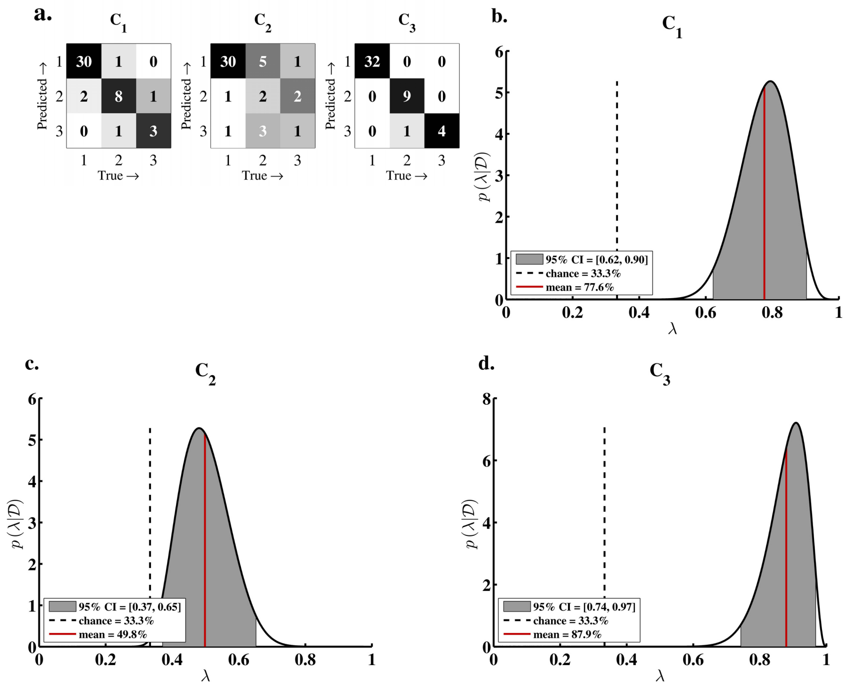

**Fig. 2.** (a) Confusion matrix of the classifiers studied in Section 3. (b)(c)(d) Posterior distribution of the classifier's multiclass balanced accuracy. For each distribution the posterior mean and 95% confidence interval is indicated. Figure generated by `IllustrativeExample0.m`, available online [9].

the prior and a growing influence from the data as its posterior interval tightens. The posterior mean and the sample balanced accuracy would agree in the limit of infinite data.

**Ranking Competing Classifiers.** The procedure described in Section 4.5 makes it possible to rank the competing classifiers $C_1, C_2$ and $C_3$ according to their posterior distributions. We begin by computing the pairwise distribution $p(\delta \mid \mathcal{D})$ for each competing classifier. The three resulting distributions are depicted in Fig. 3.

It is worth recalling that the posterior mean of $\delta$ is not simply the difference between the involved posterior means. The posterior mean of $\delta$ takes into account the uncertainty of the $\lambda$s. Continuing with the next steps of the proposed method, the resulting rank is: 1. $C_3$, 2. $C_1$ and 3. $C_2$.

## 5.2 Action Detection

There has been increasing momentum in studying models for visual recognition of human actions from images. One recent study [23] proposed a system for action modelling

**Table 1.** This table summarizes performance statistics of the classifiers studied in Section 3. $C_1, C_2$ and $C_3$ are the confusion matrices of the classifiers considered in section 3. $C_4, C_5$ and $C_6$ provide the same information scaled by a factor of 10. Likewise, $C_7, C_8$ and $C_9$ are scaled by a factor of 100. The statistics computed for all matrices is the mean and the 95% confidence interval. The table compares the PDBAC with the sample balanced accuracy, the sample accuracy, the $F_\mu$-score, and the Kappa coefficient. The grey figures represent 95% C.I. whose limits escape the domain $[0, 1]$.

