Performance Estimation bias in Class Imbalance with Minority Subconcepts

Colin Bellinger COLIN.BELLINGER@NRC-CNRC.GC.CA
National Research Council of Canada, Ottawa, Canada, K1A 0R6
Roberto Corizzo RCORIZZO@AMERICAN.EDU
Department of Computer Science, American University, Washington D.C. 20016, USA
Nathalie Japkowicz JAPKOWIC@AMERICAN.EDU
Department of Computer Science, American University, Washington D.C. 20016, USA

Editors: Nuno Moniz, Paula Branco, Luis Torgo, Nathalie Japkowicz, Michał Wozniak and Shuo Wang.

Abstract

Learning classifiers from imbalanced data is known to be a challenging and important problem in machine learning. As a results, the topic has been studied from a wide variety of angles. This includes the choice of evaluation measures and understanding the implications of minority class subconcepts on model learning. In this work, however, we argue that the community may not be using precise enough evaluation measures when assessing the performance of imbalanced learning pipelines on data that includes an imbalance in the minority class subconcepts. We show that the performance estimates from standard measures used in imbalance learning are biased towards the largest minority subconcepts, and that standard imbalance correction techniques can exacerbate the bias. Finally, we demonstrate that the bias can, in part, be corrected by applying instance weighting in the evaluation measures.

Keywords: Class imbalance; Minority subconcepts; Performance evaluation; Bias; Group fairness

1. Introduction

The class imbalance problem has been a challenge for machine learning scientists for over two decades (Kubát et al., 1998; Ling and Li, 1998; Japkowicz and Stephen, 2002; Prati et al., 2004). Many papers have been published on the subject proposing different methods to address the issue. SMOTE alone (Chawla et al., 2002) was spun into, at least, 85 variants (Kovács, 2019), and that is only one approach—albeit the most famous one—within the traditional oversampling category of methods. Deep learning oversampling approaches have also been proposed (Bellinger et al., 2018; Mullick et al., 2019; Dablain et al., 2021) along with undersampling approaches (Liu et al., 2009; Yen and Lee, 2009), cost-sensitive approaches (Ling and Sheng, 2008), one-class learning ones (Bellinger et al., 2012, 2017), and loss function optimization (Li et al., 2022). Despite all the research activity, including reviews (He and Garcia, 2009; Branco et al., 2016; Fernández et al., 2018; Johnson and Khoshgoftaar, 2019), comparisons of methods (Hulse et al., 2007; Zhu et al., 2017), *etc.*, the best method to remedy the class imbalance problem remains an open question.

In this work, we put forward the hypothesis that *we may not be using precise enough evaluation measures when assessing the quality of our class-imbalance correction methods.* In

particular, the standard evaluation measures used to estimate the performance of imbalance classification pipelines may be biased towards large subconcepts in the minority class. Our hypothesis is motivated the observation *a)* that rare class are often very important, but overlook by models (Branco et al., 2016), and *b)* the minority class should not be considered in a blanket fashion but that, instead, the separate disjuncts (or subconcepts) constituting it should be distinguished (Jo and Japkowicz, 2004; Santos et al., 2015; Zhang and Chen, 2019). In addition, this problem is connected to group fairness (Mehrabi et al. (2021)) and is a growing concern as automated decision making systems increasingly impact people in applications, such as healthcare, hiring, university acceptance, security and policing, *etc.* In particular, standard approaches to evaluation in imbalanced classification risk smoothing over poor predictive performance on rare and/or poorly sampled subgroups leading to a biased performance estimate. This paper serves to demonstrate this point on a series of examples and, in doing so, continue a conversation on evaluation of imbalanced learning started a decade or so ago by (Raeder et al., 2012).

Putting these observations together in the context of evaluation for class imbalances, we argue that popular evaluation metrics used in class imbalance problems such as the AUC, F-measure, Balanced Accuracy, and so on, although less biased toward the majority class than accuracy, fail to distinguish the subconcepts making up the minority class and take their relevance into account. Indeed, while these metrics treat the classes in a more balanced way, the performance on the larger subconcepts may skew the evaluation. Thus, we hypothesize that the less biased method themselves are biased towards subconcepts in the minority class with a higher prior probability. Consequently, independently of their prior probability, we propose to treat each of the minority subconcepts with equal importance.

In order to study evaluation bias and model performance on imbalance minority subconcepts, we engineer datasets for which we know each instance's parent subconcept and the subconcepts' priors. We compare model performance in terms of standard measures (AUC, balanced accuracy and the F-measure) to model performance in terms of instance weighted versions of these measures. In the instance weighted versions, each minority instances is weighted according to the ratio between the majority class size and the instance's subconcept size in the training set. This analysis illuminates four important research questions:

Research Question 1: Are the so-called standard less biased measures biased towards larger subconcepts in the minority class?

Research Question 2: Can instance weighting the evaluation measure correct the bias towards the larger minority subconcept(s)?

Research Question 3: What impact do common class imbalance correction approaches have on evaluation bias involving minority class subconcept?

Research Question 4: Are there other factors, such as subconcept complexity, that limit the potential of instance weighting the evaluation measure to correct the bias?

