Combined Effects of Class Imbalance and Class Overlap on Instance-based Classification

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

[1] Dept. Llenguatges i Sistemes Informàtics, Universitat Jaume I
Av. Sos Baynat s/n, 12071 Castelló de la Plana, Spain
[2] Lab. Reconocimiento de Patrones, Instituto Tecnológico de Toluca
Av. Tecnologico s/n, 52140 Metepec, México

Abstract. In real-world applications, it has been often observed that class imbalance (significant differences in class prior probabilities) may produce an important deterioration of the classifier performance, in particular with patterns belonging to the less represented classes. This effect becomes especially significant on instance-based learning due to the use of some dissimilarity measure. We analyze the effects of class imbalance on the classifier performance and how the overlap has influence on such an effect, as well as on several techniques proposed in the literature to tackle the class imbalance. Besides, we study how these methods affect to the performance on both classes, not only on the minority class as usual.

1 Introduction

The common assumption that the naturally occurring class distribution (i.e., the relative frequency of examples of each class in the data set) is best for learning is now being questioned. This is because of the increasingly common need to limit the size of large data sets and because classifiers built from data sets with high class imbalance perform poorly on minority-class instances.

There is a considerable amount of research on how to build "good" learning algorithms when the class distribution of data in the training set is imbalanced. For simplicity, and consistently with the common practice [1, 3, 5, 10], only two-class problems are here considered. 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 (the majority) class. This issue is particularly important in those applications where it is costly to misclassify minority-class examples. High imbalance occurs in real world domains where the decision system is aimed to detect a rare but important case, such as fraudulent telephone calls [6], diagnosis of an infrequent disease [18], or text categorization [15].

Most of the research addressing this problem can be classified into three categories. One consists of assigning distinct costs to the classification errors for positive and negative examples [4, 7, 13]. The second is to resample the original training set, either by over-sampling the minority class [3, 11] and/or under-sampling the majority class [10] until the classes are approximately equally represented. The third focuses on internally biasing the discrimination-based process

so as to compensate for the class imbalance [1, 6, 13]. Other techniques consist of combining several of these general methods [2, 1, 10].

Although it is often assumed that class imbalance is responsible for significant loss of classifier performance, recent investigations have directed their efforts to question whether class imbalance is directly correlated to the loss of performance or whether the class imbalance is not a problem by itself. For example, some authors have focused on the small disjuncts problem [8, 9, 16], whereas others on the problem of class overlap [12, 14]. These works suggest that there exists a connection between such problems, stating that the loss of performance when learning from unbalanced data is potentiated by other factors.

The aim of the present paper is to analyze the relation between class imbalance and class overlap, and their effects on the classification performance. We are also interested in investigating how these factors affect to both classes, since most of the proposals deal with the patterns of the minority class only, thus overlooking the consequences over the majority class. This study is performed in the framework of the Nearest Neighbor (NN) algorithm, as one of the most significant representatives of the instance-based learning.

2 Algorithms to Handle the Class Imbalance

In the present section, the algorithms used for dealing with the class imbalance are briefly described. Specifically, we focus on three different strategies: downsizing the majority class, over-sampling the minority class and, internally biasing the discrimination-based process.

Within the category of handling the imbalance by means of under-sampling the majority class, the simplest technique randomly selects a number of negative patterns to be further removed from the training set. Nevertheless, since downsizing the majority class can result in throwing away some useful information, this must be done carefully. Accordingly, other schemes employ some filtering and/or condensing algorithms to pick out and eliminate a number of negative examples [1, 10]. The method we adopt in this work consists of iteratively removing noisy and atypical patterns belonging to the majority class [1] by using the well-known Wilson's editing algorithm (WE) [17].

In the case of over-sampling the minority class, one of the most popular techniques refers to the SMOTE algorithm [3]. This consists of taking each positive pattern and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, the neighbors from the k nearest neighbors are randomly chosen. Synthetic samples are generated by taking the difference between the sample under consideration and its nearest neighbor. This difference is multiplied by a random number between 0 and 1, and added to the corresponding feature vector.

