COST SENSITIVE AND PREPROCESSING FOR CLASSIFICATION WITH IMBALANCED DATA-SETS: SIMILAR BEHAVIOUR AND POTENTIAL HYBRIDIZATIONS

Victoria López1, Alberto Fernández2, María José del Jesus2 and Francisco Herrera1
1Dept. of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada, 18071 Granada, Spain
2Dept. of Computer Science, University of Jaén, Jaén, Spain

Keywords: Classification, Imbalanced data-sets, Preprocessing, Sampling, Cost-sensitive learning, Hybridizations.

Abstract: The scenario of classification with imbalanced data-sets has supposed a serious challenge for researchers along the last years. The main handicap is related to the large number of real applications in which one of the classes of the problem has a few number of examples in comparison with the other class, making it harder to be correctly learnt and, what is most important, this minority class is usually the one with the highest interest.

In order to address this problem, two main methodologies have been proposed for stressing the significance of the minority class and for achieving a good discrimination for both classes, namely preprocessing of instances and cost-sensitive learning. The former rebalances the instances of both classes by replicating or creating new instances of the minority class (oversampling) or by removing some instances of the majority class (undersampling); whereas the latter assumes higher misclassification costs with samples in the minority class and seek to minimize the high cost errors. Both solutions have shown to be valid for dealing with the class imbalance problem but, to the best of our knowledge, no comparison between both approaches have ever been performed.

In this work, we carry out a full exhaustive analysis on this two methodologies, also including a hybrid procedure that tries to combine the best of these models. We will show, by means of a statistical comparative analysis developed with a large collection of more than 60 imbalanced data-sets, that we cannot highlight an unique approach among the rest, and we will discuss as a potential research line the use of hybridizations for achieving better solutions to the imbalanced data-set problem.

1 INTRODUCTION

In many supervised learning applications, there is a significant difference between the class prior rates, that is the probability a particular example belongs to a particular class. This situation is known as the class imbalance problem (Chawla et al., 2004; Sun et al., 2009; He and Garcia, 2009) and it is dominant in a high number of real problems including, but not limited to, telecommunications, WWW, finances, ecology, biology, medicine and so on; for which it is considered as one of the top problems in data mining (Yang and Wu, 2006). Furthermore, it is worth to point out that the positive or minority class is usually the one that has the highest interest from the learning point of view and it also implies a great cost when it is not well classified (Elkan, 2001).

The hitch with imbalanced data-sets is that standard classification learning algorithms are often biased towards the majority classes and therefore there is a higher misclassification rate in the minority class instances. Therefore, throughout the last years, many solutions have been proposed to deal with this problem, which can be categorized into two major groups:

1. Data Sampling: in which the training instances are modified in such a way as to produce a more balanced class distribution that allow classifiers to perform in a similar manner to standard classification (Batista et al., 2004; Chawla et al., 2002).

2. Algorithmic Modification: this procedure is oriented towards the adaptation of base learning methods to be more attuned to class imbalance issues (Zadrozny and Elkan, 2001). We must also stress in this case the use of cost-sensitive learning solutions, which basically assume higher misclassification costs with samples in the rare class and seek to minimize the high cost errors (Domingos,
Works in imbalanced classification usually focus on the development of new algorithms along one of the categories previously mentioned. However, there is not a study that exhaustively compares solutions from one category to another making difficult the selection of one kind of algorithm when classifying. The aim of this contribution is to develop a thorough experimental study to analyze the possible differences between preprocessing techniques and cost-sensitive learning for addressing classification with imbalanced data. In addition, we also present in the comparison a hybrid procedure that combines those two approaches to check whether there is a synergy between them.

As baseline classifier, we will use the C4.5 decision tree generating algorithm (Quinlan, 1993); firstly because it has been widely used to deal with imbalanced data-sets (Su and Hsiao, 2007; Drown et al., 2009; García et al., 2009), and secondly since it has been included as one of the top-ten data-mining algorithms (Wu and Kumar, 2009).

In order to analyze the oversampling and undersampling methodologies, we will focus on two of the most robust approaches such as the “Synthetic Minority Over-sampling TEstheneu (SMOTE) (Chawla et al., 2002) and its variant with the Wilson’s Edited Nearest Neighbour (ENN) rule (Wilson, 1972), as suggested by their performance among many different situations (Batista et al., 2004; Fernández et al., 2008). Regarding cost-sensitive methods, we have selected the C4.5-CS algorithm (Ting, 2002), which modifies the computation of the split criteria for C4.5 (normalized information gain) to take into account the a priori probabilities according to the number of samples for each class.