| | PDBAC | | bac | | acc | | $F_\mu$-score | | $K_C$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mu$ | 95%C.I | $\mu$ | 95%C.I | $\mu$ | 95%C.I | $\mu$ | 95%C.I | $\mu$ | 95%C.I |
| $C_1$ | 0.776 | [0.62 - 0.90] | 0.829 | [0.73 - 0.92] | 0.891 | [0.79 - 0.98] | 0.821 | - | 0.769 | [0.57 - 0.96] |
| $C_2$ | 0.498 | [0.37 - 0.65] | 0.462 | [0.32 - 0.59] | 0.717 | [0.58 - 0.85] | 0.457 | - | 0.331 | [0.02 - 0.63] |
| $C_3$ | 0.879 | [0.74 - 0.97] | 0.966 | [0.92 - 1.01] | 0.978 | [0.93 - 1.02] | 0.945 | - | 0.953 | [0.86 - 1.04] |
| $C_4$ | 0.822 | [0.77 - 0.87] | 0.829 | [0.80 - 0.85] | 0.891 | [0.86 - 0.92] | 0.821 | - | 0.769 | [0.70 - 0.82] |
| $C_5$ | 0.468 | [0.42 0.52] | 0.462 | [0.42 - 0.50] | 0.717 | [0.67 - 0.75] | 0.457 | - | 0.331 | [0.23 - 0.42] |
| $C_6$ | 0.955 | [0.93 - 0.98] | 0.966 | [0.95 - 0.98] | 0.978 | [0.96 - 0.99] | 0.945 | - | 0.953 | [0.92 - 0.98] |
| $C_7$ | 0.828 | [0.81 - 0.85] | 0.829 | [0.82 - 0.83] | 0.891 | [0.88 - 0.90] | 0.821 | - | 0.769 | [0.75 - 0.78] |
| $C_8$ | 0.463 | [0.45 - 0.48] | 0.462 | [0.44 - 0.47] | 0.717 | [0.70 - 0.73] | 0.457 | - | 0.331 | [0.30 - 0.36] |
| $C_9$ | 0.966 | [0.96 - 0.97] | 0.966 | [0.96 - 0.97] | 0.978 | [0.97 - 0.98] | 0.945 | - | 0.953 | [0.94 - 0.96] |

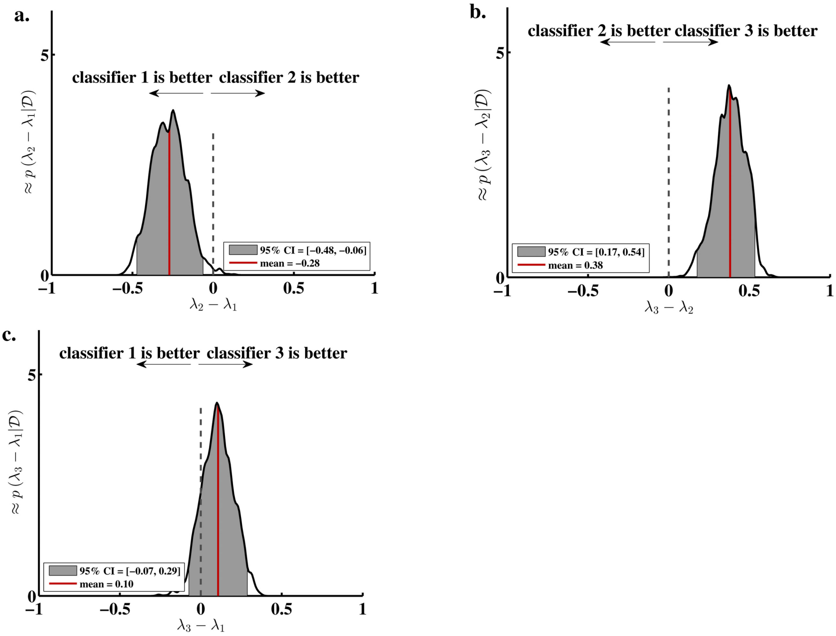

**Fig. 3.** (a)(b)(c) Posterior distribution of the difference of two posterior balanced accuracies computed as explained in Section 4.5. The posterior balanced accuracies $\lambda_1, \lambda_2$, and $\lambda_3$ stem from $C_1, C_2$, and $C_3$, respectively. For each distribution the posterior mean and the 95% confidence interval are indicated. Figure generated by `IllustrativeExample0.m`, available online [9].

