# Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions*

V. García[1,2], R.A. Mollineda[2], and J.S. Sánchez[2]

[1] Lab. Reconocimiento de Patrones, Instituto Tecnológico de Toluca
Av. Tecnológico s/n, 52140 Metepec, México
[2] Dept. Llenguatges i Sistemes Informàtics, Universitat Jaume I
Av. Sos Baynat s/n, 12071 Castelló de la Plana, Spain

**Abstract.** This paper introduces a new metric, named *Index of Balanced Accuracy*, for evaluating learning processes in two-class imbalanced domains. The method combines an unbiased index of its overall accuracy and a measure about how dominant is the class with the highest individual accuracy rate. Some theoretical examples are conducted to illustrate the benefits of the new metric over other well-known performance measures. Finally, a number of experiments demonstrate the consistency and validity of the evaluation method here proposed.

## 1 Introduction

Many learning approaches assume that the problem classes share similar prior probabilities. However, in many real-world tasks this assumption is grossly violated. Often, the ratios of prior probabilities between classes are significantly skewed. This situation is known as the imbalance problem. A two-class data set is said to be imbalanced when one of the classes (the minority one) is heavily under-represented as regards the other class (the majority one) [6]. This topic is particularly important in those applications where it is costly to misclassify examples from the minority class. Because of examples of the minority and majority classes usually represent the presence and absence of rare cases, respectively, they are also known as positive and negative examples.

As pointed out by many authors, the performance of a classification process over imbalanced data sets should not be expressed in terms of the plain accuracy and/or error rates [2, 3, 4, 8, 9]. The use of these simple measures might produce misleading conclusions since they do not take into account misclassification costs, are strongly biased to favor the majority class, and are sensitive to class skews.

Alternative measures have been proposed to evaluate classifiers in imbalanced scenarios. Some widely-known examples are Receiver Operating Characteristic (ROC) curve, the area under the ROC curve (AUC) [1], the geometric mean of class accuracies [7] and the $f$-measure [3]. Another measure, less renowned but possibly more powerful than those previously cited, refers to the optimized precision [9]. All these measures are combinations of error/accuracy rates measured separately on each class, thus alleviating biased results of the classification performance. Nevertheless, most of

* Partially supported by grants DPI2006–15542 and CSD2007–00018 from the Spanish Ministry of Education and Science, and P1–1B2007–60 from Fundació Caixa Castelló - Bancaixa.

H. Araujo et al. (Eds.): IbPRIA 2009, LNCS 5524, pp. 441–448, 2009.

these measures do not consider how dominant is the accuracy on an individual class over another. Hence their results do not reflect the contribution of each class to the overall performance. In some cases, it could be interesting (and complementary) to know whether the accuracies on each class are balanced and if not, to find out which is the 'dominant class' (the class with the highest accuracy rate).

This paper introduces a new method to evaluate the performance of a classification system in two-class imbalanced data sets. It quantifies a trade-off between an unbiased measure of overall accuracy and an index of how balanced are the two class accuracies. This relationship is represented by means of a two-dimensional graph whose axes correspond to the square of the geometric mean of class accuracies and the signed difference between the accuracies on positive and negative classes. The second term is intended to favor those cases with higher accuracy rate on the positive class. Some illustrative examples are simulated to better explain the differences between the measure here proposed and other well-known metrics. Final experiments on real-world problems are designed to demonstrate the consistency and validity of the new performance evaluation method here introduced.

## 2 Evaluation of Classifier Performance in Imbalanced Domains

Typical metrics for measuring the performance of learning systems are classification accuracy and/or error rates, which for a two-class problem can be easily derived from a $2 \times 2$ confusion matrix as that given in Table 1. These measures can be computed as $Acc = (TP+TN)/(TP+FN+TN+FP)$ and $Err = (FP+FN)/(TP+FN+TN+FP)$.

However, empirical evidence shows that these measures are biased with respect to the data imbalance and proportions of correct and incorrect classifications. Shortcomings of these evaluators have motivated search for new measures.

