Theoretical Analysis of a Performance Measure for Imbalanced Data[*]

V. García, R.A. Mollineda and J.S. Sánchez
Institute of New Imaging Technologies - Dept. Llenguatges i Sistemes Informàtics, Universitat Jaume I
Av. Sos Baynat s/n, 12071 Castelló de la Plana, Spain
{jimenezv,mollined, sanchez}@uji.es

Abstract

This paper analyzes a generalization of a new metric to evaluate the classification performance in imbalanced domains, combining some estimate of the overall accuracy with a plain index about how dominant the class with the highest individual accuracy is. A theoretical analysis shows the merits of this metric when compared to other well-known measures.

1. Introduction

Most of learning methods assume that the classes of the problem share similar prior probabilities. However, in many real-world tasks the ratios of prior probabilities between classes are significantly skewed. This is known as the *imbalance problem* [10]. A two-class data set is said to be imbalanced when one of the classes is heavily under-represented as regards the other class [6]. 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 claimed by many authors [2, 3, 8], the use of plain accuracy (error) rates to evaluate the classification performance in imbalanced domains might produce misleading conclusions, since they do not take misclassification costs into account, are strongly biased to favor the majority class, and are sensitive to class skews.

Most of alternative metrics are formulated as combinations of accuracy (error) rates measured separately on each class, to alleviate biased results. Nevertheless, none of these show up how dominant the accuracy of an individual class is over another, nor distinguish the contribution of each class to the overall performance.

This paper discusses a generalization of a new metric to estimate the classifier performance on two-class imbalanced data sets. It quantifies a trade-off between an index of how balanced both class accuracies are and some (unbiased) measure of overall accuracy. The first term is intended to explain the balance degree between class accuracies, and it favors those cases with higher accuracy rate on the positive class. Some illustrative examples and an extensive theoretical study are performed to better understand the differences between the measure here proposed and other well-known metrics.

2. Performance evaluation measures

Traditionally, classification accuracy (Acc) and/or error rates have been the standard metrics used to estimate the performance of learning systems. For a two-class problem, they can be easily derived from a 2 × 2 confusion matrix as that given in Table 1.

Table 1. Confusion matrix.

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

However, empirical and theoretical evidences show that these measures are biased with respect to data imbalance and proportions of correct and incorrect classifications. These shortcomings have motivated a search for new metrics based on simple indexes, such as the *true positive rate* (TPrate) and the *true negative rate* (TNrate). The TPrate (TNrate) is the percentage of positive (negative) examples correctly classified.

One of the most widely-used evaluation methods in the context of class imbalance is the ROC curve, which is a tool for visualizing and selecting classifiers based on their trade-offs between benefits (true positives) and costs (false positives). A quantitative representation of a ROC curve is the area under it (AUC) [1]. For just one run of a classifier, the AUC can be computed as [9] $AUC = (TPrate + TNrate)/2$.

[*] Partially supported by the Spanish Ministry of Education and Science under grants CSD2007–00018, AYA2008–05965–C04–04/ESP and TIN2009–14205–C04–04, and by Fundació Caixa Castelló–Bancaixa under grant P1–1B2009–04.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.156

IEEE computer society

Kubat et al. [7] use *the geometric mean* of accuracies measured separately on each class, with the aim of maximizing the accuracies of both classes while keeping them balanced, $Gmean = \sqrt{TPrate \cdot TNrate}$.

Both AUC and Gmean minimize the negative influence of skewed distributions of classes, but they do not show up the contribution of each class to the overall performance, nor which is the prevalent class. This means that different combinations of TPrate and TNrate may produce the same result for those metrics.

Recently, Ranawana and Palade [8] introduced the *optimized precision*, which can be computed as,

$$OP = Acc - \frac{|TNrate - TPrate|}{TNrate + TPrate} \qquad (1)$$

This represents the difference between the global accuracy and a second term that computes how balanced both class accuracies are. High OP values require high global accuracy and well-balanced class accuracies. However, OP can be strongly affected by the biased influence of the global accuracy.

3. Generalizing a new performance metric

This section provides a generalization of a primary index reported in [4], named *Index of Balanced Accuracy* (IBA). The main purpose of the generalized IBA will be to weight a measure suitable to evaluate the performance in imbalanced domains. The weighting factor will aim at favoring those results with better classification rates on the minority class.

The generalized IBA can be formulated as follows:

$$IBA_\alpha(\mathcal{M}) = (1 + \alpha \cdot Dom) \cdot \mathcal{M} \qquad (2)$$

where $(1 + \alpha \cdot Dom)$ is the weighting factor and $\mathcal{M}$ represents any performance metric.

