Proceedings of Machine Learning Research 74:64–78, 2017 Full Paper - LIDTA 2017

Sampling a Longer Life: *Binary versus One-class classification Revisited*

Colin Bellinger CBELLING@UALBERTA.CA
University of Alberta, Computing Science
Alberta Machine Intelligence Institute
Edmonton, Canada

Shiven Sharma SHIVEN.CHEEMA@GMAIL.COM
Fluent Solutions Inc.
Ottawa, Canada

Osmar R. Zaïane ZAIANE@UALBERTA.CA
University of Alberta, Computing Science
Alberta Machine Intelligence Institute
Edmonton, Canada

Nathalie Japkowicz NATHALIE.JAPKOWICZ@AMERICAN.EDU
American University, Department of Computer Science
Washington, USA

Editors: Luís Torgo, Bartosz Krawczyk, Paula Branco and Nuno Moniz.

Abstract

When faced with imbalanced domains, practitioners have one of two choices; if the imbalance is manageable, sampling or other corrective measures can be utilized in conjunction with binary classifiers (BCs). Beyond a certain point, however, the imbalance becomes too extreme and one-class classifiers (OCCs) are required. Whilst the literature offers many advances in terms of algorithms and understanding, there remains a need to connect our theoretical advances to the most practical of decisions. Specifically, given a dataset with some level of complexity and imbalance, which classification approach should be applied? In this paper, we establish a relationship between these facets in order to help guide the decision regarding when to apply OCC versus BC. Our results show that sampling provides an edge over OCCs on complex domains. Alternatively, OCCs are a good choice on less complex domains that exhibit unimodal properties. Class overlap, on the other hand, has a more uniform impact across all methods.

Keywords: one-class classification, imbalanced data, class imbalance, sampling, data complexity, class overlap, data modality

1. Introduction

It is well-known that the performance of binary (discriminatory) classifiers can suffer when trained on imbalanced datasets. At prohibitive levels of imbalance, the user is forced to

resort to one-class classifiers, which learn solely from the majority class (Japkowicz (2000)). However, the fact that discriminatory classifiers are generally perceived to be superior has spawned a rich body of research dedicated to managing the impacts of imbalance. Prominent methods in this area include sampling techniques that adjust the training distribution (Chawla et al. (2002); Han et al. (2005); Wallace et al. (2011)) and cost-adjustment methods, which alter the costs associated with erroneous prediction to skew the training bias (Elkan (2001); Wu et al. (2008)).

Regardless of the technique applied, the intention is to prolong the useful lifespan of binary classifiers trained over increasingly imbalanced settings. Doing so is putting the classifier on life-support, so as to harness its discriminatory power as long as possible. In this work, we explore the questions: How long can we extend the lifespan of a binary classifier with sampling before a one-class classifier must be applied, and how is this outcome impacted by the complexity of the data? We hypothesize that the length of the lifeline offered by a sampling method heavily depends on the properties of the data to which it is applied.

The performance of any classification method will be impacted by both the imbalance ratio as well as the properties of the two classes. Therefore, we delve into this question by explicitly looking into how these properties impact the lifespan of binary classifiers as imbalance increases. This is studied in relation to the performance of one-class classifiers. Our experiments over artificial and benchmark datasets reveal that for complex domains, sampling methods make binary classifiers more robust to increasing imbalance. We find that sampling can help binary classifier outperforming even one-class classifiers at high levels of imbalance. The impact of sampling is less prominent over simpler domains. Here, one-class classifiers become superior at moderate levels of imbalance.

Section 2 presents an overview of work done in rectifying the imbalance in data for use with binary classifiers. We proceed with the formalization of our question in Section 3. In Section 4, we outline our framework for conducting the experiments over artificial and benchmark datasets, followed by the presentation and analysis of the results in Section 5. Finally, we conclude our study in Section 6.

2. Related Work

Early literature on the topic found that class imbalance causes poor predictive performance in standard classifiers. As a result, a great deal of research has been devoted to designing algorithms that are capable of dealing with imbalanced classification tasks, such as by modifying the training distribution by adding or removing instances (Chawla et al. (2002); He et al. (2008); Wallace et al. (2011); Bellinger et al. (2016)), or by adjusting the cost associated with errors made by training (Domingos (1999); Elkan (2001); Wu et al. (2008)).

