# Restricted Decontamination for the Imbalanced Training Sample Problem*

R. Barandela[1,2], E. Rangel[1], J.S. Sánchez[3], and F.J. Ferri[4]

[1] Lab for Pattern Recognition, Inst. Tecnológico de Toluca
52140 Metepec, México rbarandela@hotmail.com
[2] Instituto de Geografía Tropical, La Habana, Cuba
[3] Dept. Llenguatges i Sistemes Informàtics, U. Jaume I, 12071 Castelló, Spain
[4] Dept. Informàtica. Universitat de València. 46100 Burjassot, Spain

**Abstract.** The problem of imbalanced training data in supervised methods is currently receiving growing attention. Imbalanced data means that one class is much more represented than the others in the training sample. 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 the present paper, we report experimental results that point at the convenience of correctly downsizing the majority class while simultaneously increasing the size of the minority one in order to balance both classes. This is obtained by applying a modification of the previously proposed Decontamination methodology. Combination of this proposal with the employment of a weighted distance function is also explored.

## 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. Performance of the resulting classification system depends on the quantity and the quality of the information contained in the TS. Recently, concern has arisen about the complications produced by imbalance in the TS. A TS is said to be imbalanced when one of the classes (the minority one) is heavily underrepresented in comparison to the other (the majority) class. For simplicity, and consistently with the common practice [7,13], we consider here only two-class problems. It has been observed that imbalanced training samples may cause a significant deterioration in the performance attainable by standard supervised methods. 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 [9], oil spills in satellite images of the sea surface [12], an infrequent disease [17], or text categorization [14].

* Work partially supported by grants 32016-A (Mexican CONACyT), 744.99-P (Mexican Cosnet), TIC2000-1703-C03-03 (Spanish CICYT) and P1-1B2002-07 (Fundació Caixa Castelló-Bancaixa).

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 424–431, 2003.

Most of the attempts for dealing with this problem can be categorized as [7]:

a) Over-sampling the minority class to match the size of the other class [5].
b) Downsizing the majority class so as to match the size of the other class [13].
c) Internally biasing the discrimination based process so as to compensate for the class imbalance [8,12].

As pointed out by many authors, overall accuracy is not the best criterion to assess the classifier's performance in imbalanced domains. For instance, consider a practical application where only 2% 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 98%. 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 an earlier study [4], we provide preliminary results of several techniques addressing the class imbalance problem. In such a work, we focused on under-sampling the majority class and also on internally biasing the discrimination process, as well as on a combination of both approaches. In the present paper, we introduce a new proposal for balancing the TS through reduction of the majority class size and, at the same time, an increase in the amount of prototypes in the minority class. To this aim, we employ a modification of the Decontamination methodology [2] that will be referred to as Restricted Decontamination. We also explore the convenience of using this technique in combination with a weighted distance measure aimed at biasing the classification procedure. These ideas are evaluated over four 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 its characteristics: 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 resampling 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. [5] proposed a technique for over-sampling the minority class. Instead of merely replicating prototypes of the minority class, they form new minority "synthetic" instances. This is done by taking each minority class instance and creating synthetic instances along the line segments joining any/all of the $k$ minority class nearest neighbors.

Barandela et al. [3] explore the convenience of designing a multiple classification system for working in imbalanced situations. Instead of using a single classifier, an ensemble is implemented. The idea is to train each one of the individual components of the ensemble with a balanced TS. In order to achieve this, each individual component of the ensemble is trained with a subset of the TS. As many subsets of the TS as required to get balanced subsets are generated. The number of subsets is determined by the difference between the amount of prototypes from the majority class and that of the minority class.

Pazzani et al. [15] 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. [8] bias the classifier in favor of certain feature relationships. Kubat et al. [12] use some counter-examples to bias the recognition process.

## 3 Proposed Strategies

In several practical applications, class identification of prototypes is a difficult and costly task. There is another source of distortion in the training data: prototypes with errors in some attribute values and instances that are atypical or exceptional. Generalization accuracy of the supervised method may be degraded by the presence of incorrectness or imperfections in the TS. Particularly sensitive to these facts are nonparametric classifiers whose training is not based upon any assumption about probability density functions. This explains the emphasis given to the evaluation of procedures used to collect and to clean the TS.

