Impact of Sampling on Learning Asymmetric-Entropy Decision Trees from Imbalanced Data

Completed Research Paper

Ikram Chaabane **Radhouane Guermazi**

Mohamed Hammami

Abstract

Learning from imbalanced data is still a challenging problem in spite of more than two decades of continuous development in this field. To deal with this problem, several data-level and algorithmic-level methods are proposed. Hybrid methods, which combine the advantages of the two previous groups, are also gaining increasing popularity. Therefore, in this paper, we put our focus on new hybrid approaches combining different sampling strategies with adapted decision trees to tackle the binary imbalanced problems. Our experiments consider five preprocessing methods and three asymmetric split criteria, which results in fifteen evaluated combinations. Unlike the majority of the studies, we take into account the intrinsic data characteristics in the analysis of each finding in order to gain a deeper understanding in the field of imbalanced data. The achieved findings, supported by statistical tests, end up to learn the extent to which sampling can be advantageous when combined with algorithmic solutions.

Keywords: Sampling, imbalanced data, asymmetric entropy

Introduction

The class imbalance is observed when the classes are unequally represented. In binary class problems, we are frequently faced with rare objects drawing the minority class versus a large number of objects belonging to the majority class. Such imbalance scenarios are ubiquitous at a lot of real applications like software defects (Rodriguez et al. 2014), cancer gene expressions (Yu et al. 2012) and fraud detection (Olszewski 2012). When the rare objects are the most interesting, which is the common case, classifiers should reach high accuracies in classifying the minority class. Nevertheless, the lack of data under this class makes it difficult to find out regularities within it and then, weakens its learning performance. Various approaches are suggested to deal with poor performances in imbalanced data situations. Some studies balance data so that to fit the assumptions of standard learning algorithms.

Such methods belong to the external approach category. Other studies reveal that poor performances, particularly regarding the minority class, may be due to adopting inappropriate inductive bias inside the learning algorithm, which refers to inefficient heuristics provided to the classifier in order to guide the learning process (Zhengxin 2000). Hence, many researchers opt for adapting or introducing more suitable bias in the presence of uneven misclassification errors. Such strategy is recognized as internal approach since made changes lie inside the learning algorithm. To take advantages from the two later strategies, most of the recent trends lay more focus on hybrid approaches which combine internal and external methods. Recent studies prove that the class imbalance cannot be the main cause of the performance degradation; however that can be due to other factors in conjunction with the skewed data distribution. Thus, new challenges arise to explore methods which can deal with each data complexity problem. Studies on this area are still lacking and need more investigations.

In this paper, we explore a new hybridization of asymmetric-entropy decision trees with various strategies of sampling data to deal with binary imbalanced data. As far as we know, we present the first trial of combining data and algorithmic solutions on decision trees. The purpose from the evaluation of this combination is to point out when preprocessing data before learning asymmetric decision trees can be fruitful in dealing with each recognized data-complexity factor. An extensive experimental study using nineteen data-sets was carried out to (1) compare the performances of over-, under- and hybrid sampling approaches on four data complexity factors, (2) analyze the impact of each sampling strategy and (3) draw conclusions about the best performing combinations for each studied data type. Our key findings demonstrate that the performed combinations of sampling strategies with asymmetric decision trees have provided promising results in comparison with algorithmic approaches in addition to important directives to handle some specific data-complexity factors.

The remainder of the paper was arranged as follows. We first review the related works describing internal and external approaches. We then describe our hybridization methodology associating sampling with asymmetric decision trees. We present our results next, discuss our findings, and offer helpful directives to classify our imbalanced data with taking into account the nature of data. Finally, we conclude and present future works for our study.

Related works

The literature on class imbalance solutions refers to two main categories, namely external approaches along with internal ones.

External approaches

The methods belonging to this category act out of the learning algorithm, which explain their nomination. Sampling ranks among the external approaches adopted highly to cope with imbalanced data. Its wide-spread use is due to its independence of the underlying classifier. The basic idea is to reduce the class imbalance either by over-sampling the minority class or under-sampling the majority one. Non-heuristic under and over-sampling may lead to discard useful patterns and replicate exact copies, respectively. Therefore, advanced approaches are investigated to avoid the loss of important patterns on the one hand, and reduce the learning time and the risk of over-fitting on the other hand. In each approach, the authors introduce, in some way, intelligence in the process of adding or removing examples. SMOTE (Chawla et al. 2002) for example, as one of the most common over-sampling advanced techniques, creates synthetic instances by selecting randomly a minority class instance and its k nearest minority class neighbors. The artificial new instances will be introduced along the line segments joining any/all of the selected neighbors, depending upon the amount of over-sampling (Lopez et al. 2013). Several later studies are based on SMOTE to produce more sophisticated algorithms such as borderline-SMOTE variants (Han et al. 2005) which over-samples only the minority class examples close to the borders, SMOTE-IPF lying more focus on both borderline and noisy rare patterns (Saez et al. 2015) and cluster-based over-sampling strategies (Puntumapon et al. 2016; Lim et al. 2017) dealing with the over-generalization problem.

For large data-sets, over-sampling may be expensive in terms of time. More importantly, the introduction of new minority examples may be carried out based on noisy examples. Under-sampling can be here an appropriate solution to reduce the class imbalance and discard misleading (noise, borderline or redundant) examples. Several approaches in this area are based on the means of a neighborhood, designated by cleaning methods like Tomek Links (Tomek 1976), the Wilson's Edited Nearest Neighbor (ENN) rule (Wilson 1972), Neighborhood Cleaning Rule (NCR) (Laurikkala 2001) and the One Side Selection (OSS) (Kubat and Matwin 1997). Another alternative to carry out an intelligent under-sampling is to build clusters including majority class examples; then each cluster may be represented either by its centroid (Cluster Centroid under-sampling) or by another representative one (i.e. Nearest Cluster Centroid approaches). Some representative works include (Yen and Lee 2009; Santos et al. 2010; Rayhan et al. 2017; Arafat et al. 2017). When data-sets are highly skewed, some investigations state the necessity of combining under and over-sampling techniques in order to enhance the learner generalization (Kotsiantis and Pintelas 2003; Han et al. 2005). In this context, SMOTE-ENN is a reference hybrid sampling approach associating SMOTE with the ENN cleaning rule (Batista et al. 2004).