In this work, we focus on imbalanced binary classification problems, having selected a benchmark of 66 problems from KEEL data-set repository\(^1\) (Alcalá-Fdez et al., 2011). We perform our experimental study focusing on the precision of the models using the Area Under the ROC curve (AUC) (Huang and Ling, 2005). This study is carried out using nonparametric tests to check whether there exist significant differences among the obtained results (Demšar, 2006; García and Herrera, 2008).

This contribution is organized as follows: first, Section 2 presents the problem of imbalanced data-sets and the metric we have employed in this context whereas Section 3 describes the main methodologies to address the problem: the preprocessing methods used, cost-sensitive classification and a wrapper approach to combine both. In Section 4 an analysis of preprocessing techniques versus cost-sensitive learning approaches can be found. Finally, the conclusions of this work are commented in Section 5.

2 IMBALANCED DATA-SETS IN CLASSIFICATION

In this section, we first introduce the problem of imbalanced data-sets and then we present the evaluation metrics for this type of classification problem which differs from usual measures in classification.

2.1 The Problem of Imbalanced Data-sets

In the classification problem field, the scenario of imbalanced data-sets appears frequently. The main property of this type of classification problem is that the examples of one class outnumber the examples of the other one (Japkowicz and Stephen, 2002; Guo et al., 2008; Sun et al., 2009; He and García, 2009). The minority class usually represents the most important concept to be learnt; since it might be associated with exceptional and significant cases (Weiss, 2004), or because the data acquisition of these examples is costly (Weiss and Tian, 2008).

Since most of the standard learning algorithms consider a balanced training set, this situation may cause the obtention of suboptimal classification models, i.e. a good coverage of the majority examples whereas the minority ones are misclassified frequently; therefore, those algorithms which obtain a good behaviour in the framework of standard classification do not necessarily achieves the best performance for imbalanced data-sets (Fernández et al., 2010). There are several reasons behind this behaviour which are enumerated below:

1. The use of global performance measures for guiding the search process, such as standard accuracy rate, may benefit the covering of the majority examples.

2. Classification rules that predict the positive class are often highly specialized and thus their coverage is very low, hence they are discarded in favour of more general rules, i.e. those that predict the negative class.

3. It is always difficult to distinguish between noise examples and minority class examples and they can be completely ignored by the classifier.

In recent years, the imbalanced learning problem has received a high attention in the machine learn-
ing community. Specifically, regarding real world domains the importance of the imbalance learning problem is growing, since it is a recurring problem in many applications. As a few examples, we may find very high resolution airborne imagery (Chen et al., 2011), face recognition (Kwak, 2008) and especially medical diagnosis (Lo et al., 2008; Mazurowski et al., 2008). It is important to remember that the minority class usually represents the concept of interest and it is the most difficult to obtain from real data, for example patients with illnesses in a medical diagnosis problem; whereas the other class represents the counterpart of that concept (healthy patients).

2.2 Evaluation in Imbalanced Domains

The evaluation criteria is a key factor in both assessing the classification performance and guiding the classifier modelling. In a two-class problem, the confusion matrix (shown in Table 1) records the results of correctly and incorrectly recognized examples of each class.

<table>
<thead>
<tr>
<th>Positive prediction</th>
<th>Negative prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive class</td>
<td>True Positive (TP)</td>
</tr>
<tr>
<td>Negative class</td>
<td>False Positive (FP)</td>
</tr>
</tbody>
</table>

Traditionally, accuracy rate (Eq. (1)) has been the most commonly used empirical measure. However, in the framework of imbalanced data-sets, accuracy is no longer a proper measure, since it does not distinguish between the number of correctly classified examples of different classes. Hence, it may lead to erroneous conclusions, i.e., a classifier achieving an accuracy of 90% in a data-set with an IR value of 9, is not accurate if it classifies all examples as negatives.

\[
\text{Acc} = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}
\]

According to the previous issue, in this work we use the Area Under the Curve (AUC) metric (Huang and Ling, 2005), which can be defined as

\[
\text{AUC} = \frac{1 + TP_{\text{rate}} - FP_{\text{rate}}}{2} \tag{2}
\]

where \(TP_{\text{rate}}\) is the percentage of positive cases correctly classified as belonging to the positive class and \(FP_{\text{rate}}\) is the percentage of negative cases misclassified as belonging to the positive class.