In a previous work [2], a methodology for correcting a TS while employing nonparametric classifiers has been presented. The Decontamination procedure can be regarded as a cleaning process removing some elements of the TS and correcting the label of several others while retaining them. Experimental results with both simulated and real datasets have shown that the Decontamination methodology allows to cope with all types of imperfections (mislabeled, noisy, atypical or exceptional) in the TS, improving the classifier's performance and lowering its computational burden. The Decontamination methodology is based on two previously published editing techniques.

### 3.1 Basic Editing Techniques

Editing techniques are mainly aimed at improving the performance of the NN rule by filtering the training prototypes. As a byproduct, they also obtain a decrease in the TS size and, consequently, a reduction of the computational cost of the classification method. The first work of editing corresponds to Wilson [16] and several others have followed.

**Wilson's Editing procedure.** This technique consists of applying the $k$-NN ($k > 1$) classifier to estimate the class label of every prototype in the TS. Those instances whose class label does not agree with the class associated to the majority of the $k$ neighbors are discarded. The procedure is:

1. Let $S = X$ ($X$ is the original TS and $S$ will be the edited TS)
2. For each $x$ in $X$ do:
   a) Find the $k$ nearest neighbors of $x$ in $X - \{x\}$
   b) Discard $x$ from $S$ if its label disagrees with the class associated with the largest number of the $k$ neighbors.

**Generalized Editing (GE: Koplowitz and Brown [11]).** This is a modification of the Wilson's algorithm. Out of concern with the possibility of too many prototypes being removed from the TS because of Wilson's editing procedure, this approach consists of removing some suspicious prototypes and to change the class labels of some other instances. Accordingly, it can be regarded as a technique for modifying the structure of the TS (through re-labeling of some prototypes and not only for eliminating atypical instances). In GE, two parameters have to be defined: $k$ and $k'$ in such a way that $(k + 1)/2 \leq k' \leq k$. This editing algorithm can be written as follows:

1. Let $S = X$ ($X$ is the original training set and $S$ will be the processed TS)
2. For each $x$ in $X$ do:
   a) Find the $k$ nearest neighbors of $x$ in $X - \{x\}$.
   b) If a class has at least $k'$ representatives among those $k$ neighbors, then label $x$ according to that class (independently of its original class label). Otherwise, discard it from $S$.

### 3.2 The Decontamination Methodology in Brief

The Decontamination methodology involves several applications of the GE technique, followed by the employment, also repeatedly, of the Wilson's Editing algorithm. Repetition in the application of each one of these techniques stops if one of the following criteria is fulfilled:

1. Stability in the structure of the TS has been reached (no more removals and no more re-labeling).
2. Estimate of the misclassification rate (leave-one-out method; see [10]) has begun to increase.

3. One class has resulted emptied (all its representatives in the TS have been removed or transferred to another class) or has resulted with too few prototypes (less than five training instances for each attribute).

### 3.3 Proposed Modification of the Decontamination Methodology

In the present paper, we present a modification of the Decontamination methodology: the Restricted Decontamination. In this restricted way, the Decontamination process is applied only to the majority class. That is, changes of label or removal from the TS affect only to those prototypes representing the majority class. In this way, a decrease in the amount of prototypes of the majority class is obtained. At the same time, some prototypes, originally in the majority class, are incorporated (by changing their labels) to the minority class, increasing the size of this latter class.

In the present work, Restricted Decontamination is employed for the first time for handling imbalance. This restricted procedure was initially designed to handle situations when information about the particular application area could imply existence of contamination in only some of the classes [1]. The source for this information could be given by some characteristics of the process used to collect the TS or by the intrinsic nature of the problem at hand.

### 3.4 The Weighted Distance Function

As a technique for internally biasing the discrimination procedure, we have experimented with a modification of the Euclidean metric that can be regarded as a weighted distance function [4]. With this modification, when classification of a new pattern $y$ is attempted, and in the search through the TS of its nearest neighbor, the following quantity must be computed for each training instance $x$:

$$d_W(y,x) = (n_i/n)^{1/m} d_E(y,x)$$

where $i$ refers to the class of instance $x$, $n_i$ is the number of training patterns from this class, $n$ is the TS size, $m$ is the dimensionality of the feature space and $d_E(\cdot)$ is the Euclidean metric.