By re-sampling data, absolutely, we reduce the class imbalance. Nevertheless, the learning process will assume the equivalence of the misclassification costs, which is improper for imbalanced scenarios.

Internal approaches

Several investigations are proposed to manipulate classifiers internally so as to adapt the inductive bias towards the minority class. kNN, for example was adjusted through assigning imbalanced weights to the classes so that the distance to the positive class prototypes becomes much lower than the distance to the ones of the majority class (Barandela et al. 2003). This will therefore foster the minority class prototypes to be the nearest neighbor of new patterns. Similarly, for support vector machines, solutions biased the kernel function so as to push the hyperplane closer to the positive class. Different ways for adjusting this bias are presented in (Veropoulos et al. 1999; Wu and Chang 2003). Thanks to their simplicity and ability to deal with big data, decision trees were also the focus of several researchers to take advantage of them in imbalanced scenarios. To this end, three main axes were investigated to adapt (1) the split criterion (Guermazi et al. 2018), (2) the pruning scheme of the decision tree (Chaabane et al. 2017) and (3) the assignment rule of a class to each example (Marcellin 2008). The main developed axe concerns the adjustment of the uncertainty measure used to assess the impurity and select the best split at each node. In this context, we highlighted three asymmetric entropies:

- Off-Centered Entropy (OCE) (Lenca et al. 2010) is defined by expr. 1 which respects the Shannon entropy definition; nevertheless, it uses a transformation function in order to convert the maximum uncertainty from an imbalanced to a balanced situation.

$$\eta(p) = -\pi\, log_2(\pi) - (1-\pi) log_2(1-\pi) \quad (1)$$

  Where $\pi = \begin{cases} \frac{p}{2w} & if\ 0 \le p \le w \\ \frac{p+1-2w}{2(1-w)} & if\ w \le p \le 1 \end{cases}$

  According to the OCE definition, $(\pi, 1-\pi)$designates the uniform distribution associated with the prior imbalanced class distribution $(p, 1-p)$.

- Weighted Information Entropy (IEW) (Zhao and Li 2017) which also uses the same shape of Shannon's entropy (cf. expr. 2). However, the probabilities are estimated on account of a weighted coefficient of each class.

$$IEW(p_w) = -p_w log_2 p_w - (1-p_w) log_2(1-p_w) \quad (2)$$

  Where $p_w(D_i) = \frac{w_i * |D_i|}{\sum_{i=1}^{|D|} w_i * |D_i|}$ and w_i is the weighted coefficient of the class D_i defined as:

$w_i = \frac{|U|}{|U_{D_i}|}$ where $|*|$ represents the number of elements in *, U is the set of objects and U_{D_i} represents the elements of the D_i-class.

- Asymmetric entropy (AE) (Zighed et al. 2010) is defined by expr. 3. It represents a new entropy definition affecting directly the class imbalance problem by introducing bias fostering the minority class. If Y denote the class variable and p (resp.$(1-p)$) the probability of $Y = 1$ (resp.$Y = 0$) The maximum uncertainty is reached when $p = w$, considering that the distribution $(w; 1-w)$ may be adjusted to the prior class distribution or any other distribution taking into account the misclassification cost (Singh et al. 2010).

$$h_w(p) = \frac{p(1-p)}{p(1-p)+(p-w)^2} \quad (3)$$

Methodology

In the asymmetric scenarios, the real data is so complex that using only one strategy to deal with, cannot tackle all difficulties. For example, sampling can be advantageous when it enhances the learning of minority patterns by oversampling and discard unsafe instances hindering the classification performance through under-sampling. Nevertheless, in some cases, over-sampling may lead to an over-fitting of the minority class as under-sampling can cause the loss of reliable patterns. This may be due to a defective capture of the specificities of the problem in query and so inadequate adjustment of sampling parameters. Here, algorithmic solutions discharged from a lot of adjustments may be more efficient. In order to take advantage from the two strategies, we thought of a hybrid solution which firstly pre-processes data by a sampling strategy and then applies a special decision tree algorithm considering the class imbalance during the tree building. Therefore, on the one hand, the input data is better filtered so that the model will be more reliable. On the other hand, the classifier captures directly the class imbalance by using heuristics fostering the minority class. In Figure 1, we illustrate the general process of the classification followed in this research. We focus our interest in the class imbalance on data-sets with binary target class. Hence, the first to do in the preprocessing step is to turn multi-class data-sets into binary ones if this is not the case.

Next, if train and test sets are not available, then we proceed with k fold cross validation depending on the size of the data-set. The final and most important stage is to re-sample the training set for each fold. We have selected approaches from each sampling strategy (over-, under- and hybrid sampling) in order to specify which one is the best suitable for each data characteristics. Therefore, we preprocessed data by two over-sampling strategies namely SMOTE (Chawla et al. 2002) and Borderline2-SMOTE (Han et al. 2005) (B2-SMOTE), two under-sampling strategies namely the cluster and the Nearest cluster-based under-sampling (CC and NCC) (Yen and Lee 2009) and the SMOTE-ENN, noted SE, as a hybrid sampling approach. For accurate conclusions, we choose to distinguish four problem categories according to the nature of the minority instances. We are based on Krystina and Jerzy' classification which differentiates between safe and unsafe minority examples classified in themselves into borderline, rare and outlier examples (Napierala and Stefanowski 2016). In other words, we explore the impact of each sampling strategy on each data-set category according to the characteristics of their minority class examples. Hence, it can be concluded in which case of data-sets the under or hybrid sampling combined with an asymmetric decision tree can be more reliable than the use of over sampling and vice versa. The grown asymmetric decision trees are using one of these entropies: AE (Zighed et al. 2010), OCE (Lenca et al. 2010) or IEW (Zhao and Li 2017) proposed in the context of algorithmic solutions for the class imbalance problem.

