Exploring the Performance of Resampling Strategies for the Class Imbalance Problem

V. García, J.S. Sánchez, and R.A. Mollineda

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,sanchez,mollined}@uji.es

Abstract. The present paper studies the influence of two distinct factors on the performance of some resampling strategies for handling imbalanced data sets. In particular, we focus on the nature of the classifier used, along with the ratio between minority and majority classes. Experiments using eight different classifiers show that the most significant differences are for data sets with low or moderate imbalance: over-sampling clearly appears as better than under-sampling for local classifiers, whereas some under-sampling strategies outperform over-sampling when employing classifiers with global learning.

1 Introduction

Class imbalance constitutes one of the problems that has recently received most attention in research areas such as Machine Learning, Pattern Recognition and Data Mining. A two-class data set is said to be imbalanced if one of the classes (the minority one) is represented by a very small number of instances in comparison to the other (the majority) class [1]. Besides, the minority class is the most important one from the point of view of the learning task. It has been observed that class imbalance may cause a significant deterioration in the performance attainable by standard learners because these are often biased towards the majority class [2]. This issue is particular important in real-world applications where it is costly to misclassify examples of the minority class, such as diagnosis of an infrequent diseases [3], detection of fraudulent telephone calls [4], detection of oil spills in radar images [5], text categorization [6], and credit assessment [7]. 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.

Research on this topic can be categorized into three groups. One has primarily focused on the implementation of solutions for handling the imbalance both at the data and algorithmic levels. Another group has addressed the problem of measuring the classifier performance in imbalanced domains. The third consists of analyzing the relationship between class imbalance and other data complexity characteristics. From these three general topics in class imbalance, data level methods are the most investigated. These methods consist of balancing the original data set, either by over-sampling the minority class [8–11] and/or by under-sampling the majority class [12–14], until the problem classes are approximately equally represented.

Conclusions about what is the best data level solution for the class imbalance problem are divergent. In this sense, Hulse et al. [15] suggest that the utility of the resampling methods depends on a number of factors, including the ratio between positive and negative examples, other characteristics of data, and the nature of the classifier.

In the present work, we study the influence of the imbalance ratio (i.e., ratio between minority and majority classes) and the nature of the classifier used on the effectiveness of some popular resampling techniques for handling the class imbalance problem. To this end, we will carry out experiments over real databases with two different levels of imbalance, employing eight classifiers and four performance measures.

The rest of the paper is organized as follows. Section 2 provides some performance measures especially useful for class imbalance problems. Section 3 reviews several well-known resampling strategies. Next, in Sect. 4 the experimental set-up is described. Section 5 reports the results and discusses the most important findings. Finally, Sect. 6 concludes the present study and outlines possible directions for future research.

2 Performance Evaluation in Class Imbalance Problems

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

Table 1. Confusion matrix for a two-class problem

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

However, empirical evidence shows that most of these commonly used measures are biased towards the majority class. Shortcomings of these evaluators have motivated the search for alternative measures, such as the geometric mean of class accuracies [10] and the area under the ROC curve (AUC) [16].

Given the true positive rate, $TPrate = TP/(TP + FN)$, and the true negative rate, $TNrate = TN/(TN + FP)$, the *geometric mean* of $TPrate$ and $TNrate$ is computed as

$$Gm = \sqrt{TPrate \cdot TNrate} \tag{1}$$

This measure can be seen as a sort of correlation between both rates, because a high value occurs when they both are also high, while a low value is related to at least one low rate.

More recently, Garcia et al. proposed a new measure called *index of balanced accuracy* [17], which is computed as

$$IBA = (1 + 0.1 \cdot (TPrate - TNrate)) \cdot TPrate \cdot TNrate \tag{2}$$

The IBA metric quantifies a certain trade-off between an unbiased measure of overall accuracy and an index of how balanced are the two class accuracies. Unlike most performance 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).

3 Resampling

Data level methods for balancing the classes consists of resampling the original data set, either by over-sampling the minority class or by under-sampling the majority class, until the classes are approximately equally represented. Both strategies can be applied in any learning system, since they act as a preprocessing phase, allowing the learning system to receive the training instances as if they belonged to a well-balanced data set. Thus, any bias of the system towards the majority class due to the different proportion of examples per class would be expected to be suppressed.

However, resampling methods have shown important drawbacks. Under-sampling may throw out potentially useful data, while over-sampling artificially increases the size of the data set and consequently, worsens the computational burden of the learning algorithm.

