When Overlapping Unexpectedly Alters the Class Imbalance Effects

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

[1] Lab. Reconocimiento de Patrones, Instituto Tecnológico de Toluca
Av. Tecnologico 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 makes use of several performance metrics to extend the understanding of a challenging imbalanced classification task. More specifically, we refer to a problem in which the minority class is more represented in the overlap region than the majority class, that is, the overall minority class becomes the majority one in this region. The experimental results demonstrate that the use of a set of appropriate performance measures allows to figure out such an atypical case.

1 Introduction

The class imbalance problem has received considerable attention in areas such as Machine Learning and Pattern Recognition. A data set is said to be imbalanced when one of the classes (the minority one) is heavily under-represented in comparison to the other class (the majority one). This issue is particularly important in real-world applications where it is costly to misclassify examples from the minority class, such as the diagnosis of rare diseases and the detection of fraudulent telephone calls, among others. 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.

The research in this topic has been mainly addressed to find solutions for learning from imbalanced data [1,2,4,9]. This constitutes a challenging task because standard discriminant learning tends to bias towards the most represented class [9]. A closely related issue that has also received much attention refers to the evaluation of the classifier performance in these domains [5,6]. The usual method consists of measuring the fraction of test examples correctly (or incorrectly) classified. Numerous investigations have demonstrated that this metric is not the most appropriate in imbalance problems because it may produce good overall performance, but ignoring (and hiding) results on the minority (and usually the most important) class [6,8,10,11].

Alternative measures have been proposed to evaluate classifiers, which are especially useful in the presence of two-class imbalanced data [5,8,9,11]. Their common characteristic is that they are based upon performance indexes over each individual class, being able to find out skewed behavior of classifiers in

J. Martí et al. (Eds.): IbPRIA 2007, Part II, LNCS 4478, pp. 499–506, 2007.

favor of a specific class. Some widely known examples are Receiver Operating Characteristic (ROC) curve, area under the ROC curve, g-mean, sensitivity, specificity, and precision. Apart from being useful for classifier evaluation, these measures could help to characterize the data complexity so as to find out the reasons that affect the classifier behavior.

It is generally accepted that imbalance is the main responsible for a significant degradation of the performance on individual classes, even under the presence of other difficulties, such as overlapping. It seems to be true in cases where the imbalance ratio in the overlap region(s) is similar to the overall imbalance ratio. In these common situations, alternative metrics have been exhaustively analyzed and their values, easily interpreted. Nevertheless, less frequent but possible cases in which the minority class is more represented than the majority class in the overlap region, have not been studied enough. Considering that classification errors come mostly from overlap, how would the performance measures evaluate those atypical scenarios? How can they be explained by these measures?

The ultimate aim of this paper is to answer those questions. For such a purpose, we have designed two classification experiments over two-class synthetic data sets with a fixed overall imbalance ratio in order to make results not dependent on this parameter. The first experiment considers a typical situation in which both the imbalance in the overlap region and the overall imbalance are identical while overlapping changes. This will establish a baseline to analyze the results of the next part. The second experiment operates on data sets where the minority class is locally denser than the majority class in the overlap region. This situation leads to obtain values different from those expected, considering the only a priori knowledge (the overall imbalance). Discussion of new results will focus on the difficulty of figure out such particular data complexity, which is not usually taken into account in general studies.

2 Performance Measures in Class Imbalance Problems

Most of performance measures for two-class problems are built over a 2×2 confusion matrix as illustrated in Table 1. From this, four simple measures can be directly obtained: TP and TN denote the number of positive and negative cases correctly classified, while FP and FN refer to the number of misclassified positive and negative examples, respectively.

The most widely used metrics for measuring the performance of learning systems are the *error rate* and the *accuracy*, which can be computed as $(TP + TN)/(TP+FN+TN+FP)$. Nevertheless, researchers have demonstrated that,

Table 1. Confusion matrix for a two-class problem

	Positive prediction	Negative prediction
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

when the prior class probabilities are very different, these measures are not appropriate because they do not consider misclassification costs, are strongly biased to favor the majority class, and are sensitive to class skews [6,8,10,11].

