PERGAMON

Pattern Recognition 36 (2003) 849–851

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Rapid and Brief Communication

Strategies for learning in class imbalance problems☆

R. Barandela[a], J.S. Sánchez[b,*], V. García[a], E. Rangel[a]

[a] *Instituto Tecnológico de Toluca, Av. Tecnológico s/n, 52140 Metepec, Mexico*
[b] *Department de Llenguatges i Sistemes Informàtics, Universitat Jaume I, 12071 Castellón, Spain*

Received 1 August 2002; accepted 21 August 2002

1. Introduction

A set of examples or training set (TS) is said to be imbalanced if one of the classes is represented by a very small number of cases compared to the other classes. Following the common practice [1,2], we consider only two-class problems and therefore, the examples are either positive or negative (that is, either from the minority class or the majority class, respectively). High imbalance occur in applications where the classifier is to detect a rare but important case, such as fraudulent telephone calls, oil spills in satellite images, failures in a manufacturing process, or a rare medical diagnoses. It has been observed that class imbalance may cause a significant deterioration in the performance attainable by standard supervised methods.

Most of the attempts at dealing with this problem can be grouped into three categories [2]. One is to assign distinct costs to the classification errors. The second is to resample the original TS, either by over-sampling the minority class and/or under-sampling the majority class until the classes are approximately equally represented. The third consists in internally biasing the discrimination-based process so as to compensate for the class imbalance.

As pointed out by many authors, the performance of a classifier in applications with class imbalance must not be expressed in terms of the average accuracy. For instance, consider a domain where only 2% examples are positive. In such a situation, labeling all new samples as negative would give an accuracy of 98%, but failing on all positive cases. Consequently, in environments with imbalanced classes, other measures have been proposed. The ROC curve and the geometric mean are good indicators of the classifier performance in these domains because they are independent of the distribution of examples between classes [1]. The geometric mean is defined as $g = \sqrt{a^+ a^-}$, where a^+ denotes the accuracy on the positive examples, and a^- is the accuracy on the negative examples. This measure tries to maximize the accuracy on each of the two classes while keeping these accuracies balanced.

In this work, we present preliminary results of a more extensive research that we are conducting to explore all the issues related to the class imbalance problem. Initially, we focused on resampling the TS and also on internally biasing the discrimination-based process, as well as on a combination of them. These approaches are evaluated over four real datasets using a nearest neighbor (NN) classifier and the geometric mean as performance measure.

2. Proposed strategies

Replicating the minority class to eliminate imbalance in the TS does not add new information to the system. Moreover, working in that direction means to worsen the known computational burden of some learning algorithms, such as the NN rule and the Multi-Layer Perceptron.

Thus, in the case of resampling the TS, the first strategy we adopted consists in downsizing the majority class by selecting a representative subset of the negative examples. For this, the only requirement is that all positive examples must be kept in the TS, even knowing that some of them can be noisy. A second resampling strategy consists in downsizing both classes (i.e., removing positive and negative examples from the TS).

Nevertheless, since downsizing the majority class can result in throwing away some useful information, this process must be done carefully. Accordingly, editing and condensing schemes offer a good alternative for removing noisy and redundant examples. We have tried three selection algorithms.

☆ Partially supported by grants TIC2000-1703-C03-03 from the Spanish CICYT, SAB2001-0004 from the Spanish MECD, and 32016-A from the Mexican CONACyT.

* Corresponding author. Universitat Jaume I, Av. Vincent Sos Baynat s/n, Castellón, 12006, Spain. Tel.: +34-964-728350; fax: +34-964-728435.

E-mail address: sanchez@uji.es (J.S. Sánchez).

PII: S0031-3203(02)00257-1

Table 1
Averaged values of the geometric mean

Original TS		Phoneme		Satimage		Glass		Vehicle	
		A	B	A	B	A	B	A	B
		73.8	76.0	70.9	75.9	86.7	88.2	55.8	59.6
WE	1st application	74.9	75.7	72.6	76.1	86.2	87.9	62.8	64.9
(majority class)	2nd application	74.8	75.5	72.9	76.2			62.8	64.8
	3rd application	74.6	75.3	73.0	76.2			64.0	65.8
WE+MS	1st application	74.5	72.4	74.0	74.2	86.1	86.2	65.7	65.6
(majority class)	2nd application	74.6	72.6	74.2	74.3			65.6	65.2
	3rd application	74.5	72.5	74.2	74.3			65.8	65.7
k-NCN	1st application	75.0	75.9	71.9	76.2	86.2	87.9	62.1	64.5
(majority class)	2nd application	74.7	75.4	72.3	76.2	86.2	87.9	64.1	66.3
	3rd application	74.7	75.4	72.3	76.2			65.2	67.4
k-NCN + MS	1st application	74.9	71.9	73.7	74.2	85.7	86.2	65.6	64.8
(majority class)	2nd application	71.6	71.9	74.4	74.8	85.7	86.2	66.2	65.8
	3rd application	74.6	72.1	74.4	74.8			66.5	66.8
MS (majority class)		74.0	70.0	72.3	73.0	86.2	86.6	60.3	60.3
MS (both classes)		72.2	72.8	70.1	73.3	86.4	87.1	59.7	59.9
WE (both classes)		73.8	76.7	66.4	68.8			47.5	51.5
WE + MS (both classes)		72.4	72.8	65.7	67.1			50.1	51.5

