ON THE SUITABILITY OF NUMERICAL PERFORMANCE MEASURES FOR CLASS IMBALANCE PROBLEMS

Vicente García, J. Salvador Sánchez and Ramón A. Mollineda
Institute of New Imaging Technologies, Universitat Jaume I, Av. Sos Baynat, s/n, 12071, Castellón de la Plana, Spain

Keywords: Performance measure, Class imbalance problem, Classification.

Abstract: The class imbalance problem has been reported as an important challenge in various fields such as Pattern Recognition, Data Mining and Machine Learning. A less explored research area is related to how to evaluate classifiers on imbalanced data sets. This work analyzes the behaviour of performance measures widely used on imbalanced problems, as well as other metrics recently proposed in the literature. We perform two theoretical analysis based on Pearson correlation and operations for a 2×2 confusion matrix with the aim to show the strengths and weaknesses of those performance metrics in the presence of skewed distributions.

1 INTRODUCTION

A problem that has received considerable attention is when the data sets show heavily skewed ratios of prior probabilities between classes, what has been usually called the imbalance problem (Sun et al., 2009). This may affect standard learning algorithms which assume that the classes of the problem share similar prior probabilities. A two-class data set is said to be imbalanced when the instances of a class (the majority one) heavily outnumbers the instances of the other (the minority) class. This topic is particularly important in those applications where it is costly to misclassify examples from the minority class (Kennedy et al., 2010; Khalilia et al., 2011; Kamal et al., 2009).

Several works have shown that the use of plain accuracy and/or error rates to evaluate the classification 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 (Daskalaki et al., 2006; Fatourechi et al., 2008; Folleco et al., 2008; Ferri et al., 2009; Gu et al., 2009; Huang and Ling, 2005; Seliya et al., 2009).

In this paper, we review and analyse the most popular performance measures used to evaluate classifiers on imbalanced problems, as well as recently introduced but less renowned metrics. Hereafter the paper is organized as follows. Section 2 reviews numerical performance metrics used for the evaluation of classifiers on two-class imbalanced data sets. Section 3 shows two theoretical studies based on the computation of Pearson correlation coefficients and the assessment of invariance properties with respect to changes to a confusion matrix. Finally. Section 4 remarks the main conclusions of this work.

2 PERFORMANCE METRICS

Traditionally, classification accuracy and/or error rates have been the standard performance metrics used to evaluate the performance of learning systems. For a two-class problem, these can be easily derived from a 2×2 confusion matrix (Table 1), $Acc = (TP + TN)/(TP + FN + TN + FP)$ and $Err = 1 - Acc$.

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)

Empirical and theoretical evidences show that these measures are strongly 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 indices, such as the *true positive rate* (TPr) , the *true negative rate* (TNr), and the precision (or purity). The TPr (TNr) is the percentage of positive (negative) examples correctly classified. The precision is defined as the percentage of examples that are correctly labeled as positive, $Prec = TP/(TP + FP)$.

García V., Salvador Sánchez J. and A. Mollineda R. (2012).
ON THE SUITABILITY OF NUMERICAL PERFORMANCE MEASURES FOR CLASS IMBALANCE PROBLEMS.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 310-313
DOI: 10.5220/0003783303100313

One of the most popular techniques for the evaluation of classifiers in imbalanced problems is the Receiver Operating Characteristic (ROC) curve. A quantitative representation of a ROC curve is the area under it, which is known as AUC (Bradley, 1997). When only one run is available from a classifier, the AUC can be computed as the arithmetic mean (macro-average) of TPr and TNr (Sokolova et al., 2006), $AUC = (TPr + TNr)/2$.

Kubat and Matwin (1997) use the geometric mean of accuracies measured separately on each class, $Gmean = \sqrt{TPr \cdot TNr}$. This metric is associated to a point on the ROC curve, and the idea is to maximize the accuracies of both classes while keeping them balanced. Although *AUC* and *Gm* minimize the negative influence of skewed class distributions, they cannot distinguish between the contribution of each class to the overall performance, nor which is the dominant class. This means that different combinations of TPr and TNr may produce the same result *AUC* and *Gm*.

