Knowledge-Based Systems 59 (2014) 66–74

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A bias correction function for classification performance assessment in two-class imbalanced problems

Vicente García *, Ramón A. Mollineda, J. Salvador Sánchez

Institute of New Imaging Technologies, Department of Computer Languages and Systems, Universitat Jaume I, Av. Sos Baynat, s/n, 12071 Castellón de la Plana, Spain

ARTICLE INFO

Article history:
Received 20 June 2013
Received in revised form 10 December 2013
Accepted 22 January 2014
Available online 3 February 2014

Keywords:
Class imbalance
Performance measure
Classification
Geometric mean
Accuracy
Evaluation

ABSTRACT

This paper introduces a framework that allows to mitigate the impact of class imbalance on most scalar performance measures when used to evaluate the behavior of classifiers. Formally, a correction function is defined with the aim of highlighting those classification results that present moderately higher prediction rates on the minority class. Besides, this function punishes those scenarios that are biased towards the majority class, but also those that are strongly biased to favor the minority class. This strategy assumes a typical imbalance task, in which the minority class contains the most relevant samples to the research purposes. A novel experimental framework is designed to show the advantages of our approach when compared to the standard use of well-established measures, demonstrating its consistency and validity.

1. Introduction

Most of traditional learning methods assume that the classes of the problem share similar prior probabilities and/or misclassification costs. However, in many real-world tasks the ratios of prior probabilities between classes are significantly skewed. This situation is typically 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 regarding the other class (the majority one) [1]. Paradoxically, the minority class is often the most important and usually the one with the highest misclassification costs. Some typical real-life applications where this problem arises are prediction of microarray gene expression [2], prediction of corporate bankruptcy [3] and credit risk [4], fraud detection in mobile telephone communications [5] and text categorization [6]. Because of examples of the minority and majority classes usually represent the presence and absence of rare cases respectively, they are also referred to as positive and negative examples.

As pointed out by many authors [7–10], the use of plain accuracy and/or error rates to evaluate the performance of classifiers 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 non-sensitive to class skews.

A plethora of alternative scalar and graphical methods have been proposed to properly assess classification performance on imbalanced scenarios. Graphical approaches depict trade-offs between two or more evaluation perspectives, allowing a richer analysis of results but making the comparison of learning algorithms a non-trivial issue. Some well-studied examples are the Receiver Operating Characteristic (ROC) curve [11,12], the Precision–Recall (P–R) curve [13], cost curves [14] and the Bayesian Receiver Operating Characteristic (B-ROC) curve [15]. A state-of-the-art of graphical evaluation methods can be found in [16].

Conversely, scalar methods summarize the entire performance information in a single measurement, which makes easier the comparison of different classifiers although it could mask subtle details of their behaviors. Three representative examples are the area under the ROC curve [11], the geometric mean of class accuracies [17] and the *f*-measure [18]. Despite their merits, these metrics present some weaknesses that could lead to incorrect conclusions [19,20]. A common one is that they do not reflect the sign of the bias of a classification result. For instance, some of the most popular measures used in imbalanced binary domains produce identical outcomes when evaluating two opposite classification scenarios (the two class performances are swapped with each other), which generally becomes a critical shortcoming due to the asymmetric misclassification costs for the different classes. Since the bias correction function proposed here fine-tunes numerical performance assessments, this paper focuses on scalar measures.

* Corresponding author. Tel.: +34 964 729072; fax: +34 964 728730.
E-mail addresses: jimenezv@uji.es (V. García), mollined@uji.es (R.A. Mollineda), sanchez@uji.es (J.S. Sánchez).

http://dx.doi.org/10.1016/j.knosys.2014.01.021
0950-7051/

Lately, García et al. [21] introduced a performance assessment method that pursues to correct the aforementioned effects of imbalance. This consists of a weighting factor that rewards outcomes with higher prediction rates on the minority class. However, the use of a fixed weight in each classification task may entail two important limitations: (i) the resulting static solution is unable to adapt to specific classification scenarios and (ii) it rewards indiscriminately any classifier bias that favors the minority class.

The present paper reformulates the performance evaluation method proposed by García et al. [21] with the aim of overcoming its weaknesses. In brief, the new proposal consists of a bias correction function that modulates the operation of any standard scalar measure by means of an adaptive weighting factor. More specifically, the main contributions of this model are:

1. The correction function favors only moderate biased results towards the minority class, but penalizes any strong bias.
2. The weighting factor is now a dynamic parameter specifically computed to fit each particular classification scenario.