The idea behind this distance proposal is to compensate for the imbalance in the TS without actually altering the imbalance. Weights are assigned, unlike in the usual weighted $k$-NN rule proposals, to the respective classes and not to the individual prototypes. In that way, since the weighting factor is greater for the majority class than for the minority one, distance values to training instances of the minority class are much more reduced than the distance values to the training examples of the majority class. This produces a tendency for the new patterns to find their nearest neighbor among the cases of the minority class, increasing the accuracy in that class.

**Table 1.** Characterization of the datasets employed in the experiments

| Datasets | Attributes | Training Sample | | Test Sample | |
|---|---|---|---|---|---|
| | | class 1 | class 2 | class 1 | class 2 |
| Phoneme | 5 | 1268 | 3054 | 318 | 764 |
| Satimage | 36 | 500 | 4647 | 126 | 1162 |
| Glass | 9 | 24 | 150 | 5 | 35 |
| Vehicle | 18 | 170 | 508 | 42 | 126 |

## 4 Experimental Results

The Restricted Decontamination proposal, and its combination with the weighted distance in the classification stage, are assessed through experiments carried out with four real datasets taken from the UCI Database Repository [6]. In each dataset, five-fold cross validation was employed (80% for the TS and 20% for a test set). Results to be presented hereafter represent the averaged values of the five replications. To facilitate comparison with other published results [13], in the Glass set the problem was transformed for discriminating class 7 against all the other classes, and in the Vehicle dataset the task is to classify class 1 against all the others. Satimage dataset was also mapped to configure a two-class case, 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. These modified datasets are described in Table 1. As can be seen, now class 2 is the majority class and class 1 is the minority one.

The results are shown in Table 2. The average $g$ values obtained when classifying with the original TSs, and with these TSs after we have processed them with the idea of Kubat and Matwin [13], are also included for comparison purposes. For a better illustration, results produced by the usual Decontamination procedure [2] are reported too. The Restricted Decontamination proposed here yields an improvement in performance (as measured by the $g$ criterion), in comparison to all the other methods. This improvement is more remarkable when the weighted distance is employed for classifying new patterns. It is also important to note that the results from the procedure of Kubat and Matwin are excelled in all datasets. The usual Decontamination methodology has been shown to produce important benefits [2] when considering the general accuracy, but it is not convenient in those cases when imbalance in the TS is present.

**Table 2.** Averaged mean values (and standard deviations) of the $g$ criterion

| Procedure | Phoneme | Satimage | Glass | Vehicle |
|---|---|---|---|---|
| Original TS | 73.8 | 70.9 | 86.7 | 56.0 |
| Decontamination & Euclidean classif. | 69.6 | 67.3 | 84.6 | 46.8 |
| **Restricted Decontam. & Euclidean classif.** | **73.8** | **75.4** | **86.2** | **66.4** |
| Decontamination & Weighted classif | 73.6 | 68.9 | 84.6 | 49.7 |
| **Restricted Decontam. & Weighted classif.** | **74.6** | **77.4** | **87.9** | **66.3** |
| Kubat and Matwin | 68.3 | 72.9 | 79.0 | 65.4 |

The effects of the Restricted Decontamination can be better analyzed by considering the balance obtained in the TS after its application (see Table 3). Results in this table indicate a decrease in the size of the majority class (number 2), while the minority class (number 1) size is increased. On the other hand, the usual Decontamination procedure deteriorates the imbalance in the TS, when compared with the original TS. The proposal of Kubat and Matwin, by aggressively under-sampling the majority class, produces an imbalance in the other direction, very remarkable in Phoneme and Glass datasets.