3 ADDRESSING CLASSIFICATION WITH IMBALANCED DATA: PREPROCESSING AND COST-SENSITIVE LEARNING

A large number of approaches have been proposed to deal with the class imbalance problem. These approaches can be categorized into two groups: the internal approaches that create new algorithms or modify existing ones to take the class imbalance problem into consideration (Barandela et al., 2003; Sun et al., 2007; Ducange et al., 2010) and external approaches that preprocess the data in order to diminish the effect of their class imbalance (Batista et al., 2004; Estabrooks et al., 2004).

Regarding this, in this section we first introduce the main features of preprocessing techniques, focusing on SMOTE (Chawla et al., 2002) and SMOTE+ENN (Batista et al., 2004), which will be used along the experimental study. Next, we describe cost-sensitive learning and the C4.5-CS methodology (Ting, 2002). Finally, we present a framework to automatically detect a threshold for preprocessing using an underlying algorithm, in this case, a cost-sensitive approach.

3.1 Preprocessing Imbalanced Data-sets: Resampling Techniques

In the specialized literature, we can find some papers about resampling techniques studying the effect of changing the class distribution to deal with imbalanced data-sets.

Those works have proved empirically that, applying a preprocessing step in order to balance the class distribution, is usually an useful solution (Batista et al., 2004; Fernández et al., 2008; Fernández et al., 2010). Furthermore, the main advantage of these techniques is that they are independent of the underlying classifier.

Resampling techniques can be categorized into three groups or families:

1. **Undersampling Methods**, which create a subset of the original data-set by eliminating instances (usually majority class instances).
2. **Oversampling Methods**, which create a superset of the original data-set by replicating some instances or creating new instances from existing ones.
3. **Hybrids Methods**, which combine both sampling approaches.
Among these categories, there are several proposals where the simplest preprocessing are non-heuristic methods such as random undersampling and random oversampling. In the first case, the major drawback is that it can discard potentially useful data, that could be important for the induction process. For random oversampling, several authors agree that this method can increase the likelihood of occurring overfitting, since it makes exact copies of existing instances.

According to the previous facts, more sophisticated methods have been proposed. Among them, SMOTE (Chawla et al., 2002) has become one of the most renowned approaches in this area. In brief, its main idea is to create new minority class examples by interpolating several minority class instances that lie together for oversampling the training set.

With this technique, the positive class is oversampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the $k$ minority class nearest neighbours. Depending upon the amount of over-sampling required, neighbours from the $k$ nearest neighbours are randomly chosen. This process is illustrated in Figure 1, where $x_i$ is the selected point, $x_{1i}$ to $x_{4i}$ are some selected nearest neighbours and $r_1$ to $r_4$ the synthetic data points created by the randomized interpolation.

![Figure 1: An illustration of how to create the synthetic data points in the SMOTE algorithm.](image)

However, in oversampling techniques, and especially for the SMOTE algorithm, the problem of over generalization is largely attributed to the way in which it creates synthetic samples. Specifically, SMOTE generates the same number of synthetic data samples for each original minority example and does so without consideration to neighboring examples, which increases the occurrence of overlapping between classes (Wang and Iapakowicz, 2004). For this reason we also consider a hybrid approach in this work, “SMOTE + ENN”, where the Wilson’s ENN Rule (Wilson, 1972) is used after the SMOTE application to remove from the training set any example misclassified by its three nearest neighbours.

3.2 Cost-sensitive Learning

Cost-sensitive learning takes into account the variable cost of a misclassification of the different classes (Domingos, 1999; Zadrozny et al., 2003). In this case, a cost matrix codifies the penalties $C(i,j)$ of classifying examples of one class as a different one; if we use the notation 1 for minority and 0 for majority class, $C(i,1) = TN$ or $TP$. These misclassification cost values can be given by domain experts, or learned via other approaches (Sun et al., 2009; Sun et al., 2007). Specifically, when dealing with imbalanced problems it is usually of most interest to recognize the positive instances rather than the negative ones and therefore, the cost when misclassifying a positive instance is higher than the cost of misclassifying a negative one.