The F-measure (Rijsbergen, 1979) is used to integrate the true positive rate and precision into a single metric, $F = ((1+\beta^2) \cdot (TPr \cdot Prec))/(\beta^2 \cdot Prec + TPr)$. The non-negative real β is a tunable parameter to control the influence of the true positive rate and precision separately. Typically, β is set to 1, thus obtaining the F_1–measure ($F_1 \in [0,+1]$), which can be viewed as a harmonic mean of the true positive rate and precision, $F_1 = (2 \cdot TPr \cdot Prec)/(Prec + TPr)$.

The kappa statistic ($\kappa \in [-1,+1]$) (Cohen, 1960) measures pairwise agreement among a set of classifications, correcting for expected chance agreement, $\kappa = (P_A - P_e)/(1 - P_e)$. Here, P_A denotes the accuracy and P_e is the proportion of times that an agreement by chance could be expected, $P_e = ((N_p \cdot R_p) + (N_n \cdot R_n))/N^2$, N_p (N_n) and R_p (R_n) represent the total number of actual and predicted positive (negative) instances respectively, and N is the total number of instances in the data set.

Ranawana and Palade (2006) proposed a new measure called optimized precision, which is computed as $OP = Acc - |TNr - TPr|/(TNr + TPr)$. High *OP* performances require high global accuracies and well-balanced class accuracies. However, *OP* can be strongly affected by the bias of the global accuracy.

Cohen et al. (2006) proposed the mean class-weighted accuracy, which is defined for a two-class problem as weighted mean between TPr and TNr, $cwA = w \cdot TPr + (1-w) \cdot TNr$. The coefficient w is a normalized weight assigned to TPr, such that $0 \leq w \leq 1$.

The weighted AUC measure (Weng and Poon, 2008) has been proposed to give more weights to the areas close to the top of the ROC graph, which is the region with higher TPr. The idea is to move a certain percentage of weights from the bottom areas to the upper areas of the ROC curve. This can be performed by means of a recursive formula that computes the new weight of an area using the weight of the preceding area; thus for n areas to sum, the next weight w can be computed as follows:

$$w(x) = \begin{cases} \rho & x = 0 \\ w(x-1) \cdot \rho + (1-\rho) & 0 < x < n \\ \frac{w(x-1) \cdot \rho + (1-\rho)}{1-\rho} & x = n \end{cases} \quad (1)$$

where $\rho \in [0,+1]$ is the percentage of weight to transfer to the next area towards the top of the ROC curve. When ρ is 0, the weighted AUC is equal to the conventional AUC. $w(i)$ represents the weight for the bottom area, which is used to compute the new AUC-based measure by adding up the successive weighted areas, $wAUC = \sum_{i=0}^{n} area(i) \cdot w(i)$.

Batuwita and Palade (2009) showed that some performance measures could lead to sub-optimal classification models, i.e, with a higher true positive rate and a lower true negative rate. Thus they proposed to combine *Gm*, TNr and the proportion of the negative examples (P_n) into one measure called the adjusted geometric mean, $AGm = (Gm + TNr \cdot P_n)/(1 + P_n)$, which is more sensitive to the changes in TNr than to changes in TPr.

More recently, García et al. (2010) proposed a new measure called generalized index of balanced accuracy, which can be expressed in terms of *Gm* as follows: $IBA_\alpha(Gm) = (1 + \alpha \cdot Dom) \cdot Gm$, where *Dom*, called *dominance*, is defined as $Dom = TPr - TNr$, and it is weighted by $\alpha \geq 0$ to reduce its influence on the result of the particular metric. The IBA function not only takes care of the overall accuracy but also intends to favor classifiers with better results on the positive class. In the present paper, we will use $\alpha = 0.05$.

