Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data

V. García[a,*], J.S. Sánchez[b], A.I. Marqués[c], R. Florencia[a], G. Rivera[a]

[a] *División Multidisciplinaria en Ciudad Universitaria, Universidad Autónoma de Ciudad Juárez, Av. José de Jesús Delgado 18100, Ciudad Juárez, Chihuahua (Mexico)*
[b] *Institute of New Imaging Technologies, Department of Computer Languages and Systems Universitat Jaume I, Av. de Vicent Sos Baynat, s/n 12071 Castelló de la Plana (Spain)*
[c] *Department of Business Administration and Marketing Universitat Jaume I, Av. de Vicent Sos Baynat, s/n 12071 Castelló de la Plana (Spain)*

Abstract

Data plays a key role in the design of expert and intelligent systems and therefore, data preprocessing appears to be a critical step to produce high-quality data and build accurate machine learning models. Over the past decades, increasing attention has been paid towards the issue of class imbalance and this is now a research hotspot in a variety of fields. Although the resampling methods, either by under-sampling the majority class or by over-sampling the minority class, stand among the most powerful techniques to face this problem, their strengths and weaknesses have typically been discussed based only on the class imbalance ratio. However, several questions remain open and need further exploration. For instance, the subtle differences in performance between the over- and under-sampling algorithms are still under-comprehended, and we hypothesize that they could be better explained by analyzing the inner structure of the data sets. Consequently, this paper attempts to investigate and illustrate the effects of the resampling methods on the inner structure of a data set by exploiting local neighborhood information, identifying the sample types in both classes and analyzing their distribution in each resampled set. Experimental results indicate that the resampling methods that pro-

[*] Corresponding author. Tel.: +52 688 2100 ext. 6700

Email addresses: vicente.jimenez@uacj.mx (V. García), sanchez@uji.es (J.S. Sánchez), imarques@uji.es (A.I. Marqués), rogelio.florencia@uacj.mx (R. Florencia), gilberto.rivera@uacj.mx (G. Rivera)

duce the highest proportion of safe samples and the lowest proportion of unsafe samples correspond to those with the highest overall performance. The significance of this paper lies in the fact that our findings may contribute to gain a better understanding of how these techniques perform on class-imbalanced data and why over-sampling has been reported to be usually more efficient than under-sampling. The outcomes in this study may have impact on both research and practice in the design of expert and intelligent systems since a priori knowledge about the internal structure of the imbalanced data sets could be incorporated to the learning algorithms.

Keywords: Class imbalance, Sample types, Resampling, Local neighborhood

1. Introduction

Organizations are nowadays focused on exploiting the vast amounts of data generated from many sources and with multiple formats for competitive advantage. To this end, expert and intelligent systems are developed to make decisions based on insights extracted from the data sets. Since the potential of these systems relies on the quality of data, preprocessing becomes one of the most critical and effort-inducing stages in their development.

In many real-life applications, the data sets are typically imbalanced, which has been described as a challenging problem and the subject of several research efforts. A binary data set is said to be imbalanced if one of the classes is represented by a very small number of examples compared to the other class. By convention, the examples of the minority class are labeled as positive and those of the majority class are called negative.

It has been observed that class imbalance may cause an important deterioration of the performance attainable by most standard classifiers because they are strongly biased towards the classification of the negative examples and are not competent enough to classify the minority class correctly (Branco et al., 2016). However, the poor accuracy of existing models on positive examples could be attributed not only to class imbalance but also to a variety of factors, such as noisy data, class overlapping, lack of density and small disjuncts (He & Garcia, 2009; Jo & Japkowicz, 2004; López et al., 2013). This means that the class imbalance may be not a problem by itself and countering class imbalance will not always lead to an improvement in performance (García et al., 2008; Japkowicz, 2003).

A large number of strategies have been proposed to deal with the class imbalance problem, which can be mainly grouped into three categories (Haixiang et al., 2017; Krawczyk, 2016). One is to assign distinct costs to the misclassifications on each class in such a way that an error made on the minority class will be more costly than an error made on the majority class. The second strategy is to preprocess the imbalanced data, either by enlarging (over-sampling) the minority class and/or shrinking (under-sampling) the majority class until the classes are approximately equally represented. The third group consists in internally biasing the discrimination-based process to compensate for the class imbalance.

The resampling techniques have probably been the most investigated because they are independent of the underlying classifier and can be easily implemented for any problem (Estabrooks et al., 2004; García et al., 2016; Weiss, 2004). The level of imbalance is reduced in both over- and under-sampling algorithms, with the assumption that a more balanced set should provide better classification results. However, these methods also present some weaknesses due to the artificial alteration of the original distribution of classes. For instance, under-sampling may throw out potentially useful data (leading to information loss) and augment the variance of the classifier, while over-sampling increases the population of the data set by generating synthetic examples and increases the likelihood of overfitting and the computational burden of any learning model (Kang et al., 2017; López et al., 2013; Vuttipittayamongkol & Elyan, 2020; Wong et al., 2018).

Though conclusions about what is the most efficient resampling strategy for the class imbalance problem are divergent, many studies have reported that over-sampling usually performs better than under-sampling (Bach et al., 2017; Batista et al., 2004; García et al., 2012; Prati et al., 2015; Van Hulse et al., 2007; Yin & Gai, 2015). These conclusions have been drawn from mere experimental comparisons of a collection of resampling techniques to evaluate their performance, while the reasons why over-sampling is generally superior to under-sampling have not been properly investigated. Moreover, many of those studies have considered the imbalance ratio (the ratio of the majority class size to the minority class size) as the unique data difficulty factor[1], thus neglecting other relevant data characteristics that could help to explain the behavior of each of the three resampling strategies.

Taking into account the limitations just mentioned, the motivation of this paper is to provide further insight into the underlying causes of the apparent superiority of over-sampling. In pursuing this objective, the contribution of this present pa-

[1] Data difficulty factors refer to internal and local characteristics of class distributions that may degrade the performance of standard classifiers.

per is a large-scale experimental analysis with 22 resampling methods across six articial data sets and 73 real-life data sets to understand the superiority of over-sampling based on the distribution of safe and unsafe samples. To this end, we address the following questions:

(i) What effect do the resampling algorithms have on the inner structure of the class-imbalanced data sets?,

(ii) Can the superiority of over-sampling algorithms be explained in terms of safe and unsafe samples?,

(iii) Does there exist a close link between the amount of safe and unsafe and the performance of the strategies?

Unlike the common procedure that focuses only on the minority class, we assume that the majority one also deserves to be analyzed because the distribution of negative samples may provide meaningful information. Our hypothesis is that over-sampling often outperforms under-sampling because the former leads to a distribution of sample types with more safe examples and less unsafe cases than the latter. Hopefully, this will allow us to expand our understanding of how the performance of the resampling strategies is related to their effects on the structure of a data set. The findings of this study can serve as a valuable guideline to design expert and intelligent systems for many real-life applications that have to deal with class-imbalanced data such as fraud detection, cancer malignancy grading, fault detection in industrial machinery and software defect prediction, among many others.

Henceforward, this paper is organized as follows. Section 2 summarizes a pool of works concerned with analyzing the possible relationships between class imbalance and other data difficulty factors. Section 3 provides a summary of representative resampling techniques, which will be further used for the experimental analysis. Section 4 presents a neighborhood-based categorization of the different sample types that can be found in an imbalanced data set. Next, Section 5 describes the research methodology that we have adopted to conduct this study and presents the thorough experimentation carried out. Finally, Section 6 remarks on the main findings and outlines possible directions for further research.

2. Class imbalance and other data difficulty factors

As already remarked, the imbalanced distribution of classes itself is not the only data difficulty factor, but there exist other intrinsic data characteristics that combined with class imbalance can be even more critical and lead to a severe loss of classification performance, especially for the minority class. Das et al. (2018) proposed a categorization of the intrinsic data characteristics into two groups: (i) distribution-based data irregularities, and (ii) feature-based data irregularities. The first group covers class imbalance, outliers and noisy data, class overlapping, small disjuncts, data set shift and small data set size, whereas the second group includes missing, noisy, irrelevant and redundant features. Next, we summarize a representative collection of recent publications where the class imbalance appears as the intersection factor between both groups.

2.1. Distribution-based data irregularities

One of the first papers that intended to discover any links between class imbalance and data complexity corresponds to the one by Japkowicz & Stephen (2002), in which the authors concluded that imbalance is a relative problem that depends on both the difficulty of the data and the overall size of the training set. After this seminal work, numerous studies have explored the influence of other complexity factors in class-imbalanced data. For instance, Prati et al. (2004b) investigated how class imbalance and error-prone small disjuncts are related to each other, whereas Jo & Japkowicz (2004) claimed that the degradation of classification accuracy is more due to the presence of small disjuncts than to the class imbalance problem. A similar conclusion was drawn by Weiss (2010), who also showed that class imbalance is partly responsible for the problem with small disjuncts.

Prati et al. (2004a) showed that there exists a strong correlation between the degree of class overlapping and class imbalance. Similarly, the experimental results in two papers by García et al. (2006, 2007) suggested that the local imbalance in the overlap region has an impact on the performance of classifiers stronger than the global imbalance, especially when there exists strong overlap and synthetic examples are generated with SMOTE. On the other hand, García et al. (2008) stated that the nearest neighbor classifier was more sensitive to the size of the class overlap than to the overall imbalance ratio. Vorraboot et al. (2015) proposed some modified hybrid algorithms to improve the classification performance of highly imbalanced large data sets with overlapped regions.

Dal Pozzolo et al. (2015) showed that the benefits of using an under-sampling algorithm strongly depends on the number of samples, the variance of the classifier, the degree of imbalance and the value of the posterior probability. García et al. (2015) compared the behavior of three linear classifiers modeled on both the feature space and the dissimilarity space when the class imbalance of data sets interweaved with small disjuncts and noise; they showed that small disjuncts could be much better overcome on the dissimilarity space than on the feature space, whereas noise in imbalanced data sets cannot be completely solved through the dissimilarity-based representation. Luengo et al. (2011) evaluated the behavior of three resampling methods (SMOTE, SMOTE-ENN, and an evolutionary under-sampling algorithm) by using three data complexity measures (F1, N4, and L3) (Ho & Basu, 2002) computed over the imbalanced data sets and then, they derived two descriptive rules to identify the data sets in which the C4.5 and PART decision trees could perform well.

Napierala et al. (2010) analyzed how the noisy and borderline positive examples hindered the classification performance and concluded that focused preprocessing methods outperformed both random and cluster-based over-sampling algorithms. Stefanowski (2013) observed that the degradation of classification performance was more related to the decomposition of the minority class into small sub-groups than to the class imbalance, and also that the amount of borderline and rare examples in the minority class had an even stronger influence on the classifiers.

Sáez et al. (2016) proposed a general methodology to decide which types of positive samples should be processed by an over-sampling algorithm when facing with multi-class imbalanced distributions; the types of samples were characterized by using the local neighborhood-based procedure that will be further introduced in Section 4. Following the same line, Skryjomski & Krawczyk (2017) analyzed the structure of the minority class to transform the SMOTE algorithm into a selective over-sampling method focused on certain types of positive examples. Using two artificial data sets with different dimensions and imbalance ratios, Wojciechowski & Wilk (2017) found out that the critical factor affecting the true-positive rate was the distribution of sample types, while the impact of dimensionality and imbalance ratio was limited. Similarly, Stefanowski (2016) concluded that the performance of the most representative preprocessing approaches depends on the dominating type of minority examples.

2.2. Feature-based data irregularities

Bak & Jensen (2016) studied the imbalance problem concerning the classification of high-dimensional binary data. Blagus & Lusa (2013) observed that SMOTE (Synthetic Minority Oversampling TEchnique) did not alleviate the bias towards the classification in the majority class when the imbalanced data set was also high-dimensional. Wasikowski & Chen (2010) showed that feature selection could tackle the class imbalance problem better than some preprocessing algorithms in high-dimensional data sets.

Tomašev & Mladenić (2013) suggested that minority class hubs might be responsible for most misclassifications of the majority class in high-dimensional imbalanced data sets. Zheng et al. (2004) investigated the usefulness of common feature selection metrics (information gain, chi-square, correlation coefficient, and odds ratios) to handle imbalanced data. Van Hulse & Khoshgoftaar (2009) discussed the effect of noise resulting from the corruption of positive examples, which was the type of noise with most deterioration of the classification performance; moreover, they observed that simple classifiers such as naive Bayes and nearest neighbor were often more robust than more complex models such as support vector machines or random forests.

Zhang et al. (2017) argued that the problems of high-dimensional data and imbalance are intertwined, and therefore they should not be solved separately. Lin & Chen (2013) reported the benefits of using some feature selection algorithm as a previous step to the application of the SMOTE over-sampling technique. Other authors, however, proposed first to resample the data set and then apply a feature selection procedure (Lachheta & Bawa, 2016).