Thus, several metrics that measure the classification performance on positive and negative classes independently can be derived from Table 1. The *true positive rate*, also referred to as *recall* or *sensitivity*, $TPrate = TP/(TP + FN)$, is the percentage of correctly classified positive examples. The *true negative rate* (or *specificity*), $TNrate = TN/(TN + FP)$, is the percentage of correctly classified negative examples. The *false positive rate*, $FPrate = FP/(FP + TN)$ is the percentage of misclassified positive examples. The *false negative rate*, $FNrate = FN/(TP+FN)$ is the percentage of misclassified negative examples. Finally, the *precision* (or *purity*), $Precision = TP/(TP + FP)$, is defined as the proportion of positive cases that are actually correct.

A way to combine the TP and FP rates is by using the ROC curve. The ROC curve is a two-dimensional graph to visualize, organize and select classifiers based on their performance. It also depicts trade-offs between benefits (true positives) and costs (false positives) [7,11]. In the ROC curve, the TP rate is represented on the Y-axis and the FP rate on the X-axis. To assess the overall performance of a classifier, one can measure the fraction of the total area that falls under the ROC curve (AUC) [8]. AUC varies between 0 and +1. Larger AUC values indicate generally better classifier performance.

Kubat et al. [9] use *the geometric mean* (g-mean) of accuracies measured separately on each class, $g-mean = \sqrt{recall \times specificity}$. This measure relates to a point on the ROC curve and the idea is to maximize the accuracy on each of the two classes while keeping these accuracies balanced. An important property of the g-mean is that it is independent of the distribution of examples between classes. Another property is that it is nonlinear, that is, a change in recall (or specificity) has a different effect on this measure depending on the magnitude of recall (or specificity). An alternative metric that does not take care of the performance on the majority class corresponds to the geometric mean of precision and recall, which is defined as $gpr = \sqrt{precision \times recall}$. Like the g-mean, this measure is higher when both precision and recall are high and balanced.

3 Experimental Results and Discussion

Here, we try to show the utility of several performance measures as a tool to characterize the data complexity in class imbalance domains. To this end, we employ two distinct overlapping scenarios, both using two-dimensional synthetic data sets. Pseudo-random bivariate patterns have been generated following a uniform distribution in a square of length 100. There are 400 negative examples and 100 positive patterns, in all cases keeping the overall majority/minority ratio equal to 4. It should be pointed out that, although only one dimension appears as discriminant, inclusion of two dimensions is with the aim of making easier the interpretation of the results.

From the two scenarios employed in the experiments, the first constitutes a typical class imbalance problem with overlapping, in the sense that imbalance equally affects to the whole representation space. The second experiment refers to a more challenging situation, where the imbalance ratio in the overlap region is inverse to the overall imbalance ratio, that is, the majority and minority classes have interchanged their roles.

We have adopted a 10-fold cross-validation method: each data set was divided into ten equal parts, using nine folds as the training set and the remaining block as an independent test set. This process has been repeated ten times. The experiments consist of computing the performance metrics reported in Sect. 2, when using several classifiers of distinct natures: a nearest neighbor (1-NN) classifier, a multilayer perceptron (MLP), a naïve Bayes (NBS) classifier, a radial basis function (RBF), and a C4.5 decision tree.

3.1 Experiment I: A Typical Class Imbalance Situation

The first experiment has been over a series of six data sets with increasing class overlap. In all cases, the positive examples are defined on the X-axis in the range [50..100], while those belonging to the majority class are generated in [0..50] for 0% of class overlap, [10..60] for 20%, [20..70] for 40%, [30..80] for 60%, [40..90] for 80%, and [50..100] for 100% of overlap. Note that the overall imbalance ratio matches the imbalance ratio corresponding to the overlap region.

In Fig. 1, we have plotted the average values of g-mean, gpr, TN rate, TP rate, precision and AUC obtained by each classifier when varying the overlapping degree. First, we concentrate our analysis on the mid case of 40% of class overlap, supposing that the only a priori knowledge refers to the presence of class imbalance, ignoring the overlapping degree. From the results, it is possible to remark some observations. In particular, while the TN rates for all classifiers (except 1-NN) are 97-100%, the TP rates are close to 60%, relation that can be expected in an imbalance scenario. This, jointly with the fact that there are not relevant differences in the behavior of the distinct classifiers (i.e., the results are independent of the classifiers), suggest that measures are revealing a certain level of overlapping between both classes and more importantly, that the percentage of positive examples in the overlap region has to be approximately equal to the error on the minority class (i.e., $100\% - \text{TP rate} \approx 40\%$). This hints that the TP rate and the TN rate can be viewed as good descriptors of the data complexity.