**Table 1.** Confusion matrix for a two-class problem

| | *Predicted positive* | *Predicted negative* |
|---|---|---|
| *Positive class* | True Positive (TP) | False Negative (FN) |
| *Negative class* | False Positive (FP) | True Negative (TN) |

Some straightforward examples of alternative measures are: (i) *True positive rate* (also referred to as *recall* or *sensitivity*) is the percentage of positive examples which are correctly classified, $TPrate = TP/(TP + FN)$; (ii) *True negative rate* (or *specificity*) is the percentage of negative examples which are correctly classified, $TNrate = TN/(TN+FP)$; (iii) *False positive rate* is the percentage of negative examples which are misclassified, $FPrate = FP/(TN + FP)$; (iv) *False negative rate* is the percentage of positive examples which are misclassified, $FNrate = FN/(TP + FN)$; (v) *Precision* (or *purity*) is defined as the percentage of samples which are correctly labeled as positive, $Precision = TP/(TP + FP)$.

One of the most widely-used techniques for the evaluation of classifiers in imbalanced domains is the ROC curve, which is a tool for visualizing, organizing and selecting classifiers based on their trade-offs between benefits (true positives) and costs (false positives). Furthermore, a quantitative representation of a ROC curve is the area under it, which is known as AUC [1,5]. When only one run is available from a classifier, its AUC can be computed as $AUC = (TPrate + TNrate)/2$ [10].

Kubat et al. [7] use *the geometric mean* of accuracies measured separately on each class, $Gmean = \sqrt{TPrate \cdot TNrate}$. This measure is associated to a point on the ROC curve, and the idea is to maximize the accuracies of both classes while keeping them balanced.

Both AUC and Gmean minimize the negative influence of skewed distributions of classes, but they do not distinguish the contribution of each class to the overall performance, nor which is the prevalent class. This means that different combinations of TPrate and TNrate produce the same value of the corresponding metric (AUC or Gmean).

More recently, Ranawana and Palade [9] proposed a new measure called *optimized precision* which is computed as $OP = Acc - (|TNrate - TPrate|/(TNrate + TPrate))$. This represents the difference between the global accuracy and a second term that computes how balanced the two class accuracies are. High OP performances require high global accuracies and balanced class accuracies. Nevertheless, it has to be pointed out that it can be strongly affected by the biased influence of the global accuracy.

## 3 The New Performance Evaluation Method

This section introduces a new measure, named *Index of Balanced Accuracy* (IBA), whose expression results from the computation of the area of a rectangular region in a two-dimensional space here called *Balanced Accuracy Graph*. This space is defined by the product of the accuracies on each class ($Gmean^2$), which is a suitable measure of the overall accuracy in imbalanced domains, and by a new simple index here introduced, the *Dominance*, which measures how prevalent is the dominant class rate with respect to the other. A final simulated example illustrates the benefits of the IBA with respect to some well-known classifier performance metrics.

### 3.1 The Dominance Index

As previously pointed out, AUC and Gmean are unable to explain the contribution of each class to the overall performances, giving the same result for many different combinations of $(TPrate, TNrate)$.

A new simple index called *Dominance* is here proposed for evaluating the relationship between the TPrate and TNrate. The expected role of this index is to inform about which is the dominant class and how significant is its dominance relationship. The *Dominance* can be computed as follows:

$$Dominance = TPrate - TNrate \tag{1}$$

This measure can take on any value between $-1$ and $+1$, since both the TPrate and the TNrate are in the range $[0, +1]$. A Dominance value of $+1$ represents a situation of perfect accuracy on the positive class, but failing on all negative cases; a value of $-1$ corresponds to the opposite situation. The closer the Dominance is to 0, the more balanced both individual rates are. In practice, the Dominance can be interpreted as an indicator of how balanced the TPrate and the TNrate are.

