The Imbalanced Training Sample Problem: Under or over Sampling?

Ricardo Barandela[1,2], Rosa M. Valdovinos[1],
J. Salvador Sánchez[3], and Francesc J. Ferri[4]

[1] Instituto Tecnológico de Toluca, Ave. Tecnológico s/n, 52140 Metepec, México
{rbarandela,li_rmvr}@hotmail.com
[2] Instituto de Geografía Tropical, La Habana, Cuba
[3] Dept. Llenguatges i Sistemes Informàtics, U. Jaume I, 12071 Castelló, Spain
sanchez@uji.es
[4] Dept. d'Informàtica, U. Valencia, 46100 Burjassot (Valencia), Spain
ferri@uv.es

Abstract. The problem of imbalanced training sets in supervised pattern recognition methods is receiving growing attention. Imbalanced training sample means that one class is represented by a large number of examples while the other is represented by only a few. It has been observed that this situation, which arises in several practical domains, may produce an important deterioration of the classification accuracy, in particular with patterns belonging to the less represented classes. In this paper we present a study concerning the relative merits of several re-sizing techniques for handling the imbalance issue. We assess also the convenience of combining some of these techniques.

1 Introduction

Design of supervised pattern recognition methods is usually based on a training sample (TS): a collection of examples previously analyzed by a human expert. There is a considerable amount of recent research on how to build "good" classifiers when the class distribution of the data in the TS is imbalanced. A TS 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. For simplicity, and consistently with the common practice [8,13], only two-class problems are here considered. 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 [10], oil spills in satellite images of the sea surface [14], an infrequent disease [20], or text categorization [15].

Basic methods for reducing class imbalance in the TS can be sorted in 3 groups [12]:

a) Over-sampling (replicates examples in) the minority-class
b) Under-sampling (eliminates examples in) the majority class
c) Internally biasing the discrimination based process so as to compensate for the class imbalance [8,14]

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 806–814, 2004.
© Springer-Verlag Berlin Heidelberg 2004

As pointed out by many authors, overall accuracy is not the best criterion to assess the classifier's performance in imbalanced domains. For instance, in the thyroid data set used in [1], only 5% of the patterns belong to the minority class. In such a situation, labeling all new patterns as members of the majority class would give an accuracy of 95%. Obviously, this kind of system would be useless. Consequently, other criteria have been proposed. One of the most widely accepted criterion is the geometric mean, $g = (a^+ \cdot a^-)^{1/2}$, where a^+ is the accuracy on cases from the minority class and a^- is the accuracy on cases from the majority one [13]. This measure tries to maximize the accuracy on each of the two classes while keeping these accuracies balanced.

In previous studies [3-5], we have provided results of several techniques addressing the class imbalance problem. We have focused on under-sampling the majority class and also on internally biasing the discrimination process, as well as on combinations of both approaches. In the present paper, we present an experimental comparison of our results with those obtained with one method for over-sampling the minority class [6]. Our purpose is to illustrate the relative benefits of both basic techniques and to draw some conclusions about those situations in which one of them could be more useful than the other. We also present experimental results obtained with a combination of both resizing approaches. The experiments have been done with five real datasets using the Nearest Neighbor (NN) rule for classification and the geometric mean as the performance measure.

The NN rule is one of the oldest and better-known algorithms for performing supervised nonparametric classification. The entire TS is stored in the computer memory. To classify a new pattern, its distance to each one of the stored training patterns is computed. The new pattern is then assigned to the class represented by its nearest neighboring training pattern. Performance of NN rule, as with any nonparametric method, is extremely sensitive to incorrectness or imperfections in the TS. Nevertheless, the NN rule is very popular because of: a) conceptual simplicity, b) easy implementation, c) known error rate bounds, and d) potentiality to compete favorably in accuracy with other classification methods in real data applications.

2 Related Works

The two basic methods for resizing the TS cause the class distribution to become more balanced. Nevertheless, both strategies have shown important drawbacks. Under sampling may throw out potentially useful data, while over sampling increases the TS size and hence the time to train a classifier. In the last years, research has focused on improving these basic methods. Kubat and Matwin [13] proposed an under sampling technique that is aimed at removing those majority prototypes that are "redundant" or that "border" the minority instances. They assume that these bordering cases are noisy examples. However, they do not use any of the well-known techniques for cleaning the TS.