The remainder of the paper is divided into 4 sections. Section 2 describes the methodology used for our experiments. Section 3 presents the results we obtained on all experiments. Section 4 discusses these results and assesses how they address our three research questions, and Section 5 concludes the paper and suggests avenues for future work.

2. Experimental Methodology

We evaluate our hypothesis with an experimental setup involving imbalanced binary classification datasets including three or more subconcepts of differing prior probability in the minority class. The data sets are classified using random forest classification and the results are reported in terms of both the non-weighted "regular" version of the less biased metrics as well as their weighted version.

Data Set	**Maj.**	**Min.**	**Maj.Class**	**Min.Class**	**I.R.**
Abalone	3,292	15	6, 7, 8, 9, 10, 11, 12	1, 2, 3, 21, 22, 25, 26, 29	219.47
Automobile	123	22	0, 1, 2	-2, -1, 3	5.59
Cleveland	214	22	0, 1	2, 3, 4	9.73
Dermatology	242	35	1, 2, 3	4, 5, 6	6.91
Ecoli	307	3	CP, IM, PP, IMU	IMS, OML	102.33
Glass	175	28	1, 2, 7	3, 5, 6	6.25
Led7digit	271	107	3, 4, 5, 7, 8	0, 1, 2, 6, 9	2.53
Letter	5,585	3,253	U, D, P, T, M, A, X	Z, E, F, S, B, W, Y, N, R, L, O, H, K, C, Q, I, G, J	1.72
Penbased	5,716	1,104	0, 1, 2, 4, 7	3, 5, 6, 8, 9	5.18
Satimage	4,399	2,036	1, 3, 7	2, 4, 5	2.16
Segment	990	308	5, 6, 7	1, 2, 3, 4	3.21
Shuttle	22,170	92	1, 4, 5	2, 3, 6, 7	240.98
Texture	2,500	490	4, 6, 7, 8, 9	2, 3, 10, 12, 13, 14	5.10
Vowel	450	86	5, 6, 7, 8, 9	0, 1, 2, 3, 4, 10	5.23
Wine quality (White)	4,535	210	5, 6, 7	3, 4, 8, 9	21.60
Yeast	1,350	68	CYT, NUC, MIT, ME3, ME2	ME1, POX, EXC, VAC, ERL	19.85

Table 1: This table indicates the multiclass datasets selected for our experiments and shows the way in which they were transformed into binary classification problems.

2.1. Data Preparation

We selected 16 multiclass numerical imbalanced datasets in the Keel repository with a wide range of imbalance ratios, instances, and feature sizes. We created binary problems from these multiclass problems to simulate the type of situation that occurs in data sets with complex but unknown class distributions. In particular, we focus on domains whose classes have multimodal distributions. Although there is not always information about the composition of each class in natural domains, here, we have the advantage to know exactly what

subclasses, i.e., subconcepts, are present in each class. Furthermore, we can manipulate the size of these concepts in order to test extreme conditions where the subconcepts of the minority class are very small.

Each synthetic binary dataset is composed from a parent multiclass dataset using the same procedure. Given a multiclass dataset with C classes, we sort the class labels from most to least frequent, and partition the classes into two groups. The majority class group includes each of the $\lfloor C/2 \rfloor$ largest classes, and the minority class contains the remaining classes. In addition, we engineered the minority class of the newly created binary domains so that each successive subconcept in the minority class is decreased in size by half.

The composition of each of the generated binary data sets is shown in Table 1. The first column indicates the name of the data set; the second and third columns indicate the number of majority (Maj.) and minority (Min.) instances, respectively; the fourth and fifth columns indicate the composition of the majority (Maj.Class) and minority classes (Min.Class), respectively by listing the original classes that were joined together to form each new binary class; the last column shows the imbalance ratio (I.R) of the dataset.

2.2. Evaluation Measures

In the following experiments, we assess the minority subconcept bias in three evaluation measures AUC, balanced accuracy (BA) and the F-measure (F1). These are commonly used in binary imbalanced classification. We also assess instances weighed versions of each of these measures (denoted wAUC, wBA and wF1.) In the instances weighting, each majority class instances receives a weight equal to 1. Alternatively, each minority instances is weighted according to the ratio between the size of the largest majority subconcept and the size of the minority class instance's parent subconcept in the training set. This serves to give each minority subconcept equal influence on the final score. The instances weights are assigned via the *sample_weight* parameter in the scikit learn evaluation measure functions.

2.3. Classification

For each dataset, we ran two series of experiments: one series where no action was taken to counter the class imbalance problem and a second series where class imbalance was addressed in different ways. We used the Random Forests (RF) classifier as it is one of the mostly widely used in practice and has been adapted for imbalanced learning. The following common methods were used for imbalance correction: Random Oversampling (ROS), Random Undersampling (RUS), SMOTE and cost sensitive RF (weight). The models and imbalance correction hyper-parameters are optimized for the standard AUC, balanced accuracy, and F-Measure and compared with respect to both the standard measures and instance weighted measures on 5x2-fold stratified cross-validation runs. We use the classifier implementations from scikit-learn and imbalance correction techniques implemented in the imblearn package.