On the other hand, for internally biasing the discrimination procedure, we make use of a weighted distance function to be applied in the classification of new patterns [1]. Let $d_E(\cdot)$ be the Euclidean metric, and let Y be a new sample

to classify. Let x_i be a training example belonging to class i, let n_i be the number of examples from class i, let n be the training set size, and let d be the dimensionality of the feature space. Then, the weighted distance measure is defined as:

$$d_W(Y, x_i) = (n_i/n)^{1/d} d_E(Y, x_i) \tag{1}$$

The idea is to compensate for the imbalance in the training set without actually altering the class distribution. Weights are assigned, unlike in the usual weighted k-NN rule, to the respective classes and not to the individual examples. In that way, since the weighting factor is greater for the majority class than for the minority one, the distance to positive examples is reduced much more than the distance to negative examples. This produces a tendency for the new patterns to find their neighbor among the positive examples.

2.1 Classifier Performance for Imbalanced Data Sets

The average predictive accuracy is the standard performance measure in Pattern Recognition and Machine Learning research. However, using this form evaluation metric assumes that the error costs (the cost of a false positive and false negative) are equal, which can been criticized as being unrealistic [6, 10]. It has to be noted that highly unbalanced problems generally have highly non-uniform error costs that favor the minority class (often the class of primary interest). Therefore, classifiers that optimize average accuracy are of questionable value in these cases since they rarely will predict the minority class.

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)

Alternative methods for evaluating the classifier performance are ROC analysis and the geometric mean. For a two-class problem, these can be described using the confusion matrix as plotted in Table 1. In the present work, we are primarily interested in analyzing the classification performance on positive and negative classes independently. From the confusion matrix, these measures can be defined as $a^+ = TP/(TP + FN)$ and $a^- = TN/(TN + FP)$, respectively.

3 Experimental Results on Synthetic Data Sets

In this section, we run a number of experiments on several artificial data sets whose characteristics can be fully controlled, allowing to better interpret the

results. Pseudo-random bivariate patterns have been generated following a uniform distribution in a square of length 100, centered at $(50, 50)$. There are 400 patterns from the majority class and 100 in the minority class. Six different situations of increasing overlap have been considered, always keeping the majority/minority ratio equal to 4. In all cases, positive examples are generated in the range [50..100], while those belonging to the majority class are as follows: in [0..50] for 0% of class overlapping, in [10..60] for 20%, in [20..70] for 40%, in [30..80] for 60%, in [40..90] for 80%, and in [50..100] for 100% of overlap. Fig. 1 illustrates two examples of these data sets.

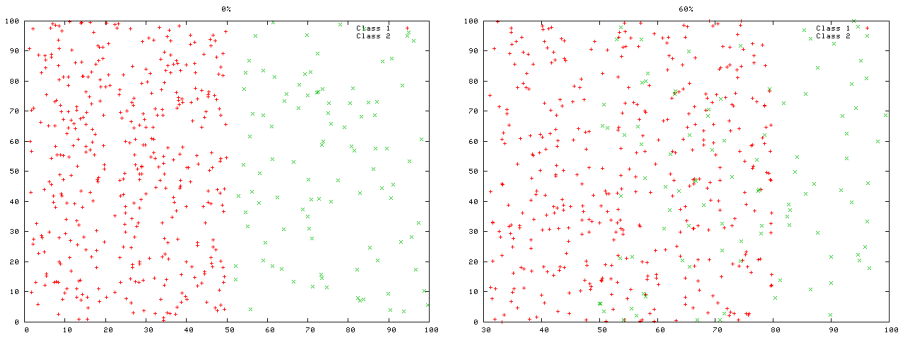
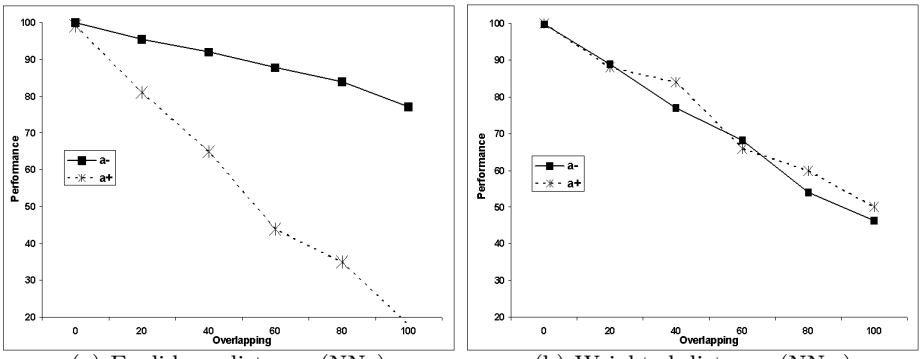