Yin et al. (2013) studied the difficulties of feature selection when applied to high-dimensional imbalanced data with Bayesian learning, and proposed two new algorithms to overcome the drawbacks: one is based on the decomposition of the majority classes into relatively smaller sub-classes, whereas the other one uses the Hellinger distance. Maldonado et al. (2014) proposed a feature selection technique using support vector machine and backward elimination in the context of high-dimensional imbalanced data sets. Viegas et al. (2018) developed a feature selection strategy for high-dimensional skewed data using genetic programming. Shahee & Ananthakumar (2019) introduced a distance-based feature selection method in order to tackle simultaneous occurrence of between-class and within-class imbalance.

3. The resampling techniques

This section presents the resampling algorithms that will be used in the experiments. As pointed out in Section 1, the resampling methods can be grouped into two main categories: under-sampling and over-sampling. In addition, some hybrid techniques combine the general ideas of under- and over-sampling to transform the skewed class distribution into a more balanced distribution. Table 1 summarizes these algorithms, which are briefly described in Appendix A.

Table 1: Summary of resampling algorithms used in the experiments

Strategy	Method	Reference
Under-sampling	Random under-sampling (RUS)	
	Hart's condensing (CNN)	Hart (1968)
	Tomek links (TL)	Tomek (1976)
	One-sided selection (OSS)	Kubat & Matwin (1997)
	Hart's condensing + Tomek links (CNN-TL)	Batista et al. (2004)
	Neighborhood cleaning (NCL) rule	Laurikkala (2001)
	Under-Sampling based on clustering (SBC)	Yen & Lee (2006)
	Class purity maximization (CPM)	Yoon & Kwek (2005)
Over-sampling	Random over-sampling (ROS)	Batista et al. (2004)
	Synthetic minority over-sampling technique (SMOTE)	Chawla et al. (2002)
	Borderline-SMOTE (B-SMOTE)	Han et al. (2005)
	Safe-Level-SMOTE (S-L-SMOTE)	Bunkhumpornpat et al. (2009)
	Nearest centroid neighborhood-based SMOTE (NCN-SMOTE)	García et al. (2012)
	Gabriel graph-based SMOTE (GG-SMOTE)	García et al. (2012)
	Relative neighborhood graph-based SMOTE (RNG-SMOTE)	García et al. (2012)
	Adaptive synthetic over-sampling (ADASYN)	He et al. (2008)
	Adjusting the direction of the synthetic minority class (ADOMS)	Tang & Chen (2008)
	Agglomerative hierarchical clustering (AHC)	Cohen et al. (2006)
Hybrid	SMOTE + Wilson's editing (SMOTE-ENN)	Batista et al. (2004)
	SMOTE + Tomek links (SMOTE-TL)	Batista et al. (2004)
	Selective preprocessing and resampling algorithm (SPIDER)	Stefanowski & Wilk (2008)
	SPIDER extension (SPIDER2)	Napierala et al. (2010)

4. Exploiting local neighborhood for the identification of sample types

When dealing with imbalanced data sets, a remarkable issue that deserves some special attention is the identification of the dominating types of examples because it can support interpretations of performance differences between the application of different resampling algorithms and can be useful in evaluating the data difficulty (Napierala & Stefanowski, 2012; Napierala et al., 2010; Stefanowski, 2016).

Several authors have proposed to distinguish two main types of samples according to their neighborhood: *safe* and *unsafe* (Kubat & Matwin, 1997; Napierala & Stefanowski, 2016; Sáez et al., 2016). The safe samples are placed in homogeneous regions with data from a single class and are sufficiently separated from examples belonging to any other classes, whereas the remaining samples are deemed unsafe. Most models classify the safe samples correctly, but the unsafe samples may make their learning especially difficult and more likely to be misclassified.

The common property of the unsafe samples is that they are located close to examples that belong to the opposite class. However, the unsafe samples can be further divided into three subtypes: *borderline*, *rare* and *outlier* (Krawczyk et al., 2014; Napierala & Stefanowski, 2016). The borderline samples are located closely to the decision boundary between classes. The rare samples form small data structures or clusters located far from the core of their class. Finally, the outliers are single samples that are surrounded by examples from the other class.

A straightforward method to identify each sample type consists of analyzing the local distribution of the data, which can be modeled either by computing their k-neighborhood or through a kernel function (this consists in setting a local area around the example and estimating the number of neighbors and their class labels within it). It has been claimed that analyzing a local distribution of examples is more appropriate than using global approaches because the minority class is often formed by small sub-groups with difficult, nonlinear borders between the classes (Napierala & Stefanowski, 2016; Sáez et al., 2016).

Suppose we have a data set, $S = \{z_i = (x_i, y_i)\}$, where $x_i \in X \subset \mathbb{R}^d$ is a vector of attributes describing the i-th example and y_i is its class label. Thus the type of a sample z_i is often determined by comparing the number of its k nearest neighbors that belong to the class of z_i with the number of neighbors of the opposite class. Following the procedure described in Algorithm 1, which is a generalization for multi-class data of the procedure proposed by Stefanowski & Wilk (2008), a safe sample is characterized by having a neighborhood dominated by examples that belong to its same class, rare samples and outliers are mainly surrounded by examples from different classes, and the borderline samples are surrounded by examples both from their same class and also from a different class. Here we have introduced two functions: *computeNeighbors* and *countSameClass*. The first one searches for the k nearest neighbors of a sample z_i and stores them in a vector named $neighbors$, while the second function counts how many of the k nearest neighbors belong to the class of z_i.

Most authors choose a fixed size of $k = 5$ because smaller values may poorly distinguish the nature of examples and higher values would violate the assumption of the local neighborhood (Bagherpour et al., 2018; Błaszczyński & Stefanowski, 2015; Fernández et al., 2018a; Krawczyk et al., 2014; Napierala et al., 2010; Napierala & Stefanowski, 2012, 2016; Ren et al., 2019; Sáez et al., 2016; Skryjomski & Krawczyk, 2017; Stefanowski, 2016; Tomašev & Mladenić, 2013). Moreover, Napierala & Stefanowski (2016) carried out a sensitivity analysis to check whether or not the parameter k could affect the results of assigning a sample type to the minority examples, and they observed that the proportion of each sample type was quite stable while changing the value of k. Thus, using $k = 5$, an example z_i will be considered as: (i) safe if at least 4 neighbors are from the class y_i; (ii) borderline if 2 or 3 neighbors belong to the class y_i; (iii) rare if only one neighbor belongs to the class y_i, and this has no more than one neighbor from

Algorithm 1 Identification of sample types for multi-class data

```
Input:
$S$ {Input data set}
$k$ {Neighborhood size}

Output:
$safe$ {Set of safe samples}
$borderline$ {Set of borderline samples}
$rare$ {Set of rare samples}
$outlier$ {Set of outlier samples}

for all $z_i \in S$ do
    $neighbors \leftarrow$ computeNeighbors($z_i, S - \{z_i\}, k$)
    $sameClass \leftarrow$ countSameClass($y_i, neighbors$)
    if $sameClass \geq \lfloor 0.8k \rfloor$ then
        $safe \leftarrow safe \cup \{z_i\}$
    else
        if $sameClass \geq \lfloor 0.5k \rfloor$ then
            $borderline \leftarrow borderline \cup \{z_i\}$
        else
            if $sameClass \geq \lfloor 0.2k \rfloor$ then
                $rare \leftarrow rare \cup \{z_i\}$
            else
                $outlier \leftarrow outlier \cup \{z_i\}$
            end if
        end if
    end if
end for
```

its same class; and (iv) outlier if all its neighbors are from the opposite class. A simple, illustrative example of this categorization is displayed in Figure 1.

The identification of the different sample types has mainly been applied to the minority class because this often constitutes the most important class for most applications with imbalanced data sets. However, the percentage of samples in each category for the majority and minority classes may differ heavily from each other and therefore, we believe that it could be useful and more informative to analyze the true distribution of sample types for both classes present in class-

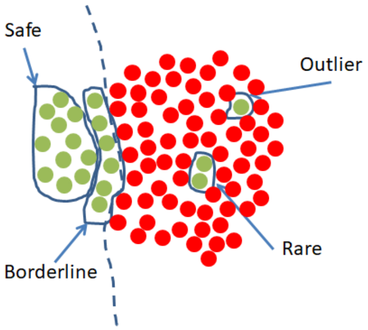


Figure 1: Example of sample types using the procedure given in Algorithm 1

imbalanced data. In this sense, the computation may resemble a means of data set evaluation that characterizes the overlap in terms of a scalar value. Considering that the class overlapping is defined as the data space where there exists a similar quantity of training samples of both classes (Chen et al., 2018; López et al., 2013), we argue that the presence of borderline samples (2 or 3 out of the 5-nearest neighbors belong to the same class) is closely related to the concept of overlapping and therefore, it seems possible to estimate the size of the overlapping regions by computing the proportion of borderline samples in a data set.

5. Experiments

Two groups of experiments on binary problems were carried out to investigate the effect of each of the three resampling strategies on the distribution of sample types in both classes, and also to discover any possible link between such distribution and the classification performance. The experiments in the first block were performed on artificial data sets taken from the paper by Napierala et al. (2010) because using synthetic data allows us to know their characteristics a priori and analyze the effects of resampling in a fully controlled environment. The second group of experiments was on a well-known benchmark suite of real-life databases widely used for class imbalance problems (Chen et al., 2019; Jing et al., 2019; Kovács, 2019; Kuncheva et al., 2019; Lopez-Garcia et al., 2019), which are all available at the KEEL database repository (Alcalá-Fdez et al., 2011). The results of both experiments were estimated by 5-fold stratified cross-validation in order to have a sufficient amount of positive examples in the test partitions.

In binary classification problems, the quite common method for evaluating the predictive performance is based on a 2 × 2 matrix confusion as shown in Table 2. Here, columns represent the predicted class and rows indicate the actual class, whereas the main diagonal contains the number of correct predictions. For estimating the effectiveness of a classifier on the positive and negative classes separately, two plain metrics can be easily obtained: the true positive rate, $TPR = TP/(TP+FN)$, which is the proportion of positive examples correctly classified, and the true negative rate, $TNR = TN/(TN + FP)$, which is the proportion of negative examples correctly classified.

Table 2: Confusion matrix for a two-class problem.

	Predicted positive (R_p)	*Predicted negative* (R_n)
Positive class (N_p)	True Positive (TP)	False Negative (FN)
Negative class (N_n)	False Positive (FP)	True Negative (TN)

In the context of class imbalance problem, the performance evaluation is carried out using more powerful metrics derived from straightforward indexes. Some examples are the geometric mean (Kubat & Matwin, 1997; Branco et al., 2016; Fernández et al., 2018b), the $F_\beta-$measure (Rijsbergen, 1979; Branco et al., 2016; Fernández et al., 2018b), and the area under the receiver operating characteristic curve (AUC) (Bradley, 1997; Branco et al., 2016; Fernández et al., 2018b). Although these performance metrics are used extensively under imbalanced domains, several studies have shown the limitations of these measures.

García et al. (2014) have documented that the geometric mean shows an invariance behavior under the change of TP with TN and FN with FP. Therefore, different combinations of TPR and TNR may produce the same values of the geometric mean. The $F_\beta-$measure combines into a single scalar value both TPR and precision ($precision = TP/TP + FP$), where the β parameter favors precision when $\beta > 1$, and TPR otherwise. Even though β allows to adjust the importance of TPR or precision, the studies of Daskalaki et al. (2006), Japkowicz (2006), Sokolova & Lapalme (2009), and Landgrebe et al. (2006) have showed that precision ignores the relative size of the negative class and displays a strong dependence upon the imbalance ratio; hence, in heavily imbalance problems (1% positives samples), any raise of FP will result in low precision and consequently, in low $F_\beta-$measure, even with high TPR values (Forman & Scholz, 2010). In the case of AUC, there may exist situations that produce the same AUC value but different accuracies (Huang & Ling, 2005). Hand & Till (2001) and Hand (2009) also have reported some limitations of the AUC such as the fact that it ignores misclassification costs and assumes that these costs depend on the classifier.

Bearing in mind that this paper aims to analyze the effects of the resampling methods on each class and that each performance measure evaluates different properties, here we will use the both straightforward TPR and TNR indexes.