Given the cost matrix, an example should be classified into the class that has the minimum expected cost. This is the minimum expected cost principle. The expected cost $R(i|x)$ of classifying an instance $x$ into class $i$ (by a classifier) can be expressed as:

$$R(i|x) = \sum_j P(j|x) \cdot C(i,j)$$

where $P(j|x)$ is the probability estimation of classifying an instance into class $j$. That is, the classifier will classify an instance $x$ into positive class if and only if:

$$P(0|x) \cdot (C(1,0) - C(0,0)) \leq P(1|x) \cdot C(0,1) - C(1,1))$$

Therefore, any given cost-matrix can be converted to one with $C(0,0) = C(1,1) = 0$. Under this assumption, the classifier will classify an instance $x$ into positive class if and only if:

$$P(0|x) \cdot C(1,0) \leq P(1|x) \cdot C(0,1)$$

As $P(0|x) = 1 - P(1|x)$, we can obtain a threshold $p^*$ for the classifier to classify an instance $x$ into positive if $P(1|x) \geq p^*$, where

$$p^* = \frac{C(1,0)}{C(1,0) - C(0,1)} = \frac{FP}{FP + FN}$$

Another possibility is to “rebalance” the original training examples the ratio of:

$$p(1)FN : p(0)FP$$

where $p(1)$ and $p(0)$ are the prior probability of the positive and negative examples in the original training set.

In summary, two main general approaches have been proposed to deal with cost-sensitive problems:
1. **Direct Methods**: The main idea of building a direct cost-sensitive learning algorithm is to directly introduce and utilize misclassification costs into the learning algorithms.

For example, in the context of decision tree induction, the tree-building strategies are adapted to minimize the misclassification costs. The cost information is used to: (1) choose the best attribute to split the data (Ling et al., 2004; Riddle et al., 1994); and (2) determine whether a subtree should be pruned (Bradford et al., 1998). On the other hand, other approaches based on genetic algorithms can incorporate misclassification costs in the fitness function (Turney, 1995).

2. **Meta-learning**: This methodology implies the integration of a “preprocessing” mechanism for the training data or a “postprocessing” of the output, in such a way that the original learning algorithm is not modified. Cost-sensitive meta-learning can be further classified into two main categories: *thresholding* and *sampling*, which are based on expressions (4) and (5) respectively:

- **Thresholding** is based on the Bayes decision theory that assign instances to the class with minimum expected cost, as introduced above. For example, a typical decision tree for a binary classification problem assigns the class label of a leaf node depending on the majority class of the training samples that reach the node. A cost-sensitive algorithm assigns the class label to the node that minimizes the classification cost ( Domingos, 1999; Zadrozny and Elkan, 2001).

- **Sampling** is based on modifying the training data-set. The most popular technique lies in resampling the original class distribution of the training data-set according to the cost decision matrix by means of undersampling/oversampling (Zadrozny et al., 2003) or assigning instance weights (Ting, 2002). These modifications have shown to be effective and can also be applied to any cost insensitive learning algorithm (Zhou and Liu, 2006).

In this work, we will make use of the cost-sensitive C4.5 decision tree (C4.5-CS) proposed in (Ting, 2002). This method changes the class distribution such that the tree induced is in favor of the class with high weight/cost and is less likely to commit errors with high cost. Specifically, the computation of the split criteria for C4.5 (normalized information gain) is modified to take into account the a priori probabilities according to the number of samples for each class.

The standard greedy divide-and-conquer procedure for inducing minimum error trees can then be used without modification, except that $W_j(t)$ (6) is used instead of $N_j(t)$ (number of instances of class $j$) in the computation of the test selection criterion in the tree growing process and the error estimation in the pruning process.

$$ W(j) = C(j) \frac{N}{\sum C(i)N_i} $$

C4.5-CS also introduces another optional modification that alters the usual classification process after creating the decision tree. Instead of classifying using the minimum error criteria, it is advisable to classify using the expected misclassification cost in the last part of the classification procedure. The expected misclassification cost for predicting class $i$ with respect to the instance $x$ is given by

$$ EC_i(x) = \sum_j W_j(t(x)) cost(i, j) $$

where $t(x)$ is the leaf of the tree that instance $x$ falls into and $W_j(t)$ is the total weight of class $j$ training instances in node $t$.