2.4. Assessment Procedure

The following experimental procedure is used to explore our 4 research questions. As described above, we run 5x2-fold stratified cross-validation for each dataset. We train RF

models on the full training sets for each engineered dataset contained in this study. In order to better understand the performance estimates and how they related to the long-tail priors of the minority class subconcepts, we extract performance statistics from multiple subsets of the testing partitions.

For each of the standard and instances weighed evaluation measures, we record the performance estimates over the full test set, X_{tst}^{full}. In addition, we record the performance estimates over all of the majority class test instances and the instances of each minority subconcept separately, $X_{tst}^{sub_i} = X_{tst}^{majoriy} \cup X_{tst}^{minority_i}$, $i \in \{1, ..., |S|\}$, where $|S|$ is the number of subconcepts in the minority class. For convenience, the subconcepts are sorted in descending order by size. We denote the subsets with the largest ($i = 1$) and smallest ($i = |S|$) minority subconcepts, as $X_{tst}^{largest}$ and $X_{tst}^{smallest}$, respectively.

3. Experimental Results

In this section, we proceed to explore the answer to each question posed in the introduction.

Table 2: The Pearson correlation (p-values following in brackets) for each evaluation measure between X_{tst}^{full} and $X_{tst}^{largest}$ and between X_{tst}^{full} and $X_{tst}^{smallest}$. The results show that performances estimates on the full test set are biased towards perform one the largest minority subconcept. The instances weighted measures partially correct the bias.

Subconcpet	**AUC**	**BA**	**F1**	**wAUC**	**wBA**	**wF1**
Largest	0.975 (7.96e-12)	0.970 (3.88e-11)	0.963 (1.42e-10)	0.900 (3.81e-07)	0.883 (1.21e-06)	0.822 (2.91e-05)
Smallest	0.548 (0.019)	0.555 (0.017)	0.491 (0.039)	0.656 (0.003)	0.688 (0.002)	0.643 (0.004)

3.1. RQ1: Are the standard measures biased towards larger minority subconcepts.

We explore this question by calculating the correlation between the overall test performance (calculated on X_{tst}^{full}) and the performance on the test subset with only the largest minority subconcept ($X_{tst}^{largest}$) and only the smallest minority subconcepts ($X_{tst}^{smallest}$). We hypothesize that if the standard measures are biased, we will find a much stronger correlation with the largest minority subconcept than with the smallest minority subconcept.

We present the results of this assessment in the first three columns of Table 2. For example, the (AUC, Largest) cell in the table shows the Pearson correlation between AUC(X_{tst}^{full}) and AUC($X_{tst}^{largest}$), and the (AUC, smallest) cell shows the Pearson correlation between AUC(X_{tst}^{full}) and AUC($X_{tst}^{smallest}$). The results show a very strong positive correlations between the scores on X_{tst}^{full} and $X_{tst}^{largest}$ for each measure. The Pearson correlations are 0.975, 0.970 and 0.963 for AUC, BA and F1, respectively. Alternatively, the Pearson correlations between the scores on X_{tst}^{full} and $X_{tst}^{smallest}$ are 0.548, 0.555 and 0.491 for AUC, BA

and F1, respectively. This indicates a much weaker positive correlation with respect to the smallest minority subconcept. The mean difference between the correlations on the largest and smallest minority subconcepts is 0.438.

These results suggest that the estimates of the models' overall performances by the standard measures are largely dictated by the predictions made on the majority class and the largest subconcept in the minority class. This could be very problematic in applications where the model is required to treat all subgroups equally. Thus, we find that it is clear from these results that the assessment of a model's performance on imbalanced data may be flawed if an imbalance, or long-tail, also occurs within the minority class subconcepts.

3.2. RQ2: can instance weighting the evaluation measure correct the bias towards the larger minority subconcept(s)

Having seen that the performance estimates of standard measures are biased by the model's effectiveness on the largest minority subconcept, in this section we proceed to consider if instance weighting the standard evaluations measures based on the relative frequency of minority subconcept in the training data serves to ameliorate the bias. To assess this, we repeat the analysis carried out in RQ1, but replace the standard measures (AUC, BA, and F1) with the instance weighted versions (wAUC, wBA, and wF1). If the weighted measures correct the bias, we expect to find the gap between the correlation of the performance estimates on X_{tst}^{full}, and $X_{tst}^{smallest}$, and X_{tst}^{full} and $X_{tst}^{largest}$ to be significantly reduced.

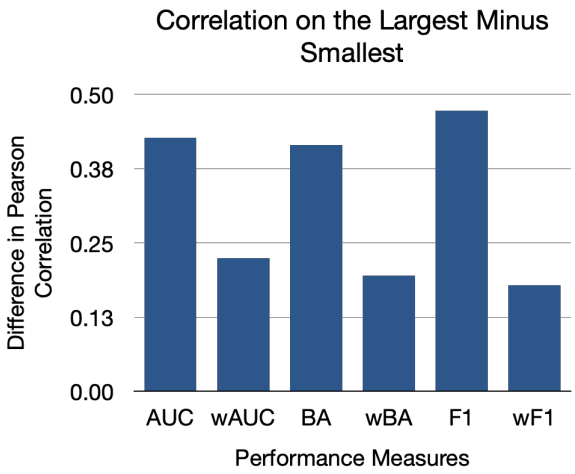

Figure 1: The difference in the Pearson correlation between performance on X_{tst}^{full} and $X_{tst}^{largest}$ and on X_{tst}^{full} and $X_{tst}^{smallest}$ for the standard and instances weighted measure.