5.1. Experiments with artificial data sets

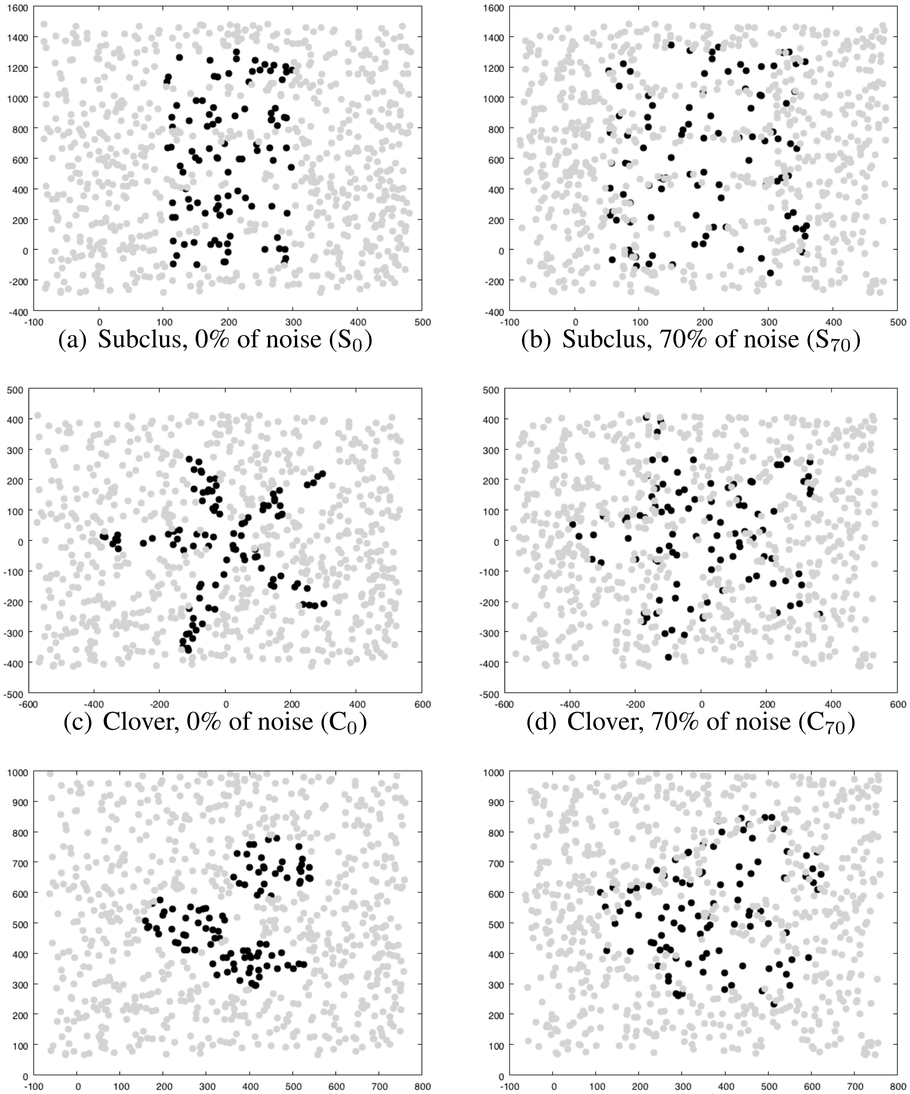
The experiments on artificial data were conducted on three databases with different shapes of the minority class (subclus, clover, and paw) whose examples are randomly and uniformly distributed in a two-dimensional feature space. In all cases, the examples of the minority class are uniformly surrounded by the majority class.

In subclus, the positive examples are located inside rectangles that form small disjuncts. Clover represents a more complex, non-linear situation, where the minority class resembles a flower with elliptic petals. In paw database, the minority class is decomposed into three elliptic sub-regions of varying cardinalities, where two sub-regions are located close to each other, and the remaining smaller sub-region is separated.

From the multiple data sets that were generated with different settings in the original paper (Napierala et al., 2010), we chose a group of databases with 800 examples, an imbalance ratio of 7, and two different levels of noise (0% and 70%). This means that the experiments were carried out over a total of 6 artificial data sets (3 shapes $\times$ 1 imbalance ratio $\times$ 2 levels of noise), which are illustrated in Figure 2.

The experiments consisted of applying the resampling techniques described in Section 3 to the original data sets and record the proportion of each sample type for both the minority class and the majority class. This will allow to analyze how each strategy affects the distribution of sample types in a data set, which may contribute to gain some insight into the behavior of these techniques when they are used in imbalanced data sets that are also characterized by other data difficulties, such as the presence of noisy samples that can largely impair the predictive results of classifiers.

Figures 3 and 5 display the proportion of safe, borderline, rare and outlier samples in the positive and negative classes after resampling the imbalanced data sets, respectively. The results that correspond to the under-sampling algorithms (U) are represented by red squares, those from the over-sampling methods (O) are indicated by blue circles, and the proportions given by the hybrid techniques (H) are depicted by green triangles. The black markers are for the proportions in the original (imbalanced) data sets with no resampling (I), which should be



(a) Subclus, 0% of noise (S_0) (b) Subclus, 70% of noise (S_{70})

(c) Clover, 0% of noise (C_0) (d) Clover, 70% of noise (C_{70})

(e) Paw, 0% of noise (P_0) (f) Paw, 70% of noise (P_{70})

Figure 2: The artificial data sets

interpreted as a reference value. Note that each scatterplot has a total of 138 points: (8 under-sampling algorithms + 10 over-sampling algorithms + 4 hybrid algorithms + 1 no resampling) × 6 databases. Moreover, Figures 4 and 6 show the proportions averaged over all algorithms of each resampling strategy.

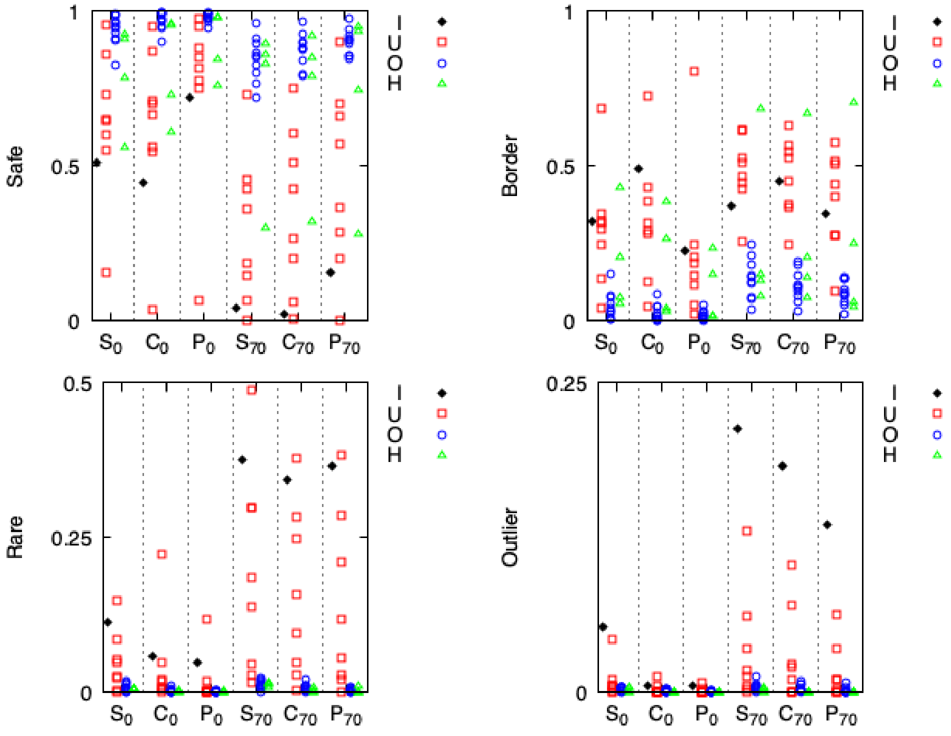


Figure 3: Proportion of sample types in the positive class for the artificial databases

These scatterplots reveal that the three resampling strategies increased the proportion of safe samples and decreased the percentage of unsafe samples in the positive class when compared to the imbalanced data sets. What is more interesting though is that, for the databases with 70% of noise (S_{70}, C_{70} and P_{70}), the over-sampling and hybrid techniques achieve a higher (lower) proportion of safe (unsafe) samples than under-sampling. While all over-sampling algorithms augmented the number of safe samples and diminished the number of unsafe samples very substantially, the proportions of safe, borderline and rare samples produced by some under-sampling methods were even worse than those in the original data sets. Regarding the hybrid techniques, these and over-sampling were not distant, except in the case of the proportion of safe and borderline samples given by SPIDER whose results were similar to those achieved by the under-sampling strategy.

When analyzing the proportion of each sample type in the negative class, the graphs in Figures 5–6 show that the proportion of safe samples after resampling

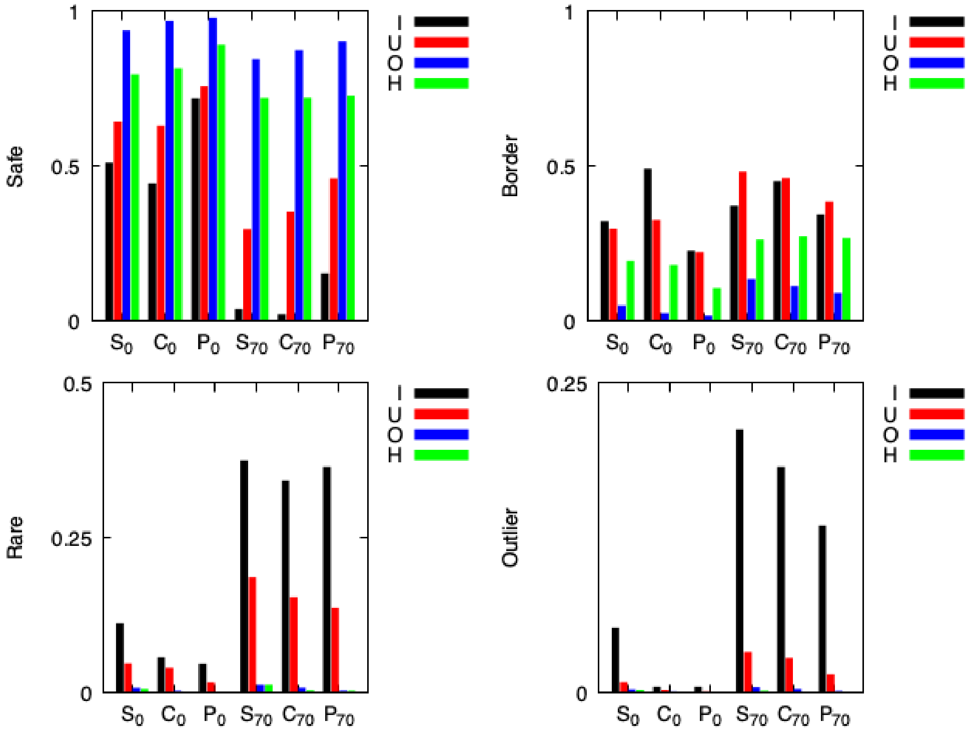


Figure 4: Average proportion of sample types in the positive class for the artificial databases

the data sets using the over-sampling and hybrid algorithms was not far from that in the original data sets. Here the proportion of unsafe samples produced by these methods increased, especially in the case of the rare and outlier types. The under-sampling techniques usually performed in an unstable behavior, serious decrease of safe samples and an evident increase of borderline and rare.

5.1.1. Classification of the artificial data sets

The results of the experiments on the proportion of each sample type identified under-sampling as an inferior choice to make up for the class imbalance, especially for the data sets with a large proportion of noisy examples (70%). This resonates well with the general conclusions drawn from numerous comparative studies available in the literature, which designate over-sampling as a usually more effective strategy than under-sampling.

To fairly assess whether or not there exists any link between the proportions of

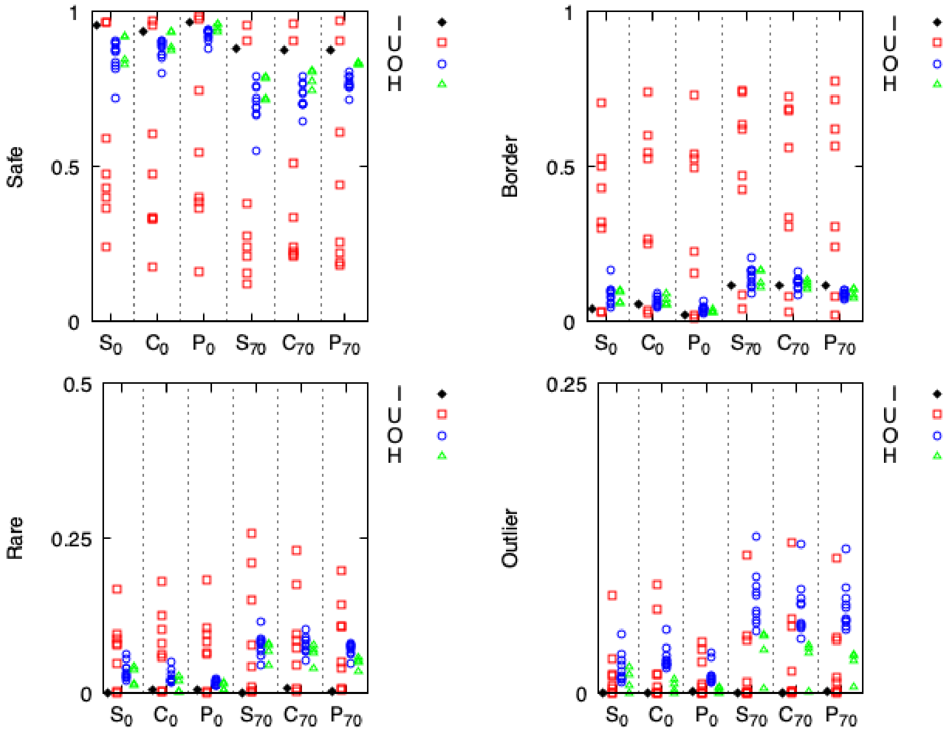


Figure 5: Proportion of sample types in the negative class for the artificial databases

safe and unsafe samples and the classification performance, a C4.5 decision tree classifier was applied to both the original (imbalanced) data sets and the collection of resampled (balanced) data sets. We chose a decision tree because it is a common learner when dealing with class-imbalanced data (Boonchuay et al., 2017; Lee, 2019; Sanz et al., 2017; Sardari et al., 2017; Sun et al., 2018). Besides C4.5 as a decision tree provides an accurate and easily interpretable model where the classification decisions can be represented in the form of "if-then" rules (Quinlan, 1993; Witten et al., 2016), while other classifiers such as neural networks are generally perceived as being a black box whose specific predictions are extremely hard to understand.