The Dom term, called *dominance*, is defined as $Dom = TPrate - TNrate$ within the range $[-1, +1]$, and it is here used to estimate the relationship between the TPrate and TNrate. The closer the dominance is to 0, the more balanced both individual rates are. If $TPrate > TNrate$, then $Dom > 0$; otherwise, $Dom < 0$.

The value of Dom is weighted by $\alpha \geq 0$ to reduce its influence on the result of the particular metric $\mathcal{M}$. Thus the weighting factor in Eq. 2 is within the range $[1-\alpha, 1+\alpha]$. Note that if $\alpha = 0$ or $TPrate = TNrate$, the IBA$_\alpha$ turns into the measure $\mathcal{M}$. In practice, one should select a value of α depending on the metric used.

3.1. Formulating IBA$_\alpha$ with Gmean

As a representative example, this paper will use Gmean because this is a suitable, well-known performance measure for class imbalanced problems. Hence IBA$_\alpha$ can now be rewritten in terms of Gmean as:

$$IBA_\alpha(Gmean) = (1 + \alpha \cdot Dom) \cdot Gmean \qquad (3)$$

Since α will depend on the metric $\mathcal{M}$, the following study is devoted to empirically set an appropriate value of α for the particular case of IBA$_\alpha$(Gmean). Also, this example will allow to clear up the behavior differences of IBA$_\alpha$ with respect to other metrics.

Let $f(\theta)$ be a classifier that depends on a set of parameters θ. Suppose that θ 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, seven possible configurations $(\theta_1, \theta_2, \ldots, \theta_7)$ have been obtained from T, and then the corresponding classifiers $f(\theta_i)$ have been run over V. Table 2 reports the results of some measures used to evaluate each classifier $f(\theta_i)$. The last step in learning should be to pick up the best configuration θ^* according to the performance measure adopted.

Table 2. A synthetic example.

	TPrate	TNrate	Acc	Gmean	AUC	OP	IBA$_{0.05}$	IBA$_{0.1}$	IBA$_{0.2}$
θ_1	0.550	0.950	0.914	0.723	0.750	0.647	0.708	0.694	0.665
θ_2	0.650	0.850	0.832	0.743	0.750	0.698	0.736	0.728	0.714
θ_3	0.700	0.800	0.791	0.748	0.750	0.724	0.745	0.741	0.733
θ_4	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750
θ_5	0.800	0.700	0.709	0.748	0.750	0.642	0.752	0.756	0.763
θ_6	0.850	0.650	0.668	0.743	0.750	0.535	0.751	0.758	0.773
θ_7	0.950	0.550	0.586	0.723	0.750	0.320	0.737	0.752	0.781

Note that configurations θ_1 and θ_7 correspond to cases with a clearly biased behavior, whereas θ_4 produces a perfect balance between TPrate and TNrate. The rest of configurations θ_2, θ_3, θ_5 and θ_6 produce less differences between TPrate and TNrate.

For this example, AUC is of no value at all since all configurations give the same value. Accuracy would select the biased θ_1 because it strongly depends on the majority class rate. Both Gmean and OP suggest the most balanced configurations (θ_3, θ_4, θ_5), ignoring the fact that the minority class is usually the most important. While Gmean does not distinguish between θ_3 and θ_5, OP would prefer θ_3 rather than θ_5 because its computation is affected by the accuracy. These drawbacks can be overcome when using the IBA$_\alpha$ measure by appropriately tuning the parameter α. One can see that IBA$_{0.05}$ and IBA$_{0.1}$ select θ_5 or θ_6, which correspond to the moderate cases with the highest TPrate.

Results of IBA$_{0.2}$ show a biased tendency of IBA$_\alpha$ towards TPrate for high and moderate values of α. This effect is due to the strong influence of Dom on IBA$_\alpha$, what justifies the need of α to weight its importance. This study suggests that the use of small values of α allow to correct this effect and thus, we propose $\alpha =$

0.05 for the calculation of IBA$_\alpha$(Gmean).

4. The theoretical analysis of IBA$_\alpha$

Two theoretical studies are performed to explore the possible advantages of IBA$_\alpha$ (with $\alpha = 0.05$ and $\mathcal{M} = Gmean$) over other metrics. One computes Pearson correlation coefficients in order to devise how IBA$_\alpha$ is correlated with other metrics that might be deemed as good or bad choices to tackle the imbalance. The second study analyzes how sensitive the metrics are under different types of changes to the confusion matrix.