**Table 3.** Percentage of patterns in each class

| Procedure | Phoneme | | Satimage | | Glass | | Vehicle | |
|---|---|---|---|---|---|---|---|---|
| | class 1 | class 2 | class 1 | class 2 | class 1 | class 2 | class 1 | class 2 |
| Original TS | 29.34 | 70.66 | 9.71 | 90.29 | 13.79 | 86.21 | 25.07 | 74.93 |
| Decontamination | 25.68 | 74.32 | 10.04 | 89.96 | 11.68 | 88.32 | 15.24 | 84.76 |
| **Restricted Decont.** | **35.04** | **64.98** | **15.09** | **84.91** | **15.06** | **84.94** | **43.97** | **56.03** |
| Kubat and Matwin | 85.76 | 14.24 | 52.40 | 47.60 | 75.00 | 25.00 | 57.75 | 42.25 |

## 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 moderate by using a procedure that allows to under-sample the majority class while over-sampling the minority class. In this direction, a new approach has been proposed in this paper. The Restricted Decontamination has been shown to improve the balance in the TS. Classification with the weighted distance, after preprocessing the TS with the Restricted Decontamination has produced important progress in the resulting g value, when compared with the original TS. These results have also excelled those obtained by the proposal of Kubat and Matwin. This can be explained because the proposal of Kubat and Matwin is based upon techniques for eliminating redundant instances and not for cleaning the TS from noisy or atypical prototypes.

Benefits of the proposal are shown even in the Glass dataset. This dataset suffers not only of the imbalance problem, but also the minority class is too small. Adequacy of the TS size must be measured by considering the number of prototypes in the smaller class and not in the whole TS. For the minority class in Glass dataset, the size/dimensionality ratio is very low: only 2.7 instances for each attribute. Restricted Decontamination and weighted distance have been able to handle this critical situation.

A more extensive research is currently being conducted 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.

## References

1. R. Barandela. *The Nearest Neighbor rule: an empirical study of its methodological aspects.* PhD thesis, Univ. Berlin, 1987.
2. R. Barandela, E. Gasca, and R. Alejo. Correcting the training data. In D. Chen and X. Cheng, editors, *Pattern Recognition and String Matching.* Kluwer, The Netherlands, 2003.
3. R. Barandela, J. S. Sánchez, and R. M. Valdovinos. New applications of ensembles of classifiers. *Pattern Analysis and Applications*, to appear, 2003.
4. R. Barandela, J.S. Sánchez, V. García, and E. Rangel. Strategies for learning in class imbalance problems. *Pattern Recognition*, 36:849–851, 2003.
5. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2000.
6. Merz C.J. and P. M. Murphy. Uci repository of machine learning databases. Technical report, University of California at Irvine, Department of Information and Computer Science, 1998.
7. T. Eavis and N. Japkowicz. A recognition-based alternative to discrimination-based multi-layer perceptrons. In *Workshop on Learning from Imbalanced Data Sets.* TR WS-00-05, AAAI Press, 2000.
8. K.J. Ezawa, M. Singh, and S.W. Norton. Learning goal oriented bayesian networks for telecommunications management. In *Proc. 13th Int. Conf. on Machine Learning*, pages 139–147, 1996.
9. T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1:291–316, 1996.
10. D. J. Hand. *Construction and assessment of classification rules.* John Wiley and Sons, Chichester, 1997.
11. J. Koplowitz and T. A. Brown. On the relation of performance to editing in nearest neighbor rules. In *Proceedings of the 4th International Joint Conference on Pattern Recognition*, 1978.
12. M. Kubat, R. Holte, and S. Matwin. Detection of oil-spills in radar images of sea surface. *Machine Learning*, 30:195–215, 1998.
13. M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the 14th International Conference on Machine Learning*, pages 179–186, 1997.
14. D. Mladenic and M. Grobelnik. Feature selection for unbalanced class distribution and naive bayes. In *Proc. 16th Int. Conf. on Machine Learning*, pages 258–267, 1999.
15. M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk. Reducing misclassification costs. In *Proc 11th Int. Conf. on Machine Learning*, pages 217–225, 1994.
16. D. L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. 2(3):408–421, 1972.
17. K. Woods, C. Doss, K.W. Bowyer, J. Solka, C. Priebe, and W.P. Kegelmeyer. Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, 7:1417–1436, 1993.