Indeed, all these comments can be now corroborated by making use of the whole knowledge about the artificial data sets. Thus it is possible to see that in the previous case of study, about 40% of positive examples are inside the overlap region. Even, we can observe that very similar effects appear on the rest of cases. On the other hand, focusing on the geometric means (see Fig. 1(c-d)), one can observe that both decrease as the overlapping degree increases, despite the imbalance ratio does not vary along the different data sets. When analyzing Fig. 1(e), the high values of precision indicate that almost all classifiers produce very few false positives. In a scenario with class imbalance, this should be fully expected because most of the negative examples will result correctly classified

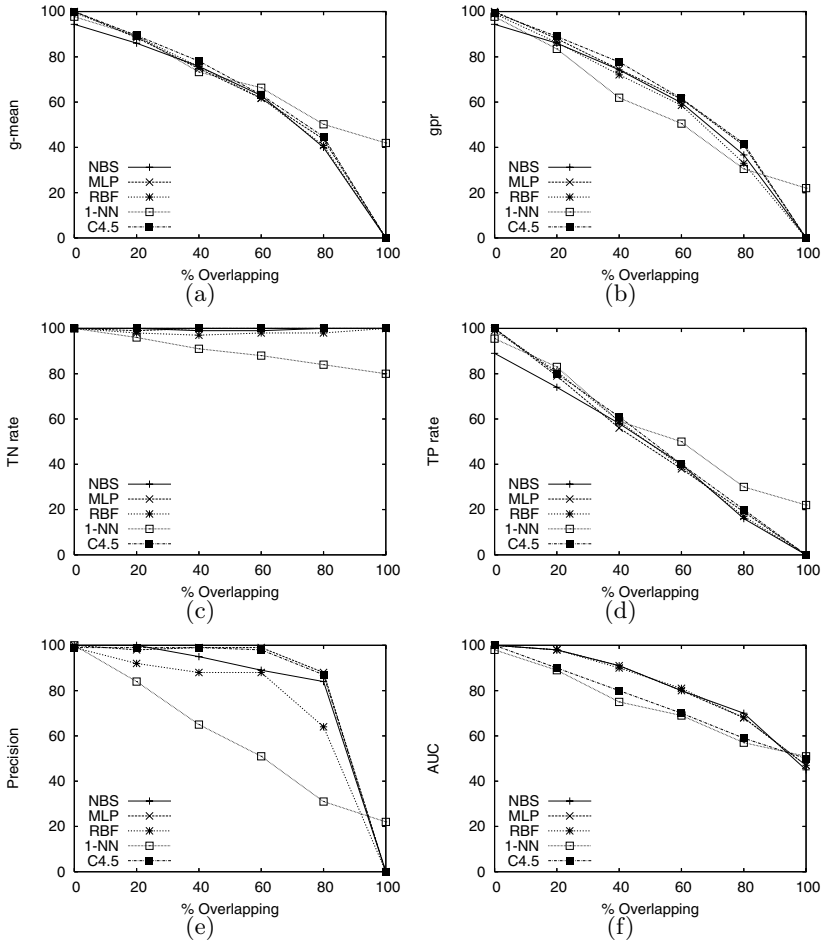

Fig. 1. Classifier performance metrics for Experiment I: (a) g-mean, (b) gpr, (c) TN rate, (d) TP rate, (e) precision, and (f) AUC

(in fact, this can also be observed in the TN rates). The close to 0% of precision in the case of 100% of class overlap means that almost all positive examples have been misclassified, thus corroborating the previous results of the TP rate.