Visualizing TPR and TNR in Figure 7 and comparing these graphs with those in Figures 3 and 5 can help us discover and interpret the possible relationships between the structure of resampled data sets and the performance of the classifier.

Our discussion of Figure 7 focuses on the results over the data sets with 70%

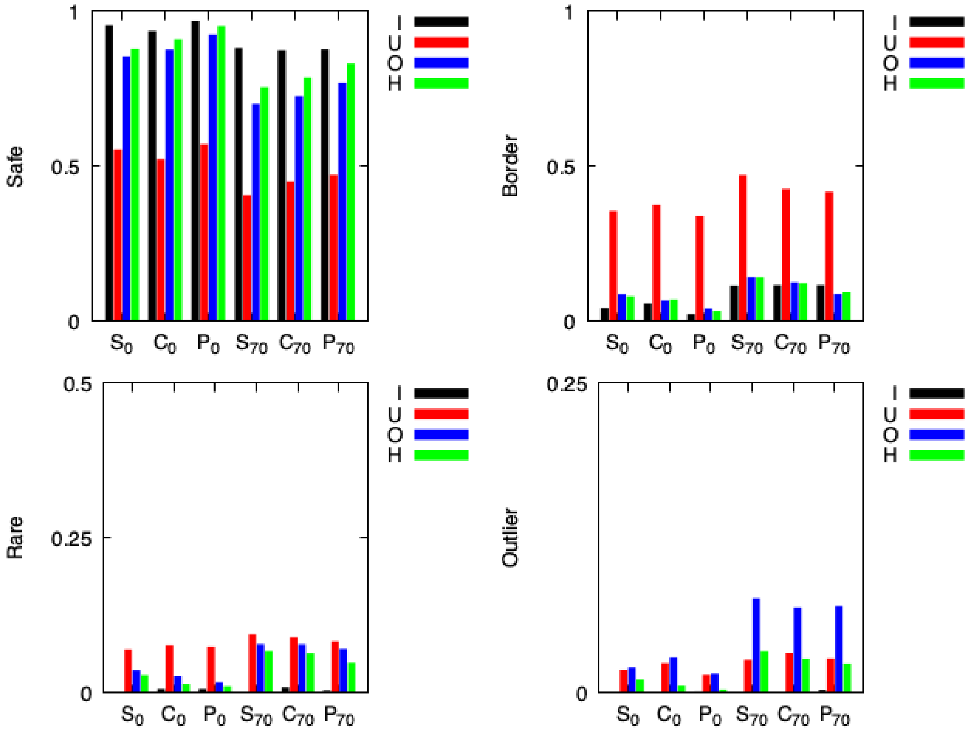


Figure 6: Average proportion of sample types in the negative class for the artificial databases

of noisy examples because these represent a more challenging problem combining imbalance and noise. As can be observed, when the classifier was applied to the class-imbalanced data, the TPR was 0 or close to 0 (i.e., all or almost all the positive samples were misclassified) and the TNR was equal to 1 (i.e., all negative samples were classified correctly). The most interesting feature of these graphs, however, is that both over-sampling and the hybrid sampling algorithms exhibited a good trade-off between high TPR and high TNR, whereas some under-sampling techniques produced high TNR but at the cost of yielding very low values of TPR (even less than 0.5).

In summary, the graphs in Figure 7 confirm that the performance of classifiers is related to the proportions of safe and unsafe samples, and these depend on the resampling strategy applied to the class-imbalanced data. A qualitative comparison between these graphs and those in Figures 3–6 suggests that over-sampling mostly performs better than under-sampling because the former increases the pro-

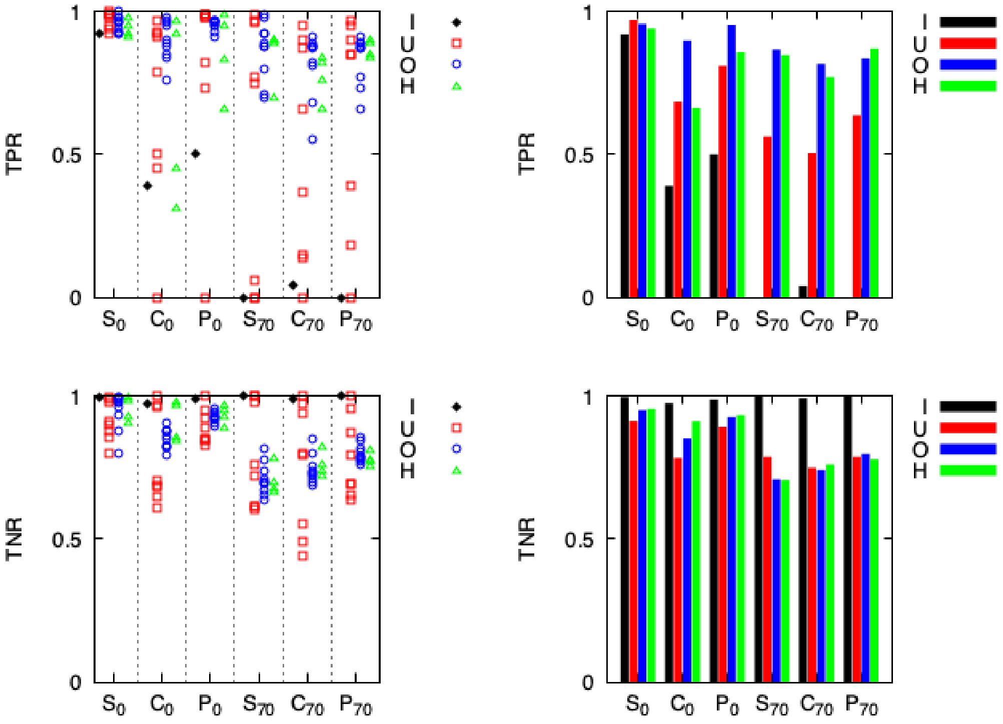


Figure 7: TPR and TNR over the artificial databases. Graphs on the right are for the averaged values

portion of safe samples and also decrease the proportion of unsafe samples much more important than the latter does.

5.2. Experiments with real-life data sets

The objective of the second series of experiments was to check whether or not the behavior of the resampling strategies over real-life data agrees with the results discussed in the previous experiments over artificial data sets. In total, we chose a collection of 73 databases, which correspond to lowly, mildly and highly imbalanced data sets. Table 3 summarizes the number of features (F), the number of examples (E) and the imbalance ratio (IR) for each database; the data sets are placed in ascending order of imbalance ratio.

Figures 8–10 plot the proportion of each sample type in the positive and negative classes for both the resampled data sets and the original data sets. The data

Table 3: Summary of the real-life data sets

	F	E	IR		F	E	IR
glass1	9	214	1.82	ecoli-0-6-7_vs_5	6	220	10.00
ecoli-0_vs_1	7	220	1.86	glass-0-1-6_vs_2	9	192	10.29
wisconsin	9	683	1.86	ecoli-0-1-4-7_vs_2-3-5-6	7	336	10.59
pima	8	768	1.87	led7digit-0-2-4-5-6-7-8-9_vs_1	7	443	10.97
iris0	4	150	2.00	glass-0-6_vs_5	9	108	11.00
glass0	9	214	2.06	ecoli-0-1_vs_5	6	240	11.00
yeast1	8	1484	2.46	glass-0-1-4-6_vs_2	9	205	11.06
haberman	3	306	2.78	glass2	9	214	11.59
vehicle2	18	846	2.88	ecoli-0-1-4-7_vs_5-6	6	332	12.28
vehicle1	18	846	2.90	cleveland-0_vs_4	13	177	12.62
vehicle3	18	846	2.99	ecoli-0-1-4-6_vs_5	6	280	13.00
glass-0-1-2-3_vs_4-5-6	9	214	3.20	shuttle-c0_vs_c4	9	1829	13.87
vehicle0	18	846	3.25	yeast-1_vs_7	7	459	14.30
ecoli1	7	336	3.36	glass4	9	214	15.47
new-thyroid1	5	215	5.14	ecoli4	7	336	15.80
new-thyroid2	5	215	5.14	page-blocks-1-3_vs_4	10	472	15.86
ecoli2	7	336	5.46	dermatology-6	34	358	16.90
segment0	19	2308	6.02	glass-0-1-6_vs_5	9	184	19.44
glass6	9	214	6.38	shuttle-6_vs_2-3	9	230	22.00
yeast3	8	1484	8.10	yeast-1-4-5-8_vs_7	8	693	22.10
ecoli3	7	336	8.60	glass5	9	214	22.78
page-blocks0	10	5472	8.79	yeast-2_vs_8	8	482	23.10
ecoli-0-3-4_vs_5	7	200	9.00	yeast4	8	1484	28.10
yeast-2_vs_4	8	514	9.08	winequality-red-4	11	1599	29.17
ecoli-0-6-7_vs_3-5	7	222	9.09	poker-9_vs_7	10	244	29.50
ecoli-0-2-3-4_vs_5	7	202	9.10	yeast-1-2-8-9_vs_7	8	947	30.57
glass-0-1-5_vs_2	9	172	9.12	yeast5	8	1484	32.73
yeast-0-3-5-9_vs_7-8	8	506	9.12	winequality-red-8_vs_6	11	656	35.44
yeast-0-2-5-7-9_vs_3-6-8	8	1004	9.14	yeast6	8	1484	41.40
yeast-0-2-5-6_vs_3-7-8-9	8	1004	9.14	winequality-white-3_vs_7	11	900	44.00
ecoli-0-4-6_vs_5	6	203	9.15	winequality-white-3-9_vs_5	11	1482	58.28
ecoli-0-1_vs_2-3-5	7	244	9.17	poker-8-9_vs_6	10	1485	58.40
ecoli-0-2-6-7_vs_3-5	7	224	9.18	shuttle-2_vs_5	9	3316	66.67
glass-0-4_vs_5	9	92	9.22	winequality-red-3_vs_5	11	691	68.10
ecoli-0-3-4-6_vs_5	7	205	9.25	poker-8-9_vs_5	10	2075	82.00
ecoli-0-3-4-7_vs_5-6	7	257	9.28	poker-8_vs_6	10	1477	85.88
vowel0	13	988	9.98				

sets on the axis X are arranged in ascending order of the proportion of safe samples in the positive class. For the sake of clarity, the results of each resampling strategy have been plotted in a different graph. In this case, we have a total number of 657 points that correspond to under-sampling (73 databases × (8 under-sampling algorithms + 1 no resampling)), 803 points to over-sampling (73 databases × (10 over-sampling algorithms + 1 no resampling)) and 365 points to hybrid sampling (73 databases × (4 hybrid algorithms + 1 no resampling)).