3 ANALYSIS OF METRICS

Two theoretical comparisons are carried out to study the behaviour of the metrics previously described in Sect 2. These measures are TPr, TNr, *Prec*, *Acc*, *AUC*, *Gm*, κ, F_1, *OP*, $IBA_{0.05}(Gm)$, *cwA*, *AGm* and *wAUC*[1]. The first one consists of the computation of Pearson correlation coefficients between all pairs of measures, considering several collections of synthetic classifier outputs randomly drawn from different levels of imbalance. This analysis focuses on how the

[1]With ρ = 0.10 like in Weng and Poon (2008).

performance measures are correlated with both TPr and TNr when dealing with imbalance. The second study concerns the ability of a measure to preserve its value under a change in the confusion matrix.

3.1 Pearson Correlation Analysis

For this study, five collections of classifier output tuples based on different imbalance degrees were generated as in Huang and Ling (2007). 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 Huang and Ling (2007), this main tuple was defined considering a particular imbalance level in the assignment of true probabilities. Given an imbalance true tuple, a perturbed tuple was generated by randomly fluctuating the true probabilities p of *negative* instances within the range $[max(0, p-\varepsilon_n), min(1, p+\varepsilon_n)]$, and the true probabilities p of *positive* instances within the range $[max(0, p-\varepsilon_p), min(1, p+\varepsilon_p)]$. 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 of classifiers.

The five collections of classifier output tuples used in the analysis were respectively drawn from five different imbalance degrees expressed in terms of the percentage of samples from the positive class: 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 $(\varepsilon_n = 0.6, \varepsilon_p = 0)$ to $(\varepsilon_n = 0, \varepsilon_p = 0.6)$ with steps $(-0.05, 0.05)$ and satisfying $\varepsilon_n + \varepsilon_p = 0.6$. An independent correlation matrix between all pairs of performance measures was built for each collection, regarding those metrics presented in this paper. From matrix correlations, several interesting conclusion can be drawn:

- As expected, *Acc* shows negative correlation with TPr but strong positive correlation with TNr. It proves that *Acc* is not appropriate for imbalanced domains.
- *AUC*, *Gm*, IBA, *wAUC* and *cwA* show positive correlation with TPr, which represents the classifier performance on the most important class (the minority one). *cwA* with $w > 0.5$ presents strong positive and negative correlation values with TPr and TNr. Note that this metric may show a biased behaviour when $w \neq 0.5$.
- Despite *OP*, κ, *AGm* and F_1 have been proposed as metrics especially suitable for imbalanced domains, the correlation analysis indicates that they are strongly correlated with TNr.

3.2 Invariance Properties

This second analysis deals with the assessment of invariance properties of the measures with respect to five changes to the confusion matrix (Sokolova and Lapalme, 2009; Tan et al., 2002). In general, a robust performance measure should detect any matrix transformation. We have used five invariance properties, which can be defined as follows:

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.

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

p4: invariance under a change in TP, while all other matrix entries remain the same.

p5: invariance under a change in FN, while all other matrix entries remain the same.

A straightforward analysis on each performance measure lets us know whether or not it meets the above invariance properties. Table 2 illustrates these results, where '–' and '+' indicate invariance and non-invariance, respectively.

Table 2: Invariance properties of the measures.

	TPr	*TNr*	*Prec*	*Acc*	*Gm*	*AUC**	F_1	*OP*	$IBA_\alpha(Gm)$	κ	*AGm*	*cwA*	$wAUC_\rho$
p1	+	+	+	–	–	–	+	–	+	–	+	+	+
p2	–	+	–	+	+	+	–	+	+	+	+	+	+
p3	–	+	+	+	+	+	+	+	+	+	+	+	+
p4	+	–	+	+	+	+	+	+	+	+	+	+	+
p5	+	–	–	+	+	+	+	+	+	+	+	+	+

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

The four more sensitive measures (columns) appear to be *IBA*, *AGm*, *cwA* and *wAUC*, which give different values for the five different changes. From **p1**, which represents the inversion of class performances, the results show that *Acc* does not distinguish TP from TN and FN from FP. Besides, one can observe that *Gm*, *AUC*, *OP* and κ, four performance measures used in imbalanced domains, are insensitive to this transformation, so they may not recognize the skew of class rate. Although F_1 is able to detect this transformation, it remains invariant under a change in TN (i.e., **p2**).