### 3.3 Hybridization. Automatically Countering Imbalance

The different solutions used to deal with the imbalanced problem have been presented in the previous subsections. So the question now is “Can we use both techniques together and achieve better results?”.

In this section we describe a procedure to integrate the cost-sensitive learning and preprocessing approaches into one, quite similar to the one proposed in (Chawla et al., 2008), which consists in a wrapper paradigm that discovers the amount of resampling needed for a data-set based on optimizing evaluation functions which can include the cost associated to the classification. This wrapper infrastructure applies cross-validation to first discover the best amounts of undersampling and oversampling, applies the preprocessing algorithms with the amounts estimated and finally runs the algorithm used over the preprocessed data-set. Figure 2 shows the algorithm procedure.

The undersampling estimation starts with no undersampling for all majority classes and obtains baseline results on the training data. Then it traverses through the search space of undersampling percentages in decrements of Sample Decrement (in this case 10%), in a greedy iterative fashion, to increase performance over the minority classes without sacrificing performance on the majority class.
The oversampling algorithm evaluates different amounts of SMOTE at steps of 100% (the number of examples from the minority class). This is a greedy search, and at each step the new performance estimates become the new baseline. That is, the initial baseline is the performance obtained via the Wrapper Undersample. If SMOTE=100% improves the performance over that baseline by some margin Increment Min, then the performance achieved at SMOTE=100% becomes the new baseline. The amount of SMOTE is then incremented by Sample Increment, and another evaluation is performed to check if the performance increase at new SMOTE amount is at least greater than Increment Min. This process repeats, greedily, until no performance gains are observed.

However, there is an important caveat to the search to avoid being trapped in a local maximum. If the average does not improve by 5% we have to verify that we have not settled on a local maximum. In order to do so, we look ahead two more steps at increasing amounts of SMOTE. If the look-ahead does not result in an improvement in performance, then the amount of SMOTE is reset to the value discovered prior to the look-ahead. This is done to allow SMOTE to introduce additional examples with the aim of improving performance. However, if the addition of examples does not help, then we go back to using the lesser amount of SMOTE discovered prior to the look-ahead.

We can use different measures to evaluate the performance of the classifier to estimate the sampling parameters. In our case, different from (Chawla et al., 2008), we use cost-sensitive learning algorithms as base classifiers, and therefore a logical evaluation criteria is the cost itself. Cost is calculated as shown in Equation 8 when we assume $C(\cdot \mid \cdot) = C(\cdot \mid 
abla)$ = 0 (as it is usual in imbalanced classification).

$$cost = FNRate \times C(\cdot \mid \nabla) + FPRate \times C(\cdot \mid \nabla)$$ (8)

4 EXPERIMENTAL STUDY

In this section, we will perform an analysis to determine the performance of the different alternatives used for imbalanced classification. Our aim is to analyze three different issues:

1. The improvement obtained by preprocessing data-sets and cost-sensitive learning over the original algorithm.
2. The possible differences between the rebalancing techniques versus cost-sensitive learning and in which cases.
3. Whether a hybrid methodology that combines a preprocessing approach and a cost-sensitive learning algorithm supposes a positive synergy and enables the achievement of more accurate results.

First, we present our experimental framework with the data-set employed in our analysis and the statistical tests that will allow us to support the extracted findings. Then, we will show the results from our study and we will discuss the main issues that will arise from the aforementioned analysis.

4.1 Experimental Framework

In order to analyze the preprocessing approach against the cost-sensitive learning strategy, we have selected 66 data-sets from the KEEL data-set repository² (Alcalá-Fdez et al., 2011). These data-sets are summarized in Table 2, where we denote the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (positive and negative), class distribution and IR.

To develop the different experiments we consider a 5-folder cross-validation model, i.e., five random partitions of data with a 20% and the combination of

²http://www.keel.es/data-sets.php
4 of them (80%) as training and test. For each data-set we consider the average results of the five partitions. The data-sets used in this study use the partitions provided by the repository in the imbalanced classification data-set section.

Furthermore, we have to identify the misclassification costs associated to the positive and negative class for the cost-sensitive learning versions. If we assign a positive sample as a negative one the associated cost is 1 (Cost(+,-) = 1) whereas if we classify a negative sample as a positive one the associated cost is 0 (Cost(-,+)) = 0). The cost of classifying correctly is the IR of the data-set (Cost(+,-) = IR) whereas if we guess the correct class should not penalize the built model.