Apart from sampling and cost-based methods, other researchers have focused their attention to understanding the nature of the imbalanced learning problem itself (Japkowicz and Stephen (2002); Prati et al. (2004); Batista et al. (2004); García et al. (2007); Denil and Trappenberg (2010)). An important early finding to come out of this work was that imbalance alone is not the problem; the properties of the data have a major impact on predictive performance as well. Japkowicz and Stephen (2002), for example, demonstrated that classifiers are not sensitive to any level of imbalance when the target classes do not

overlap. As complexity, in terms of modality and overlap, increases, systems become more sensitive to imbalance. Similarly, Prati *et al.* also studied the impact of imbalance and complexity; by altering the level of overlap as well as imbalance over synthetic datasets, they found overlap to be the major complicating factor.

In cases of extreme imbalance, research looked into the advantage that one-class classifiers could provide over binary classifiers. Japkowicz (2000), for example, compared binary to a one-class neural network with and without under and oversampling on a one-dimensional backbone dataset, and found sampling+binary to be superior. However, the results reported by (Lee and Cho (2006)) and (Bellinger et al. (2012)) contradicted these earlier findings. In both cases, one-class classification was found to be preferable on highly imbalanced data.

Whilst these studies have considered many aspects of the relationship between binary versus one-class classification on imbalanced datasets, a more comprehensive study is needed. The aforementioned work applied very few classification and sampling methods, and in the synthetic datasets, the individual aspects of complexity were not controlled separately. As a result, it is unclear how the choice of binary versus one-class classification is impacted by factors which are known to have a great impact on performance.

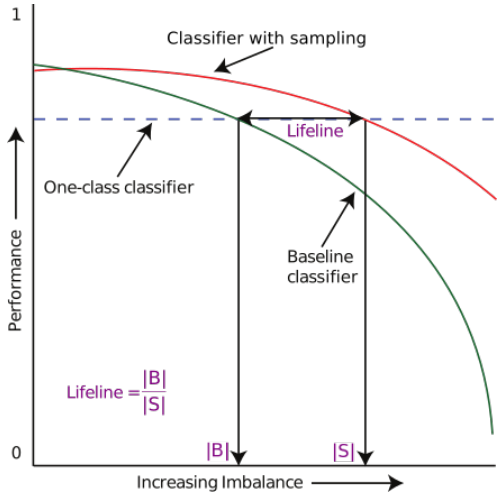
3. Assessing Imbalance, Complexity and Classification Paradigms

Existing research gives us a number of insights about class imbalance. We know, for example, that binary classifiers are often preferable to one-class classifiers even when there is imbalance (Bellinger et al. (2012)), and performance is impacted by domain complexity as well. Likewise, we know that by choosing a good corrective strategy, such as resampling or cost-adjustment, the negative impact of imbalance can be mitigated, at least to some extent (He et al. (2008)). On the extreme imbalance end of the spectrum, it has been shown that in practice one-class classifiers often perform poorly on complex, multi-modal distributions (Sharma (2016)).

However, what are the ties that bind all these facts? Specifically, given a binary classification dataset D with class prior probabilities p and $1 - p$, and a theoretical domain complexity score c, which classification paradigm should one apply? Although the existing literature clearly shows that a number of factors impact this decision, the lack of a unifying framework makes it hard to make a clear decision.

In this paper, we assess the relationship between imbalance, complexity and the choice of classification paradigm (binary, sampling+binary or one-class). Specifically, we study the degree to which sampling methods can extend the life of binary classifiers relative to one-class classifier on imbalanced domains, and relate it to the domain complexity in terms of class overlap and modality, and the performance of one-class classifiers. We introduce the *lifeline* measure, ℓ , which represents how much more imbalance sampling can allow a classifier to tolerate, relative to the performance of one-class classifiers. This allows us to establish and analyze the relationship between imbalance, complexity and the choice of classification paradigm. The idea is illustrated in Figure 1. ℓ is calculated as the ratio of the size of the minority class for the baseline and the sampled classifier when their respective performances fall below the best one-class classifier. For example, $\ell = 2$ implies that a

Figure 1: An illustration of the concept of the *lifeline* offered by sampling over increasing imbalance.



binary classifier with sampling can handle imbalance ratios twice as high as an unsampled binary classifier before its performance falls below a one-class classifier.