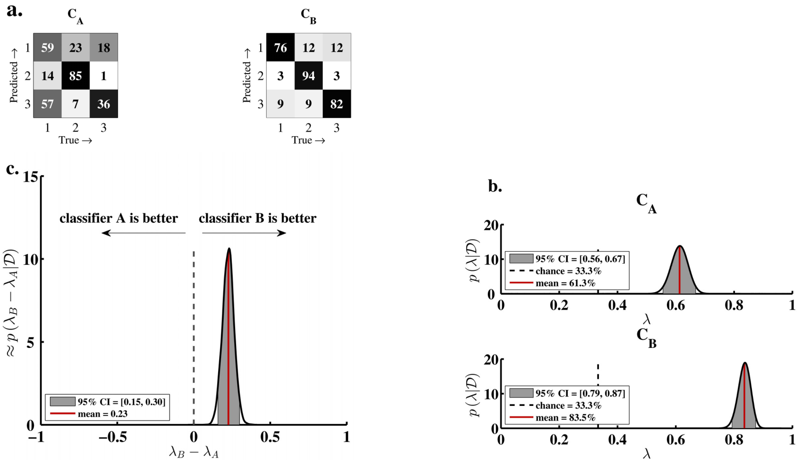

**Fig. 4.** (a) Confusion matrices of two classifiers [2, 23] for visual action detection (see Section 5.2). (b) Posterior distributions of the two classifiers' multiclass balanced accuracies. (c) Posterior distribution of the difference in balanced accuracy between the two classifiers. Figure generated by `IllustrativeExample1.m`, available online [9].

based on a classification method for human actions from image sequences. The authors compared their classifier to an alternative approach [2] using a dataset composed of three classes representing the acts of 'moving an object', 'making a sandwich', and 'opening a book', respectively.

We revisited the reported results and, based on the published confusion matrices (Fig. 4a), computed the posterior multiclass balanced accuracies along with several summary statistics (Fig. 4b). Our results show that, in contrast to conventional sample accuracies, posterior balanced accuracies provide a rich representation of our knowledge about each classifier's performance. Critically, using the approach outlined in Section (Sec. 4.5), we can infer on the difference between the two performances. Specifically, given the two competing classifiers , $C_B$ is ranked number one with a posterior expectation of the difference of 0.23. We illustrate the full posterior distribution of $\delta$ using a kernel density estimator (bandwidth 0.0125), as shown in Fig. 4c.

## 6 Discussion

Classification algorithms frequently form a critical part of complex systems for pattern recognition and machine learning, such as those found in the domain of robotics. However, evaluating the performance of a given system, and comparing it to others, is often subject to methodological limitations. Reporting overall classification accuracy, for example, is statistically unwarranted because it can only be interpreted in relation to a baseline level that depends on the degree of class imbalance at hand. Sample statistics, such as the sample balanced accuracy for instance, rectify this problem but do not readily embrace our uncertainty about the performance metric of interest.

In this paper, we described how the above limitations can be overcome in a natural way using a Bayesian framework for inferring on the balanced accuracy. The approach generalizes our previous work on balanced accuracies for binary classification problem. More importantly, a method of ranking competing classifiers using their posterior balanced accuracy is proposed. The main advantage of this methods is that it permits to account for the uncertainty in the classifiers' performance during evaluation.

One critical feature of this approach is the flexibility with which posterior inferences can be summarized. In particular, we can obtain derivative inferences, such as our confidence with which one classifier is better than another. In this context, it is worth noting that, in order to ensure a fair comparison, it is important that the algorithms whose performances are compared were applied to the same dataset.

To facilitate its widespread use, we provide an open-source MATLAB toolbox which we have made available for download [9]. With this toolbox we hope that balanced accuracies may help improve the correct evaluation and comparison of multiclass classifiers in future classification systems.

**Acknowledgments.** This work was supported by the MICINN-FEDER projects DPI2009-13710 and DPI2012-36070 as well as by DGA-FSE (grupo T04).