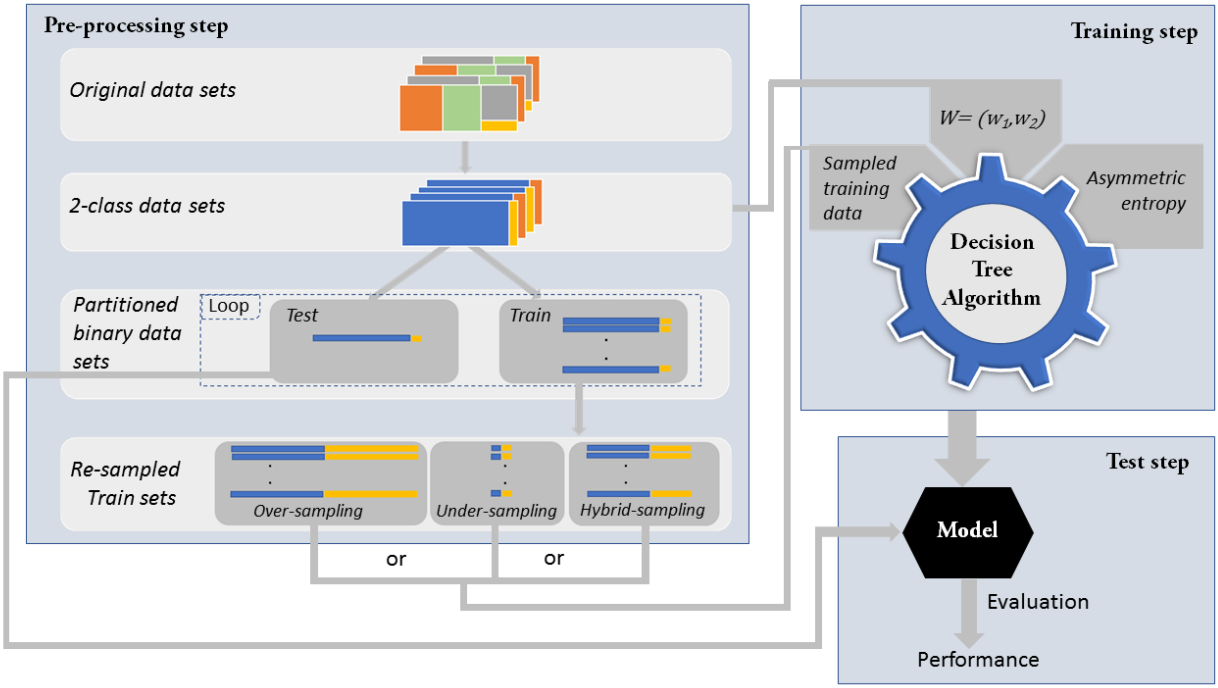

Figure1. General Process of the Hybrid Classification

Experiments

In this section, we first described the real-world two-class data-sets used in our experiments. Then, we detailed the sampling methods along with the asymmetric split criteria used to generate the experiments. After that, we presented the experimental framework including the description of the used algorithms along with their adjustments, the evaluation measures and the performed statistical tests used to assess the best classifier. Finally, we presented and analyzed the results of the performed experimental series.

Data-sets

The experimental setup used nineteen data-sets selected from different repositories. Two data-sets [4] come from the UCI repository (Hettich and Bay 1999). Their partitions into train and test sets are already available and have been used as they are partitioned. Whereas, two data-sets [5] come from Open ML (Vanschoren et al. 2014) and the other fifteen are selected from KEEL data-set repository (Alcala-Fdez et al. 2011). Since we are tackling the binary class imbalance problems, the multi-class data-sets were modified to be binary ones where each class represents the joint of one or more classes of the original data-sets. Table 1 shows the details of these data-sets in a descendant order of the imbalance level. Other than these usual data-set characteristics, there are other data features which may add the complexity of the classification task. They include the presence of outliers, borderline and rare minority class examples. The three later columns of Table 1 specify the percentage of each type in each data-set according to the categorization approach in (Napierala and Stefanowski 2016). According to this approach, borderline examples are located in the regions around decision boundary between classes. Outliers represent single minority examples located inside the majority class region. They can be valid or noisy. Whereas, rare examples are isolated few minority class examples located in the majority class region. Any researcher can be referred to (Blaszczynski and Stefanowski 2016) to label and analyze his data-set before proceeding with the learning algorithm.

Table 1. Summary of Empirical Data-Sets

Id	Name	#Ex.	#Att.	Class (min./maj.)	Min. class(%)	Safe (%)	Border (%)	Outlier (%)	Rare (%)
Id1	ecoli0vs1	220	7	(0/1)	34.96	88.3	7.8	3.9	0.0
Id2	yeast1	1484	8	(1/rest)	28.90	21.0	60.6	11.2	7.2
Id3	vehicle2	846	18	(2/rest)	25.77	89.9	9.2	0.0	0.9
Id4	vehicle1	846	18	(1/rest)	25.64	23.5	69.1	4.6	2.8
Id5	Statlog	6435	36	(1/rest)	24.04	96.7	2.6	0.4	0.3
Id6	Ecoli2	336	7	(2/ rest)	15.48	76.9	15.4	7.7	0.0
Id7	Bank	4521	16	(y/n)	11.52	13.4	47.0	26.9	12.7
Id8	Mfeat	2000	6	(P/ N)	10	98.5	0.0	1.5	0.0
Id9	Optdigits	5620	64	(0/ rest)	9.86	99.5	0.5	0.0	0.0
Id10	glass016vs2	192	9	(0,1,6/ 2)	8.86	0.0	41.2	35.3	23.5
Id11	led7digit02456789vs1	443	7	(0,2,4-9/ 1)	8.35	10.8	18.9	70.3	0.0
Id12	Glass2	214	9	(2/ rest)	7.94	0.0	35.3	29.4	35.3
Id13	pageblocks13vs4	472	10	(1,3/ 4)	5.93	78.6	14.3	0.0	7.1
Id14	Dermatology6	358	34	(6/ rest)	5.59	100	0.0	0.0	0.0
Id15	yeast1458vs7	693	8	(1,4,5,8/ 7)	4.33	0.0	10.0	50.0	40.0
Id16	Yeast5	1484	8	(5/ rest)	2.96	34.1	61.4	4.5	0.0
Id17	poker89vs6	1485	10	(8,9/ 6)	1.68	4.0	64.0	16.0	16.0
Id18	abalone20vs8910	1916	8	(20/8,9,10)	1.36	3.8	30.8	61.5	3.8
Id19	poker8vs6	1477	10	(8/ 6)	1.15	5.9	47.1	23.5	23.5