### 3.2 The Balanced Accuracy Graph

In order to take advantage of the good properties of the Gmean and the Dominance and to avoid their shortcomings, this section introduces the *Balanced Accuracy Graph* (BAG) as a tool to visualize and measure the behavior of a classifier from the joint perspective of global accuracy and Dominance. With the aim of simplicity, Gmean$^2$ is used instead of Gmean.

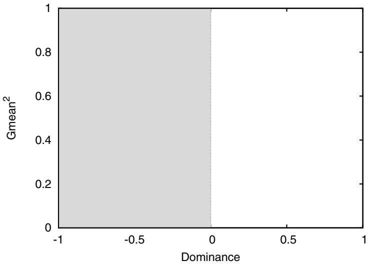

**Fig. 1.** The balanced accuracy graph. The plot represents the optimal point ($Dominance = 0$ and $Gmean^2 = 1$), producing the maximum area.

Fig. 1 illustrates the BAG as a two-dimensional coordinate system where the X axis corresponds to Dominance and the Gmean$^2$ is plotted on the Y axis. The BAG depicts relative trade-offs between Dominance and Gmean$^2$. The performance of a classifier measured as a pair (*Dominance*, $Gmean^2$) corresponds to a single point in this graph. The upper central point $(0, +1)$ represents a perfect classifier where $TPrate = TNrate = 1$, while points $(-1, 0)$ and $(+1, 0)$ match with the useless cases of $TPrate = 0, TNrate = 1$ and $TPrate = 1, TNrate = 0$, respectively.

The upper left $(-1, +1)$ and upper right $(+1, +1)$ points correspond to 'unfeasible cases' because when Dominance is $-1$ or $+1$, one of the two class rates is 0 what makes impossible for the Gmean to achieve values greater than 0. Actually, there is an infinite number of points in the BAG which represents unfeasible cases.

### 3.3 Index of Balanced Accuracy

Given a point $(d, g)$ in a BAG, it would be interesting to quantify the trade-off between Dominance and Gmean$^2$ represented by that point. To this end, we propose to use the rectangular area whose vertices are at points $(-1, 0)$, $(-1, g)$, $(d, g)$ and $(d, 0)$ (see

Fig. 1 for the case with $d = 0$ and $g = 1$). The area of such a rectangle is here named *Index of Balanced Accuracy* (IBA), which can be formalized as:

$$IBA = (1 + Dominance) \cdot Gmean^2 \tag{2}$$

When substituting Dominance and Gmean$^2$, the resulting function provides useful details for a better understanding about how IBA supports the trade-off. Besides, we add $0 \leq \alpha \leq 1$ to weight the value of Dominance. Significant effects are obtained for $\alpha \leq 0.5$. However, note that if $\alpha = 0$, the IBA turns into the Gmean$^2$.

$$IBA_\alpha = (1 + \alpha \cdot (TPrate - TNrate)) \cdot TPrate \cdot TNrate \tag{3}$$

The IBA can take on any value between 0 and $+1$, which is the area of the greatest possible rectangle, that is, the one corresponding to a point with Dominance$= 0$ and Gmean$^2 = 1$ (optimal classification). Fig. 2 illustrates the surface of the IBA (with $\alpha = 1$) as a function of TPrate and TNrate, showing that its maximum is +1 and that it occurs for $TPrate = TNrate = 1$. These facts can also be demonstrated by analytically optimizing the mathematical expression of IBA (Eq. 3).

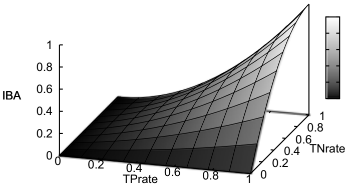

**Fig. 2.** The IBA function ($\alpha = 1$)

### 3.4 A Theoretical Example

Finally, a theoretical example is designed to clear up the benefits/advantages of the IBA with respect to other well-known metrics for classifier evaluation.