The simplest method to increase the size of the minority class corresponds to random over-sampling, that is, a non-heuristic method that balances the class distribution through the random replication of positive examples. Nevertheless, since this method replicates existing examples in the minority class, overfitting is more likely to occur. Chawla et al. [18] proposed an over-sampling technique that generates new synthetic minority instances by interpolating between several positive examples that lie close together. This method, called SMOTE, allows the classifier to build larger decision regions that contain nearby instances from the minority class. From the original SMOTE algorithm, several modifications have been proposed in the literature. For example, García et al. [19] developed three alternatives based upon the concept of surrounding neighborhood with the aim of taking into account both proximity and spatial distribution of the instances.

Random under-sampling [2, 20] aims at balancing the data set through the random removal of negative examples. Despite its simplicity, it has empirically been shown to be one of the most effective resampling methods. Unlike the random approach, many other proposals are based on a more intelligent selection of the negative examples to eliminate. For example, Kubat and Matwin [10] proposed the one-sided selection technique, which selectively removes only those negative instances that are redundant or that border the minority class examples (they assume that these bordering cases are noise). The border examples were detected using the concept of Tomek links [21]. On the other hand, Barandela et al. [8] introduced a method that eliminates noisy instances of the majority class by means of Wilson's editing [22], as well as redundant examples through the modified selective subset condensing algorithm [23].

4 Experiments

The experiments here carried out are directed to empirically evaluate several resampling strategies, pursuing to determine the influence of the imbalance ratio and the nature of classifier on the performance of over-sampling and under-sampling.

Table 2. Data sets used in the experiments

Data Set	Positive Examples	Negative Examples	Classes	Majority Class	Source
Breast	81	196	2	1	UCI[1]
Ecoli	35	301	8	1,2,3,5,6,7,8	UCI
German	300	700	2	1	UCI
Glass	17	197	9	1,2,4,5,6,7,8,9	UCI
Haberman	81	225	2	1	UCI
Laryngeal$_2$	53	639	2	1	Library[2]
Letter$_a$	789	19211	26	$2, \ldots, 26$	UCI
Phoneme	1586	3818	2	1	UCI
Optidigts	554	5066	10	1,2,3,4,5,6,7,8,10	UCI
Pendigits	1055	9937	10	1,2,3,4,5,7,8,9,10	UCI
Pima	268	500	2	1	UCI
Satimage	626	5809	7	1,2,3,5,6,7	UCI
Scrapie	531	2582	2	1	Library
Segmentation	330	1980	6	1,2,3,4,6	UCI
Spambase	1813	2788	2	1	UCI
Vehicle	212	634	4	2,3,4	UCI
Yeast	429	1055	10	1,3,4,5,6,7,8,9,10	UCI

[1] UCI Machine Learning Database Repository `http://archive.ics.uci.edu/ml/`
[2] Library `http://www.vision.uji.es/~sanchez/Databases/`

Experiments were conducted as follows:

Data sets: Seventeen real data sets (summary of whom is given in Table 2) were employed in the experiment. All data sets were transformed into two-class problems by keeping one original class and joining the objects of the remaining classes. The fifth column in Table 2 indicates the original classes that have been joined to shape the majority class. For example, in Vehicle database the objects of classes 2, 3, and 4 were combined to form a unique majority class and the original class 1 was left as the minority class.

Partitions: For each database, a 10-fold cross-validation was repeated 5 times.

Resampling strategies: random under-sampling (RUS), one-sided selection (OSS), Wilson's editing over the negative examples (WE$^-$), the combination of this with the modified selective subset condensing over the negative instances (WE$^-$+MSS$^-$), SMOTE, and the Gabriel-graph-based SMOTE (gg-SMOTE) were employed.

Classifiers: the nearest neighbor ($1, 7, 13$-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) were used, all of them taken from the Weka toolkit [24]. In order to run the NBC on the data sets here considered, the numeric attributes were modeled by a normal distribution.

Performance metrics: TPrate, TNrate, IBA and the geometric mean (Gm) were calculated to measure the classification performance.

Classifiers were applied to each original training set and also to sets that were preprocessed by the different resampling strategies. Results obtained in terms of the four performance metrics were evaluated by a paired $t-$test between each pair of methods, for each data set. Based on these values, we computed an index of performance, which is calculated as the difference between *wins* and *losses*, where *wins* is the total number of times that a method A has been significantly better than another method and *losses* is the total number of times that A has been significantly worse than another method. With the aim of summarizing the values of this index of performance, we ranked the position of the resampling algorithms: as there are 7 competing methods, the ranks were from 1 (best) to 7 (worst).