4.1. Correlation analysis

Five collections of classifier output tuples based on different imbalance ratios were generated as in [5]. A classifier output tuple consists of a list of n numeric values between 0 and 1 which represent, for n hypothetical samples, the probabilities of belonging to the positive class (classifier outputs). All tuples were generated from a main ranked list where the i-th component is the "true" probability p_i of belonging the instance i to the positive class. However, in contrast to [5], this list was defined considering a particular imbalance level in the assignment of true probabilities. For example, for an imbalance ratio of 1:3, the first 75% of instances in the list were linked to probabilities within the range $[0, 0.5]$ (negative class), while the rest were associated to probabilities in $(0.5, 1]$ (positive class). Given an imbalance true tuple as the one just described, a perturbed tuple was generated by randomly fluctuating the true probabilities p of negative samples within the range $[max(0, p - \epsilon_n), min(1, p + \epsilon_n)]$, and the true probabilities p of positive samples within the range $[max(0, p - \epsilon_p), min(1, p + \epsilon_p)]$. The use of two distortion terms, ϵ_n for the negative class and ϵ_p for the positive class, allows to simulate different scenarios of biased learning: for $\epsilon_n > \epsilon_p$, a greater proportion of negative samples should be "misclassified", while for $\epsilon_n < \epsilon_p$ the positive class should be the most affected.

Table 3 is an example of a true tuple (T) with 12 samples and an imbalance ratio of $1 : 3$, along with two derived perturbed tuples, P1 and P2, obtained from $(\epsilon_n = 0.3, \epsilon_p = 0)$ and $(\epsilon_n = 0, \epsilon_p = 0.3)$, respectively. Items typed in bold face represent misclassified samples. P1 simulates the outputs of a classifier focused on the positive class, while P2 contains the results of a biased classifier that favors the negative class.

The five collections of classifier output tuples used in the analysis were drawn from five different imbalance ratios expressed in terms of the percentage of positive samples: 5%, 10%, 15%, 20% and 25%. Each collection was composed of 130 tuples distributed in 10 per each of the 13 combinations of distortion terms ranging from $(\epsilon_n = 0.6, \epsilon_p = 0)$ to $(\epsilon_n = 0, \epsilon_p = 0.6)$ with steps $(-0.05, 0.05)$ and satisfying $\epsilon_n + \epsilon_p = 0.6$.

Table 3. An example of a true and two perturbed tuples for an imbalance ratio 1:3.

T	0.06	0.11	0.17	0.22	0.28	0.33	0.39	0.44	0.5	0.67	0.83	1.0
	−	−	−	−	−	−	−	−	−	+	+	+
P1	0.0	0.0	0.23	**0.52**	0.26	0.49	**0.55**	0.24	**0.61**	0.67	0.83	1.0
	−	−	−	**+**	−	−	**+**	−	**+**	+	+	+
P2	0.06	0.11	0.17	0.22	0.28	0.33	0.39	0.44	0.5	**0.45**	0.57	1.0
	−	−	−	−	−	−	−	−	−	**−**	+	+

An independent correlation matrix between all pairs of metrics was built for each collection. Correlation coefficients were plotted in Figure 1 to make easier the understanding of results. The axes X and Y correspond to the correlation values in the range $[-1, +1]$ and the percentage of positive samples, respectively.

Several comments related with IBA$_{0.05}$(Gmean) can be drawn from Figure 1:

- IBA$_\alpha$ shows a very low (negative) correlation with Acc, which has been proven not to be appropriate for imbalanced domains. Besides, the correlation coefficient of IBA$_\alpha$ in terms of absolute value is slightly lower than those of AUC and Gmean, which are even positive.
- IBA$_\alpha$ has a very high (positive) correlation with AUC and Gmean, suggesting that IBA$_\alpha$ can be suitable for imbalanced distributions.
- IBA$_\alpha$ appears to be clearly the most correlated measure with TPrate, which represents the classifier performance on the most important class (the minority one).
- IBA$_\alpha$ presents a very low (negative) correlation with TNrate. Although AUC and Gmean show very low correlations with TNrate, their coefficients are positive.

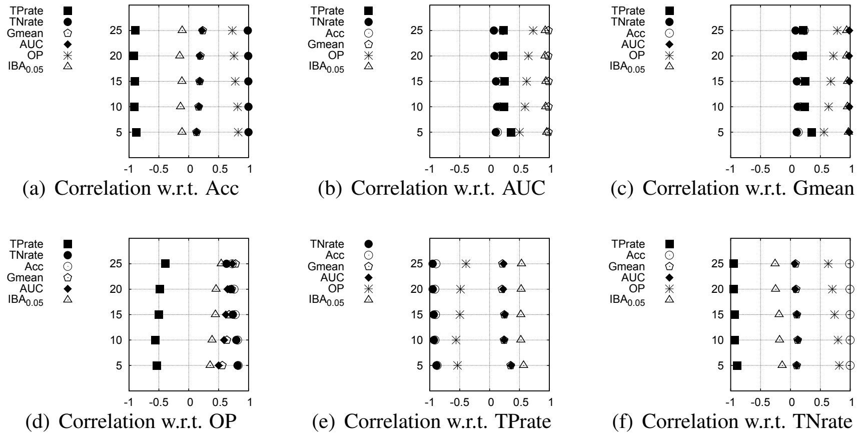
Despite OP was defined in the context of class imbalance, it is strongly correlated with Acc, TPrate and TNrate, due to the great influence of accuracy on it.