Fig. 1. Two different levels of class overlapping: 0% and 60%

For each data set (degree of overlap), we have studied the techniques described in Sect. 2. In the case of downsizing the majority class, the Wilson's editing algorithm has been applied with both the Euclidean metric (WE) and the weighted distance (WEW). Moreover, WE and WEW have been also applied to both classes (not only to the majority class). Wilson's editing has been always run with $k = 3$. On the other hand, the SMOTE algorithm (with $k = 5$) has been used to over-sampling the minority class. After preprocessing the training set, we have employed the NN rule with the Euclidean (NNe) and the weighted (NNw) distances to classify patterns from an independent test set.

Fig. 2 shows the classification performance on positive (a^+) and negative (a^-) patterns when using the NNe and NNw rules directly, that is, without preprocessing the training set. From this, it is worth remarking several issues. First, when there exists no overlapping (0%), we obtain the same performance on both classes. Second, although the majority/minority ratio keeps constant along the distinct situations, the accuracies degrade as overlap increases. Both of these results suggest that the class imbalance by itself does not strongly affect to the classifier performance. Finally, the use of the weighted distance in the NN rule allows to increase the performance on the minority class but at the same time, the accuracy on the majority class suffers from an important degradation. In

fact, from 40% of overlap, it can be noted that the performance on the minority class becomes even better than that on the majority class.



(a) Euclidean distance (NNe) (b) Weighted distance (NNw)

Fig. 2. Performance evaluation on each class when using the NN rule – synthetic data

Table 2 reports the performances on each class when preprocessing the training set by means of the techniques previously described and then using the NNe or the NNw classifiers. The values corresponding to the original training set (without any preprocessing) are also included as a baseline. The first comment refers to the results over the data set with no overlap (0%), in which all methods obtain similar performance on both classes. As already pointed out, it seems that the class imbalance does not constitute an important difficulty for the learning system under such a "simple" situation. In contrast, as class overlap increases, the effect of the imbalance on the performance becomes clear enough.

When comparing the preprocessing methods, one can observe a different behavior depending on the use of the Euclidean distance (NNe) or the weighted distance (NNw) for the classification of new patterns. In the case of NNe, except the application of Wilson's editing (WE) to both classes, all algorithms improve the performance on the minority class (a^+), but at the cost of reducing the performance on the majority class (a^-). It has to be noted that WEW (editing only the majority class) and SMOTE allow the highest increase in performance on the minority class, independently of the degree of class overlap. Nevertheless, both of these techniques produce an important loss of performance on negative examples. On the other hand, editing both classes does not seem to be an appropriate alternative: it is able to "clean" the class overlap but, at the same time, it clearly increases the imbalance due to the removal of patterns from the majority and the minority classes.