Chawla et al. [6] proposed a technique for over sampling the minority class and, instead of merely replicating prototypes of the minority class, they form new minority instances by interpolating between several minority examples that lie close together.

Pazzani et al. [16] take a slightly different approach when learning from an imbalanced TS by assigning different weights to prototypes of the different classes. On the

other hand, Ezawa et al. [9] bias the classifier in favour of certain attribute relationships. Kubat et al. [14] use some counter-examples to bias the recognition process.

In an earlier study [3], we provided preliminary results of several techniques addressing the class imbalance problem. In that work, we focused on under sampling the majority class by using several editing and pruning techniques, conveniently adapted to the imbalance case. We proposed also a mechanism for internally biasing the discrimination-based process, and we evaluated the combination of this biasing mechanism with some under sampling methods. In [4], we have extended this idea with a modification of the Wilson's Editing [19] technique. This modification, that biases the editing procedure, allows a better and higher decrease in the number of prototypes of the majority class. We have also explored [5] the convenience of designing a multiple classification system for working in imbalanced situations. Instead of using a single classifier, an ensemble has been implemented. The idea is to train each one of the individual components of the ensemble with a balanced TS. In order to achieve this, as many training sub-samples as required to get balanced subsets are generated. The number of sub-samples is determined by the difference between the amount of prototypes from the majority class and that of the minority class.

3 Techniques to Be Evaluated

The main purpose of the present paper is to experimentally compare several techniques for handling the imbalance situation. Some of these techniques, corresponding to the under sampling and biasing approaches, have already shown important increases in the g value obtained in classification tasks. The experiments to be reported below, include now an over sampling method. All these techniques are explained hereafter.

3.1 Under Sampling Approach

As already explained in Section 2, we have experimented with several methods [3] aimed at reducing the size of the majority class. Out of concern for the possibility of eliminating useful information, we have employed well-known editing algorithms, in particular the already classical Wilson's proposal [19]. One of the contributions of [3] has been the application of this editing technique only to the majority class.

Wilson's Editing. Wilson's Editing corresponds to the first proposal to edit the NN rule. In a few words, it consists of applying the k-NN classifier to estimate the class label of all prototypes in the TS and discard those samples whose class label does not agree with the class associated with the largest number of the k neighbors.

Weighted Editing. Despite the important obtained results, it was observed in [3] that the editing technique did not produce significant reductions in the size of the majority class. Accordingly, the imbalance in the training sample is not diminished in an important way.

It is worthy to remember that Wilson's technique consists essentially in a sort of classification system. The corresponding procedure works by applying the k-NN clas-

sifier to estimate the class label of all prototypes in the TS, as explained above. Of course, this *k*-NN classifier is also affected by the imbalance issue. When applied to prototypes of the majority class, the imbalance in the TS will cause a tendency to find most of their *k* nearest neighbors into that majority class. Consequently, only a few of the majority class prototypes will be removed. This means that the majority class is not completely cleaned of atypical cases and also that the balance in the TS is far from being reached.

To cope with this difficulty, in [4] we introduced the employment of the weighted distance below mentioned, not only in the classification phase but also in editing the majority class in the TS. That is, we apply the Editing algorithm, but using the weighted distance instead of the Euclidean metric. In that way, the already explained tendency has been overturned.

A Pruning Technique: The Modified Selective Subset. The NN rule generalizes accurately for many real applications. However, since it must store all the available training patterns and search through all of them to identify a new pattern, it has large memory requirements and works slowly in the classification phase. Many proposals have been done to reduce the TS size, while trying to maintain accuracy rate in the classification phase of the NN rule. Hart's [11] idea of a *consistent* subset has become a milestone in this research line. But his algorithm to obtain this consistent subset suffers for several well-known drawbacks. That has stimulated a sequel of new algorithms attempting to remedy these faults. Particularly remarkable is the approach of Ritter et al. [17] with a clear and precise formulation of the desired goals and of the way to reach them (the Selective Subset).

According to Hart's statement, the Condensed Subset (CS) is a subset *S* of the TS such that every member of TS is closer to a member of *S* of the same class than to a member of *S* of a different class. Ritter et al. have changed this concept in their Selective Subset (SS) by defining it as that subset *S* such that every member of TS must be closer to a member of *S* of the same class than to a member of TS (instead of *S*) of a different class. Their purpose is to eliminate the order-dependence of the building algorithm. Instead of using a greedy algorithm, Ritter et al. use a kind of branch and bound algorithm that implicitly considers every solution. In fact, they define the SS as the smallest subset containing at least a *related* prototype for each of the original ones. In this context, related means that it is able to correctly classify the corresponding prototype. As Ritter et al. have recognized, their algorithm does not necessarily conduct to a unique solution. Moreover, although they stated the importance of selecting "samples near the decision boundaries", this requisite is not included in the criteria serving as a basis for their SS.