5 Results

Since we are interested in analyzing the performance of the resampling strategies with respect to the degree of imbalance, we computed an imbalance ratio as

$$IR = \frac{n^+}{n^-} \tag{3}$$

where n^+ and n^- denote the number of positive and negative examples in the data set, respectively.

From the imbalance ratio IR, we divided the data sets into two categories. A first group consists of the databases that can be deemed as strongly imbalanced ($IR \leq 0.27$): Ecoli, Glass, Laryngeal$_2$, Letter$_a$, Optdigits, Pendigits, Satimage, Scrapie, and Segmentation. In the second group, we find the databases with a low/moderate imbalance ($IR > 0.27$): Breast, German, Haberman, Phoneme, Pima, Spambase, Vehicle, and Yeast.

5.1 Results on Data Sets with Severe Imbalance

Regarding to the data sets with a severe imbalance, several interesting conclusions can be drawn from performance rankings plotted in Fig. 1. First, one can observe that all the resampling techniques improve the accuracy on the minority class (TPrate), but entailing a certain reduction of the TNrate (see Fig. 1(a) and 1(b)). However, it is worth noting that the deterioration of the TNrate produced by over-sampling is much less significant than that of under-sampling. This is due to the large number of negative examples removed by the under-sampling algorithms.

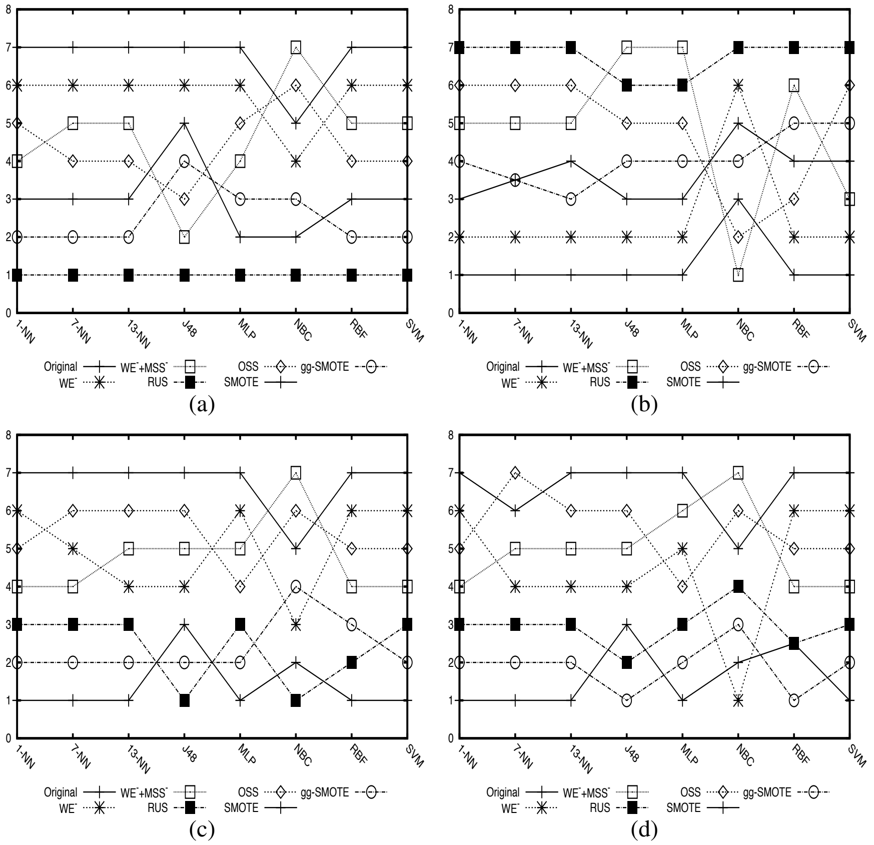


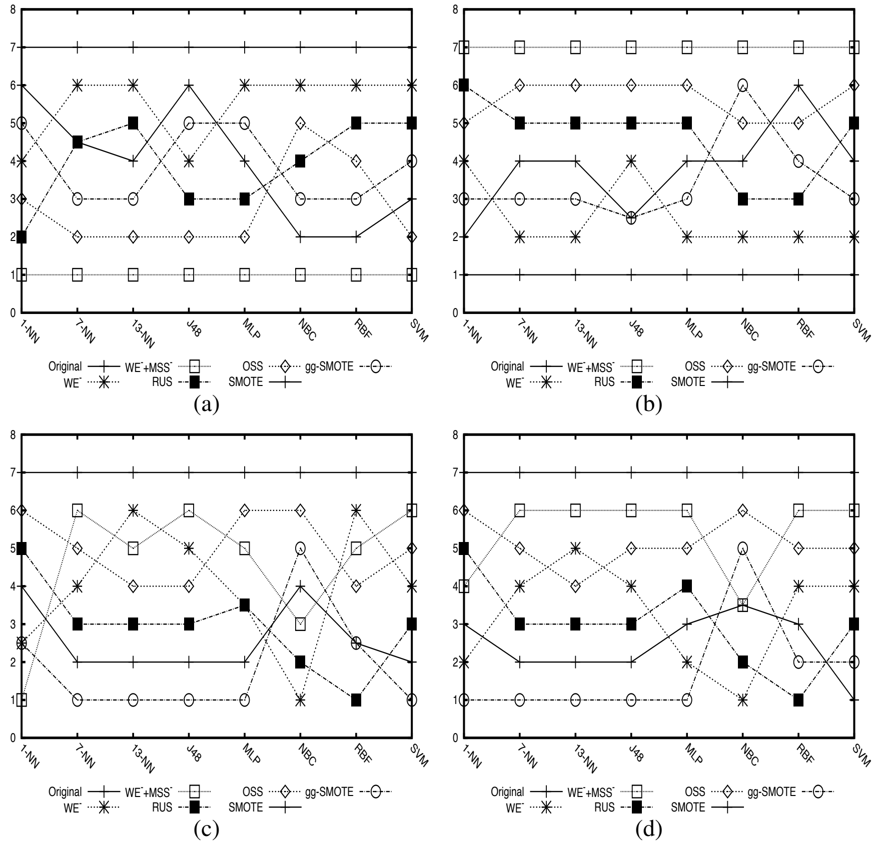
Fig. 1. Rankings for the strongly imbalanced data sets when evaluated with (a) TPrate, (b) TNrate, (c) IBA and, (d) Gm

When employing the global performance metrics, that is, IBA (Fig. 1(c)) and Gm (Fig. 1(d)), the over-sampling techniques are significantly better than the under-sampling algorithms, except for the NBC and J48 classifiers.

It is also interesting to remark that the random under-sampling appears as one of the best strategies in terms of IBA and Gm. Paradoxically, the ”intelligent” under-sampling techniques generally show the worst ranks, independently of the classifier used; only the WE$^-$ approach obtains good performance when applied with the NBC classifier.