We commence our analysis by formalizing the questions we want to answer in this paper: *1).* Given a domain with a certain level of complexity, to what extent does sampling binary classifiers prolong their life on imbalanced domains? *2).* To what extent does sampling increase the robustness of binary classifiers to increasing imbalance? *3).* How does the lifeline and robustness relate to the performance of a one-class classifier over the same domain? and *4).* Does one form of complexity impact the robustness, lifeline or one-class classifier performance more than the other?

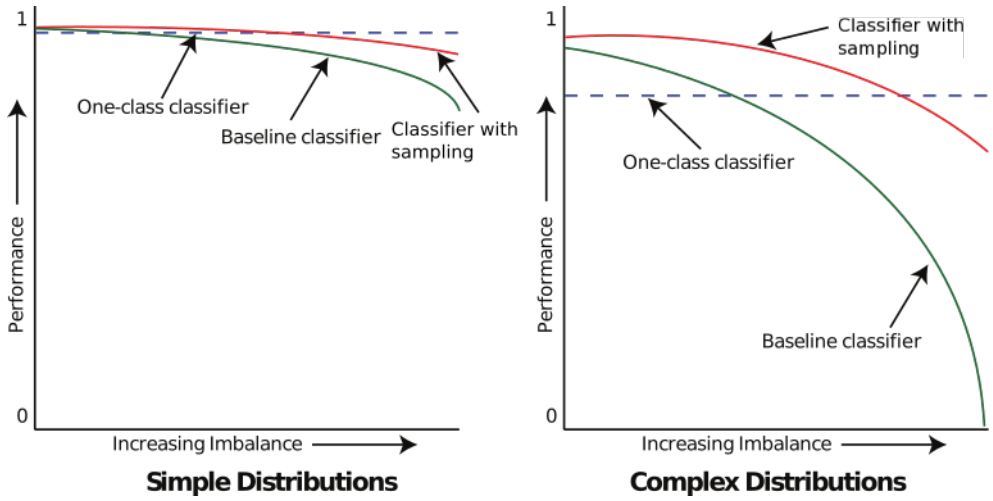
3.1. The Ties That Bind

As a brief recap from the previous section, existing literature gives us the following facts: *a)* binary classifiers (BCs) are preferable to one-class classifiers (OCCs), *b)* high absolute imbalance impacts the performance of BCs, *c)* sampling methods can improve the performance of BCs in imbalanced situations, and *d)* domain complexity impacts both OCCs and BCs.

Given these facts, it follows that as imbalance increases, the rate at which the performance of a BC decreases will be directly proportional to the complexity of the domain. Furthermore, the performance of a OCC will be higher over less complex domains than over more complex domains. We graphically illustrate these observations in Figure 2.

From Figure 2, we can formulate broad answers to the questions we postulated. Regarding the lifeline offered by sampling, we hypothesize that, as imbalance increases, the impact of sampling is more prominent over a complex domain than over a simpler domain. This implies that sampling can allow us to harness the power of a BC for longer over complex domains. Naturally, this would depend on how quickly the performance drops as imbalance increases, *i.e.*, how robust it is. This leads us to the next question; given that domain

Figure 2: Classifier performance trends over increasing imbalance on a simple and complex domain



complexity impacts BCs at high levels of imbalance, we would expect sampling to make them far more robust. The implication of this hypothesis directly relates to the first question. Specifically, the more robustness a sampler provides, the longer the expected lifeline. Relating these hypotheses to the performance of OCCs, we hypothesize that over a complex domain, sampling could potentially provide an advantage over a OCC; if complexity impacts a OCC negatively, sampling can extend the life of a BC such that even at extreme imbalance, it outperforms a OCC.

Here, we offer hypotheses to three of our four questions based on current literature. In subsequent sections, we conduct a series of experiments to test these hypotheses empirically, and provide an answer to our fourth question: does one aspect of complexity, or both, have an impact on classification over increasing imbalance for both paradigms of classification. We begin by detailing our experimental framework in the following section.