4 CONCLUSIONS

In this paper, we reviewed a number of performance metrics typically used to evaluate classifiers on imbalanced data sets. Two theoretical experiments were

designed to analyse the behaviour and sensitivity of these measures in imbalanced problems. The results obtained suggest that *AUC*, *Gm*, *IBA*, *wAUC* and *cwA* are more suitable for dealing with imbalance. However, two of these metrics, *AUC* and *Gm*, do not detect the exchange of positive and negative values in the confusion matrix, so they may not recognize the asymmetry of class results. In the case of F-measure, which has been also proposed as a favorable metric on imbalanced data sets, the previous analysis showed that this measure is highly correlated with the results on the negative class.

ACKNOWLEDGEMENTS

This work has partially been supported by the Spanish Ministry of Education and Science under grants CSD2007–00018, AYA2008–05965–0596 and TIN2009–14205, the Fundació Caixa Castelló-Bancaixa under grant P1–1B2009–04, and the Generalitat Valenciana under grant PROMETEO/2010/028.

REFERENCES

Batuwita, R. and Palade, V. (2009). A new performance measure for class imbalance learning: application to bioinformatics problems. In *the 8th ICMLA'09*, pages 545–550.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Cohen, G., Hilario, M., Sax, H., Hugonnet, S., and Geissbuhler, A. (2006). Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence Medicine*, 37(1):7–18.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ and Psychol Meas*, 20(1):37–46.

Daskalaki, S., Kopanas, I., and Avouris, N. (2006). Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence*, 20(5):381–417.

Fatourechi, M., Ward, R., Mason, S., Huggins, J., Schlogl, A., and Birch, G. (2008). Comparison of evaluation metrics in classification applications with imbalanced datasets. In *the 7th ICMLA'08*, pages 777 –782.

Ferri, C., Hernández-Orallo, J., and Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38.

Folleco, A., Khoshgoftaar, T. M., and Napolitano, A. (2008). Comparison of four performance metrics for evaluating sampling techniques for low quality class-imbalanced data. In *the 7th ICMLA'08*, pages 153–158.

García, V., Mollineda, R. A., and Sánchez, J. S. (2010). Theoretical analysis of a performance measure for imbalanced data. In *the 20th ICPR'2010*, pages 617–620.

Gu, Q., Zhu, L., and Cai, Z. (2009). Evaluation measures of the classification performance of imbalanced data sets. In *the 4th ISICA'09*, pages 461–471. Springer-Verlag.

Huang, J. and Ling, C.-X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans Knowl Data Eng*, 17(3):299–310.

Huang, J. and Ling, C.-X. (2007). Constructing new and better evaluation measures for machine learning. In *the 20th IJCAI'07*, pages 859–864.

Kamal, A., Zhu, X., Pandya, A., Hsu, S., and Shoaib, M. (2009). The impact of gene selection on imbalanced microarray expression data. In *the 1st BICoB'09*, pages 259–269.

Kennedy, K., Mac Namee, B., and Delany, S. (2010). Learning without default: A study of one-class classification and the low-default portfolio problem. In *the AICS'09*, pages 174–187.

Khalilia, M., Chakraborty, S., and Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, 11(1):51.

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

Ranawana, R. and Palade, V. (2006). Optimized Precision - a new measure for classifier performance evaluation. In *the IEEE CEC'09*, pages 2254–2261.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworths, London, UK.

Seliya, N., Khoshgoftaar, T., and Van Hulse, J. (2009). A study on the relationships of classifier performance metrics. In *the 21st ICTAI'09*, pages 59 –66.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In *the AJCAI'06*, pages 1015–1021.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Inf Process & Manag*, 45(4):427–437.

Sun, Y., Wong, A., and Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719.

Tan, P.-N., Kumar, V., and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *the 8th ACM SIGKDD'02*, pages 32–41.

Weng, C. G. and Poon, J. (2008). A new evaluation measure for imbalanced datasets. In *the 7th AusDM'08*, pages 27–32.