Let $f(\theta)$ be a classifier that depends on a set of parameters $\theta$. Suppose that $\theta$ should be optimized so that $f(\theta)$ can discriminate between the two classes of a particular imbalanced problem (with a ratio 1:10). Let $T$ and $V$ be the training and validation sets, respectively. During learning, four possible configurations ($\theta_1$, $\theta_2$, $\theta_3$, $\theta_4$) have been obtained from $T$, and then the corresponding classifiers $f(\theta_i)$ have been run over $V$. Table 2 reports the results of several performance measures used to evaluate each particular classifier $f(\theta_i)$. The last step in learning should be to pick up the best configuration $\theta^*$ according to the performance measure adopted.

First of all, note that configurations $\theta_1$ and $\theta_4$ correspond to cases with a clearly biased behavior, whereas $\theta_2$ and $\theta_3$ produce less differences between TPrate and TNrate. Both accuracy and AUC would select one of those biased $\theta_1$ and $\theta_4$. In the case of accuracy, this is because it strongly depends on the majority class rate. The geometric mean and OP suggest one of the moderate configurations $\theta_2$ and $\theta_3$, ignoring the fact that the minority class is usually the most important. While the former does not distinguish between them, OP would prefer $\theta_2$ rather than $\theta_3$ because its computation is affected

**Table 2.** Several performance measures for the theoretical example (highlighted are the best results for each metric)

| | TPrate | TNrate | Acc | Gmean | AUC | OP | IBA$_1$ | IBA$_{0.5}$ | IBA$_{0.1}$ |
|---|---|---|---|---|---|---|---|---|---|
| $\theta_1$ | 0.550 | 0.950 | **0.914** | 0.723 | **0.750** | 0.647 | 0.314 | 0.418 | 0.502 |
| $\theta_2$ | 0.680 | 0.810 | 0.798 | **0.742** | 0.745 | **0.711** | 0.479 | 0.515 | 0.544 |
| $\theta_3$ | 0.810 | 0.680 | 0.692 | **0.742** | 0.745 | 0.605 | 0.622 | 0.587 | **0.558** |
| $\theta_4$ | 0.950 | 0.550 | 0.586 | 0.723 | **0.750** | 0.320 | **0.732** | **0.627** | 0.543 |

by the accuracy. These drawbacks can be overcome when using the IBA measure by appropriately tuning the parameter $\alpha$ (see Eq. 3). One can see that IBA$_{0.1}$ selects $\theta_3$, which corresponds to the moderate case with the highest TPrate.

## 4 Experiments

In this section we present an experiment with the aim of validating usefulness and consistency of the IBA measure. To this end, we will compare IBA with a number of representative metrics: accuracy, geometric mean, AUC, and optimized precision. The experiment is carried out on 17 real data sets taken from the UCI Machine Learning Database Repository (`http://archive.ics.uci.edu/ml/`) and a private library (`http://www.vision.uji.es/~sanchez/Databases/`). All data sets were transformed into two-class problems by keeping one original class and joining the objects of the remaining classes. From these databases, here we have included the results of the four cases representing the most diversity of situations. The majority/minority ratio for each of these databases is: Breast (2.42), Glass (11.59), Satimage (9.28), and Laryngeal-2 (12.06). The results for the rest of databases are available at (`http://www.vision.uji.es/~sanchez/Results`).

For each database, we have estimated the performance measures mentioned above by repeating 5 times a 10–fold cross–validation when using different classifiers: the nearest neighbor (1-NN) rule, a multi-layer perceptron (MLP), a support vector classifier (SVC), the naïve Bayes classifier (NBC), a decision tree (J48), and a radial basis function network (RBF).

### 4.1 The Results

For each database, the six classifiers have been used and their results (TPrate and TNrate) have been evaluated in terms of the five performance metrics. From this, each measure will suggest the best classifier. Our study will consist of judging the decision inferred from each measure with the aim of remarking the merits of IBA when compared to other metrics.