As obtaining a more accurate decision boundary is more important that achieving true minimality, the SS procedure has been modified in two main ways. First, the minimality criterion has been partially substituted by an explicit boundary proximity criterion. And second, the procedure has been converted into a greedy algorithm that ends scanning the TS only twice. This Modified Selective Subset (MSS [2]) turns out to be much simpler and usually obtains subsets with improved quality boundaries and with slightly larger sizes than the corresponding SS solutions.

In [3], we have discussed the usefulness of the MSS technique for handling the imbalanced situation. Here, this pruning algorithm in included only for reducing the TS size after it has been considerable increased by the over sampling method.

3.2 Biasing Mechanism

For internally biasing the discrimination procedure, we proposed in [3] a weighted distance function to be used in the classification phase. Let $d_E(\cdot)$ be the Euclidean metric, and let Y be a new pattern to be classified. Let x_0 be a training prototype from class i, let N_i be the number of prototypes from class i, let N be the TS size, and let m be the dimensionality of the feature space. Then, the weighted distance measure is defined as:

$$d_W(Y,x_0) = (N_i/N)^{1/m} \cdot d_E(Y,x_0) \quad (1)$$

The basic idea behind this weighted distance is to compensate for the imbalance in the TS without actually altering the class distribution. Thus, weights are assigned, unlike in the usual weighted k-NN rule [7], to the respective classes and not to the individual prototypes. In such a way, since the weighting factor is greater for the majority class than for the minority one, the distance to positive minority class prototypes becomes much lower than the distance to prototypes of the majority class. This produces a tendency for the new patterns to find their nearest neighbor among the prototypes of the minority class.

3.3 Over Sampling Approach

Most of the proposed techniques for increasing the size of the minority class merely replicate some of the minority class prototypes. Inclusion of exact copies of some minority class examples means to raise the requirement in computational resources. Moreover, with this procedure, overfitting is likely to occur, particularly in some learning models like the decision trees [17]. To avoid the overfitting problem, Chawla et al. [6] form new minority class prototypes by interpolating between minority class prototypes that lie close together. The technique these authors proposed, takes each minority class prototype and introduces "synthetic" prototypes along the line joining any/all of the minority class nearest neighbors. Depending upon the amount of over sampling required, neighbors from the k nearest neighbors are randomly chosen. In the experiments they reported, k is set to five. When, for instance, the amount of over sampling needed is 200%, only two neighbors from the five nearest neighbors are chosen and one prototype is generated in the direction of each of these two neighbors. Synthetic prototypes are generated in the following way: take the difference between the feature vector (prototype) under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration.

4 Experimental Results

All these techniques, as well as combinations of some of them, were assessed with experiments that were carried out with five datasets. Four of these datasets have been taken from the UCI Database Repository (http://www.ics.uci.edu/~mlearn/). The Mammography dataset was kindly provided by N. V. Chawla and it was reported in

[6] and in [20]. Five-fold cross validation was employed to obtain averaged results of the g criterion. To facilitate comparison with other published results, in the Glass dataset the problem was transformed for discriminate class 7 against all the other classes and in the Vehicle dataset the task was to classify class 1 against all the others. Satimage dataset was also mapped to configure a two-class problem: the training patterns of classes 1, 2, 3, 5 and 6 were joined to form a unique class and the original class 4 was left as the minority one. Phoneme and Mammography are two-class datasets.

Table 1. Mean values of the geometric mean.