Figure 2 illustrates the four aforementioned data types. We classify the data-sets of Table 1 into four groups depending on the dominating type of identified minority examples (cf. Table 2). The considered thresholds are assigned, in accordance with (Stefanowski 2016), to 50% for borderline data-sets and 20% for both of the rare and outlier data-set categories.

Table 1. Classification of Data-Sets According to the Presence of Each Type of Examples

Data-set category	Safe	Border	Outlier	Rare
Id	14, 9, 8, 5, 3, 1, 13, 6	4, 17, 16, 2, 19	11, 18, 16, 10, 12, 7, 19	15, 12, 19, 10

In our conducted experiments, we separate the comparative results on each category of data-set in order to explore to which extent each sampling strategy can be beneficial.

---

[4] The data-sets' Ids are in {5; 9}

[5] The data-sets' Ids are in {7; 8}

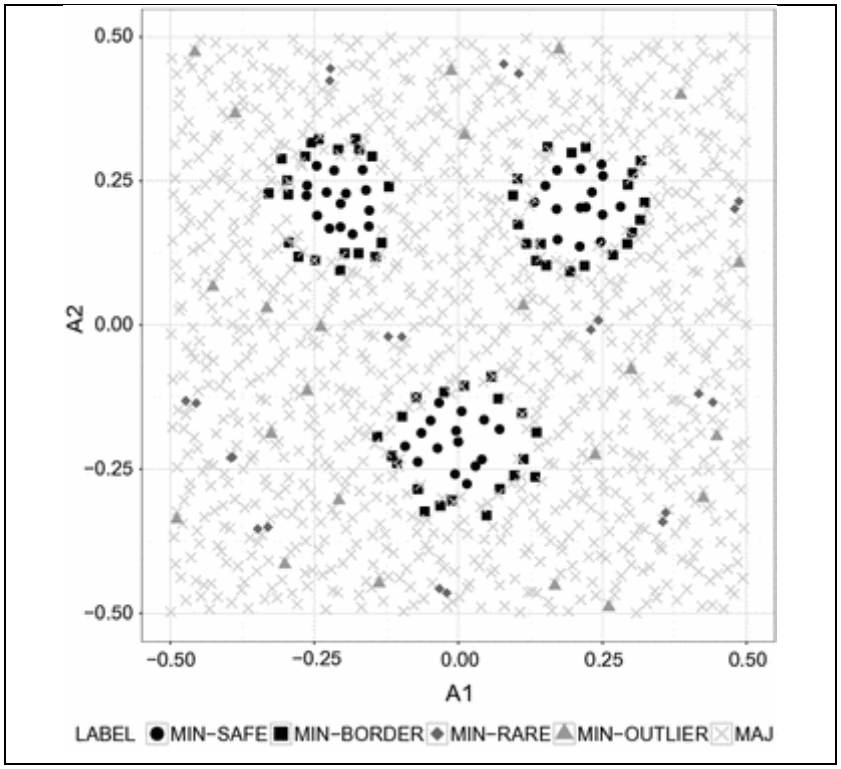

Figure 2. Visualization Example of the Four Types of Minority Class Patterns (Blaszczynski and Stefanowski 2018)

Experimental framework

Considering adapted approaches to the class imbalance problem, our main objective is to explore when it is beneficial to introduce a sampling step on the training data. Hence, the used base learners represent decision trees relying on asymmetric entropies to split the examples of each of their nodes. We have varied the asymmetric split criteria, namely AE (Zighed et al. 2010), OCE (Lenca et al. 2010) and IEW (Zhao and Li 2017), in order to generalize conclusions. For the three opted split criteria, we set the maximum uncertainty vector $(w; 1 - w)$ to the prior class distribution so that to be in the same experimental conditions.

Pure decision trees are built on the overall experimental study, in accordance with the finding of (Stefanowski 2016) which encourages the employ of unpruned decision trees with unsafe data. On each category of split-criterion decision trees, we have conducted several series of experiments which can be classified into three groups depending on the used sampling strategy: (i) the over-sampling approaches cover SMOTE (Chawla et al. 2002) and Borderline2-SMOTE (Han et al. 2005) denoted by SM and B2-SM, respectively, (ii) the under-sampling approaches include both of the Cluster Centroid and the Nearest Cluster centroid approaches (Yen and Lee 2009), designated by CC and NCC, respectively, and (iii) the hybrid sampling approaches represented by SMOTE-ENN (Batista et al. 2004), noted SE on our experiments.

It is worth mentioning that preprocessing data was performed through the imbalanced-learn API (Lemaitre et al. 2017) since it offers the implementation of different solutions for imbalanced data along with the required evaluation measures to assess the prediction models. For all the *k*-based neighborhood methods, *k* is set to 5; the sampling ratio was set to 1 and the Euclidean distance is used in all cases to capture the neighbors.