Since the "best" evaluation performance measure depends on many factors [22,20], their comparison becomes a complex task and there does not exist a standard methodology. A novel two-level experimental framework is here proposed. First, a correlation study between a number of performance metrics and a series of misclassification costs [23] is carried out. Second, a non-parametric statistical test is addressed to demonstrate that the measures perform significantly different. To the best of our knowledge, no previous work has exploited this methodology to compare performance measures in imbalanced domains.

The correction function has been tested over two simple measures that can be deemed as opposite ends of the performance evaluation spectrum on imbalanced problems. These are the classification accuracy, which appear to be the paradigm of biased behavior, and the geometric mean of individual class rates (one of the simplest way of unbiased measuring a classification result). The underlying hypothesis is that if the correction function properly tunes these two extremes, it will also be able to correct any other midway approach.

2. Performance evaluation

For a two-class problem, the decision made by a classification model over a set of objects can be expressed in the form of a 2×2 confusion matrix where columns represent the predicted class and rows indicate the actual class (see Table 1).

Several straightforward indices can be easily formulated from such a confusion matrix, revealing results on the positive and negative classes separately. Some examples are: (i) *True positive rate* (also referred to as *recall* or *sensitivity*) is the percentage of positive examples that are correctly classified, $TPr = TP/(TP + FN)$; (ii) *True negative rate* (or *specificity*) is the percentage of negative examples that are correctly classified, $TNr = TN/(TN + FP)$; (iii) *False positive rate* is the percentage of negative examples that are misclassified, $FPr = FP/(TN + FP)$; (iv) *False negative rate* is the percentage of positive examples that are misclassified, $FNrate = FN/(TP + FN)$; and (v) *Precision* (or *purity*) is defined as the percentage of samples that are correctly labeled as positive, $Prec = TP/(TP + FP)$.

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)

From these simple indices, it is possible to derive more powerful metrics based on combinations of those error/accuracy rates measured separately on each class. For example, the classification accuracy (Acc) evaluates the effectiveness of the learner by its percentage of correct predictions, $Acc = (TP + TN)/(TP + FN + TN + FP)$. However, empirical and theoretical evidences show that this measure is strongly biased with respect to data imbalance [10,24,25]. This has raised a key question about how to evaluate the performance of classifiers affected by the class imbalance problem. Early attempts have consisted of taking well-established performance evaluation methods from several research domains, such as signal decision theory, image retrieval, and medical decision making. Some representative examples are the *f*-measure ($f_1 = (2 \cdot TPr \cdot Prec)/(Prec + TPr)$ [18]), the geometric mean of the true positive and true negative rates ($Gm = \sqrt{TPr \cdot TNr}$ [26]), and the area under the ROC curve ($AUC = (TPr + TNr)/2$ [27]).[1]

Apart from these widely-known metrics, many other methods have more recently been proposed with the aim of reducing the bias of the accuracy. Ranawana and Palade [28] introduce the optimized precision, which estimates the difference between the accuracy and a relationship index that computes how balanced both class accuracies are. Based on this work, Hossin et al. [29] propose the optimized accuracy with extended recall-precision, which builds the relationship index by using precision and recall. Cohen et al. [30] formulate the mean class-weighted accuracy, which assigns different weights to the true positive and true negative rates in order to compensate for class imbalance. In the same fashion of weighting metrics, Timotius and Miaou [31] introduce the arithmetic means. Weng and Pong [32] propose a method to compute AUC with a cost bias, which gives more weights to the areas close to the top of the ROC graph.

In the credit scoring realm, Kennedy et al. [33] propose a particular adaption of the *f*-measure called harmonic mean, which employs specificity instead of precision. Batuwita and Palade point out that in biomedical imbalanced data sets is more important to increase the true negative rate than the true positive rate [34]. Accordingly, they propose the adjusted geometric mean by combining the geometric mean and the proportion of negative examples of the data set, thus allowing to achieve high true negative rates while keeping low reductions of the true positive rates. Recently, Maratea et al. [35] introduce the adjusted *f*-measure, which can be viewed as a geometric mean of *f*-measure values computed for the two classes by considering different weights.

Although these measures have demonstrated to be suitable for problems with skewed class distributions and unequal classification errors costs [8,12,36–38], most of these performance measures do not distinguish between contributions of individual classes to the overall performance. This means that they do not take into consideration the magnitude and direction of differences between the accuracies measured separately on each class (the classifier bias). The importance of this comes from the different misclassification costs, which are very usual in the context of imbalance. As already said, in many real-life problems, it may be convenient to promote classification results with higher performance on the minority class, provided that the importance of the majority class is not underestimated.

3. An adaptive correction function for performance evaluation

This section proposes a meaningful improvement of the *Index of Balanced Accuracy* (IBA) introduced by García et al. [21]. IBA weights a standard performance measure, in order to compensate

[1] This formula of AUC, which is also called balanced accuracy, is valid when only one run is available.