Training sets	Phoneme	Satimage	Glass	Vehicle	Mammography
Original TS Euclidean Classif.	73.8	70.9	86.7	55.8	60.2
Original TS Weighted. Classif.	76.0	75.9	88.2	59.6	75.8
Under-sampling majority class					
Euclidean Editing & Classif.	74.9	73.0	86.2	64.0	63.9
Euclid. Edit.+Weighted Classif.	75.7	76.2	87.9	65.8	76.2
Weighted+Edit.+Euclid. Classif.	75.0	74.5	86.2	65.6	70.0
Weighted Editing & Classif.	75.3	77.8	87.9	67.2	78.7
Over-sampling minority class and processing both classes					
Synthetic prototypes	73.6	77.1	88.7	59.7	83.4
Synthetic &Wilson's Editing	74.9	78.5	86.4	64.5	86.8
Synthetic & Modif. Select. Subset	70.3	74.1	88.2	57.1	80.8
Synthetic & Wilson & MSS	74.8	76.2	85.9	62.7	86.0

The obtained experimental results are shown in Table 1. This table has three parts. In the first one, the results when employing the original TS, both with Euclidean and Weighted distance, are included for comparison purposes. In the second part, we present the geometric mean values observed when the TS was under sampled through Wilson's Editing and Weighted Editing. Here also, the classification was done twice with each edited TS, using the Euclidean and the Weighted distances. In the third part of the table, results of the over sampling technique are incorporated. In this case, no weighted distance for classification has been employed since balance in the TSs has been attained by the over sampling technique.

Fom the figures in Table 1, it is evident that the over sampling approach can not compete, in most of the datasets, with the combination of the Weighted Editing (for under sampling) and the Weighted classification (the biasing mechanism). The difference in the Glass dataset (88.7 vs. 87.9) was not statistically significant. The only exception is the Mammography dataset, where results obtained after over sampling excelled to those of all the other evaluated techniques.

The explanation for these, somehow contradictory, results is to be found in the amount of imbalance present in each dataset (see Table 2). When the imbalance in the TS is not very big (say, a majority/minority ratio less than 10), then the under sampling techniques, particularly the Weighted Editing, can be useful in reducing enough the imbalance as to produce an important enhancement in the performance of the classifier. However, when this ratio is greater, the degree of balance achieved is not satisfactory. With the employed under sampling techniques, we are very careful in not throwing away potentially useful information. Accordingly, not many majority class prototypes are removed. With greater ratios, it is much better to employ the over sampling technique, even at the cost of a considerable increase in the total TS size.

Table 2. Imbalance present in each training dataset (majority/minority ratio).

Training sets	Phoneme	Satimage	Glass	Vehicle	Mammography
Original TS	2.41	9.29	6.25	2.99	44.12
After under-sampling majority class					
Euclidean Editing	2.27	8.94	6.13	2.44	43.99
Weighted Editing	2.15	8.64	6.02	2.31	43.03
Over-sampling minority class and processing both classes					
Synthetic prototypes	1.20	1.03	1.04	1.49	1.01
Synthetic &Wilson's Editing	1.18	0.94	1.03	1.31	1.04
Synthetic & Modif. Select. Subset	1.01	1.49	1.42	1.42	2.12
Synthetic & Wilson & MSS	0.93	1.35	0.94	1.24	1.00

Table 3. Size of the TSs (Original and after application of the under and over sampling).

Training sets	Phoneme	Satimage	Glass	Vehicle	Mammography
Original TS	4322	5147	174	678	10062
After under-sampling majority class					
Euclidean Editing	4150.8	4971.6	171.2	584.8	10032.9
Weighted Editing	3997.8	4820.6	168.6	562.0	9818.5
Over-sampling minority class and processing both classes					
Synthetic prototypes	5590	9147	294	848	19572
Synthetic &Wilson's Editing	5185.2	8706.0	285.0	686.8	17725.3
Synthetic & Modif. Select. Subset	1201.2	1322.8	27.6	321.2	6931.9
Synthetic & Wilson & MSS	756.0	906.4	18.6	176.2	1599.5

This concern for the huge increase in the TS size produced by over sampling (almost twice the number of original prototypes), has been the motivation for exploring the convenience of applying preprocessing techniques after the formation of new minority class prototypes (see Table 3). As usual, the combined employment of Wilson's Editing and the pruning technique, MSS, has yielded a considerable decrease in the TS size and, in general, a classification performance better than before their application. Thus, another recommendation: in those cases where over sampling the TS is a must, it is convenient, afterwards, to try to clean the TS and to reduce its size.

5 Concluding Remarks

In many real-world applications, supervised pattern recognition methods have to cope with highly imbalanced TSs. Traditional learning systems such as the NN rule can be misled when applied to such practical problems. This effect can become softer by using procedures to resize (under sampling or over sampling) the TS. In the present paper we have assessed the relative merits of these two approaches for re-sampling the TS. Our results indicate that, when the imbalance is not very severe, techniques for appropriately under sampling the majority class are the best option. Only when the majority/minority ratio is very high it is required to over sampling the minority class. Convenience of using combinations of some techniques is also established. In particular, this combination is remarkable in those cases where over sampling is unavoidable. In these situations, cleaning of the TS and reduction of its size, after the over

sampling is done, allows for a considerable decrease in the computational burden of the NN rule and for an increase in the classification performance of the system.