A close look at these scatterplots shows that the discussion of the results for

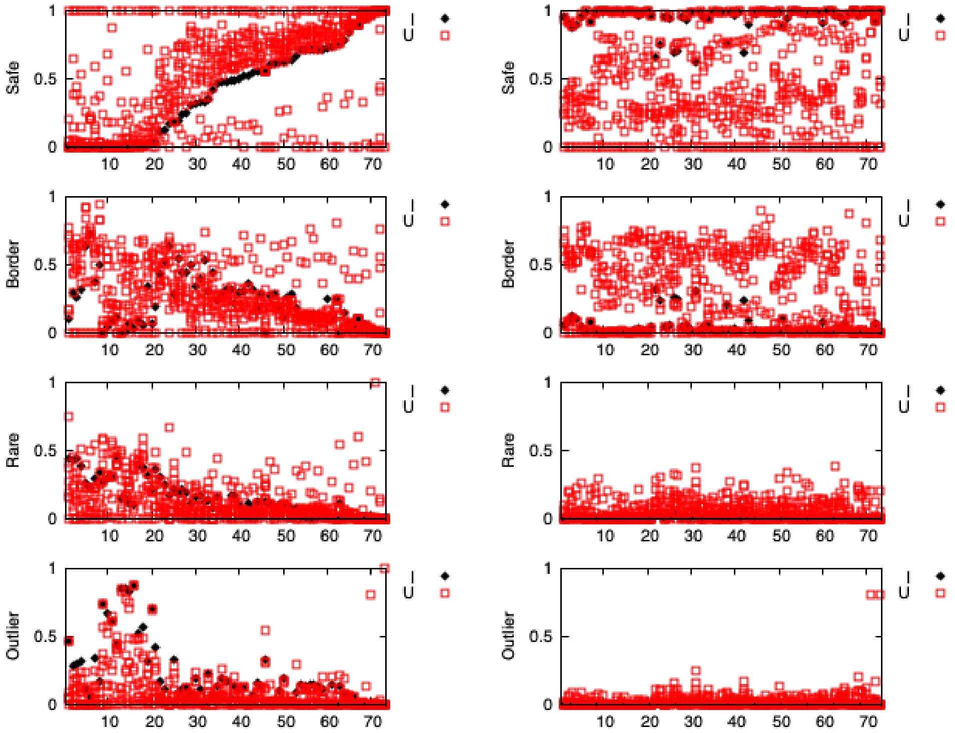


Figure 8: Proportion of sample types in the positive (left) and negative (right) classes for the real-life databases preprocessed by under-sampling algorithms

the synthetic data also apply to those for the real-life databases. Indeed, as one can observe in Figure 8, the proportion of safe samples in many sets that were preprocessed by some under-sampling algorithms was even inferior to that in the original data sets. Similarly, the amount of unsafe samples in many under-sampled data sets was greater than that in the original data sets. As to over-sampling (Figure 9) and the hybrid strategy (Figure 10), the graphs show that most algorithms increased the number of safe samples and also decreased the proportion of unsafe samples, which is especially remarkable for the positive class.

To summarize the results of the graphs in Figures 8–10, we averaged the proportions of each sample type over all algorithms for each resampling strategy. The most interesting features of the graphs depicted in Figure 11 is that under-sampling produced a proportion of safe samples in both classes clearly lower than over-sampling and hybrid sampling, whereas the amount of unsafe samples was

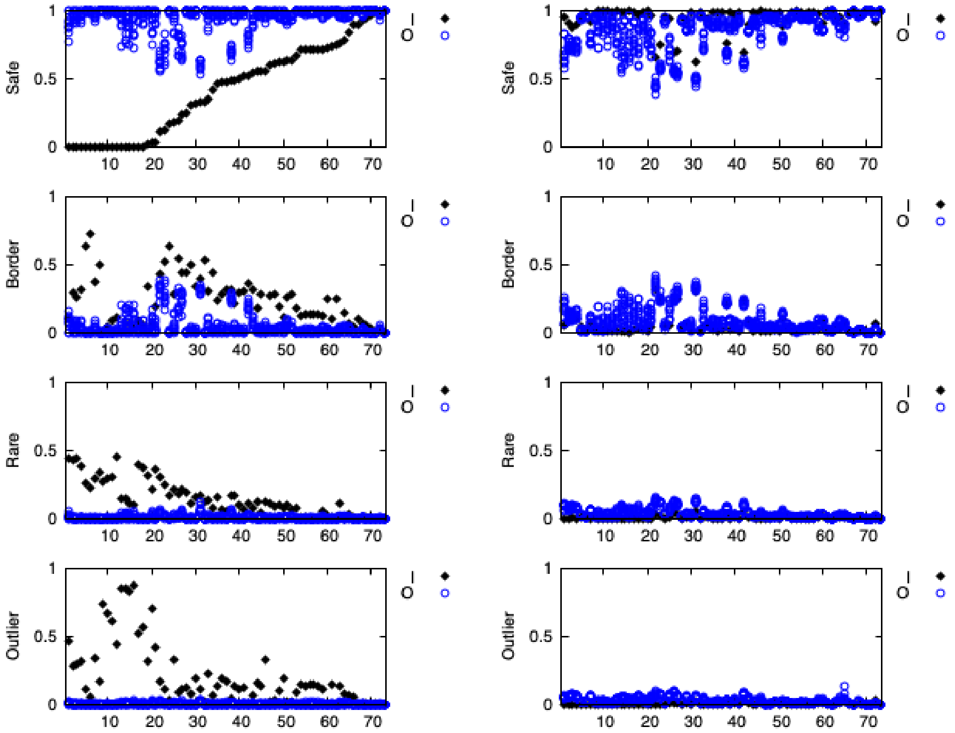


Figure 9: Proportion of sample types in the positive (left) and negative (right) classes for the real-life databases preprocessed by over-sampling algorithms

higher in the under-sampled sets than in the sets preprocessed by the other two resampling strategies.

As further evidence, Table 4 reports an index of improvement. For each resampling algorithm A, the index of improvement is calculated as the difference between *wins* and *losses*, where *wins* (*losses*) is the total number of times (databases) that the proportion of samples produced by A has been better (worse) than that in the original data set. Note that *better* means that the proportion of safe samples in the resampled data set is higher than that in the original data set, while for the unsafe sample types it means that the proportion of samples in the resampled data set is lower than that in the original data set. Such an index provides a means of estimating the benefits of using a resampling technique to face the imbalance problem. For each resampling strategy, the averaged index across all their algorithms has also been included in this table.

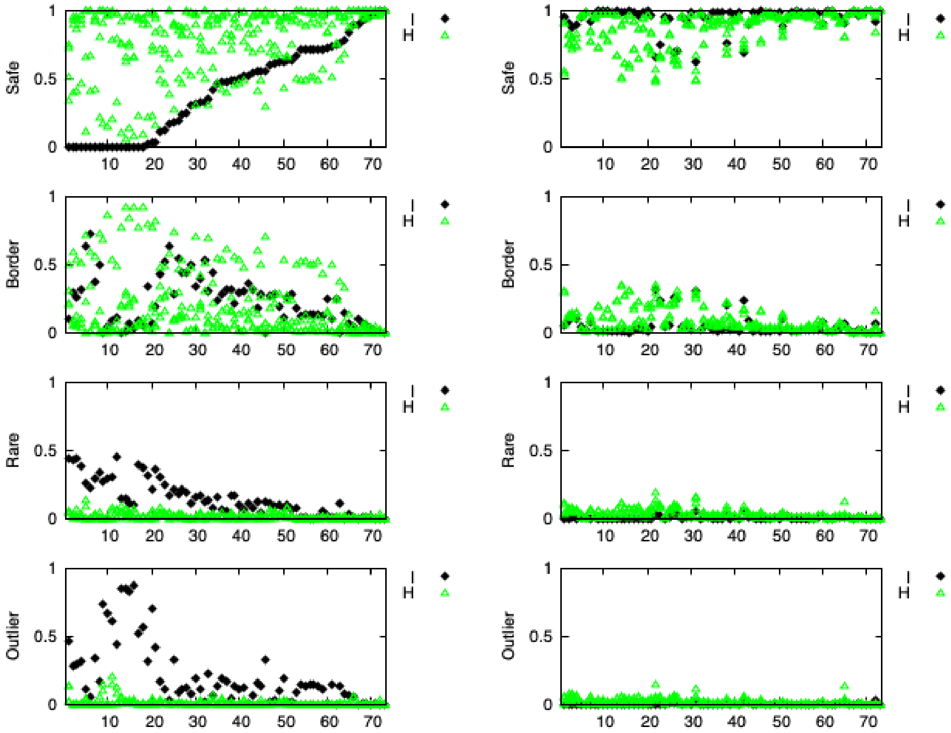


Figure 10: Proportion of sample types in the positive (left) and negative (right) classes for the real-life databases preprocessed by hybrid algorithms

Regarding the positive class, the index of improvement reported in Table 4 shows that the over-sampling strategy produced the best outputs for the safe and borderline types, whereas the hybrid methods achieved the highest averaged index when analyzing the proportion of rare and outlier samples. Nevertheless, the superiority of the hybrid techniques over the over-sampling methods came from the poor behavior of the AHC algorithm in processing the unsafe samples. As already observed in the experiments with synthetic data, under-sampling was the worst strategy regarding the improvement of the balanced data over the original (imbalanced) data, revealing that it yielded the lowest proportion of safe samples and also the highest proportion of unsafe samples.

For the majority class, most algorithms achieved a negative score of the index of improvement, which means that the balanced data sets consist of less safe samples and more unsafe samples than the original data sets. Note that this result is

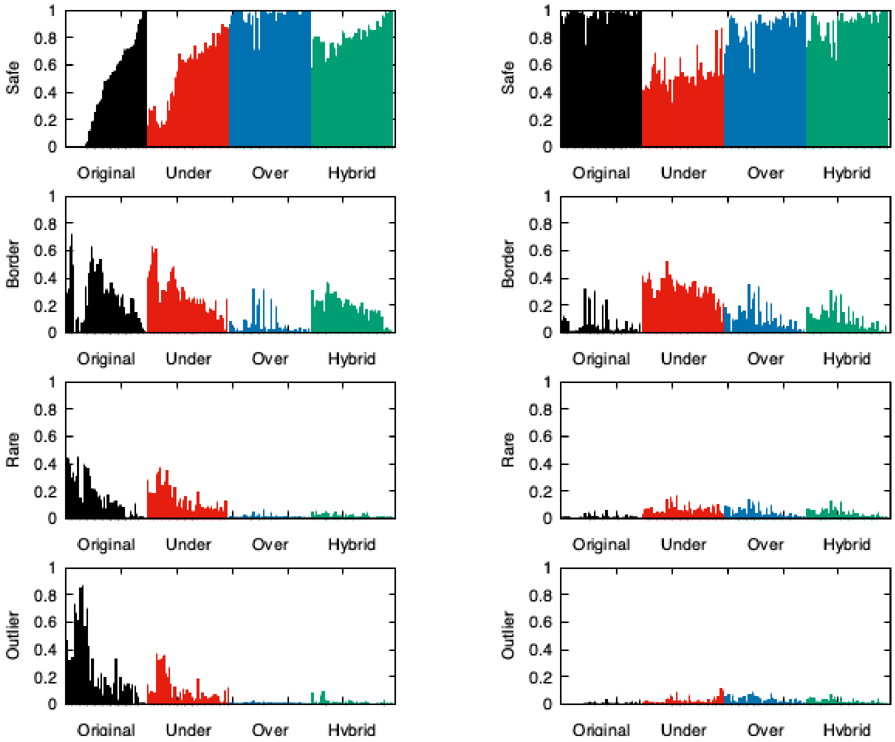


Figure 11: Proportion of sample types in the positive (left) and negative (right) classes for the real-life databases averaged over all algorithms

consistent with the ultimate objective of the resampling techniques as they mainly concentrate on improving the minority class.

In summary, the numerical indices of improvement agree with the results depicted in the scatterplots of Figures 8–10. On the other hand, the conclusions drawn from the experiments over the real-life data closely resemble those reached in the experiments over the synthetic databases.

5.2.1. Classification of the real-life data sets

Like in the experiments on the synthetic data, a C4.5 decision tree was applied to both the imbalanced and the resampled data sets to check for any link between the proportions of sample types and the resulting classification performance.

As the characteristics of the 73 experimental databases may differ from each other considerably, we firstly categorized them into three groups according to the

Table 4: Index of improvement

	Positive class				Negative class			
	Safe	Border	Rare	Outlier	Safe	Border	Rare	Outlier
RUS	69	15	39	66	-68	-66	-66	-57
OSS	64	10	21	68	-70	-70	-65	-38
CNN	62	-31	17	68	-69	-69	-67	-38
TL	41	15	20	43	54	56	13	2
CNN-TL	65	12	38	67	-70	-70	-68	-58
NCL	52	27	25	54	52	52	-5	0
CPM	-53	-52	-65	15	-73	-71	-52	-19
SBC	70	59	64	67	-71	58	24	13
Under-sampling	46	7	20	56	-39	-23	-36	-24
ROS	70	65	66	68	-67	-20	-68	-66
ADASYN	70	53	65	67	-69	-44	-69	-67
ADOMS	70	57	62	68	-69	-59	-67	-66
AHC	73	-70	-62	-67	73	-70	-70	-67
SMOTE	70	53	63	68	-69	-44	-66	-65
B-SMOTE	70	62	64	68	-69	-22	-68	-66
S-L-SMOTE	70	66	64	68	-68	-32	-67	-66
NCN-SMOTE	70	50	62	68	-68	-49	-65	-65
GG-SMOTE	70	50	62	68	-68	-50	-65	-65
RNG-SMOTE	70	55	64	68	-68	-44	-66	-65
Over-sampling	70	44	51	54	-54	-43	-67	-66
SMOTE-ENN	71	51	63	67	-52	-36	-59	-56
SMOTE-TL	70	52	63	68	-56	-36	-65	-53
SPIDER	20	-38	63	67	-54	-36	-61	-18
SPIDER2	66	53	61	62	-56	-28	-59	-55
Hybrid	57	30	63	66	-55	-34	-61	-46

prevalent type of positive samples in the original data sets (see Appendix B): safe, borderline, and rare-outlier (databases in which the positive samples are mainly placed between the rare and the outlier types). The purpose of this categorization was to better understand the behavior of the resampling strategies as a function of the distribution of sample types in the imbalanced data sets.