The combination of each asymmetric-entropy decision tree with a non-sampling/sampling strategy is evaluated through three metrics: sensitivity, specificity along with the Index of Balanced Accuracy (IBA). When the two first measures assess the minority and the majority class accuracies, separately, IBA represents an overall assessment-performance measure. Although, AUC ranks among the most

widely appealed measures in the context of imbalanced learning, it suffers from severe incoherencies which were revealed by (Hand 2009; Thai-Nghe et al. 2011; Powers 2012; Stapor 2017). Thus, IBA can be a suitable alternative for an unbiased performance assessment. Indeed, it defines a weighted trade-off between a global performance measure ($Gmean^2$) and a new proposed signed index to reflect how balanced are the individual accuracies (Garcia et al. 2009) (cf. expr. 4). Hence, IBA rises more importantly when the improvement on the minority class is more important than on the majority one. The adopted evaluation measures are expressed as follow:

$$True\ Positive\ Rate(TPR) = \ Sensitivity\ = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate\ (TNR) = \ Specificity\ = \frac{TN}{N} = \frac{TN}{TN + FP}$$

where TP and TN respectively denote the number of positive and negative examples that are correctly classified, while FN and FP respectively denote the number of misclassified positive and negative examples.

$$IBA = (1 + 0.1 \times dominance) \times Gmean^2 \qquad (4)$$
$$= \left(1 + 0.1\ \times (TPR - TNR)\right) \times TPR \times TNR$$

Apart from the data-sets whose training and test sample are available (Id5 and Id9), the evaluation on the remaining ones was performed through 10 stratified cross-validation except for Id7 in which we proceeded with 3 stratified cross-validation since it is sufficiently large.

Based on the obtained results, we have carried out the Friedman aligned ranks test in order to show at first glance how good a method is when compared to its competitors. The best algorithm is the one that achieves the lowest average rank. Then, a statistic value according to each test is computed. More details about their formulas may be found in (Derrac et al 2011). Once one of these tests rejected the null hypothesis H_0 affirming the similarity of means of two or more algorithms, the Holm post-hoc test should be performed in order to find which algorithms reject the hypothesis of equality with respect to a selected control method in a 1* n comparison. We computed, therefore, the adjusted p-value (APV) associated with each comparison, which represents the lowest level of significance of a hypothesis that results in a rejection. Indeed, an APV lower than α results in a rejected post-hoc test's H_0, which confirms the significance of the difference with the control method. Otherwise, the post-hoc test's H_0 is accepted and no important differences are recorded. For legibility reasons, we simply record the status of the post-hoc test's H_0 ('A' for accepted and 'R' for rejected) in tables 3, 4 and 5. For all the statistical tests, we adjust the control method to the non-sampling strategy (designated by 'none') and the significance level α to the standard value of 0.05.

Results

In this section, we present and discuss the results of introducing different sampling strategies before learning asymmetric decision trees, taking account the various categories of data-sets. Tables 3, 4 and 5 summarize the obtained results in terms of sensitivity, specificity and IBA, respectively.

On tracking the impact of the experimental series on the minority class, Table 3 shows that for the three asymmetric-entropy decision trees, Friedman aligned ranks test accepts the null hypothesis H_0 when data is safe. Hence, sampling and no-sampling strategies perform similarly in the presence of safe data. Such claim confirms, as it is in other recent studies, that clear decision boundaries of the minority and majority classes could be a sufficient condition to achieve high classification performances regardless the class imbalance ratio and the classifier (Blaszczynski and Stefanowski 2018). In contrast, learning difficulties are observed when the high-class imbalance ratio is in conjunction with other data difficulties such as the presence of outliers, borderline or rare examples. Clearly, in Table 3, the Friedman test reveals differences on the tested algorithms' performances since each method may tackle differently the data internal characteristics. Furthermore, it is noteworthy that

ranking fosters performing sampling with unsafe data-sets. Indeed, according to the three entropy-asymmetric decision trees, under-sampling strategies (CC and NCC) and particularly CC ranks first in best performing with difficult data. The second-ranked strategy is the hybrid one (SE), whereas over-sampling techniques (SM and B2-SM) are shown to be the least improving. Although enhancement on the minority class is observed with all sampling strategies, it is significant only with under-sampling. In fact, reducing majority patterns may firstly clean the minority class regions so that to reduce the risk of overlapping on borderline data-sets rather than reducing the presence of small disjuncts (i.e. sub-clusters of the minority class consisting of few examples) so that to generalize the classifier regarding outliers and rare data-sets. Therefore, under-sampling represents a consistent solution with unsafe data-sets. On the other hand, the limited enhancement of over-sampling can be due to an over fitting of the minority class. Indeed, using a high sampling ratio (=1) may disadvantage B2-SMOTE especially with outliers and borderline data-sets as it over-strengths the boundary decision regions with borderline and noisy examples. This explains the better ranking of SM over B2-SM on the aforementioned data types. To conclude, over-sampling data without taking into account its internal characteristics can be assessed to be useless as it may amplify the learning difficulties.

Table 3. Sensitivity results: Aligned Friedman ranks with a significance level $\alpha = 0.05$ and the decision on the null hypothesis H_0 (R: rejected, A: accepted) based on the Holm test. The considered control method is the Non-Sampling ('None').

Split criterion	Sampling	Safe		Border		Outlier		Rare	
		Rank	H_0	Rank	H_0	Rank	H_0	Rank	H_0
AE	None	H_0 accepted		26.200 (6)	-	34.571 (6)	-	22.250 (6)	-
	SM			16.900 (4)	A	28.071 (5)	A	16.375 (5)	A
	B2-SM			21.000 (5)	A	26.643 (4)	A	14.250 (4)	A
	CC			4.300 (1)	R	6.571 (1)	R	2.500 (1)	R
	NCC			10.100 (2)	R	9.857 (2)	R	7.250 (2)	R
	SE			14.500 (3)	A	23.286 (3)	A	12.375 (3)	A
OCE	None	H_0 accepted		25.800 (6)	-	35.786 (6)	-	21.375 (6)	-
	SM			15.600 (4)	A	25.286 (4)	A	15.625 (5)	A
	B2-SM			20.800 (5)	A	26.714 (5)	A	15.000 (4)	A
	CC			6.500 (1)	R	7.214 (1)	R	2.500 (1)	R
	NCC			9.600 (2)	R	9.143 (2)	R	6.750 (2)	R
	SE			14.700 (3)	A	24.857 (3)	A	13.750 (3)	A
IEW	None	H_0 accepted		25.900 (6)	-	33.928 (6)	-	22.500 (6)	-
	SM			15.400 (4)	A	26.214 (4)	A	15.500 (5)	A
	B2-SM			19.900 (5)	A	27.000 (5)	A	13.750 (4)	A
	CC			4.500 (1)	R	6.428 (1)	R	2.875 (1)	R
	NCC			13.500 (2)	A	9.571 (2)	R	6.750 (2)	R
	SE			13.800 (3)	A	25.857 (3)	A	13.625 (3)	A

Taking the advantages of under-sampling along with the disadvantages of over-sampling regarding unsafe data, the hybrid approach SE ever keeps the intermediate rank (3) for all unsafe data types and asymmetric-entropy decision trees.