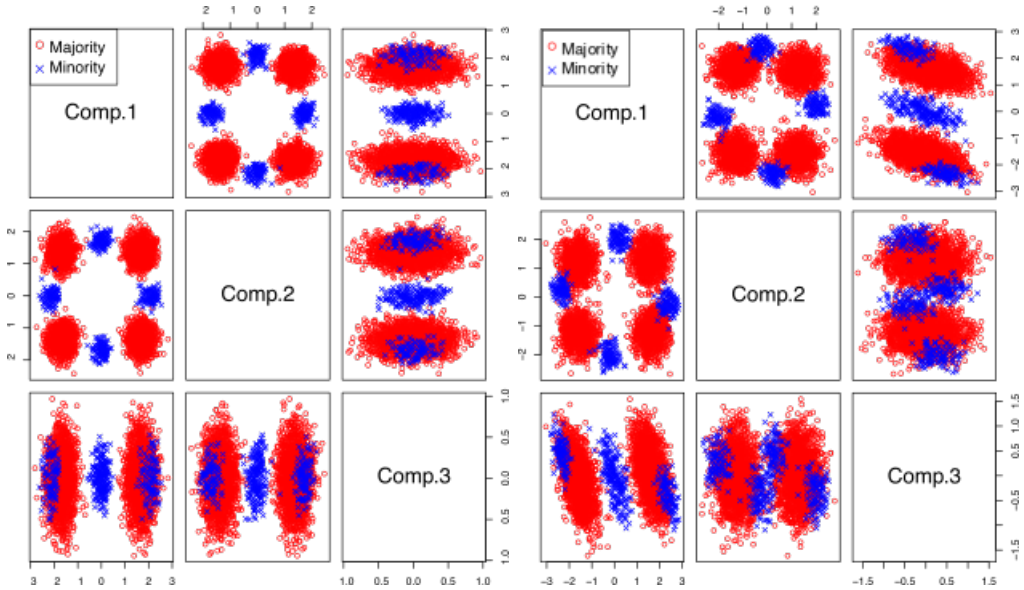
4. Experimental Framework

In this section, we describe the framework employed for our experiments over both the artificial and real-world data. We begin by first describing the datasets employed, followed by the various algorithms and evaluation metrics.

The artificial setting is composed of four 5-dimensional datasets which are various combinations of multi-modal and unimodal majority and minority distributions. Here, we are specifically focused on the modality of the majority class because of the implications for the induction of one-class classifiers. Specifically, we generated two unimodal datasets, one with and one without overlap, and two multi-modal datasets, one with and one without overlap. This latter case is illustrated in Figure 3.The specifications for these are as follows (in each $N(\mu, \sigma)$, μ represents the mean vector, and σ the standard deviation) :

Data 1 : Unimodal target and multimodal outlier distributions:

Figure 3: The first three principle components of the artificial multi-modal datasets without (left) and with (right) overlap



Target : $N([15, 15, 15, 15, 15], 2.75)$

Outlier : $N([5, 5, 5, 15, 15], 2) \cup N([25, 25, 25, 15, 15], 2)$
$\cup N([15, 15, 15, 5, 5], 2) \cup N([15, 15, 15, 25, 25], 2)$

Data 2 : Unimodal target and multimodal outlier distributions:

Target : $N([15, 15, 15, 15, 15], 2.75)$

Outlier : $N([5, 5, 5, 15, 15], 3.75) \cup N([25, 25, 25, 15, 15], 3.75)$
$\cup N([15, 15, 15, 5, 5], 3.75) \cup N([15, 15, 15, 25, 25], 3.75)$

Data 3 : Multimodal target and multimodal outlier distributions, no overlap:

Target : $N([5, 5, 5, 5, 5], 3) \cup N([25, 25, 25, 5, 5], 3)$
$\cup N([5, 5, 5, 25, 25], 3) \cup N([25, 25, 25, 25, 25], 3)$

Outlier : $N([15, 15, 15, 2.5, 2.5], 2) \cup N([27.5, 27.5, 27.5, 15, 15], 2)$
$\cup N([2.5, 2.5, 2.5, 15, 15], 2) \cup N([15, 15, 15, 15, 27.5, 27.5], 2)$

Data 4 : Multimodal target and multimodal outlier distributions, overlap:

Target : $N([10, 5, 5, 10, 5], 3) \cup N([20, 25, 25, 10, 5], 3)$
$\cup N([10, 5, 5, 20, 25], 3) \cup N([20, 25, 25, 20, 25], 3)$

Outlier : $N([17.5, 15, 15, 5, 2.5], 2) \cup N([25, 27.5, 27.5, 17.5, 15], 2)$
$\cup N([5, 2.5, 2.5, 12.5, 15], 2) \cup N([12.5, 15, 15, 15, 25, 27.5], 2)$

Table 1: Datasets employed and associated properties. The datasets marked with † are from the Pattern Recognition Laboratory: http://prlab.tudelft.nl.