Focusing on the results with NNw, the effect just described is still more evident. In this case, when there exists a very high overlapping, all schemes (except WE on both classes) invert the classifier behavior: the performance on

Table 2. Performance on each class with NNe and NNw (synthetic data sets)

	0%		20%		40%		60%		80%		100%	
	a^-	a^+	a^-	a^+	a^-	a^+	a^-	a^+	a^-	a^+	a^-	a^+
	NNe											
Original	100	99.0	95.5	81.0	92.0	65.0	87.8	44.0	84.0	35.0	77.3	18.0
WE	-	-	93.8	82.0	89.5	67.0	85.5	52.0	79.0	37.0	68.3	23.0
WE both classes	100	96.0	98.5	80.0	99.5	54.0	97.3	40.0	98.8	17.0	97.8	1.0
WEW	100	100	90.3	85.0	82.8	74.0	71.8	61.0	63.0	58.0	53.0	48.0
WEW both classes	100	99.0	93.0	83.0	91.3	69.0	83.0	53.0	74.3	42.0	69.3	34.0
SMOTE	100	99.0	94.3	86.0	84.8	75.0	78.8	58.0	69.8	47.0	62.3	34.0
	NNw											
Original	99.8	100	89.0	88.0	77.0	84.0	68.3	66.0	54.0	60.0	46.3	50.0
WE	-	-	88.6	87.0	76.8	85.0	68.5	65.0	53.8	60.0	46.5	54.0
WE both classes	100	99.0	94.3	82.0	96.0	63.0	89.5	49.0	83.0	29.0	61.5	32.0
WEW	99.3	100	85.3	92.0	70.3	88.0	60.5	78.0	48.3	72.0	37.8	59.0
WEW both classes	99.5	100	89.0	91.0	79.8	79.0	67.0	69.0	56.0	60.0	50.0	51.0

the minority class becomes even better than the performance on the majority class. As a preliminary conclusion of these experiments, one can see that most techniques to handle the class imbalance are able to improve the performance on the positive examples, although producing some decrease in performance on the negative patterns.

4 Experimental Results on Real Data Sets

We here experimented with four real data sets taken from the UCI Machine Learning Database Repository (`http://www.ics.uci.edu/~mlearn`), to validate the results obtained over the synthetic databases. All data sets were transformed into two-class problems to facilitate comparison with other published results [1]. Five-fold cross validation was employed.

Table 3 reports, for each real data set, the performance on both classes when preprocessing the original training sets by means of different techniques and classifying with NNe and NNw. Analyzing the results with NNe over the original training sets (without preprocessing), one can observe that in the Phoneme database both classes show similar and quite low performances, thus suggesting that there probably exists high overlap and low class imbalance (the majority/minority ratio is 2.41). Conversely, in the case of the Satimage database, the performance on the majority class is high enough and clearly better than the performance on the minority class, which indicates a very considerable class imbalance (the majority/minority ratio is 9.28).

It has to be remarked that in all data sets, the use of the weighted distance for classification (NNw) produces an important improvement in the performance on the minority class, and it does not lead to a significant loss of performance on the majority class. On the other hand, when the training set has been preprocessed,

Table 3. Performance on each class with NNe and NNw (real data sets)

	Phoneme		Satimage		Glass		Vehicle	
	a^-	a^+	a^-	a^+	a^-	a^+	a^-	a^+
	NNe							
Original	79.1	68.8	91.7	54.8	98.9	76.0	82.8	37.6
WE	74.5	75.3	90.4	58.3	97.7	76.0	78.2	50.5
WE both classes	79.5	68.6	91.7	48.1	99.4	72.0	89.7	25.2
WEW	73.1	77.0	89.2	60.6	97.7	76.0	76.8	54.8
WEW both classes	77.0	73.1	90.1	54.6	99.4	72.0	87.0	31.4
SMOTE	76.0	71.2	86.2	69.1	98.3	80.0	80.5	44.3
	NNw							
Original	71.1	82.3	88.4	65.1	97.1	80.0	78.5	45.2
WE	68.1	84.2	87.5	66.2	96.6	80.0	75.0	56.2
WE both classes	74.1	79.3	89.4	54.1	98.3	76.0	85.7	31.0
WEW	66.9	84.8	86.7	67.5	95.6	80.0	73.8	60.0
WEW both classes	71.8	81.2	88.0	59.2	98.3	80.0	81.7	38.6

it seems that the best results are achieved with the editing schemes applied only to the negative examples, that is, WE and WEW. In fact, as already observed in the artificial data sets, the schemes based on downsizing both classes are not able to appropriately balance the data sets.