The results for this analysis are in the last three columns of Table 2 (wAUC, wBA, wF1). The table shows that the gap between the performance estimate correlations involving $X_{tst}^{largest}$ and $X_{tst}^{smallest}$ have been narrowed, but not closed. The Pearson correlations for performance on X_{tst}^{full} and $X_{tst}^{largest}$ are 0.900, 0.883 and 0.822 for wAUC, wBA and wF1, respectively. This a mean decrease in correlation of 0.118 from the corresponding standard measures. More importantly, the correlations between X_{tst}^{full} and $X_{tst}^{smallest}$ increase to 0.656, 0.688 and 0.643 for wAUC, wBA and wF1, respectively. This represents a mean increase in correlation of 0.131. This indicates that the performance on the smallest minority subconcept has an elevated influence on the performance estimates from the complete test set X_{tst}^{full} when the weighted measure is used. We see this as a reduction in the performance estimation bias.

Finally, Figure 1 shows the difference in the Pearson correlation between performance on X_{tst}^{full} and $X_{tst}^{largest}$ and on X_{tst}^{full} and $X_{tst}^{smallest}$ for the standard and instances weighted measure. The differences are much smaller when the weighted measures are used. This provides additional evidence that the instance weighted measures provides performances estimates that are less biased towards the largest minority subconcepts.

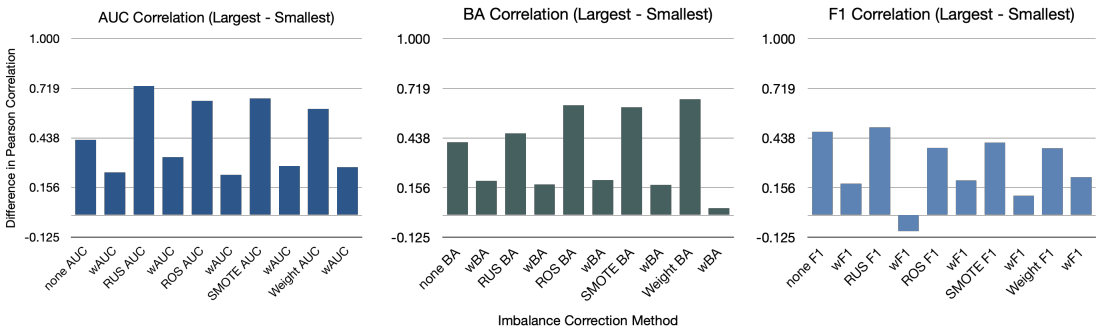

Figure 2: The difference between the Pearson correlations on X_{tst}^{full} and $X_{tst}^{largest}$ and on X_{tst}^{full} and $X_{tst}^{smallest}$ AUC and wAUC (left), BA and wBA (centre) and F1 and wF1 (right) after the application of the imbalance correction methods.

3.3. RQ3: What impact do common class imbalance correction approaches have on evaluation bias involving minority class subconcept?

In this section, we explore the impact of four common imbalance corrections techniques on the performance estimation bias in the standard measures and instance weighted measures. In particularly, we examine if the imbalance correction methods increase or decrease the bias with respect to the largest minority subconcepts.

Figure 2 plots the difference between the Pearson correlations between X_{tst}^{full} and $X_{tst}^{largest}$ and between X_{tst}^{full} and $X_{tst}^{smallest}$ for AUC and wAUC (left), BA and wBA (centre) and F1 and wF1 (right) after the application of the imbalance correction methods. The first two bars in each plot correspond to the difference according to the standard measure and weighted measure without any imbalance correction. These serve as our baseline to understand if the bias is increased or decreased. The remaining 8 bars correspond to the standard measure and weighted measure after the application of RUS, ROS, SMOTE and cost sensitive RF (weight).

The results shows a similar pattern for AUC and BA, along with similar pattern for wAUC and wBA. Specifically, when the performance is estimated with the standard measures, all imbalance correction methods cause a large increase in the correlation gap between X_{tst}^{full} and $X_{tst}^{largest}$ and between X_{tst}^{full} and $X_{tst}^{smallest}$. With AUC, for example, the correlation gap increases from approximately 0.44 to approximately 0.70 after imbalance correction is applied. Alternatively, when the correlation gaps are measured with instance weighed AUC and BA, there is little noticeable difference with and without imbalance correction. The one exception is for wBA with the cost sensitive RF. Here, the correlation gap is reduced from

the uncorrected baseline. With respect to F1 and wF1, there is minimal difference in the correlation gap, except for RUS with wF1 where the correlation between the performance estimates on X_{tst}^{full} and $X_{tst}^{smallest}$ is greater than that corresponding to X_{tst}^{full} and $X_{tst}^{largest}$. This produces a negative value in the plot.