Dataset	$\|M\|$	d	*Modality*	*Overlap*
Diabetes	66	8	unimodal	significant
Sonar	24	60	unimodal,spread	significant
Delft pump AR app.†	94	160	multi-modal	significant
Alphabets	749	15	multi-modal	significant
ForestC1	624	54	unimodal	significant
Biomed healthy†	33	5	unimodal,spread	moderate
Waveform 0†	149	21	unimodal	moderate
Heart	30	13	multi-modal	moderate
Cancer wpbc non-ret†	23	33	unimodal	significant
Spambase spam†	906	57	unimodal,spread	moderate
ForestC2C5	600	54	unimodal	moderate
Forest	742	54	bimodal	moderate
Ionosphere	31	34	unimodal	moderate
Arrhythmia normal†	91	278	unimodal	significant

Table 1 lists the benchmark datasets that were considered for our study, along with a description of their complexity with respect to overlap and modality. These insights were gained by analyzing their PCA plots of their first three components. Figures 8, 9 and 10 illustrate the first two principle components of three typical scenarios[1]. Figure 8, contains the PCA plot for the Deft pump dataset. It is shown to be multi-modal and contains class overlap. The PCA plot for the Biomed dataset is depicted as being unimodal with overlap and high variance in Figure 9. Finally, Figure 10 shows the Arrhythmia dataset to be unimodal with significant overlap.

It is important to note that, whilst PCA is commonly applied to visualize high-dimensional data, there alternative approaches the may provide addition insight. For completeness, we cross checked our categorizations with two-dimensional visualizations using T-SNE (Maaten and Hinton (2008)), which is widely applied in manifold learning and deep learning applications. Although, the distribution of the data points often looks different, we found the categorizations to be consistent. In addition to visual analysis, more quantitative methods of assessing domain complexity would also be helpful here. This is an ongoing area of research (Cano (2013); Anwar et al. (2014)), which we are incorporating into our future work.

In order to perform a comprehensive analysis we consider a multitude of binary classifiers, sampling methods and one-class classifiers. The binary classifiers used are Naïve Bayes (NB), Nearest Neighbour (IBK), Decision Trees (J48), Multilayer Perception (MLP) and Support Vector Machines (SVM). The one-class classifiers employed are the Autoencoder (AE), One-class support vector machine (ocSVM) and the Mahalanobis distance.

1. The reminder of the plots can be made available upon request.

With respect to the sampling methods, we employ random oversampling and undersampling, SMOTE (Chawla et al. (2002)) ($K = 3, 5, 7$), Borderline SMOTE (B1,B2) (Han et al. (2005)), SMOTE with one-sided selection with Tomek links (Kubat and Matwin (1997)) and adaptive synthetic sampling (He et al. (2008)).

All our experiments have been conducted in Python using the sklearn (Pedregosa et al. (2011)) library. The binary classifiers were induced with their default parameters. For ocSVM parameters, for a given target rejection rate, random searches were executed over the ν, γ and C parameters. The model that, on average, produced a rejection rate closest to the specified value, was selected as the best. A similar process was applied for the AE, with exception that the random search was applied over the number of epochs, hidden units and noise level.

The performance measure we use is the geometric mean of the per-class accuracy's, a metric immune to class imbalance. It is given by g-mean $= \sqrt{acc_1 \times acc_2}$, where acc_i is the accuracy of the classifier on instances belonging to class i. Evaluation is done by first splitting the data in half and using one half as the training set and the other half as the testing set. In order to account for variance in the samples, and ultimately the results, we ran 30 iterations of this split, and the final g-mean is the average of these runs. We split the dataset in half as some of the UCI datasets are not very large to begin with, and consequently standards like 10-fold cross validation would result in very small testing sets. To simulate the effect of imbalance, we fix the majority to its natural size in the dataset, and exponentially decrease the size of the minority set, until there are only 4 instances in the minority set. We generate 20 such minority sets by letting $o = |minority|$ be the number of minority instances. Then, the size of each minority set is 4^w, where $w \in S$, an 20-element vector of equally spaced values that range from $width$ to 1. The value of $width$ is calculated as $\frac{\lg o}{\lg 4}$.