3.2 Experiment II: An Unexpected Practical Case

The second experiment has been carried out over a collection of five artificial data sets in which the number of elements in the overlap region varies in such a way that the overall minority class becomes majority in this region. To this end, the negative examples have been defined on the X-axis to be in the range [0..100] in all data sets, while the positive cases have been generated in the ranges [75..100], [80..100], [85..100], [90..100], and [95..100]. The first means that both

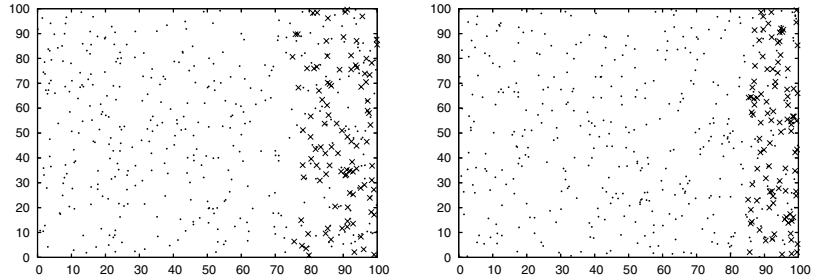

Fig. 2. Two different cases in Experiment II: [75..100] and [85..100]. For this latter case, note that in the overlap region, the majority class is under-represented in comparison to the minority class

classes have the same number of patterns (and density) in the overlap region (i.e, no imbalance in this region). The latter has 100 positive and 20 negative examples in the overlap region, that is, the minority class appears as majority in such a region. Fig. 2 illustrates two examples of these data sets.

Fig. 3 shows the averaged values of g-mean, gpr, TN rate, TP rate, precision and AUC obtained by each classifier for the five different cases described above. As in the previous experiment, we firstly discuss the results for the mid case [85..100], with the aim of finding out some data characteristics by using the performance metrics and the (only) a priori knowledge concerning the presence of a high imbalance ratio in all data sets.

Values of g-mean, gpr, TP rate and TN rate in Fig. 3 indicate significant errors (close to 10-20%) on both classes. Surprisingly, the results reveal that, despite addressing an imbalance problem, for each classifier the TP and TN rates are comparable: the TP rate is 80-100% and the TN rate is 85-90%. On the other hand, the precision, unlike the first experiment, is low enough (about 65%), mainly due to the amount of (unexpected) errors on the majority class. All these observations suggest high overlapping between the two classes. Nevertheless, the fact that the overlapping affects both positive and negative examples, can be deemed as contradictory to our a priori knowledge (the existence of an important class imbalance) since this effect is more likely to be produced in a balanced set. Finally, a deeper analysis of the comparable values of TP and TN rates in absolute terms, considering this strongly imbalance scenario, concludes that there are many more errors on the majority class than on the minority one. Thus, taking all these into account, it can be guessed that the different classifiers have identified the overlapping region as belonging to the minority class. In other words, in such an overlap region there exists a majority of positive examples.

The full knowledge of class distributions confirms again our suspects. Figure 3 shows the performance measures used while the minority class becomes denser along with the decrease of the overlapping region. Despite the full overlapping (of the minority class) and the strong imbalance, all the measures reveal an improvement of classifier performances. This is due to the change of the imbalance ratio