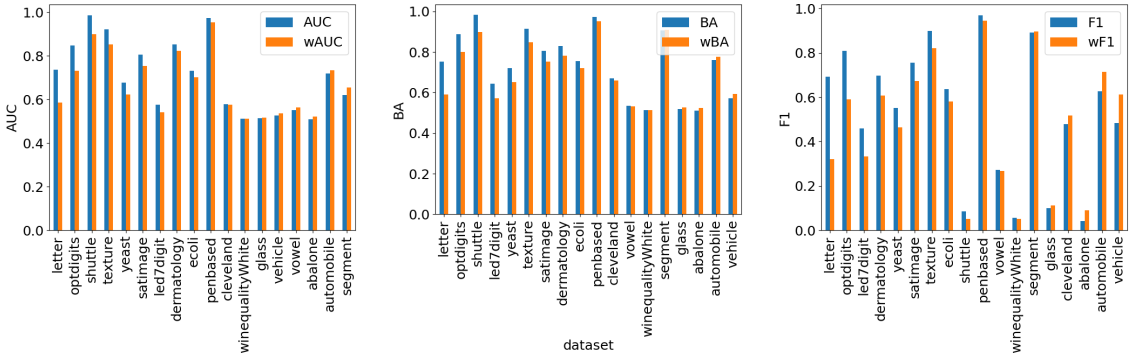

Figure 3: Comparison of the performance estimates on each dataset by the standard measures (AUC, BA, F1) and the instance weighed versions (wAUC, wBA, wF1.)

Our analysis of the detailed results revealed that the increased correlation gap for AUC and BA is due to a significant decrease in the correlation between the performance on X_{tst}^{full} and on $X_{tst}^{smallest}$ after the imbalance correction is applied. Specifically, the imbalance correction causes a mean decrease in the correlations for (X_{tst}^{full} and $X_{tst}^{largest}$) of 0.009 and 0.075 for AUC and wAUC, respectively. Alternatively, there is a mean decrease with respect to (X_{tst}^{full} and $X_{tst}^{smallest}$) of 0.243 and 0.109 for AUC and wAUC. A similar trend exists for BA and wBA. The pattern is less clear cut for F1 and wF1. Therefore, the imbalance correction methods potentially decrease the model's performance on the smallest minority subconcept, at least when performance is estimated with AUC or BA.

3.4. RQ4: Are there other factors that limit the potential of instance weighting the evaluation measure correct the bias?

In RQ2, we found that although instance weighting the evaluation measures reduces the bias towards the larger subconcepts, it does not entirely remove it. In this section, we aim to understand where and why instance weighted has a positive impact on bias reduction in performance estimate, and we explore the other factors that influence the outcomes.

Figure 3 shows the 5x2-fold stratified cross validated mean score for the standard measures and the weighted measures on X_{tst}^{full}. The blue bars correspond with the performance according to the standard measures and the orange bars correspond with the instances weighted versions. The plots are sort with respect to the difference between the two measures (*e.g.* $AUC - wAUC$). Therefore, datasets where the weighted measure is more pessimistic than the standard measure appear towards the left in each plot.

For approximate half of the datasets, the weighted measure is more pessimistic ($M > wM$) than the standard measure. A lower score by weighted measure suggests that it has

adjusted the performance estimate downward to account for relatively poor performance on the smaller minority subconcepts. The plots corresponding with each measure also show that there are a few datasets where the standard and instance weighted measures give approximately the same performance estimate ($M \approx wM$) and where the weighted measure is more optimistic than the standard measure ($M < wM$).

A close examination of the sorting of the datasets in Figure 3 reveals that there is reasonable agreement about the datasets in the ($M > wM$), ($M \approx wM$) and ($M < wM$) categories. Of the top 6 datasets in the $M > wM$ category for AUC and BA, 5 are the same and 3 of the top 6 are same across all three measures. The wine quality, glass, vowel and penbased are in the $M \approx wM$ category for all measures. Finally, abalone, automobile are in the $M < wM$ category for all measures. Given that by design all of the minority subconcepts follow a long-tail distribution, the $M \approx wM$ and $M < wM$ situations suggest that factors aside from the concept priors are influencing the outcome. With this in mind, we proceed to explore each of the three scenarios by selecting archetypal datasets to scrutinize in more detail.