The scatterplots of TPR versus TNR are displayed in Figures 12–14. The uppermost graphs correspond to the results achieved with under-sampling, the middle ones are for the over-sampling algorithms, and the lowermost ones are for the hybrid methods.

The graphs in this figure reveal that the over-sampling algorithms and the hybrid sampling methods performed similarly, irrespective of the prevalent type of positive samples. As already highlighted in the experiments on artificial data, both strategies led to a good trade-off between high TPR and high TNR with

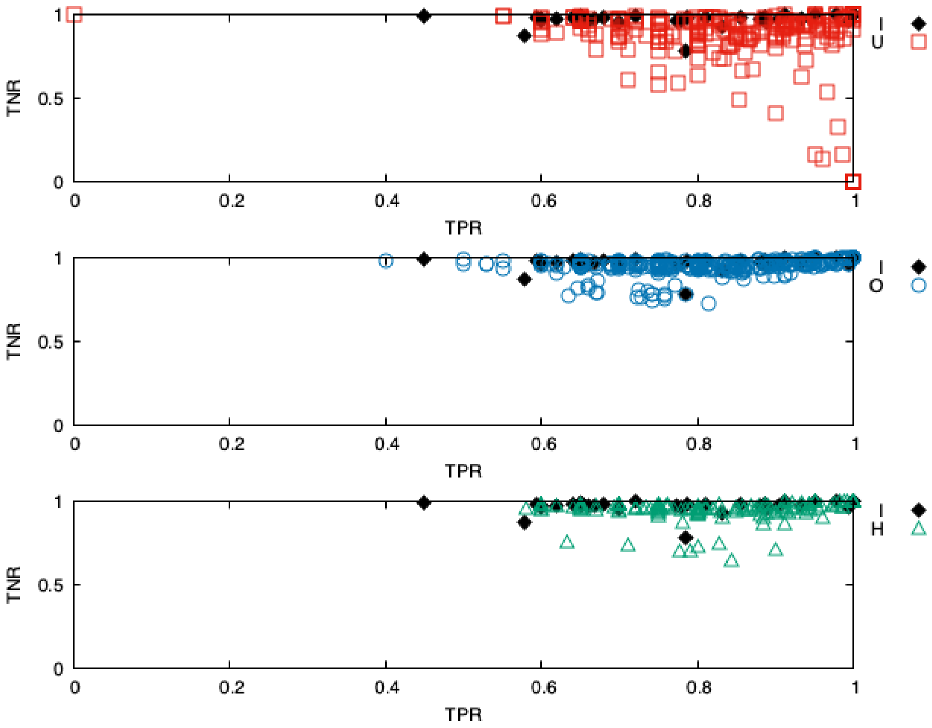


Figure 12: TPR versus TNR over the safe data sets

points located to the top right corner of the scatterplots for the safe and borderline databases. We observed, however, a different behavior pattern for the rare-outlier databases: in this case, the over-sampling and hybrid techniques still achieved very high values of TNR, but also an important degradation of accuracy on the positive class with a majority of points lying on the left side of the graphs (i.e., TPR ≤ 0.5).

Regarding the scatterplots for the under-sampling strategy, we found pretty different behaviors among methods. For the safe databases, a majority of points are located close to the top right corner of the graph (high TPR and high TNR), but a few points lie near the bottom right corner (high TPR and very low TNR). This behavior was similar to that shown for the borderline databases, although both TPR and TNR were usually lower than those achieved for the safe databases. For the rare-outlier databases, one can see that the results of under-sampling were worst than those of the over-sampling and hybrid algorithms, with many points

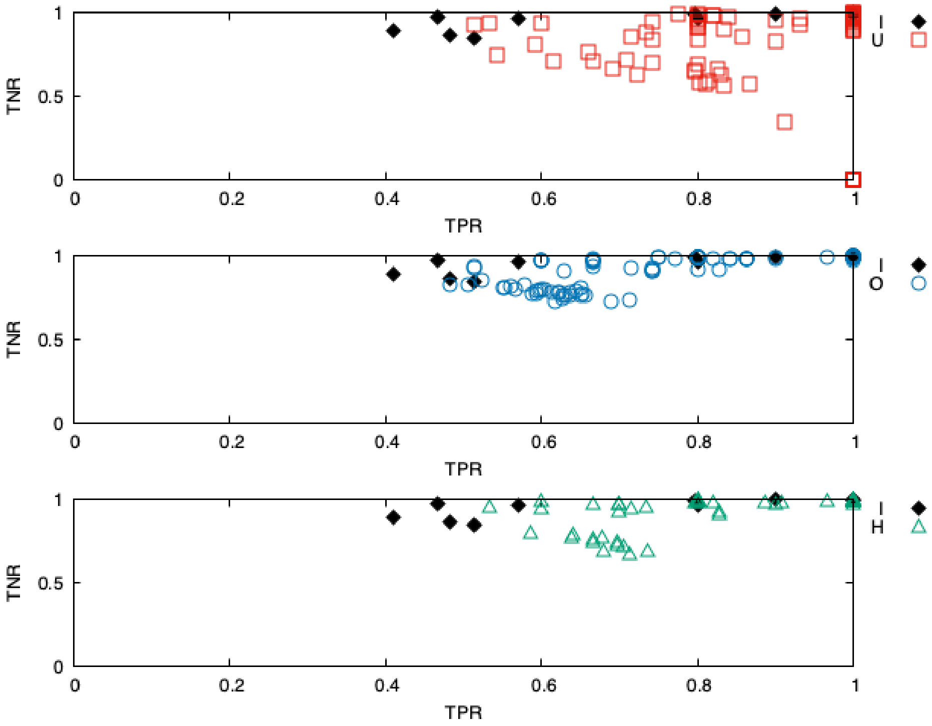


Figure 13: TPR versus TNR over the borderline data sets

representing low TPR and low TNR.

In summary, these results reveal that there exist several links between the distribution of sample types produced by the resampling strategies and the classification performance, thus suggesting that the analysis of such a distribution is indeed a useful tool to understand the behavior of each preprocessing method. In general, the over-sampling and hybrid techniques can be claimed to be more effective than under-sampling, independently of the prevalent type of positive samples in the imbalanced data set. However, the most meaningful differences appeared when under-sampling was applied to the databases with a majority of rare and outlier samples, which correspond to the most difficult cases for standard classifiers.

6. Conclusions

Our motivation for this work came from the observation that many studies on class imbalance stated that over-sampling mostly performs better than under-

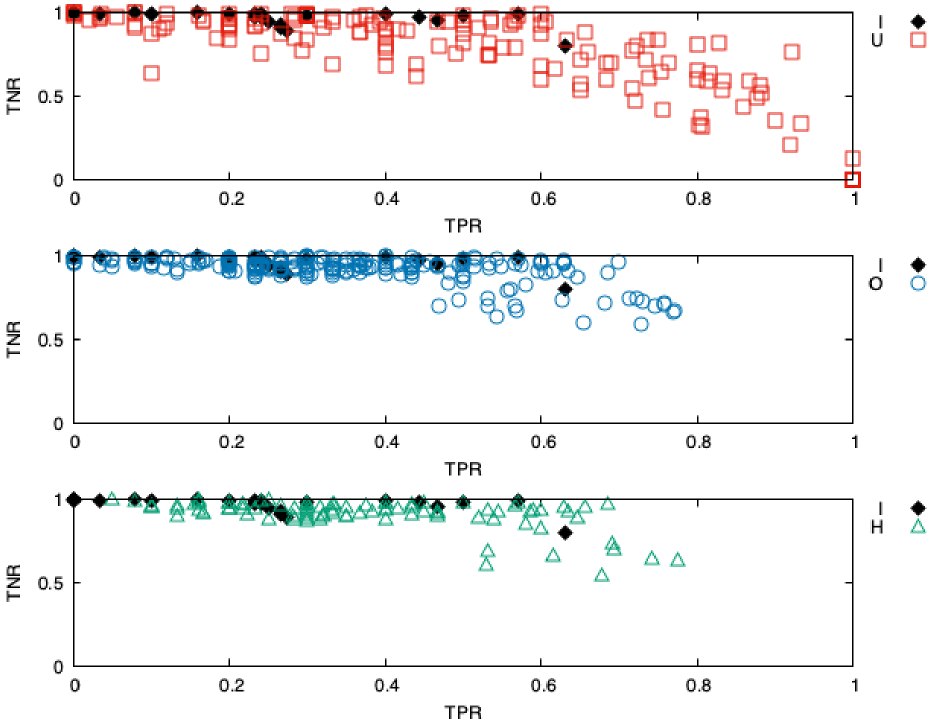


Figure 14: TPR versus TNR over the rare-outlier data sets

sampling, but the reasons for its superiority were not adequately addressed. Thus we have intended to increase understanding of the behavior of resampling strategies by analyzing the distribution of sample types in the balanced data sets. Our hypothesis was that the apparent superiority of over-sampling techniques comes from the fact that these provide a higher proportion of safe samples and a lower proportion of some subtypes of unsafe samples than the under-sampling methods.

The experiments to check whether or not our hypothesis holds have consisted in gathering the information related to the local neighborhood of both classes, calculating the proportions of each sample type and investigating for any links between these proportions and the classification performance of a decision tree. From the experiments over artificial and real-life data, we have found that the over-sampling algorithms and the hybrid resampling methods increased the proportion of safe samples and also diminished the proportion of unsafe samples much more importantly than under-sampling did. We claim that this result is already important by itself because it suggests that classification with over-sampled data sets will be presumably easier and more effective than using under-sampled data sets.

When compared the resulting distribution of sample types with the classification performance measured by the true-positive and true-negative rates, we have observed that our hypothesis mostly holds. In general, the strategies with the highest proportion of safe samples and the lowest proportion of unsafe samples corresponded to those with the highest overall performance, which may indicate that there are some relationships between the proportions of safe and unsafe samples and the performance of the classifier.

We believe that the findings of this study can be of interest for the research community in expert and intelligent systems because it allows to gain a more in-depth insight into the performance of resampling strategies for class-imbalanced data and expands the current knowledge about why over-sampling performs generally better than under-sampling. On the other hand, the conclusions drawn in this paper could provide support for the development of new preprocessing algorithms by incorporating some a priori knowledge about the internal structure of the imbalanced data sets. Another practical implication that could deserve to be further studied is the design of a meta-learning recommendation system for characterizing classification problems. This is based on the idea of using the categorization of examples as a means to guess the best performing algorithm according to the inner structure of each data set.

Despite its contributions, the results of this paper should not be interpreted without accounting some limitations that could be addressed in future works. First, the research has focused on the analysis of relatively small-sized data sets (at most 5472 examples and 34 features), and so any generalization is limited to this particular context. It would be useful to replicate this study when the number of examples is in the order of millions to billions and the number of features is in the order of thousands, where the boundary conditions are very different and much more complex. A second limitation is that the categorization of examples has been based on computing their k-neighborhood, but it would be worth comparing the results of this study with those given by the use of a kernel function. Finally, the emphasis of this paper has been on three common resampling strategies, but it could be extended to ensemble-based preprocessing methods such as RUSBoost (Seiffert et al., 2010), SMOTEBoost (Chawla et al., 2003), EasyEnsemble (Liu et al., 2009) and SMOTEBagging (Wang & Yao, 2009), which have been shown to be among the most effective techniques in many real-life applications.

Acknowledgments

This work was partially supported by the Generalitat Valenciana [PROMETEOII/2014/062]; and the Universitat Jaume I [UJI-B2018-49].

Appendix A. Resampling methods

This appendix provides a brief description of the resampling algorithms used in the experiments.

Appendix A.1. Under-sampling

Random under-sampling (RUS) balances the data set through the random removal of negative examples. Although important information can be lost when examples are discarded at random, this algorithm has empirically been shown to be one of the most effective under-sampling methods.

Many other under-sampling proposals are based on a more intelligent selection of the negative examples to be eliminated. For instance, the Hart's condensing (CNN) algorithm (Hart, 1968) has been used as an under-sampling technique by applying the concept of consistent subset to eliminate the negative examples that are sufficiently far away from the decision boundary because these examples can be considered irrelevant for learning. Analogously, the Tomek links (TL) (Tomek, 1976) have already been employed to remove the majority class examples since, if two examples form a Tomek link, then either one of these examples is noise or both examples are borderline.