Finally, given that our study is not concerned with specific classifiers and sampling methods, we do not report results specifically for each sampling method and classifier. Instead, at a particular level of imbalance for a dataset, the result we report for the sampled and unsampled classifier is the maximum for all the different classifier and sampling method combinations.

5. Results and Analysis

We first proceed by examining our hypotheses over artificial data. This is followed by a validation of UCI datasets.

5.1. Artificial Datasets

Figures 4, 5, 6 and 7 illustrate the performance trends of the sampled and unsampled binary classifiers (BCs), as well as the best one-class classifier (OCC), over the unimodal and multi-modal artificial datasets with and without overlap, respectively. The y-axis is the g-mean, whereas the x-axis corresponds to the extent of imbalance (as outlined in the previous section), with 0 implying the original distribution, and 20 implying only 4 minority class instances.

From these plots, we observe that over the unimodal datasets, the best BC performance is nearly perfect up to moderate levels of imbalance. The OCCs exhibit a similarly strong

Figure 4: Classifier trends over the artificial unimodal dataset with overlap.

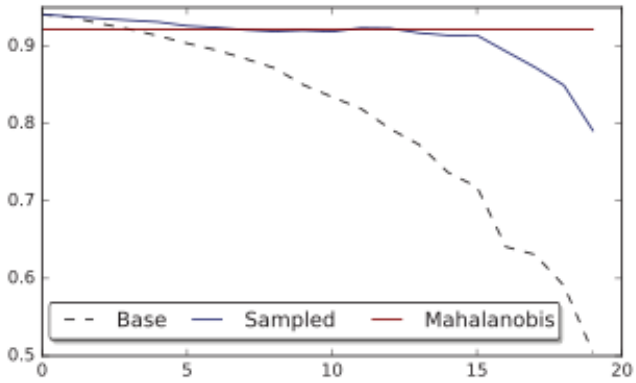


Figure 5: Classifier trends over the artificial multi-modal dataset with overlap.

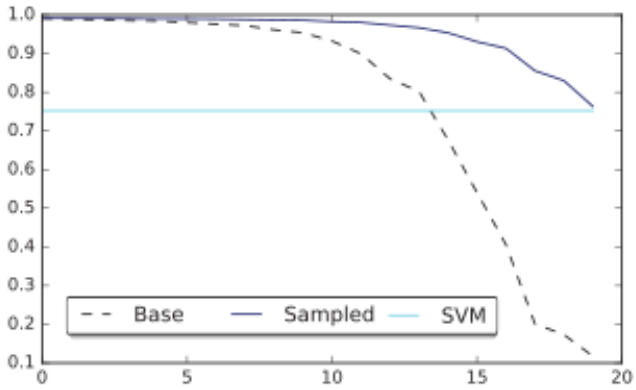


Figure 6: Classifier trends over the artificial unimodal dataset with no overlap.

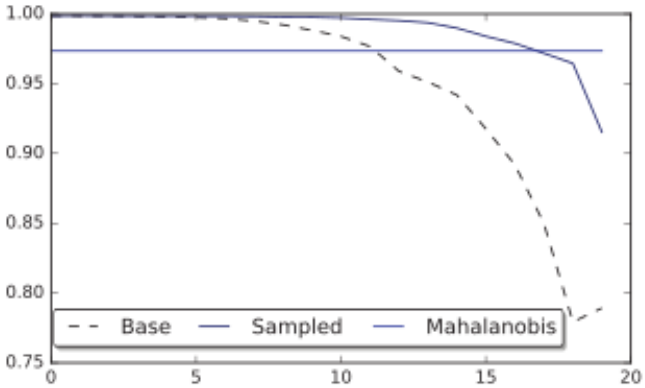
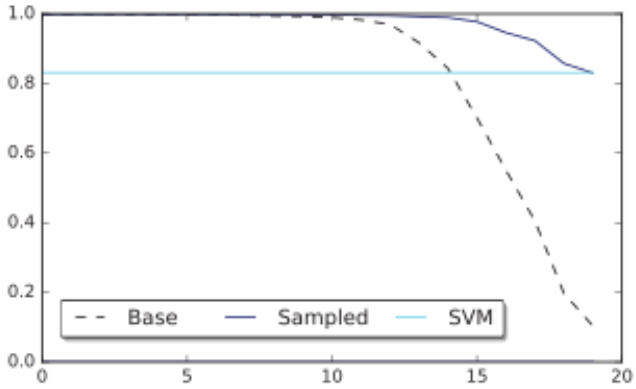