Finally, statistical analysis needs to be carried out in order to find significant differences among the results obtained by the studied methods (Demšar, 2006; Garcia et al., 2009; Garcia et al., 2010). Since the study is split in parts comparing a group of algorithms, we use non-parametric statistical tests for multiple comparisons. Specifically, we use the Mann-Whitney test (Sheskin, 2006) to detect statistical differences among a group of results and the Shaffer post-hoc test (Shaffer, 1986) in order to find out which algorithms are distinctive among an n x n comparison.

Furthermore, we consider the average ranking of the algorithms in order to show graphically how good a method is with respect to its partners. This ranking is obtained by assigning a position to each algorithm depending on its performance for each data-set. The algorithm which achieves the best accuracy in a specific data-set will have the first rank value (1); then, the algorithm with the second best accuracy is assigned rank 2, and so forth. This task is carried out for all data-sets and finally an average ranking is computed as the mean value of all rankings.

### 4.2 Contrasting Preprocessing and Cost-sensitive Learning in Imbalanced Data-sets

Table 3 shows the average results in training and test together with the corresponding standard deviation for the seven versions of the C4.5 algorithm used in the study: the base classifier, the base classifier used over the preprocessed data-sets, the cost-sensitive version of the algorithm and the hybrid versions of it. We stress in boldface the best results achieved for the prediction ability of the different techniques.

---

A table showing the summary of imbalanced data-sets.

**Table 2: Summary of imbalanced data-sets.**

<table>
<thead>
<tr>
<th>Data-set</th>
<th>#Ex.</th>
<th>#Atts.</th>
<th>Class (-,+)</th>
<th>%Class(-; +)</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glass0</td>
<td>214</td>
<td>9</td>
<td>(build-win-float-proc, remainder)</td>
<td>(9.91, 90.09)</td>
<td>9.09</td>
</tr>
<tr>
<td>Ecoli067vs35</td>
<td>222</td>
<td>7</td>
<td>(cp,omL,pp; imL,om)</td>
<td>(13.55, 86.45)</td>
<td>6.38</td>
</tr>
<tr>
<td>New-thyroid1</td>
<td>215</td>
<td>5</td>
<td>(hyper; remainder)</td>
<td>(16.28, 83.72)</td>
<td>5.14</td>
</tr>
<tr>
<td>Iris0</td>
<td>150</td>
<td>4</td>
<td>(Iris-Setosa; remainder)</td>
<td>(33.33, 66.67)</td>
<td>2.00</td>
</tr>
</tbody>
</table>

---

3http://www.keel.es/imbalanced.php
Table 4: Shaffer test for the C4.5 variety of algorithms using the AUC measure.

<table>
<thead>
<tr>
<th></th>
<th>C4.5 None</th>
<th>SMOTE+ENN</th>
<th>SENN</th>
<th>CS</th>
<th>Wr_SMOTE</th>
<th>Wr_US</th>
<th>Wr_SENN</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>x</td>
<td>-(6.404E-6)</td>
<td>-(4.058E-8)</td>
<td>-6.404E-6</td>
<td>-(7.904E-6)</td>
<td>-(0.00341)</td>
<td>= (0.37846)</td>
</tr>
<tr>
<td>SMOTE</td>
<td>+ (6.404E-6)</td>
<td>x</td>
<td>= (1.0)</td>
<td>= (1.0)</td>
<td>= (1.0)</td>
<td>= (1.0)</td>
<td>+ (0.04903)</td>
</tr>
<tr>
<td>SENN</td>
<td>+ (4.058E-8)</td>
<td>= (1.0)</td>
<td>x</td>
<td>= (1.0)</td>
<td>= (1.0)</td>
<td>= (0.22569)</td>
<td>+ (0.00152)</td>
</tr>
<tr>
<td>CS</td>
<td>+ (6.404E-6)</td>
<td>= (1.0)</td>
<td>= (1.0)</td>
<td>x</td>
<td>= (1.0)</td>
<td>= (1.0)</td>
<td>+ (0.04903)</td>
</tr>
<tr>
<td>Wr_SMOTE</td>
<td>+ (7.904E-6)</td>
<td>= (1.0)</td>
<td>= (1.0)</td>
<td>= (1.0)</td>
<td>x</td>
<td>= (1.0)</td>
<td>+ (0.04903)</td>
</tr>
<tr>
<td>Wr_US</td>
<td>+ (.00341)</td>
<td>= (1.0)</td>
<td>= (0.22569)</td>
<td>= (1.0)</td>
<td>= (1.0)</td>
<td>x</td>
<td>= (1.0)</td>
</tr>
<tr>
<td>Wr_SENN</td>
<td>- (.37846)</td>
<td>- (.04903)</td>
<td>- (.00152)</td>
<td>- (.04903)</td>
<td>- (.04903)</td>
<td>= (1.0)</td>
<td>x</td>
</tr>
</tbody>
</table>