Kubat & Matwin (1997) proposed the one-sided selection (OSS) technique, which selectively removes only those negative samples that either are redundant or border the minority class examples (assuming that these bordering cases are noise): the borderline examples are detected using the Tomek links, while the redundant ones are eliminated with Hart's condensing. A similar method corresponds to the CNN-TL algorithm (Batista et al., 2004), which firstly finds a consistent subset and then applies the procedure based on the Tomek links.

Unlike the one-sided selection technique, the neighborhood cleaning (NCL) rule (Laurikkala, 2001) concentrates more on data filtering than on data reduction; to this end, Wilson's editing (ENN) (Wilson, 1972) is employed to identify and remove noisy negative examples. According to the authors, NCL performs better than OSS and processes noisy examples more carefully. However, this method is strongly biased in favor of the minority class and leads to poor specificity and overall accuracy.

Yen & Lee (2006) presented an under-sampling algorithm based on clustering (SBC): it first clusters all the original examples into some clusters, and then selects an appropriate number of majority class samples from each cluster by considering the ratio of the number of majority class examples to the number of minority class examples in the cluster. On the other hand, Yoon & Kwek (2005) proposed the class purity maximization (CPM) algorithm, which intends to split the majority class into dense clusters. The idea is to determine majority examples that are far away from the decision boundary, that is, to find as many clusters of majority samples as possible that do not contain any positive example or at most very few minority examples.

Appendix A.2. Over-sampling

The simplest strategy to augment the minority class is random over-sampling (ROS), which corresponds to a non-heuristic method that balances the class distribution through a random replication of positive examples (Batista et al., 2004). Although effective, this method may increase the likelihood of overfitting since it makes exact copies of the minority class examples.

Chawla et al. (2002) proposed the SMOTE algorithm, which generates artificial samples of the minority class by interpolating existing examples that lie close together. It first finds the k positive nearest neighbors for each minority class example and then, the synthetic examples are generated in the direction of some or all of those nearest neighbors. Depending upon the amount of over-sampling required, a certain number of examples from the k nearest neighbors are randomly chosen.

Although SMOTE has demonstrated to be an effective method for the class imbalance problem, it may overgeneralize the minority class as it disregards the distribution of majority class neighbors and consequently, the generation of synthetic examples may increase the overlapping between classes (Maciejewski & Stefanowski, 2011). In order to address this weakness in SMOTE, the resampling process can be altered to account for the class density around the minority class examples. For instance, the borderline-SMOTE algorithm (Han et al., 2005) consists of using only positive examples close to the decision boundary since these are more likely to be misclassified.

The Safe-Level-SMOTE algorithm (Bunkhumpornpat et al., 2009) calculates a “safe level” coefficient (sl) for each minority class example, which is defined as the number of other minority class examples among its k neighbors, to generate new synthetic examples close to safe regions. If the coefficient sl is equal or close to 0, such an example is considered as noise; if sl is close to k, then this example may be located in a safe region of the minority class.

García et al. (2012) modified the original SMOTE method by using the surrounding neighborhood concept when selecting the k positive neighbors of the minority class examples. The authors proposed three variations of the algorithm, each one based on a particular surrounding neighborhood realization (Sánchez & Marqués, 2002) for over-sampling the minority class: the nearest centroid neighborhood (NCN), the Gabriel graph (GG) and the relative neighborhood graph (RNG).

He et al. (2008) introduced an adaptive synthetic over-sampling (ADASYN) approach for learning from imbalanced data sets. The rationale behind this algorithm is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, thus shifting the decision boundary to be more focused on those examples that are harder to learn.

The ADOMS algorithm proposed by Tang & Chen (2008) is based on generating artificial examples along the first principal component axis of local data distribution composed of a positive sample and its k nearest neighbors. When $k = 1$, the result of this method matches that of SMOTE.

Another exciting proposal for populating the minority class is based on the application of an agglomerative hierarchical clustering (AHC) algorithm (Cohen et al., 2006). It uses single- and complete-linkage in succession to vary the clusters produced. Then the clusters are gathered from all levels of the resulting dendrograms and their centroids are computed and concatenated with the original positive samples. This results in augmenting the number of positive examples to match the size of the negative class.

Appendix A.3. Hybrid resampling

Although SMOTE produces well-balanced class distributions, some other difficulties often present in skewed data sets are not solved. For instance, class overlapping appears to be a widespread situation: some negative examples may be located within the clusters of the minority class and some synthetic positive examples may encroach on the majority class clusters. To overcome this problem and create non-overlapped class clusters, Batista et al. (2004) proposed the SMOTE-ENN technique: it consists in applying the Wilson's editing algorithm to the over-sampled data set to remove misclassified examples of both classes.

Another straightforward hybridization technique is based on the combination of SMOTE with the Tomek links (Batista et al., 2004). This method (SMOTE-TL) removes positive and negative examples that form a link after over-sampling the data set through SMOTE.

Stefanowski & Wilk (2008) introduced a selective preprocessing and resampling algorithm (SPIDER) that firstly preprocesses the data set to identify the safe and noisy examples. After this initial stage, all the noisy negative samples are removed, and the safe negative examples are kept. On the other hand, the minority class is modified according to one of the following three strategies: weak amplification, weak amplification and relabeling, and strong amplification.

SPIDER2 is an extension of the SPIDER, which consists of two phases to preprocess the majority class and the minority class respectively (Napierala et al., 2010). Firstly, it identifies the safe and unsafe (noisy and borderline) negative examples and then, it either removes or relabels the noisy samples. In the second phase, the algorithm identifies the positive examples taking into account the changes introduced in the data set during the first phase. Next, it replicates the noisy examples of the minority class. The only difference between this technique and SPIDER is that the latter processes both classes simultaneously.

Appendix B. Safe, borderline and rare-outlier databases

Tables B.5–B.7 report the databases included in each category according to the prevalent type of positive samples in the original data sets: safe (S), borderline (B) and rare-outlier (R-O) samples.

References

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., & Herrera, F. (2011). KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, *17*, 255–287.

Bach, M., Werner, A., Żywiec, J., & Pluskiewicz, W. (2017). The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. *Information Sciences*, *384*, 174–190.

Bagherpour, S., Nebot, A., & Mugica, F. (2018). Wrapper-based fuzzy inductive reasoning model identification for imbalance data classification. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1–8).

Bak, B. A., & Jensen, J. L. (2016). High dimensional classifiers in the imbalanced case. *Computational Statistics & Data Analysis*, *98*, 46–59.

Table B.5: Safe data sets

	S	B	R-O		S	B	R-O
iris0	1.000	0.000	0.000	ecoli4	0.663	0.175	0.163
ecoli-0_vs_1	0.997	0.004	0.000	glass-0-1-2-3_vs_4-5-6	0.632	0.280	0.088
shuttle-c0_vs_c4	0.986	0.006	0.008	vehicle2	0.628	0.254	0.118
shuttle-2_vs_5	0.964	0.031	0.005	ecoli-0-1_vs_2-3-5	0.624	0.115	0.261
segment0	0.944	0.048	0.008	page-blocks0	0.623	0.186	0.191
wisconsin	0.925	0.064	0.012	new-thyroid2	0.607	0.279	0.114
shuttle-6_vs_2-3	0.900	0.100	0.000	new-thyroid1	0.593	0.271	0.136
dermatology-6	0.900	0.038	0.063	ecoli-0-3-4-7_vs_5-6	0.550	0.280	0.170
led7digit-0-2-4-5-6-7-8-9_vs_1	0.837	0.068	0.095	yeast-2_vs_8	0.550	0.000	0.450
ecoli2	0.779	0.144	0.077	yeast-2_vs_4	0.549	0.181	0.270
glass6	0.759	0.000	0.241	ecoli1	0.546	0.318	0.136
vowel0	0.747	0.247	0.006	glass0	0.521	0.357	0.121
ecoli-0-1-4-6_vs_5	0.738	0.100	0.163	ecoli-0-1-4-7_vs_5-6	0.520	0.300	0.180
vehicle0	0.729	0.245	0.026	yeast3	0.502	0.287	0.212
yeast-0-2-5-7-9_vs_3-6-8	0.717	0.116	0.167	ecoli-0-2-6-7_vs_3-5	0.488	0.214	0.297
ecoli-0-3-4_vs_5	0.713	0.138	0.150	glass1	0.481	0.312	0.207
ecoli-0-3-4-6_vs_5	0.713	0.138	0.150	ecoli-0-6-7_vs_5	0.475	0.313	0.213
ecoli-0-2-3-4_vs_5	0.713	0.138	0.150	ecoli-0-1-4-7_vs_2-3-5-6	0.473	0.277	0.250
ecoli-0-4-6_vs_5	0.713	0.138	0.150	ecoli-0-6-7_vs_3-5	0.468	0.239	0.293
ecoli-0-1_vs_5	0.713	0.125	0.163				

Table B.6: Borderline data sets

	S	B	R-O		S	B	R-O
page-blocks-1-3_vs_4	0.420	0.445	0.135	glass4	0.251	0.442	0.307
yeast-0-2-5-6_vs_3-7-8-9	0.343	0.306	0.351	yeast1	0.236	0.441	0.323
yeast5	0.330	0.528	0.142	vehicle1	0.192	0.539	0.269
pima	0.329	0.399	0.271	haberman	0.108	0.429	0.463
yeast6	0.314	0.336	0.350	glass5	0.000	0.500	0.500
ecoli3	0.307	0.500	0.193				

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*, 20–29.

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*, 106.

Błaszczyński, J., & Stefanowski, J. (2015). Neighbourhood sampling in bagging for imbalanced data. *Neurocomputing*, *150*, 529 – 542.

Boonchuay, K., Sinapiromsaran, K., & Lursinsap, C. (2017). Decision tree in-

Table B.7: Rare-outlier data sets

	S	B	R-O		S	B	R-O
yeast-0-3-5-9_vs_7-8	0.180	0.280	0.540	poker-9_vs_7	0.000	0.114	0.886
glass-0-4_vs_5	0.168	0.636	0.197	cleveland-0_vs_4	0.000	0.096	0.904
vehicle3	0.122	0.524	0.355	poker-8-9_vs_6	0.000	0.090	0.910
yeast-1_vs_7	0.033	0.192	0.775	yeast-1-2-8-9_vs_7	0.000	0.083	0.917
winequality-white-3_vs_7	0.025	0.063	0.913	winequality-red-8_vs_6	0.000	0.069	0.932
yeast4	0.024	0.339	0.637	yeast-1-4-5-8_vs_7	0.000	0.058	0.942
glass-0-6_vs_5	0.000	0.721	0.279	poker-8-9_vs_5	0.000	0.040	0.960
glass-0-1-6_vs_5	0.000	0.636	0.364	winequality-white-3-9_vs_5	0.000	0.030	0.970
glass2	0.000	0.366	0.634	winequality-red-4	0.000	0.005	0.995
glass-0-1-6_vs_2	0.000	0.310	0.690	poker-8_vs_6	0.000	0.000	1.000
glass-0-1-4-6_vs_2	0.000	0.289	0.711	winequality-red-3_vs_5	0.000	0.000	1.000
glass-0-1-5_vs_2	0.000	0.264	0.736				

duction based on minority entropy for the class imbalance problem. *Pattern Analysis and Applications*, *20*, 769–782.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*, 1145 – 1159.

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, *49*, 31:1–31:50.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for handling the class imbalanced problem. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 475–482). Bangkok, Thailand.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTE-Boost: Improving prediction of the minority class in boosting. In *Proceedings of the 7th European Conference on Knowledge Discovery in Databases* (pp. 107–119). Cavtat-Dubrovnik, Croatia.

Chen, H., Li, T., Fan, X., & Luo, C. (2019). Feature selection for imbalanced data based on neighborhood rough sets. *Information Sciences*, *483*, 1 – 20.

Chen, L., Fang, B., Shang, Z., & Tang, Y. (2018). Tackling class overlap and imbalance problems in software defect prediction. *Software Quality Journal*, *26*, 97–125.

Cohen, G., Hilario, M., Sax, H., Hugonnet, S., & Geissbuhler, A. (2006). Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine*, *37*, 7–18.

Dal Pozzolo, A., Caelen, O., & Bontempi, G. (2015). When is undersampling effective in unbalanced classification tasks? In A. Appice, P. P. Rodrigues, V. Santos Costa, C. Soares, J. Gama, & A. Jorge (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 200–215). Cham, Switzerland: Springer International Publishing.

Das, S., Datta, S., & Chaudhuri, B. B. (2018). Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition*, *81*, 674–693.

Daskalaki, S., Kopanas, I., & Avouris, N. (2006). Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence*, *20*, 381–417.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, *20*, 18–36.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018a). Imbalanced classification with multiple classes. In *Learning from Imbalanced Data Sets* (pp. 197–226). Cham: Springer International Publishing.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018b). Performance measures. In *Learning from Imbalanced Data Sets* (pp. 47–61). Cham: Springer International Publishing.

Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explorations Newsletter*, *12*, 49–57.

García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, *98*, 1–29.