4.2. Invariance properties

This second analysis intends to assess invariance properties of various metrics with respect to four basic changes to the confusion matrix of Table 1. A measure is said to be invariant to a certain change if it cannot distinguish a new configuration from the previous one. In general, a robust performance measure should detect every matrix transformation. Four invariance properties [9] are here used to demonstrate that IBA$_\alpha$ is more sensitive to changes than the remaining metrics.

p1 invariance under the exchange of TP with TN and FN with FP.

p2 invariance under a change in TN, while all other matrix entries remain the same.



(a) Correlation w.r.t. Acc (b) Correlation w.r.t. AUC (c) Correlation w.r.t. Gmean

(d) Correlation w.r.t. OP (e) Correlation w.r.t. TPrate (f) Correlation w.r.t. TNrate

Figure 1. Scatterplot of Pearson correlation coefficients.

p3 invariance under a change of FP, while the other matrix entries do not change.

p4 invariance under scaling: $TP \to k_1TP$, $TN \to k_2TN$, $FP \to k_1FP$, $FN \to k_2FN$, where $k1, k2 > 0$.

Table 4 reports the invariance properties of the measures considered in this paper. '+' and '–' indicate invariance and non-invariance, respectively. As can be observed, IBA$_\alpha$ is the only measure capable of detecting all types of changes, what suggests that it is more sensitive to changes than the other metrics.

Table 4. Invariance properties.

	Acc	TPrate	TNrate	Gmean	AUC*	OP	IBA$_\alpha$
p1	+	–	–	+	+	–	–
p2	–	+	–	–	–	–	–
p3	–	+	–	–	–	–	–
p4	–	–	–	–	–	–	–

* Valid for AUC when only one classifier run is available.

5. Conclusions

We have analyzed a generalization of a new metric, IBA$_\alpha$, to evaluate the classifier performance in two-class imbalanced problems. It is defined as a trade-off between a global performance measure and a simple signed index to reflect how balanced the individual accuracies are. High values of IBA$_\alpha$ are achieved when the accuracies of both classes are high and significantly balanced. Unlike most metrics, IBA$_\alpha$ does not take care of the overall accuracy only, but also intends to favor classifiers with better results on the positive class.

Two theoretical studies have shown the benefits of the new metric when compared to other measures. In this sense, it has been proven that IBA$_\alpha$ is strongly correlated with AUC and Gmean (generally accepted as good measures for imbalance problems). However, unlike AUC and Gmean, IBA$_\alpha$ is more correlated with TPrate and less (and negatively) correlated with accuracy. Also, a study on invariance properties has shown that IBA$_\alpha$ is more sensitive to changes to the confusion matrix than the other measures here considered.

References

[1] P. W. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Patt. Recog.*, 30(7):1145–1159, 1997.

[2] S. Daskalaki, I. Kopanas, and N. Avouris. Evaluation of classifiers for an uneven class distribution problem. *Appl. Artif. Intell.*, 20(5):381–417, 2006.

[3] C. Ferri, J. Hernández-Orallo, and R. Modroiu. An experimental comparison of performance measures for classification. *Patt. Recog. Lett.*, 30(1):27–38, 2009.

[4] V. García, R. A. Mollineda, and J. S. Sánchez. Index of balanced accuracy: A performance measure for skewed class distributions. In *4th IbPRIA*, pages 441–448, 2009.

[5] J. Huang and C. X. Ling. Constructing new and better evaluation measures for machine learning. In *20th IJCAI*, pages 859–864, 2007.

[6] N. Japkowicz and S. Stephen. The class imbalance problem: a systematic study. *Intell. Data Anal.*, 6(5):40–49, 2002.

[7] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: one-sided selection. In *14th ICML*, pages 179–186, 1997.

[8] R. Ranawana and V. Palade. Optimized Precision - A new measure for classifier performance evaluation. In *IEEE CEC*, pages 2254–2261, 2006.

[9] M. Sokolova. Assessing invariance properties of evaluation measures. In *Workshop on Testing of Deployable Learning and Decision Systems*, 2006.

[10] Y. Sun, A. K. C. Wong, and M. S. Kamel. Classification of imbalanced data: A review. *Int'l. J. Patt. Recog. Artif. Intell.*, 23(4):687–719, 2009.