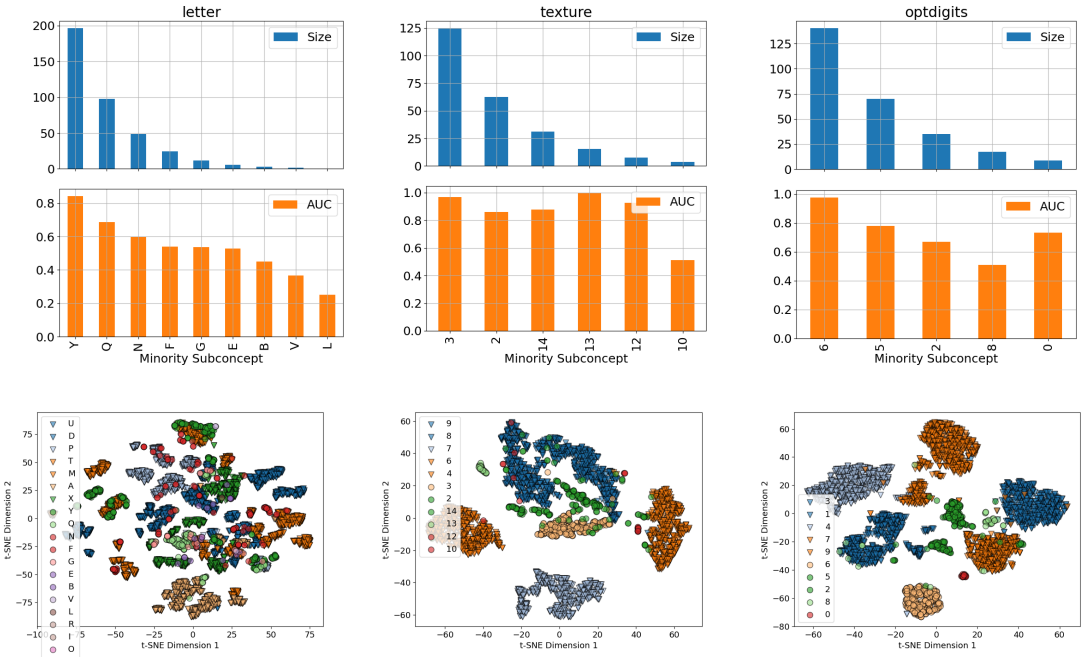

Figure 4: Analysis of three datasets in the $M > wM$ category. The top row presents the mean minority subconcept training frequency, the middle row presents the mean per-minority subconcept AUC and the bottom row shows the TSNE plots for the letter, texture and optdigits datasets.

Figure 4 highlights some of the underlying scenarios associated with a more pessimistic weighted measures. The columns in the figure are associated with the letter, texture and optdigits datasets, respectively. The top row shows the mean minority subconcept training

size, the middle row shows the per-subconcept AUC (AUCs for $X_{tst}^{sub_i}$, $i \in \{1, ..., |S|\}$) and the final row shows the corresponding TSNE plots. In the TSNE plot, majority class samples are shown as triangles and the minority class samples are shown as circles. The colour of each marker signifies the parent subconcept.

In the case of the letter dataset, we see a relatively consistent decrease in subconcept AUC (middle row) with subconcept size in the training set (top row). The mean AUC over the full test set (X_{tst}^{all}) is 0.736 whereas by inflating the importance of the tail of the subconcept distribution with wAUC, this reduces this to 0.585. This a classic case where de-biasing with instances weighting helps. For texture dataset, all of the minority subconcepts, except subconcept 10, are relatively easy. This is emphasized by the TSNE plot where the instances from most subconcepts, except subconcept 10, are separated from the majority class. Alternatively, samples of subconcept 10 overlap with the majority class. Here, we see a more modest de-biasing effect because learned model produces similar high scores on the other rare minority subconcepts, such as 12 and 13. The mean standard AUC is 0.922 and the wAUC is 0.852 for X_{tst}^{all}. This appears to be a reasonable correction in the overall performance estimate given the individual subconcept scores.

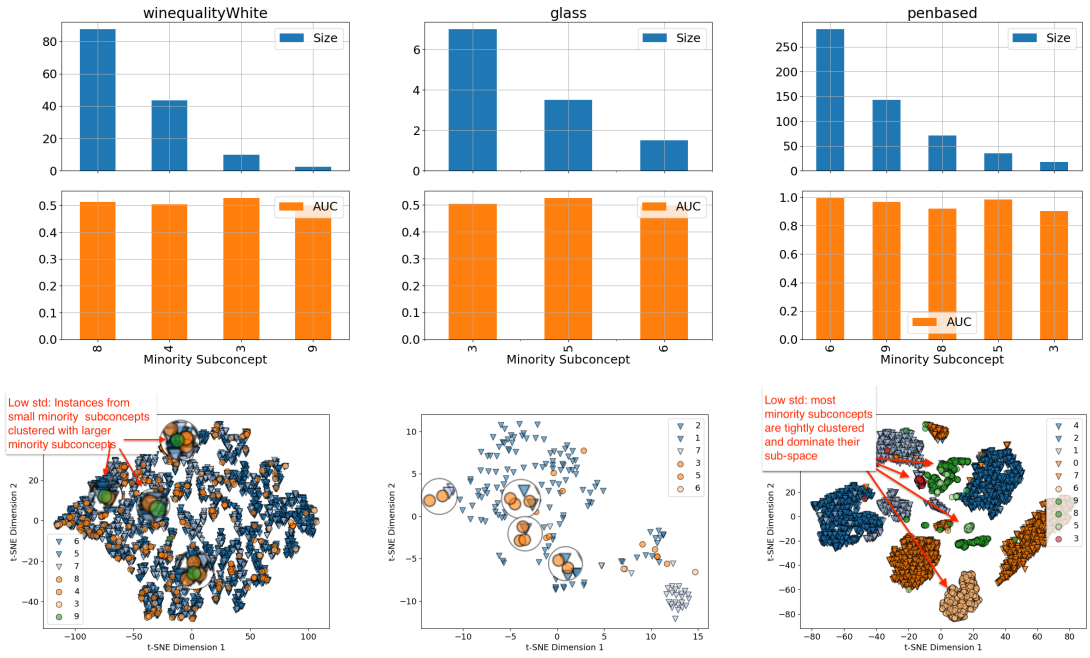

Figure 5: Analysis of three datasets in the $M \approx wM$ category. The top row presents the mean minority subconcept training frequency, the middle row presents the mean per-minority subconcept AUC and the bottom row shows the TSNE plots for the wine quality, glass and penbased datasets.