García, V., Alejo, R., Sánchez, J. S., Sotoca, J. M., & Mollineda, R. A. (2006). Combined effects of class imbalance and class overlap on instance-based classification. In E. Corchado, H. Yin, V. Botti, & C. Fyfe (Eds.), *Intelligent Data Engineering and Automated Learning* (pp. 371–378). Berlin, Heidelberg: Springer.

García, V., Mollineda, R. A., & Sánchez, J. S. (2008). On the k-NN performance in a challenging scenario of imbalance and overlapping. *Pattern Analysis and Applications*, *11*, 269–280.

García, V., Mollineda, R. A., & Sánchez, J. S. (2014). A bias correction function for classification performance assessment in two-class imbalanced problems. *Knowledge-Based Systems*, *59*, 66 – 74.

García, V., Sánchez, J., & Mollineda, R. (2007). An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In L. Rueda, D. Mery, & J. Kittler (Eds.), *Progress in Pattern Recognition, Image Analysis and Applications* (pp. 397–406). Berlin, Heidelberg: Springer.

García, V., Sánchez, J. S., Martín-Félez, R., & Mollineda, R. A. (2012). Surrounding neighborhood-based SMOTE for learning from imbalanced data sets. *Progress in Artificial Intelligence*, *1*, 347–362.

García, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, *25*, 13–21.

García, V., Sánchez, J. S., Ochoa Domínguez, H. J., & Cleofas-Sánchez, L. (2015). Dissimilarity-based learning from imbalanced data with small disjuncts and noise. In *Proceedings of the 7th Iberian Conference on Pattern Recognition and Image Analysis* (pp. 370–378). Santiago de Compostela, Spain.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239.

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A new oversampling method in imbalanced data sets learning. In *Proceedings of the International Conference on Intelligent Computing* (pp. 878–887). Hefei, China.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, *77*, 103–123.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, *45*, 171–186.

Hart, P. (1968). The condensed nearest neighbour rule. *IEEE Transactions on Information Theory*, *14*, 515–516.

He, H., Bai, Y., Garcia, E., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 1322–1328). Hong Kong, China.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1263–1284.

Ho, T. K., & Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 289–300.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*, 299–310.

Japkowicz, N. (2003). Class imbalance: Are we focusing on the right issue? In *Proceedings of the Workshop on Learning from Imbalanced Data Sets II* (pp. 17–23). Washington DC.

Japkowicz, N. (2006). Why question machine learning evaluation methods? (an illustrative review of the shortcomings of current methods. In *Proceedings of the AAAI'06 Workshop on evaluation methods for machine learning* (pp. 6–11).

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*, 429–449.

Jing, X., Zhang, X., Zhu, X., Wu, F., You, X., Gao, Y., Shan, S., & Yang, J. (2019). Multiset feature learning for highly imbalanced data classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *0*, 1–1.

Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, *6*, 40–49.

Kang, Q., Chen, X., Li, S., & Zhou, M. (2017). A noise-filtered under-sampling scheme for imbalanced classification. *IEEE Transactions on Cybernetics*, *47*, 4263–4274.

Kovács, G. (2019). An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*, *83*, 1–13.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*, 221–232.

Krawczyk, B., Woźniak, M., & Herrera, F. (2014). Weighted one-class classification for different types of minority class examples in imbalanced data. In *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining* (pp. 337–344). Piscataway, NJ.

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the 14th International Conference on Machine Learning* (pp. 179–186). Nashville, TN.

Kuncheva, L. I., Arnaiz-González, Á., Díez-Pastor, J.-F., & Gunn, I. A. D. (2019). Instance selection improves geometric mean accuracy: a study on imbalanced data classification. *Progress in Artificial Intelligence*, *8*, 215–228.

Lachheta, P., & Bawa, S. (2016). Combining synthetic minority oversampling technique and subset feature selection technique for class imbalance problem. In *Proceedings of the International Conference on Advances in Information Communication Technology & Computing* (pp. 25:1–25:6). Bikaner, India.

Landgrebe, T. C. W., Paclik, P., Duin, R. P. W., & Bradley, A. P. (2006). Precision-recall operating characteristic P-ROC curves in imprecise environments. In *18th International Conference on Pattern Recognition (ICPR'06)* (pp. 123–127). volume 4.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. In *Proceedings of the 8th Conference on Artificial Intelligence in Medicine* (pp. 63–66). Cascais, Portugal.

Lee, J. (2019). AUC4.5: AUC-based C4.5 decision tree algorithm for imbalanced data classification. *IEEE Access*, *7*, 106034–106042.

Lin, W.-J., & Chen, J. J. (2013). Class-imbalanced classifiers for high-dimensional data. *Briefings in Bioinformatics*, *14*, 13–26.

Liu, X., Wu, J., & Zhou, Z. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*, 539–550.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141.

Lopez-Garcia, P., Masegosa, A. D., Osaba, E., Onieva, E., & Perallos, A. (2019). Ensemble classification for imbalanced data based on feature space partitioning and hybrid metaheuristics. *Applied Intelligence*, *49*, 2807–2822.

Luengo, J., Fernández, A., García, S., & Herrera, F. (2011). Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling. *Soft Computing*, *15*, 1909–1936.

Maciejewski, T., & Stefanowski, J. (2011). Local neighbourhood extension of SMOTE for mining imbalanced data. In *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining* (pp. 104–111). Paris, France.

Maldonado, S., Weber, R., & Famili, F. (2014). Feature selection for high-dimensional class-imbalanced data sets using support vector machines. *Information Sciences*, *286*, 228–246.

Napierala, K., & Stefanowski, J. (2012). The influence of minority class distribution on learning from imbalance data. In *Proceedings of the 7th International Conference on Hybrid Artificial Intelligence Systems* (pp. 139–150). Salamanca, Spain.

Napierala, K., & Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, *46*, 563–597.

Napierala, K., Stefanowski, J., & Wilk, S. (2010). Learning from imbalanced data in presence of noisy and borderline examples. In *Proceedings of the 7th International Conference on Rough Sets and Current Trends in Computing* (pp. 158–167). Warsaw, Poland.

Prati, R. C., Batista, G. E. A. P. A., & Monard, M. C. (2004a). Class imbalances versus class overlapping: An analysis of a learning system behavior. In *Proceedings of the 3rd Mexican International Conference on Artificial Intelligence* (pp. 312–321). Mexico City, Mexico.

Prati, R. C., Batista, G. E. A. P. A., & Monrad, M. C. (2004b). Learning with class skews and small disjuncts. In A. L. C. Bazzan, & S. Labidi (Eds.), *Advances in Artificial Intelligence* (pp. 296–306). Berlin, Heidelberg: Springer volume 3171 of *Lecture Notes in Computer Science*.

Prati, R. C., Batista, G. E. A. P. A., & Silva, D. F. (2015). Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. *Knowledge and Information Systems*, *45*, 247–270.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ren, S., Zhu, W., Liao, B., Li, Z., Wang, P., Li, K., Chen, M., & Li, Z. (2019). Selection-based resampling ensemble algorithm for nonstationary imbalanced stream data learning. *Knowledge-Based Systems*, *163*, 705 – 722.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. (2nd ed.). Newton, MA, USA: Butterworth-Heinemann.

Sáez, J. A., Krawczyk, B., & Woźniak, M. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, *57*, 164–178.

Sánchez, J. S., & Marqués, A. I. (2002). Enhanced neighbourhood specifications for pattern classification. In D. Chen, & X. Cheng (Eds.), *Pattern Recognition and String Matching* (pp. 673–702). Boston, MA: Springer volume 13.

Sanz, J., Fernandez, J., Bustince, H., Gradin, C., Fortún, M., & Belzunegui, T. (2017). A decision tree based approach with sampling techniques to predict the survival status of poly-trauma patients. *International Journal of Computational Intelligence Systems*, *10*, 440–455.

Sardari, S., Eftekhari, M., & Afsari, F. (2017). Hesitant fuzzy decision tree approach for highly imbalanced data classification. *Applied Soft Computing*, *61*, 727 – 741.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *40*, 185–197.

Shahee, S. A., & Ananthakumar, U. (2019). An effective distance based feature selection approach for imbalanced data. *Applied Intelligence*, *0*, 1–29.

Skryjomski, P., & Krawczyk, B. (2017). Influence of minority class instance types on SMOTE imbalanced data oversampling. In L. Torgo, B. Krawczyk, P. Branco, & N. Moniz (Eds.), *Proceedings of the 1st International Workshop on Learning with Imbalanced Domains: Theory and Applications* (pp. 7–21). Skopje, Macedonia volume 74 of *Proceedings of Machine Learning Research*.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*, 427 – 437.

Stefanowski, J. (2013). Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. In S. Ramanna, L. C. Jain, & R. J. Howlett (Eds.), *Emerging Paradigms in Machine Learning* (pp. 277–306). Berlin, Heidelberg: Springer.

Stefanowski, J. (2016). Dealing with data difficulty factors while learning from imbalanced data. In J. Mielniczuk, & S. Matwin (Eds.), *Challenges in Computational Statistics and Data Mining* (pp. 333–363). Cham, Switzerland: Springer volume 605.

Stefanowski, J., & Wilk, S. (2008). Selective pre-processing of imbalanced data for improving classification performance. In *Proceedings of the 10th International Conference in Data Warehousing and Knowledge Discovery* (pp. 283–292). Turin, Italy.

Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*, *425*, 76 – 91.

Tang, S., & Chen, S. (2008). The generation mechanism of synthetic minority class examples. In *Proceedings of the 5th International Conference on Information Technology and Applications in Biomedicine* (pp. 444–447). Shenzhen, China.

Tomašev, N., & Mladenić, D. (2013). Class imbalance and the curse of minority hubs. *Knowledge-Based Systems*, *53*, 157–172.

Tomek, I. (1976). Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, *6*, 769–772.

Van Hulse, J., & Khoshgoftaar, T. (2009). Knowledge discovery from imbalanced and noisy data. *Data & Knowledge Engineering*, *68*, 1513–1542.

Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 935–942). Corvallis, OR.

Viegas, F., Rocha, L., Gonçalves, M., Mourao, F., Sá, G., Salles, T., Andrade, G., & Sandin, I. (2018). A genetic programming approach for feature selection in highly dimensional skewed data. *Neurocomputing*, *273*, 554–569.

Vorraboot, P., Rasmequan, S., Chinnasarn, K., & Lursinsap, C. (2015). Improving classification rate constrained to imbalanced data between overlapped and non-overlapped regions by hybrid algorithms. *Neurocomputing*, *152*, 429–443.

Vuttipittayamongkol, P., & Elyan, E. (2020). Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. *Information Sciences*, *509*, 47–70.

Wang, S., & Yao, X. (2009). Diversity analysis on imbalanced data sets by using ensemble models. In *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining* (pp. 324–331). Nashville, TN.

Wasikowski, M., & Chen, X. (2010). Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on Knowledge and Data Engineering*, *22*, 1388–1400.

Weiss, G. M. (2004). Mining with rarity: A unifying framework. *ACM SIGKDD Explorations Newsletter*, *6*, 7–19.

Weiss, G. M. (2010). The impact of small disjuncts on classifier learning. In R. Stahlbock, S. F. Crone, & S. Lessmann (Eds.), *Data Mining: Special Issue in Annals of Information Systems* (pp. 193–226). Boston, MA: Springer.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbour rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, *2*, 408–421.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*. (4th ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Wojciechowski, S., & Wilk, S. (2017). Difficulty factors and preprocessing in imbalanced data sets: An experimental study on artificial data. *Foundations of Computing and Decision Sciences*, *42*, 149–176.

Wong, G. Y., Leung, F. H., & Ling, S.-H. (2018). A hybrid evolutionary preprocessing method for imbalanced datasets. *Information Sciences*, *454–455*, 161–177.

Yen, S., & Lee, Y. (2006). Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In *Proceedings of the International Conference on Intelligent Computing* (pp. 731–740). Kunming, China.

Yin, H., & Gai, K. (2015). An empirical study on preprocessing high-dimensional class-imbalanced data for classification. In *Proceedings of the IEEE 17th International Conference on High Performance Computing and Communications, IEEE 7th International Symposium on Cyberspace Safety and Security, and IEEE 12th International Conference on Embedded Software and Systems* (pp. 1314–1319). New York, NY.

Yin, L., Ge, Y., Xiao, K., Wang, X., & Quan, X. (2013). Feature selection for high-dimensional imbalanced data. *Neurocomputing*, *105*, 3–11.

Yoon, K., & Kwek, S. (2005). An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics. In *Proceedings of the 5th International Conference on Hybrid Intelligent Systems* (pp. 303–308). Rio de Janeiro, Brazil.

Zhang, C., Guo, J., & Lu, J. (2017). Research on classification method of high-dimensional class-imbalanced data sets based on SVM. In *Proceedings of the IEEE 2nd International Conference on Data Science in Cyberspace* (pp. 60–67). Shenzhen, China.

Zheng, Z., Wu, X., & Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter*, *6*, 80–89.