Concerning the impact of sampling on the majority class, we rely on Table 4 which proves that avoiding sampling is almost the best strategy for the classification of majority examples. It is worth noting that enhancement on the minority class is always achieved to the detriment of the majority one. That is why, we observe a rank exchange of the best and the worst performing strategies from Table 3 to Table 4, whereas the hybrid approach SE keeps its intermediate ranking. An additional important remark states that NCC under sampling technique generates insignificant deterioration over non-sampling on the majority class, particularly for borderline data-sets; in contrast with its significant improvement on the minority class (cf. Table 3). Recalling that in the context of class imbalance problems, we are mainly interested in enhancing the minority class prediction without much disadvantaging the majority one. Therefore, NCC seems to be a good choice to tackle borderline data when faced with a class imbalance task.

Table 4. Specificity results: Aligned Friedman ranks with a significance level $\alpha = 0.05$ and the decision on the null hypothesis H_0 (R: rejected, A: accepted) based on the Holm test. The considered control method is the Non-Sampling ('None').

Split criterion	Sampling	Safe		Border		Outlier		Rare	
		Rank	H_0	Rank	H_0	Rank	H_0	Rank	H_0
AE	None	16.625 (1)	-	10.200 (1)	-	11.428 (1)	-	6.750 (1)	-
	SM	17.062 (2)	A	10.400 (2)	A	13.714 (2)	A	7.000 (2)	A
	B2-SM	26.250 (4)	A	11.200 (3)	A	15.357 (3)	A	9.750 (3)	A
	CC	37.812 (6)	R	25.000 (6)	R	37.214 (6)	R	22.125 (6)	R
	NCC	28.625 (5)	A	23.800 (5)	A	33.786 (5)	R	18.875 (5)	A
	SE	20.625 (3)	A	12.400 (4)	A	17.500 (4)	A	10.500 (4)	A
OCE	None	16.687 (1)	-	9.600 (1)	-	11.857 (1)	-	5.500 (1)	-
	SM	17.875 (2)	A	9.600 (1)	A	13.571 (2)	A	8.250 (2)	A
	B2-SM	25.687 (4)	A	12.600 (4)	A	16.357 (3)	A	8.875 (3)	A
	CC	39.375 (6)	R	25.700 (6)	R	37.071 (6)	R	22.500 (6)	R
	NCC	27.125 (5)	A	23.500 (5)	A	32.071 (5)	R	18.500 (5)	R
	SE	20.250 (3)	A	12.000 (3)	A	18.071 (4)	A	11.375 (4)	A
IEW	None	18.000 (2)	-	10.500 (2)	-	10.571 (1)	-	5.000 (1)	-
	SM	17.312 (1)	A	9.900 (1)	A	15.634 (3)	A	8.500 (2)	A
	B2-SM	22.812 (4)	A	11.600 (3)	A	14.928 (2)	A	9.000 (3)	A
	CC	38.062 (6)	R	26.000 (6)	R	37.286 (6)	R	22.500 (6)	R
	NCC	30.000 (5)	A	23.200 (5)	A	33.714 (5)	R	18.500 (5)	R
	SE	20.812 (3)	A	11.800 (4)	A	16.857 (4)	A	11.500 (4)	A

For a better evaluation of the trade-off between the minority and the majority class performances in a context of imbalanced data, we opt for IBA as a recent evaluation measure able to aggregate the two class performances with a better weighting than standard recall and precision harmonic means (i.e. F-measure and G-mean). Hence, Table 5 is considered to report the comparative study results in terms of IBA. A set of conclusions can be drawn from this table.

First, for safe, border and rare data-sets, the used sampling strategies cannot bring significant enhancements over non-sampling in terms of IBA. Second, for outlier data-sets, under-sampling and

especially NCC is ranked first among other strategies. Its significant improvement is validated through rejected H_0 when using AE and OCE decision trees. The last observation points out the accepted Friedman test's H_0 when using IEW decision trees and regardless of the type of data, which denotes a performing similarity of sampling and non-sampling strategies. This can be due to a more significant deterioration on the majority class than enhancement on the minority class.

Table 5. IBA results: Aligned Friedman ranks with a significance level $\alpha = 0.05$ and the decision on the null hypothesis H_0 (R: rejected, A: accepted) based on the Holm test. The considered control method is the Non-Sampling ('None').

Split criterion	Sampling	Safe		Border		Outlier		Rare	
		Rank	H_0	Rank	H_0	Rank	H_0	Rank	H_0
AE	None	16.875 (1)	-	H_0 accepted		32.714 (6)	-	H_0 accepted	
	SM	20.312 (3)	A			24.357 (4)	A		
	B2-SM	20.875 (4)	A			24.714 (5)	A		
	CC	37.875 (6)	R			15.286 (2)	R		
	NCC	20.062 (2)	A			11.286 (1)	R		
	SE	31.000 (5)	A			20.643 (3)	A		
OCE	None	15.875 (1)	-	H_0 accepted		33.714 (6)	-	H_0 accepted	
	SM	17.062 (2)	A			20.214 (3)	A		
	B2-SM	23.562 (3)	A			23.143 (5)	A		
	CC	36.125 (6)	R			19.643 (2)	A		
	NCC	25.375 (4)	A			10.000 (1)	R		
	SE	29.000 (5)	A			22.286 (4)	A		
IEW	None	H_0 accepted		H_0 accepted		H_0 accepted		H_0 accepted	
	SM								
	B2-SM								
	CC								
	NCC								
	SE								

Conclusion and future works

The class imbalance problem is still attracting a considerable interest due to its prevalence in several real life applications. The variety of the intrinsic data characteristics of each problem amplifies the classification difficulties in imbalanced scenarios. In fact, borderline, rare and outlier examples represent unsafe data and recognized to have a huge effect on the performance degradation (Lopez et al. 2013; Napierala and Stefanowski 2016).