The optdigits dataset presents the final situation. In it we see a decrease in performance that correlates with the minority subconept priors, until subconcept 0. Subconcept 0 has

an AUC similar to the more frequent subconcepts 2 and 5. The TSNE plot reveals that subconcept 0 is densely packed and isolated from the majority class likely allowing it to be more easily classified. Although the least frequent subconcept is not the most difficult, wAUC adjusts the AUC score downward to account for the lower performances on the other less common subconcepts. The standard AUC is 0.848 and the wAUC is 0.730 for X_{tst}^{all}. Once again, this appears to be a reasonable adjustment.

Figure 5 depicts three datasets from the $M \approx wM$ category where there is little difference between the standard measures and the weighted versions. The common pattern on these datasets is illustrated in the middle row of the figure. Specifically, the per-subconcept performance is independent of the subconcepts priors. Through an visual inspection of the corresponding TSNE plots, we see two categories of distributions that lead to this outcome. Either the smallest subconcepts are clustered with instances from the larger minority subconcepts (*e.g.* wine quality, glass), or all subconcepts are tightly clustered and dominate their subspaces. Both situations render the learning challenge homogeneous across subconcepts. The AUCs and wAUCs for X_{tst}^{all} are (0.512, 0.513), (0.514,0.517) and (0.974,0.953) for wine quality, glass and penbased, respectively. Thus, in this situations no de-biasing is needed and the weighted measures have little impact.

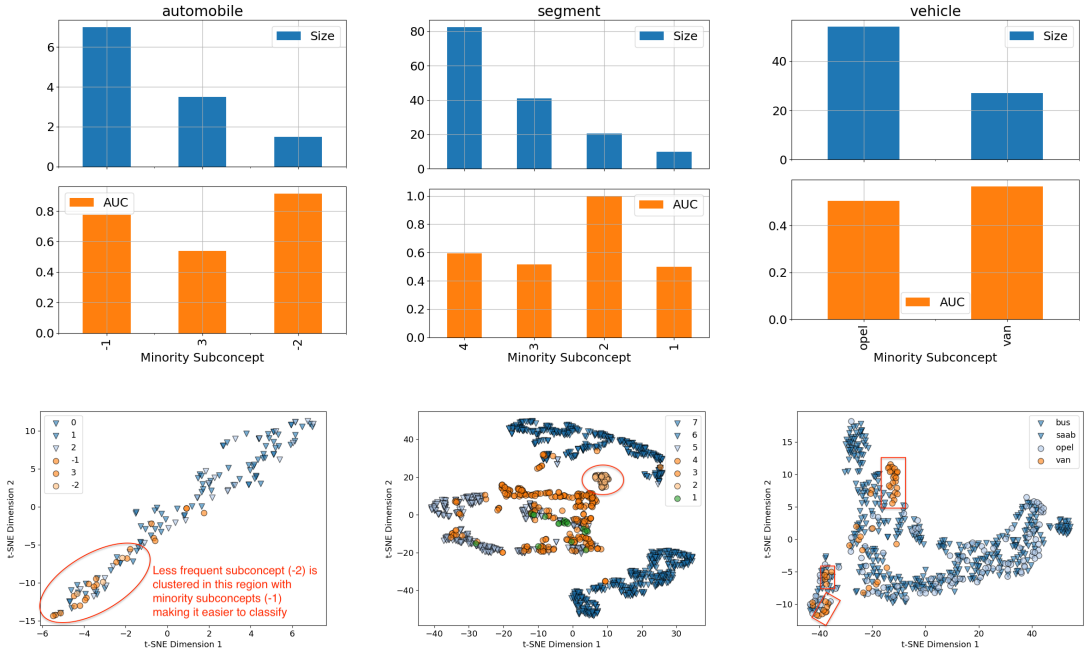

Figure 6: Analysis of three datasets in the $M < wM$ category. The top row presents the mean minority subconcept training frequency, the middle row presents the mean per-minority subconcept AUC and the bottom row shows the TSNE plots for the segment, vehicle, and automobile datasets.

Finally, Figure 6 shows the $M < wM$ category. As demonstrated by the AUC plots in the middle row, in this scenario, the performance on at least one of the smallest subconcept (subconcept 2, van, -2 in segment, vehicle and automobile, respectively) is noticibly better than the larger minority subconcepts. Here, the instance weighting pushes the performance estimate of the weighted measure above that of the standard measure. The corresponding TSNE plots show that the high performing, but less frequent, subconcept has better separation from the majority class than its more frequent counterparts. For applications where the performance estimate should equally reflect the model's proficiency on all minority subconcpts, the increase in the score by wAUC is, once again, an appropriate outcome.