## References

1. Akbani, R., Kwek, S.S., Japkowicz, N.: Applying Support Vector Machines to Imbalanced Datasets. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 39–50. Springer, Heidelberg (2004)
2. Aksoy, E., Abramov, A., Worgotter, F., Dellen, B.: Categorizing Object-action Relations from Semantic Scene Graphs. In: 2010 IEEE International Conference on Robotics and Automation (ICRA), pp. 398–405 (May 2010)
3. Andreopoulos, A., Hasler, S., Wersing, H., Janssen, H., Tsotsos, J., Korner, E.: Active 3D Object Localization Using a Humanoid Robot. IEEE Transactions on Robotics 27(1), 47–64 (2011)
4. Berger, J.O.: Could fisher, jeffreys and neyman have agreed on testing? Statistical Science 18(1), 1–32 (2003)
5. Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics. Springer-Verlag New York, Inc., Secaucus (2006)
6. Brodersen, K.H., Mathys, C., Chumbley, J.R., Daunizeau, J., Ong, C.S., Buhmann, J.M., Stephan, K.E.: Bayesian Mixed-Effects Inference on Classification Performance in Hierarchical Data Sets. Journal of Machine Learning Research 13, 3133–3176 (2012)
7. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The Balanced Accuracy and Its Posterior Distribution. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 3121–3124 (August 2010)
8. Cadena, C., Galvez-Lopez, D., Tardos, J.D., Neira, J.: Robust Place Recognition With Stereo Sequences. IEEE Transactions on Robotics 28(4), 871–885 (2012)
9. Carrillo, H.: GBAC (2013), `http://www.mloss.org/software/view/447/`
10. Carrillo, H., Latif, Y., Neira, J., Castellanos, J.A.: Fast Minimum Uncertainty Search on a Graph Map Representation. In: IEEE / RSJ International Conference on Intelligent Robots and Systems (IROS 2012), Vilamoura, Algarve, Portugal (October 2012)

11. Carrillo, H., Reid, I., Castellanos, J.A.: On the Comparison of Uncertainty Criteria for Active SLAM. In: IEEE International Conference on Robotics and Automation, pp. 2080–2087 (2012)
12. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research (JAIR) 16, 321–357 (2002)
13. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20(1), 37–46 (1960)
14. Ess, A., Schindler, K., Leibe, B., Van Gool, L.: Object Detection and Tracking for Autonomous Navigation in Dynamic Environments. The International Journal of Robotics Research 29(14), 1707–1725 (2010)
15. Galvez-Lopez, D., Tardos, J.D.: Bags of Binary Words for Fast Place Recognition in Image Sequences. IEEE Transactions on Robotics 28(5), 1188–1197 (2012)
16. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian data analysis. CRC press (2003)
17. Granstrm, K., Schn, T.B., Nieto, J.I., Ramos, F.T.: Learning to close loops from range data. The International Journal of Robotics Research 30(14), 1728–1754 (2011)
18. Japkowicz, N., Stephen, S.: The Class Imbalance Problem: A systematic Study. Intelligent Data Analysis 6(5), 429–449 (2002)
19. Kerman, J.: Neutral noninformative and informative conjugate beta and gamma prior distributions. Electronic Journal of Statistics 5, 1450–1470 (2011)
20. Kruschke, J.K.: Doing Bayesian Data Analysis: A Tutorial with R and BUGS, 1st edn. Academic Press / Elsevier, Amsterdam (2011)
21. Landgrebe, T., Duin, R.: Efficient Multiclass ROC Approximation by Decomposition via Confusion Matrix Perturbation Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(5), 810–822 (2008)
22. Leon-Garcia, A.: Probability and Random Processes for Electrical Engineers, 2nd edn. Addison-Wesley, Reading (1994)
23. Luo, G., Bergstrom, N., Ek, C., Kragic, D.: Representing Actions with Kernels. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2028–2035 (September 2011)
24. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning series. The MIT Press, Cambridge (2012)
25. Nishii, R., Tanaka, S.: Accuracy and inaccuracy assessments in land-cover classification. IEEE Transactions on Geoscience and Remote Sensing 37(1), 491–498 (1999)
26. Rijsbergen, C.J.V.: Information Retrieval, 2nd edn. Butterworth-Heinemann, Newton (1979)
27. Siagian, C., Itti, L.: Biologically Inspired Mobile Robot Vision Localization. IEEE Transactions on Robotics 25(4), 861–873 (2009)