Two of them are in the group of editing: the classical Wilson's proposal (WE) [3] and the k-NCN (nearest centroid neighborhood) scheme [4]. Both aim at filtering the TS by deletion of noisy or atypical examples, generally increasing the NN accuracy. However, they do not produce an important amount of removed examples. So, the third algorithm refers to the modified selective (MS) [5] condensing, which is based on the idea of a consistent subset and guarantees a satisfactory approximation to the decision boundaries as they are defined by the whole TS. Finally, employment of the combined editing–condensing (WE + MS and k-NCN + MS) is also proposed as a downsizing strategy.

For internally biasing the discrimination procedure, we propose a weighted distance function to be used in the classification task. Let $d_E(\cdot)$ be the Euclidean metric, and let Y be a new sample to classify. Let x_0 be a training example from class i, let n_i be the number of examples 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

$$d_W(Y,x_0) = (n_i/n)^{1/m} d_E(Y,x_0).$$

The idea is to compensate for the imbalance in the TS 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 samples to find their NN among the positive examples, increasing the value of the geometric mean.

3. Experimental results and preliminary conclusions

Experiments were carried out over four datasets from the UCI Database Repository (http://www.ics.uci.edu/∼mlearn/). Five-fold cross validation was employed. All datasets were transformed into two-class problems to facilitate comparison with other published results [1].

First, we studied the different selection algorithms for downsizing the majority class. Moreover, WE and k-NCN were also used in an iterative manner. Second, MS, WE, and WE combined with MS were applied to both classes (not only to the majority class).

After preprocessing the TS, the examples in the test portion were classified with the NN rule, using both Euclidean and weighted distance measures. Averaged results of the geometric mean are shown in Table 1. Columns A and B contain the results obtained by employing the Euclidean and the weighted distances, respectively.

The best results are always obtained when the weighted distance is employed for classification. In all datasets, this technique alone produces a considerable improvement in the

performance measure. Thus, the weighted distance shows itself as a good resource to transform the classification procedure for taking into account the class imbalance, although other weighting factors must be still analyzed.

On the other hand, benefits of editing are well known for increasing the NN accuracy. This effect is corroborated for the geometric mean. The same can be stated about the performance obtained when processing the TS with the combined editing–condensing.

Repeated application of editing shows similar or better results than those of the single editing. In our experiments, the iterative procedure was stopped when no further removals were produced or when a class became empty of examples.

Glass dataset suffers not only from the imbalance problem, but also the minority class is too small. Adequacy of the TS size must be measured by considering the number of positive examples and not that of the whole TS. For the minority class in Glass dataset, the size/dimensionality rate is very low: 2.7 examples for each attribute. Either more positive examples are added to the TS or the number of attributes is reduced by a convenient feature selection. For this reason, we did not carry out any kind of filtering of the minority class in this dataset. In the other three datasets, exploration was done to evaluate the convenience of processing the minority class too, for removing noisy and redundant examples. Improvement was obtained only in the Phoneme dataset when WE was applied in both classes.

Despite the successful results, a problem common to all these downsizing techniques is that they do not permit control on the number of examples to be removed. Consequently, eliminated examples can be too many or too few to adequately solve the imbalance problem. Genetic algorithms could be of interest, in particular those that address the simultaneous selection of examples and features.

One of the most promising research lines refers to create an ensemble of classifiers by distributing the TS to reach balance in each of the resulting training samples. This involves a great variety of possibilities that we will cover in the next future.

References

[1] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, Proceedings of the 14th International Conference on Machine Learning, Nashville, USA, 1997, pp. 179–186.

[2] T. Eavis, N. Japkowicz, A recognition-based Alternative to Discrimination-based Multi-layer Perceptrons, Advances in Artificial Intelligence, Lecture Notes in Computer Science, Vol. 1822, Springer, Berlin, 2000, pp. 280–292.

[3] D.L. Wilson, Asymptotic properties of nearest neighbor rules using edited data sets, IEEE Trans. Systems Man Cybern. 2 (1972) 408–421.

[4] F.J. Ferri, J.S. Sánchez, F. Pla, Editing Prototypes in the Finite Sample Size Case Using Alternative Neighbourhoods, Advances in Pattern Recognition, Lecture Notes in Computer Science, Vol. 1451, Springer, Berlin, 1998, pp. 620–629.

[5] R. Barandela, N. Cortés, A. Palacios, The nearest neighbor rule and the reduction of the training sample size, Proceedings of the Ninth Spanish Symposium on Pattern Recognition and Image Analysis 1, Benicàssim, Spain, 2001, pp. 103–108.