Figure 7: Classifier trends over the artificial multi-modal dataset with no overlap.



performance that is just below the BC. For the unsampled BC, however, we observe a sharp decline in performance as the imbalance increases; this decline is particularly sharp in the case of overlap. The sampled BC, on the other hand, remains robust to the increasing imbalance in both datasets, and only begins to show signs of suffering at extremely high levels of imbalance.

The gap between the g-mean of the best binary approach and best one-class approach is much greater on multi-modal datasets. The OCCs do not exhibit high performance here. Alternatively, the unsampled BC commences with a strong g-mean, but it reaches an imbalance threshold beyond with it can no longer produce an accurate model. As with the unimodal data, this occurs at a moderate levels of imbalance. The sampled BC remains robust to the increasing imbalance; indeed, it maintains a higher g-mean than the OCC even in the most extreme case of imbalance.

With these observations, we return to the questions asked previously. With respect to the *lifeline*, from the figures, we observe that sampling provides a significant boost to the life of BCs; this is particularly evident over the more complex multi-modal domains. When considering *robustness*, over all scenarios, we observe a remarkable increase in the robustness of BCs with sampling, especially at high levels of imbalance. This is particularly pronounced over the multi-modal datases, as is evident by the best and worst g-means; the baselines drop from over 0.99 to just over 0.10, whereas the sampled classifiers are able to stay well over 0.75 even at the most extreme levels of imbalance. When considering the *relationship to OCC performance*, we observe that sampling can lead to better performance even in cases of extreme imbalance over complex domains. On the multi-modal domain without overlap, the sampled BC is only slightly below the best OCC. When overlap is added, however, even at the highest level of imbalance the sampled BC has an advantage over the OCC. Finally, looking at the *relationship with domain complexity*, the results demonstrate that the OCC is more severely impacted by multi-modality than overlap. The unsampled BCs are less robust in the presence of multi-modality when the imbalance ranges from moderate to high. With sampling, however, there is a greater robustness to combined effect of imbalance and modality. Finally, overlap impacts the robustness of the unsampled BCs as well; the OCCs do not seem to suffer much from it.

Table 2: Classifier statistics for the benchmark datasets.

Dataset	B_{nat}	B_{imb}	S_{nat}	S_{imb}	OCC	ℓ	$size$
Delft pump	0.997	0.384	0.999	0.852	0.578	∞	$-/5/94$
Diabetes	0.657	0.226	0.743	0.634	0.615	∞	$-/42/66$
Sonar	0.721	0.418	0.773	0.656	0.543	∞	$-/9/24$
Alphabets	0.972	0.081	0.990	0.524	0.707	16.23	9/146/769
ForestC1	0.803	0.623	0.823	0.674	0.724	11.00	15/165/624
Waveform	0.868	0.269	0.880	0.719	0.777	5.20	5/26/149
Biomed	0.866	0.755	0.884	0.833	0.863	2.60	6/16/33
Cancer wpbc	0.566	0.330	0.611	0.490	0.496	2.30	4/9/23
Heart	0.821	0.575	0.830	0.731	0.740	2.20	4/9/30
Ionosphere	0.846	0.659	0.868	0.687	0.846	1.93	16/31/31
Spambase	0.895	0.458	0.909	0.604	0.801	1.37	16/22/906
ForestC2C5	0.827	0.474	0.851	0.604	0.754	1.28	25/32/600
Forest	0.900	0.546	0.911	0.647	0.789	1.24	12/15/742
Arrhythmia	0.680	0.379	0.703	0.413	0.650	1.20	39/47/91

These results demonstrate our hypotheses at work. In the subsequent section we proceed to validate these on domains from both the UCI repository of datasets and the one-class classification repository available from the Pattern Recognition Laboratory [2].