The results of the statistical analysis of the C4.5 family are as follows. For the sake of a visual comparison, Figure 3 shows the average ranking for these approaches. Under the AUC measure, the Iman-Davenport test detects significant differences among the algorithms, since the p-value returned (1.88673E-10) is lower than our α-value (0.05). The differences found are analyzed with a Shaffer test, shown in Table 4. In this table, a “+” symbol implies that the algorithm in the row is statistically better than the one in the column, whereas “-” implies the contrary; “=” means that the two algorithms compared have no significant differences. In brackets, the adjusted p-value associated to each comparison is shown.

Figure 3: Average rankings using the AUC measure for the C4.5 variety of algorithms.

Observing the results from Tables 3 and 4, we conclude that the standard C4.5 approach is outperformed by most of the methodologies that deal with imbalanced data. All methodologies, the hybrid version that uses only an oversampling step with SMOTE+ENN, have significant differences versus the base C4.5 classifier. This analysis answers our first question of the study, that is, the classification performance is degraded in an imbalance scenario having a bias towards the majority class examples and the use of the aforementioned techniques (preprocessing and cost-sensitive learning) allow us to obtain a better discrimination of the examples of both classes resulting in an overall good classification for all concepts of the problem (positive and negative classes).

Comparing the results when applying preprocessing we can see that the performance of these methods is not statistically different for any of its versions. In addition, the performance of those preprocessing methods is also not different to the cost-sensitive learning version of C4.5. This second part of the study has reflected that the two employed solutions are quite similar between them and it was not possible to highlight one of them as the most adequate one for classification. For that reason, the question on which approach is preferable for addressing classification with imbalanced data-sets is still unresolved.

Finally, regarding the hybridization of cost-sensitive learning and preprocessing by using a wrapper routine, it can be seen that there are significant differences both between the different hybrid versions and with the other alternatives. The hybrid version that uses an oversampling step with SMOTE+ENN is outperformed by all the other versions except the base version. The rest of the hybrid versions are not statistically different from the performance of usual approaches for imbalanced classification. Therefore, we cannot state that the hybridization in decision trees produces a positive synergy between the two techniques. According to these results, the preliminary version of the hybrid technique can be further improved both applying a finest combination of the individual approaches or by using more specific methods with a better synergy between them.

5 CONCLUSIONS

In this work we have analyzed the behaviour of preprocessing and cost-sensitive learning in the framework of imbalanced data-sets in order to determine whether there are any significant differences between
both approaches and therefore which one of them is preferred and in which cases. Additionally, we have proposed a hybrid approach that integrates both approaches together.

First of all, we have determined that both methodologies improve the overall performance for the classification with imbalanced data, which was the expected behaviour. Next, the comparison between preprocessing techniques against cost-sensitive learning hints that there are no differences among the different preprocessing techniques. The statistical study, supported by a large collection of more than 60 imbalanced data-sets, lets us say that both preprocessing and cost-sensitive learning are good and equivalent approaches to address the imbalance problem.

Finally, we have shown that our preliminary versions of hybridization techniques are truly competitive with the standard methodologies. We must stress that this is a very interesting trend for research as there is still room for improvement regarding hybridization between preprocessing and cost-sensitive learning.

ACKNOWLEDGEMENTS

This work has been supported by the Spanish Ministry of Science and Technology under Projects TIN2011-28488 and TIN2008-06681-C06-02 and the Andalusian Research Plans P10-TIC-6858 and TIC-3928. V. López holds a FPU scholarship from Spanish Ministry of Education.

REFERENCES


Zadrozny, B. and Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown. In Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining (KDD’01), pages 204–213.