From results in Table 3, some preliminary conclusions can be drawn. In general, as expected, accuracy appears ineffective in imbalanced domains. IBA usually chooses the classifier with the highest TPrate, demonstrating to be robust as regards to the parameter $\alpha$. The rest of measures are usually affected by high TNrates, thus undervaluing the relevance of TPrate. Focusing on each particular database, some comments can be remarked:

**Table 3.** Experimental results (highlighted are the best results for each metric)

Breast

| | TPrate | TNrate | Acc | Gmean | AUC | OP | IBA$_1$ | IBA$_{0.5}$ | IBA$_{0.1}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1NN | 0.454 | 0.761 | 0.671 | 0.588 | 0.608 | **0.419** | **0.240** | **0.293** | **0.335** |
| MLP | 0.368 | 0.845 | 0.705 | 0.558 | 0.606 | 0.312 | 0.163 | 0.237 | 0.296 |
| SVC | 0.328 | 0.864 | 0.708 | 0.533 | 0.596 | 0.258 | 0.132 | 0.208 | 0.269 |
| NBC | 0.402 | 0.865 | **0.730** | **0.590** | **0.634** | 0.365 | 0.187 | 0.268 | 0.332 |
| J48 | 0.321 | 0.891 | 0.724 | 0.535 | 0.606 | 0.254 | 0.123 | 0.204 | 0.270 |
| RBF | 0.390 | 0.867 | 0.728 | 0.582 | 0.629 | 0.348 | 0.177 | 0.258 | 0.322 |

Glass

| | TPrate | TNrate | Acc | Gmean | AUC | OP | IBA$_1$ | IBA$_{0.5}$ | IBA$_{0.1}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1NN | 0.318 | 0.948 | 0.898 | 0.549 | 0.633 | 0.400 | 0.111 | 0.206 | 0.282 |
| MLP | 0.141 | 0.969 | 0.903 | 0.370 | 0.555 | 0.157 | 0.024 | 0.080 | 0.125 |
| SVC | 0.000 | 0.991 | **0.912** | 0.000 | 0.495 | -0.088 | 0.000 | 0.000 | 0.000 |
| NBC | 0.753 | 0.444 | 0.468 | 0.578 | 0.598 | 0.210 | **0.437** | **0.386** | **0.344** |
| J48 | 0.353 | 0.963 | 0.915 | **0.583** | **0.658** | **0.451** | 0.132 | 0.236 | 0.319 |
| RBF | 0.000 | 0.999 | 0.920 | 0.000 | 0.499 | -0.080 | 0.000 | 0.000 | 0.000 |

Satimage

| | TPrate | TNrate | Acc | Gmean | AUC | OP | IBA$_1$ | IBA$_{0.5}$ | IBA$_{0.1}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1NN | 0.705 | 0.960 | **0.935** | 0.823 | 0.833 | 0.782 | 0.504 | 0.591 | 0.660 |
| MLP | 0.637 | 0.966 | 0.934 | 0.785 | 0.802 | 0.729 | 0.413 | 0.515 | 0.596 |
| SVC | 0.000 | 1.000 | 0.903 | 0.000 | 0.500 | -0.097 | 0.000 | 0.000 | 0.000 |
| NBC | 0.870 | 0.815 | 0.820 | **0.842** | **0.843** | **0.788** | **0.749** | **0.729** | **0.713** |
| J48 | 0.550 | 0.959 | 0.919 | 0.726 | 0.755 | 0.648 | 0.312 | 0.420 | 0.506 |
| RBF | 0.000 | 1.000 | 0.903 | 0.000 | 0.500 | -0.097 | 0.000 | 0.000 | 0.000 |