5 Conclusions and Future Work

The class imbalance by itself does not seem to constitute a crucial problem for instance-based classification. In fact, in the presence of imbalance with 0% of overlap, the NN classifier provides high performance on both classes. In contrast, the combination of class imbalance and class overlapping suppose an important deterioration of the performance. These results suggest that the imbalance in the overlap region has a strong influence on the classification performance.

The experiments carried out suggest that the application of some under-sampling technique to both classes leads to poor performance on the minority class. Conversely, the use of editing combined with NNw classification makes the performance on the majority class to become worse than that on the minority class. This is mainly due to the fact that in the overlap region, after editing the negative examples, the minority class is more represented than the majority class. On the other hand, although SMOTE has been recognized as one of the best techniques to handle the imbalance problem, the experiments demonstrate that in the presence of high overlap it is not better than editing, since the generation of synthetic patterns involves an increase of noise in the data set.

Future work is primarily addressed to characterize 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 and the class overlap situations.

Acknowledgments

This work has been partially supported by grants TIC2003-08496 from the Spanish CICYT and SEP-2003-C02-44225 from the 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. 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.
4. Domingos, P.: Metacost: a general method for making classifiers cost-sensitive. In: *Proc. 5th Intl. Conf. on Knowledge Discovery and Data Mining*, 155–164, 1999.
5. Eavis, T., Japkowicz, N.: A recognition-based alternative to discrimination-based multi-layer perceptrons, In: *Proc. Workshop on Learning from Imbalanced Data Sets*, Technical Report WS-00-05, 2000.
6. Fawcett, T., Provost, F.: Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1:291–316, 1996.
7. Gordon, D.F., Perlis, D.: Explicitly biased generalization. *Computational Intelligence*, 5:67–81, 1989.
8. Japkowicz, N.: Class imbalance: are we focusing on the right issue?. In: *Proc. Intl. Workshop on Learning from Imbalanced Data Sets II*, 2003.
9. Jo, T. Japkowicz, N.: Class imbalances versus small disjuncts. *SIGKDD Explorations*, 6:40–49, 2004.
10. Kubat, M., Matwin, S.: Adressing the curse of imbalanced training sets: one-sided selection. In: *Proc. 14th Intl. Conf. on Machine Learning*, 179–186, 1997.
11. Ling, C.X., Li, C.: Data mining for direct marketing: problems and solutions. In: *Proc. 4th Intl. Conf. on Knowledge Discovery and Data Mining*, 73–79, 1998.
12. Orriols, A., Bernardó, E.: The class imbalance problem in learning classifier systems: a preliminary study. In: *Proc. Conf. on Genetic and Evolutionary Computation*, 74–78, 2005.
13. Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., Brunk, C.: Reducing misclassification costs. In: *Proc. 11th Intl. Conf. on Machine Learning*, 217–225, 1994.
14. Prati, R.C., Batista, G.E., Monard, M.C.: Class imbalance versus class overlapping: an analysis of a learning system behavior. In: *Proc. 3rd Mexican Intl. Conference on Artificial Intelligence*, 312–321, 2004.
15. Tan, S.: Neighbor-weighted K-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications*, 28:667–671, 2005.
16. Weiss, G.M.: *The Effect of Small Disjuncts and Class Distribution on Decision Tree Learning*. PhD thesis, Rutgers University (2003)
17. Wilson, D.L.: Asymptotic properties of nearest neighbour rules using edited data. *IEEE Trans.on Systems, Man and Cybernetics*, 2:408–421, 1972.
18. Woods, K., Doss, C., Bowyer, K.W., Solka, J., Priebe, C., Kegelmeyer, W.P.: Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, 7:1417–1436, 1993.