5.2 Results on Data Sets with Low/Moderate Imbalance

Fig. 2 shows the performance rankings for the data sets with a low/moderate level of imbalance. Like in the case of the strongly imbalanced databases, all the resampling algorithms increase the TPrate (Fig. 2(a)) whereas produce a certain decrease in TNrate (Fig. 2(b)). Once again, this effect is much more obvious for the under-sampling techniques. In fact, the most evident case corresponds to the WE$^-$+MSS$^-$ strategy, which obtains the best rank for the TPrate and the worst one for the TNrate. A more detailed analysis of the results produced by this technique reveals that the classes have interchanged their respective ”roles”, thus the majority class has now become the minority one and vice versa.



(a)

(b)

(c)

(d)

Fig. 2. Rankings for data sets with low/moderate imbalance when evaluated with (a) TPrate, (b) TNrate, (c) IBA and, (d) Gm

When focused on IBA (Fig. 2(c)) and Gm (Fig. 2(d)) measures, one can observe two different situations depending on the nature of the classifier. For local classifiers such as k-NN, over-sampling is consistently better than under-sampling; in this case, gg-SMOTE stands out from the remaining strategies. Nevertheless, for classifiers with global learning, it seems that random and WE$^-$ under-sampling algorithms obtain good performance.

6 Conclusions and Further Extensions

In this paper, we have studied the effect of the classifier used and the degree of imbalance on the performance of different resampling techniques. The analysis has been based upon a number of experiments over 17 real databases with different levels of imbalance, using 8 distinct classifiers.

Experiments suggest that in fact these two factors have strong influence on the effectiveness of the resampling strategies. More specifically, the most significant differences are for data sets with low or moderate imbalance: over-sampling clearly appears as better than under-sampling for local classifiers, whereas some under-sampling strategies outperform over-sampling when employing classifiers with global learning.