Laryngeal-2

| | TPrate | TNrate | Acc | Gmean | AUC | OP | IBA$_1$ | IBA$_{0.5}$ | IBA$_{0.1}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1NN | 0.694 | 0.970 | 0.949 | 0.821 | 0.832 | 0.783 | 0.488 | 0.581 | 0.655 |
| MLP | 0.770 | 0.979 | **0.963** | 0.868 | 0.875 | **0.844** | 0.596 | 0.675 | 0.738 |
| SVC | 0.672 | 0.942 | 0.922 | 0.796 | 0.807 | 0.754 | 0.462 | 0.547 | 0.616 |
| NBC | 0.943 | 0.844 | 0.851 | **0.892** | **0.894** | 0.796 | **0.875** | **0.836** | **0.804** |
| J48 | 0.638 | 0.980 | 0.954 | 0.791 | 0.809 | 0.742 | 0.411 | 0.518 | 0.604 |
| RBF | 0.558 | 0.985 | 0.952 | 0.742 | 0.772 | 0.676 | 0.316 | 0.433 | 0.526 |

**Breast:** 1NN and NBC provide similar results in terms of TPrate and TNrate. However, OP and IBA suggest 1NN because its performances on both classes are more balanced. In contrast, the other measures select NBC, where the overall error is lower due to a greater bias to the majority class.

**Glass:** IBA prefers NBC because the TPrate is clearly the best and even it is higher than the TNrate. Gmean, AUC and OP choose J48 because they are strongly affected by the overall error, despite the low performance on the minority class makes this classifier useless.

**Satimage:** This is a straightforward case, in which all measures (except accuracy) give NBC as the best classifier. Both TPrate and TNrate are high enough and they are sufficiently balanced.

**Laryngeal-2:** Gmean, AUC and IBA select NBC, which seems to be the classifier with the best performance. The fact that OP prefers MLP is because it depends on the overall accuracy (here particularly affected by a significant imbalance ratio).

## 5 Conclusions and Further Extensions

In this paper, we have introduced a new method to evaluate the performance of classification systems in two-class problems with skewed data distributions. It is defined as a trade-off between a global performance measure (Gmean$^2$) and a new proposed signed index to reflect how balanced are the individual accuracies (Dominance). High values of the new measure IBA are obtained when the accuracies of both classes are high and balanced. Unlike most metrics, the IBA function does not take care of the overall accuracy only, but also intends to favor classifiers with better results on the positive class (generally the most important class). The most closely related measure to IBA is the optimized precision, although this is biased to the majority class.

Theoretical and empirical studies have shown the robustness and advantages of IBA with respect to some other well-known performance measures. Future work will primarily be addressed to extend the combination of Dominance with other global metrics especially useful for certain real-world applications. Also, this kind of performance evaluation methods could be designed to include misclassification costs.

## References

1. Bradley, P.W.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30, 1145–1159 (1997)
2. Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. SIGKDD Exploration Newsletters 6, 1–6 (2004)
3. Daskalaki, S., Kopanas, I., Avouris, N.: Evaluation of classifiers for an uneven class distribution problem. Applied Artificial Intelligence 20, 381–417 (2006)
4. Elazmeh, W., Japkowicz, N., Matwin, S.: Evaluating misclassifications in imbalanced data. In: Proc. 17th European Conference on Machine Learning, pp. 126–137 (2006)
5. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. on Knowledge and Data Engineering 17, 299–310 (2005)
6. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intelligent Data Analysis 6, 40–49 (2002)
7. Kubat, M., Matwin, S.: Adressing the curse of imbalanced training sets: one-sided selection. In: Proc. 14th Intl. Conf. on Machine Learning Nashville, TN, pp. 179–186 (1997)
8. Provost, F., Fawcett, T.: Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In: Proc. 3rd Intl. Conf. on Knowledge Discovery and Data Mining Newport Beach, CA, pp. 43–48 (1997)
9. Ranawana, R., Palade, V.: Optimized Precision - A new measure for classifier performance evaluation. In: Proc. IEEE Congress on Evolutionary Computation, pp. 2254–2261 (2006)
10. Sokolova, M., Japkowicz, N., Szpakowicz, S.: Beyond Accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. In: Sattar, A., Kang, B.-h. (eds.) AI 2006. LNCS, vol. 4304, pp. 1015–1021. Springer, Heidelberg (2006)