Being aware by the well performing of unpruned decision trees against unsafe data (Stefanowski 2016) along with the advantages of sampling in imbalanced contexts, we proposed to combine the aforementioned approaches to address the presented data complexity factors. Therefore, we

investigated three sampling strategies, namely under-, over- and hybrid sampling in addition to three types of decision trees distinguished by different asymmetric split criteria (i.e. AE, OCE and IEW).

Each alternative's performance is assessed regarding to each complexity factor. Our key contributions are thus:

- Propose a novel hybridization combining sampling with asymmetric decision trees, which was not investigated to the best of our knowledge.
- Take into account safe and unsafe data characteristics on the experimental study, along with the class imbalance. Such analysis offers a better understanding of the class imbalance problem by making it more challenging, but realistic.
- Capture the extent to which sampling can be beneficial in combination with asymmetric decision trees.
- Show significant sensitivity improvements of under-sampling approaches in handling unsafe data. When IBA is considered, significant improvements of under-sampling are restricted to outlier data, whereas no significant enhancements are shown when hybrid and over-sampling strategies are applied. Concerning safe data-sets, non-sampling ranks first.
- Use IBA as a recent evaluation measure to assess the trade-off between the two class performances in imbalance context. The provided results are supported by statistical tests to rank the sampling strategies and confirm the significance of their differences against asymmetric decision trees using real data (without sampling).

Our study opens interesting perspectives to investigate more sophisticated sampling approaches taking into account the type of the sampled examples. Hence, a noisy or borderline example should be under-sampled as long as that did not harm the prediction performance. The same idea may be introduced on the decision trees, so that to maintain only rules covering safe examples.

References

Alcala-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., and Garcia, S. 2011. "KEEL datamining software tool: Data-set repository, integration of algorithms and experimental analysis framework," *Multiple-Valued Logic and Soft Computing* (17:2-3), pp. 255–287.

Arafat, M. Y., Hoque, S., and Farid, D. M. 2017. "Cluster-based under-sampling with random forest for multi-class imbalanced classification," in *11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, IEEE, pp. 1–6.

Barandela, R., Sanchez, J., Garcia, V., and Rangel, E. 2003. "Strategies for learning in class imbalance problems," *Pattern Recognition* (36:3), pp. 849–851 (doi: 10.1016/s0031-3203(02)00257-1).

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. 2004. "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter* (6:1), pp. 20–29 (doi: 10.1145/1007730.1007735).

Blaszczynski, J., and Stefanowski, J. 2018. "Local Data Characteristics in Learning Classifiers from Imbalanced Data," in *Advances in Data Analysis with Computational Intelligence Methods: Dedicated to Professor Jacek Zurada*, Cham: Springer International Publishing, pp. 51–85.

Chaabane, I., Guermazi, R., and Hammami, M. 2017. "Adapted pruning scheme for the framework of imbalanced data-sets," in *Procedia Computer Science* (112), pp. 1542–1553 (doi: 10.1016/j.procs.2017.08.060).

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. 2002. "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research* (16), pp. 321–357 (doi: 10.1613/jair.953).

Derrac, J., Garcia, S., Molina, D., and Herrera, F. 2011. "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm and Evolutionary Computation* (1:1), pp. 3–18 (doi: 10.1016/j.swevo.2011.02.002).

Garcia, S., Fernandez, A., Luengo, J., and Herrera, F. 2010. "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining:

Experimental analysis of power," *Information Sciences* (180:10), pp. 2044–2064 (doi: 10.1016/j.ins.2009.12.010).

Garcia, V., Mollineda, R. A., and Sanchez, J. S. 2009. "Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions," in *Pattern Recognition and Image Analysis: 4th Iberian Conference, IbPRIA 2009 Povoa de Varzim, Portugal, June 10-12, 2009 Proceedings*, Berlin: Springer, pp. 441–448.

Guermazi, R., Chaabane, I., and Hammami, M. 2018. "AECID: Asymmetric entropy for classifying imbalanced data," *Information Sciences* (467), pp. 373–397 (doi: 10.1016/j.ins.2018.07.076).

Han, H., Wang, W., and Mao, B. 2005. "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data-sets Learning," in *ICIC* Lecture Notes in Computer Science (1st ed., Vol. 3644), Springer, pp. 878–887.

Hand, D. J. 2009. "Measuring classifier performance: a coherent alternative to the area under the ROC curve," *Machine Learning* (77:1), pp. 103–123 (doi: 10.1007/s10994-009-5119-5).

Hettich, S., and Bay, S. D. 1999. "The UCI KDD Archive [http://kdd.ics.uci.edu]," Irvine, CA: University of California, Department of Information and Computer Science.

Kotsiantis, S. B., and Pintelas, P. E. 2003. "Mixture of Expert Agents for Handling Imbalanced Data-sets," in *Annals of Mathematics, Computing and Teleinformatics* (Vol. 1), pp. 46–55.

Kubat, M., and Matwin, S. 1997. "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection," in *Proceedings of the Fourteenth International Conference on Machine Learning*, Morgan Kaufmann, pp. 179–186.

Laurikkala, J. 2001. "Improving Identification of Difficult Small Classes by Balancing Class Distribution," in *Artificial Intelligence in Medicine*, Berlin, Heidelberg: Springer, pp. 63–66.

Lemaitre, G., Nogueira, F., Aridas, C.K. 2017. "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning" in *Journal of Machine Learning Research* (18:17), pp. 1-5.