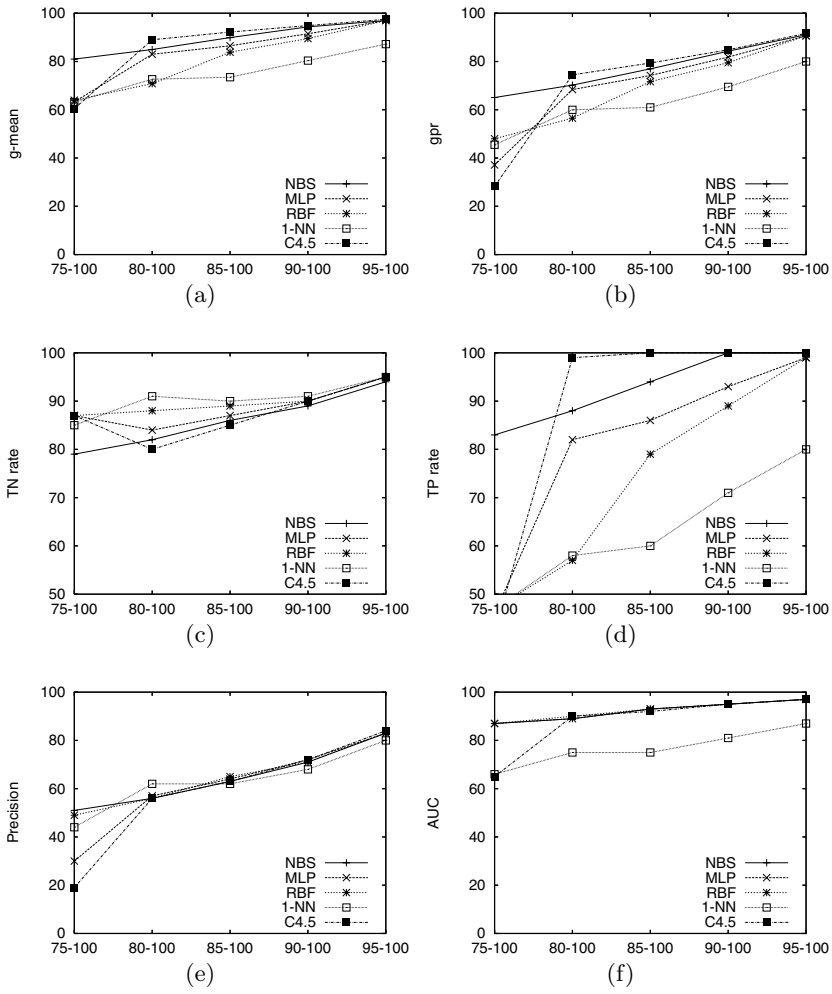

Fig. 3. Classifier performance measures for Experiment II: (a) g-mean, (b) gpr, (c) TN rate, (d) TP rate, (e) Precision, and (f) AUC

in the overlapping region which benefits both classes, that is, the increase of the density of the positive examples and the reduction of the number of (affected) negative examples.

The discussion just exposed should provide some guidelines of how to predict these rare situations through a number of classifier performance measures. In this way, when individual errors in a two-class imbalance domain are significant and similar (for more than one classifier), there likely exists an overlapping region where the minority class is more represented than the majority one.

4 Conclusion

This paper has been motivated by two main issues. First, we believe that performance measures used in imbalance domains can be suitable to characterize data complexity, besides their primary role, i.e., classifier evaluation. Second, when the imbalance ratio in the overlap region is inversely related to the overall imbalance, the classification results may be different from those expected in a typical imbalance scenario. These ideas have been validated by inferring the complexity of a challenging two-class imbalanced data set in which the minority class becomes majority in the overlap region.

After carrying out two experiments, some conclusions can be drawn. In most cases, it will be necessary to employ a set of performance measures so as to get a better understanding of the data characteristics and the classifier behavior. In this sense, the use of diverse classification models allows to find out the degree of influence of the classifiers on the performance results. When most classifiers coincide, the measures can describe the complexity of data distributions.

Acknowledgments. This work has partially been supported by grants DPI2006-15542 from Spanish CICYT and SEP-2003-C02-44225 from Mexican CONACyT.

References

1. Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognition 36, 849–851 (2003)
2. Batista, G.E., Pratti, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations 6, 20–29 (2004)
3. Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms. In: Proc. of 23rd Intl. Conf. on Machine Learning, pp.161–168 (2006)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
5. Chawla, N.V., Japokowicz, N., Kolcz, A.: Special Issue on Learning from Imbalanced Data Sets (Editorial). SIKGDD Explorations 6, 1–6 (2004)
6. Daskalaki, S., Kopanas, I., Avouris, N.: Evaluation of classifiers for an uneven class distribution problem. Applied Artificial Intelligence 20, 381–417 (2006)
7. Fawcett, T.: ROC graphs with instance-varying costs. Pattern Recognition Letters 27, 882–891 (2006)
8. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. on Knowledge and Data Engineering 17, 299–310 (2005)
9. Kubat, M., Matwin, S.: Adressing the curse of imbalanced training sets: one-sided selection. In: Proc. of 14th Intl. Conf. on Machine Learning, pp. 179–186 (1997)
10. Landgrebe, T.C.W., Paclick, P., Duin, R.P.W.: Precision-recall operating characteristic (P-ROC) curves in imprecise environments. In: Proc. of 18th Intl. Conf. on Pattern Recognition, pp. 123–127 (2006)
11. Provost, F., Fawcett, T.: Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In: Proc. of 3rd Intl. Conf. on Knowledge Discovery and Data Mining, pp. 43–48 (1997)