The present report is part of a more extensive research we are conducting to explore all the issues linked to the imbalanced TSs. At present, we are studying the convenience of applying genetic algorithms to reach a better balance among classes. We are also experimenting in situations with more than two classes, as well as doing some research about the convenience of using these procedures to obtain a better performance with other classifiers, such as the neural networks models.

Acknowledgements

This work has been partially supported by grants 32016-A from the Mexican CONACYT, 644.03-P from the Mexican Cosnet, and TIC2003-8496-C04 from the Spanish CICYT.

References

1. Aha, D., Kibler, D.: Learning Representative Exemplars of Concepts: An Initial Case Study, Proceedings of the Fourth International Conference on Machine Learning (1987) 24-30.
2. Barandela, R., Cortés, N., Palacios, A.: The Nearest Neighbor rule and the reduction of the training sample size, *Proc. 9th Spanish Symposium on Pattern Recognition and Image Analysis* **1** (2001) 103-108.
3. Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for learning in class imbalance problems, *Pattern Recognition* **36(3)** (2003) 849-851.
4. Barandela, R., Sánchez, J.S., García, V., Ferri, F.J.: Learning from Imbalanced sets through resampling and weighting, *Lecture Notes in Computer Science* **2652** (2003) 80-88.
5. Barandela, R., Valdovinos, R. M., Sánchez, J.S.: New applications of ensembles of classifiers, *Pattern Analysis and Applications* **6(3)** (2003) 245-256.
6. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research* **16** (2000) 321-357.
7. Dudani, S.A.: The distance-weighted *k*-nearest neighbor rule, *IEEE Trans. on Systems, Man and Cybernetics* **6** (1976) 325-327.
8. Eavis, T., Japkowicz, N.: A Recognition-based Alternative to Discrimination-based Multi-Layer Perceptrons, Workshop on Learning from Imbalanced Data Sets. *Technical Report WS-00-05,* AAAI Press (2000).
9. Ezawa, K.J., Singh, M., Norton, S.W.: Learning goal oriented Bayesian networks for telecommunications management, In: *Proc. 13th Int. Conf. on Machine Learning* (1996) 139-147.
10. Fawcett, T., Provost, F.: Adaptive fraud detection, *Data Mining and Knowledge Discovery* **1** (1996) 291-316.
11. Hart, PE.: The Condensed Nearest Neighbor rule. *IEEE Trans. on Information Theory* **6(4)** (1968) 515-516.
12. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study, *Intelligent Data Analysis Journal* **6(5)** (2002) 429-450.

13. Kubat, M., Matwin, S.: Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *Proceedings of the 14th International Conference on Machine Learning,* (1997) 179-186.
14. Kubat, M., Holte, R., Matwin, S.: Detection of Oil-Spills in Radar Images of Sea Surface. *Machine Learning,* **30** (1998) 195-215.
15. Mladenic, D., Grobelnik, M.: Feature selection for unbalanced class distribution and naïve Bayes, In: *Proc. 16th Int. Conf. on Machine Learning* (1999) 258-267.
16. Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., Brunk, C.: Reducing misclassification costs, In: *Proc 11th Int. Conf. on Machine Learning* (1994) 217-225.
17. Ritter, GI., Woodruff, HB., Lowry, SR., Isenhour, TL: An Algorithm for Selective Nearest Neighbor Decision Rule. *IEEE Trans. on Information Theory* **21(6) (1975)** 665-669.
18. Weiss, GM., Provost, F.: Learning when training data are costly: The effect of class distribution on tree induction, Journal of Artificial Intelligence Research **19** (2003) 315-354.
19. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data sets, *IEEE Trans. on Systems, Man and Cybernetics* **2** (1972) 408-421.
20. Woods, K., Doss, C., Bowyer, K.W., Solka, J., Priebe, C., Kegelmeyer, W.P.: Comparative evaluation of pattern recognition techniques for detection of micro-calcifications in mammography, *International Journal of Pattern Recognition and Artificial Intelligence* **7** (1993) 1417-1436.