Lenca, P., Lallich, S., and Vaillant, B. 2010. "Construction of an Off-Centered Entropy for the Supervised Learning of Imbalanced Classes: Some First Results," *Communications in Statistics - Theory and Methods* (39:3), pp. 493–507 (doi: 10.1080/03610920903140247).

Lim, P., Goh, C. K., and Tan, K. C. 2017. "Evolutionary Cluster-Based Synthetic Oversampling Ensemble (ECO-Ensemble) for Imbalance Learning," *IEEE Transactions on Cybernetics* (47:9), pp. 2850–2861 (doi: 10.1109/tcyb.2016.2579658).

Lopez, V., Fernandez, A., Garcia, S., Palade, V., and Herrera, F. 2013. "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information Sciences* (250), pp. 113–141 (doi: 10.1016/j.ins.2013.07.007).

Marcellin, S. 2008. "Arbres de décision en situation d'asymétrie," thesis, Lyon.

Napierala, K., and Stefanowski, J. 2016. "Types of minority class examples and their influence on learning classifiers from imbalanced data," *Journal of Intelligent Information Systems* (46:3), pp. 563–597 (doi: 10.1007/s10844-015-0368-1).

Olszewski, D. 2012. "A probabilistic approach to fraud detection in telecommunications," *Knowledge-Based Systems* (26), pp. 246–258 (doi: 10.1016/j.knosys.2011.08.018).

Powers, D. M. 2012. "The problem of Area Under the Curve," in *International Conference on Information Science and Technology (ICIST)* , pp. 567–573.

Puntumapon, K., Rakthamamon, T., and Waiyamai, K. 2016. "Cluster-Based Minority Over-Sampling for Imbalanced Datasets," *IEICE Transactions on Information and Systems* (E99.D:12), pp. 3101–3109 (doi: 10.1587/transinf.2016edp7130).

Rayhan, F., Ahmed, S., Mahbub, A., Jani, R., Shatabda, S., and Farid, D. M. 2017. "CUSBoost: Cluster-Based Under-Sampling with Boosting for Imbalanced Classification," in *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, pp. 1–5 (doi: 10.1109/csitss.2017.8447534).

Rodriguez, D., Herraiz, I., Harrison, R., Dolado, J., and Riquelme, J. C. 2014. "Preliminary Comparison of Techniques for Dealing with Imbalance in Software Defect Prediction," in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* EASE '14, New York, USA: ACM, pp. 1–43.

Saez, J. A., Luengo, J., Stefanowski, J., and Herrera, F. 2015. "SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with

filtering," *Information Sciences* Supplement C (291), pp. 184–203 (doi: 10.1016/j.ins.2014.08.051).

Santos, M. S., Abreu, P. H., Garcia-Laencina, P. J., Simao, A., and Carvalho, A. 2015. "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients," *Journal of Biomedical Informatics* (58), pp. 49–59 (doi: 10.1016/j.jbi.2015.09.012).

Singh, A., Liu, J., and Guttag, J. 2010. "Discretization of continuous ECG based risk metrics using asymmetric and warped entropy measures," in *2010 Computing in Cardiology*, pp. 473–476.

Stąpor, K. 2017. "Evaluating and Comparing Classifiers: Review, Some Recommendations and Limitations," Advances in Intelligent Systems and Computing Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017, pp. 12–21 (doi: 10.1007/978-3-319-59162-9_2).

Stefanowski, J. 2016. "Dealing with Data Difficulty Factors While Learning from Imbalanced Data," in *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence (605), Springer, pp. 333–363.

Thai-Nghe, N., Gantner, Z., and Schmidt-Thieme, L. 2011. "A new evaluation measure for learning from imbalanced data," *The 2011 International Joint Conference on Neural Networks*, pp. 537-542 (doi: 10.1109/ijcnn.2011.6033267).

Tomek, I. 1976. "Two Modifications of CNN," *IEEE Transactions on Systems, Man, and Cybernetics* (SMC-6:11), pp. 769–772 (doi: 10.1109/tsmc.1976.4309452).

Vanschoren, J., Rijn, J. N. V., Bischl, B., and Torgo, L. 2014. "OpenML," *ACM SIGKDD Explorations Newsletter* (15:2), pp. 49–60 (doi: 10.1145/2641190.2641198).

Veropoulos, K., Campbell, C., and Cristianini, N. 1999. "Controlling the sensitivity of support vector machines," in *Proceedings of the International Joint Conference on AI*, pp. 55–60.

Wilson, D. L. 1972. "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data," *IEEE Transactions on Systems, Man, and Cybernetics* (SMC-2:3), pp. 408–421 (doi: 10.1109/tsmc.1972.4309137).

Wu, G., and Chang, E. Y. 2003. "Class-boundary alignment for imbalanced dataset learning," in *ICML 2003 Workshop on Learning from Imbalanced Data-sets*, pp. 49–56.

Yen, S.-J., and Lee, Y.-S. 2009. "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications* (36:3), pp. 5718–5727 (doi: 10.1016/j.eswa.2008.06.108).

Yu, H., Ni, J., Dan, Y., and Xu, S. 2012. "Mining and integrating reliable decision rules for imbalanced cancer gene expression data-sets," *Tsinghua Science and Technology* (17:6), pp. 666–673 (doi: 10.1109/tst.2012.6374368).

Zhao, H., and Li, X. 2017. "A cost sensitive decision tree algorithm based on weighted class distribution with batch deleting attribute mechanism," *Information Sciences* (378), pp. 303–316 (doi: 10.1016/j.ins.2016.09.054).

Zhengxin, C. 2000. *Computational intelligence for decision support* Computational Intelligence for Decision Support (1st ed.), Boca Raton, FL: CRC Press.

Zighed, D. A., Ritschard, G., and Marcellin, S. 2010. "Asymmetric and Sample Size Sensitive Entropy Measures for Supervised Learning," in *Advances in Intelligent Information Systems* Studies in Computational Intelligence (265), Berlin, Heidelberg: Springer, pp. 27–42.