4. Discussion

Our results show that in the majority of cases, the performance of the models looks worse when considered in terms of the weighted metrics thus suggesting that the standard metrics are biased towards larger minority class subconcepts and that the models perform relatively worse on the less frequent minority class subconcepts. Like the minority class itself, the less frequent subconcepts may be critically important and/or it may be required to treat all subgroups equally. In such cases, a weighted measure should be consider or at the very least the risks of bias in the minority class subconcept should be assessed. In addition, we found that the bias in the AUC and BA estimates of the standard measures appears to become worse after the application of the standard imbalance correction techniques. We believe that this may be indicative of the imbalance correction techniques reinforcing imbalance in the minority subconcepts. On the other hand, the estimates from wAUC and wBA seem to effectively account for the additional bias after imbalance correction. Nonetheless, additional research is required on the relationship between minority subconcept bias and imbalance correction.

4.1. Limitations

The findings of this work are limited by the scope of the experiments conducted. In particular, the experiments employ a single classifier and only a subset of the possible imbalance correction techniques are considered. In order to create a controlled setting under which to study this problem, we have manipulated benchmark datasets to form imbalanced binary classification problems with long-tail minority subconcepts. Changing the hyperparameters of the setup, such as the benchmark datasets, imbalance ratios, and the selection and order of the minority and majority parent classes could lead to variations in our results. Finally, it is possible that real-world subconcepts have different (perhaps more challenging) properties than the artificial subconcepts in our evaluation. Although we hypothesize that our findings generally extend to other standard methods and data distributions, additional work is required to confirm this hypothesis.

5. Conclusion and Future Work

This work focuses on highlighting the problem of bias in standard imbalanced measures with respect to subconcepts in the minority class by creating test scenarios in which the subconcept priors are known in advance. Future work will repeat these experiments using

various kinds of techniques for addressing the class imbalance problem. We expect that such an experiment will shed light on the true value of different methods when a more precise, weighted, evaluation metric is used. Furthermore, we should consider leveraging empirical information available in the training set, such as the frequency in data clusters and instance complexity, when assigning instances weights in the evaluation. Our goal is to design a method that will do just that and thus improve upon current methods with respect to the more precise evaluation metrics.

References

Colin Bellinger, Shiven Sharma, and Nathalie Japkowicz. One-class versus binary classification: Which and when? *2012 11th International Conference on Machine Learning and Applications*, 2:102–106, 2012.

Colin Bellinger, Shiven Sharma, Osmar R Zaiane, and Nathalie Japkowicz. Sampling a longer life: Binary versus one-class classification revisited. In *LIDTA@PKDD/ECML*, 2017.

Colin Bellinger, Chris Drummond, and Nathalie Japkowicz. Manifold-based synthetic oversampling with manifold conformance estimation. *Machine Learning*, 107:605–637, 2018.

Paula Branco, Luís Torgo, and Rita P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49:1 – 50, 2016.

N. Chawla, K. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *ArXiv*, abs/1106.1813, 2002.

Damien A. Dablain, B. Krawczyk, and N. Chawla. Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE transactions on neural networks and learning systems*, PP, 2021.

Alberto Fernández, Salvador García, Mikel Galar, Ronaldo Cristiano Prati, B. Krawczyk, and Francisco Herrera. Learning from imbalanced data sets. In *Cambridge International Law Journal*, 2018.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21:1263–1284, 2009.

Jason Van Hulse, Taghi M. Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In *International Conference on Machine Learning*, 2007.

Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6:429–449, 2002.

Taeho Jo and Nathalie Japkowicz. Class imbalances versus small disjuncts. *SIGKDD Explor.*, 6:40–49, 2004.

Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6:1–54, 2019.

György Kovács. Smote-variants: A python implementation of 85 minority oversampling techniques. *Neurocomputing*, 366:352–354, 2019.

Miroslav Kubát, Robert C. Holte, and Stan Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30:195–215, 1998.

Mingchen Li, Xuechen Zhang, Christos Thrampoulidis, Jiasi Chen, and Samet Oymak. Autobalance: Optimized loss functions for imbalanced data. In *Neural Information Processing Systems*, 2022.

Charles X. Ling and Chenghui Li. Data mining for direct marketing: Problems and solutions. In *Knowledge Discovery and Data Mining*, 1998.

Charles X. Ling and Victor S. Sheng. Cost-sensitive learning and the class imbalance problem. 2008.

Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39:539–550, 2009.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021.

Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1695–1704, 2019.

Ronaldo Cristiano Prati, Gustavo E. A. P. A. Batista, and Maria Carolina Monard. Class imbalances versus class overlapping: An analysis of a learning system behavior. In *Mexican International Conference on Artificial Intelligence*, 2004.

Troy Raeder, George Forman, and N. Chawla. Learning from imbalanced data: Evaluation matters. 2012.

Miriam Seoane Santos, Pedro Henriques Abreu, Pedro J. García-Laencina, Adélia Simão, and Armando Carvalho. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *Journal of biomedical informatics*, 58: 49–59, 2015.

Show-Jane Yen and Yue-Shi Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Syst. Appl.*, 36:5718–5727, 2009.

Jue Zhang and Li Chen. Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. *Computer Assisted Surgery*, 24:62 – 72, 2019.

Bing Zhu, Bart Baesens, and Seppe vanden Broucke. An empirical comparison of techniques for the class imbalance problem in churn prediction. *Inf. Sci.*, 408:84–99, 2017.