The present work has revealed some interesting research avenues with regards to the resampling strategies for imbalanced data sets, such as: (i) The analysis of the data sets by means of data complexity (or problem difficulty) measures, thus obtaining a better description of data and allowing a more accurate application of specific techniques to tackle the class imbalance problem; (ii) The use of a larger number of resampling strategies to draw more exhaustive, precise conclusions; (iii) The application of several resampling algorithms to real-world imbalanced problems; and (iv) To extend this study to cost-sensitive learning.

Acknowledgment

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

References

1. He, H., Garcia, E.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9) (2009) 1263–1284
2. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis **6**(5) (2002) 429–449
3. Cohen, G., Hilario, M., Sax, H., Hugonnet, S., Geissbuhler, A.: Learning from imbalanced data in surveillance of nosocomial infection. Artificial Intelligence in Medicine **37**(1) (2006) 7–18
4. Fawcett, T., Provost, F.: Adaptive fraud detection. Data Minining and Knowledge Discovery **1**(3) (1997) 291–316
5. Kubat, M., Holte, R., Matwin, S.: Machine learning for the detection of oil spills in satellite radar images. Machine Learning **30**(2-3) (1998) 195–215
6. Tan, S.: Neighbor-weighted K-nearest neighbor for unbalanced text corpus. Expert Systems with Applications **28**(4) (2005) 667–671

7. Huang, Y.M., Hung, C.M., Jiau, H.: Evaluation of neural networks and data mining methods on a credit assesssment task for class imbalance problem. Nonlinear Analysis: Real World Applications **7**(4) (2006) 720–757
8. Barandela, R., Sánchez, J., García, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognition **36**(3) (2003) 849–851
9. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: A new over-sampling method in imbalanced data sets learning. In: Advances in Intelligent Computing, Lecture Notes in Computer Science. Volume 3644., Springer-Verlag (2005) 878–887
10. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. In: Proceedings of The 14th International Conference on Machine Learning. (1997) 179–186
11. Yen, S.J., Lee, Y.S., Lin, C.H., Ying, J.C.: Investigating the effect of sampling methods for imbalanced data distributions. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics. Volume 5. (Oct. 2006) 4163–4168
12. Batista, G., Prati, R., Monard, M.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter **6**(1) (2004) 20–29
13. García, S., Herrera, F.: Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. Evolutionary Computation **17**(3) (2009) 275–306
14. He, G., Han, H., Wang, W.: An over-sampling expert system for learning from imbalanced data sets. In: International Conference on Neural Networks and Brain. Volume 1. (2005) 537–541
15. Hulse, J., Khoshgoftaar, T., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: Proceedings of the 24th International Conference on Machine learning. (2007) 935–942
16. Daskalaki, S., Kopanas, I., Avouris, N.: Evaluation of classifiers for an uneven class distribution problem. Applied Artificial Intelligence **20**(5) (2006) 381–417
17. García, V., Mollineda, R.A., Sánchez, J.S.: Index of balanced accuracy: A performance measure for skewed class distributions. In et al, H.A., ed.: 4th Iberian Conference IbPRIA 2009, Lecture Notes in Computer Science 5524, Póvoa de Varzim, Portugal, Springer (June 2009) 441–448
18. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: Synthetic minority oversampling technique. Journal of Artificial Intelligence Research **16** (2002) 321–357
19. García, V., Sánchez, J.S., Mollineda, R.A.: On the use of surrounding neighbors for synthetic over-sampling of the minority class. In: 8th WSEAS International Conference on Simulation, Modelling and Optimization, Santander (Spain), WSEAS Press (September 2008) 389–394
20. Zhang, J., Mani, I.: kNN approach to unbalanced data distributions: a case study involving information extraction. In: Proceedings Workshop on Learning from Imbalanced Datasets. (2003)
21. Tomek, I.: Two modifications of cnn. IEEE Transactions on Systems, Man and Cybernetics **6**(11) (1976) 769–772
22. Wilson, D.: Asymptotic properties of nearest neighbour rules using edited data. IEEE Transactions on Systems, Man and Cybernetics **2** (1972) 408–421
23. Barandela, R., Ferri, F., Sánchez, J.: Decision boundary preserving prototype selection for nearest neighbor classification. International Journal of Pattern Recognition and Artificial Intelligence **19**(6) (2005) 787–806
24. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005)