5.2. Benchmark Datasets

The results produced on the benchmark datasets are presented in Tables 2. The g-mean of the best binary classifier (BC) for each datasets natural class distribution and the best g-mean at the maximum imbalance are denoted by B_{nat} and B_{imb} respectively. This illustrates the degradation as the imbalance increases. Similar statistics for the best sampling method are denoted by S_{nat} and S_{imb}, and the g-mean for the best one-class classifier (OCC) is given by OCC. The lifelines are denoted by ℓ, and the final column, of the form (#/#/#), shows the number of instances of the minority class when the best sampler and BC fall below the OCC, and the original size of the minority class. The ℓ value of ∞ indicates situations where the best sampled classifier is better than the OCC at every level of imbalance; thus, the lifeline is extended 'infinitely' by sampling.

The common property of the four datasets with the longest lifelines is that they have a significant degree of overlap between the classes and are at least somewhat multi-modal. Diabetes and Sonar do not have strong, clearly defined modes, but the small datasets make them additionally challenging. As an example of the strong mutli-modality, we show the PCA plot of the Delft pump data on the left in Figure 8. Furthermore, unsampled BCs exhibit a significant performance drop as imbalance increases on these datasets. This is illustrated in the results plot for Deft pump dataset on the right in Figure 8. Sampling provides a remarkable level of robustness over these datasets; for example, in the Alphabets

2. See: http://prlab.tudelft.nl

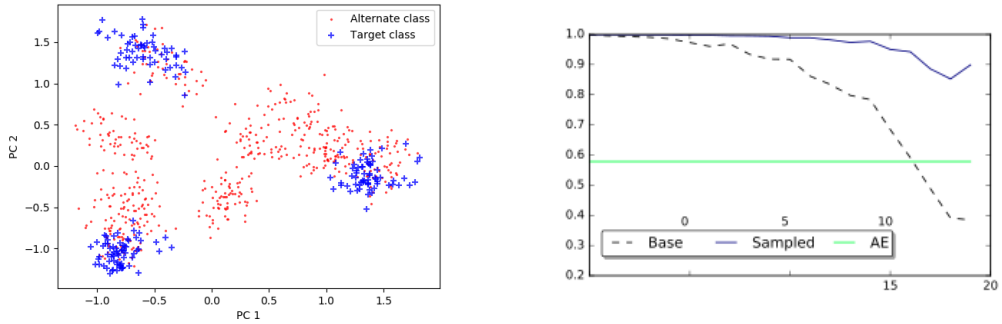


Figure 8: PCA plot and results curves for the Deft pump dataset.

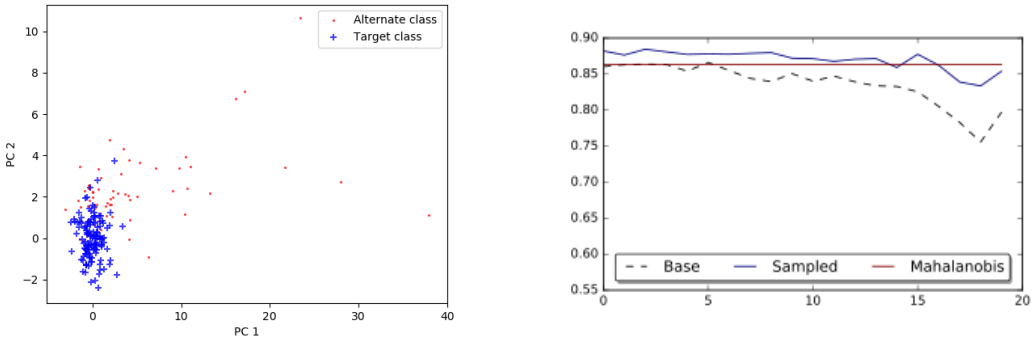


Figure 9: PCA plot and results curves for the Biomed dataset.

dataset, the BC g-mean drops to almost 0 from a high value of 0.972, whereas with sampling it only drops to 0.524.

As hypothesized, OCCs are impacted by domain complexity, as shown by the *OCC* over the top four datasets. Upon closer inspection, we see two scenarios arising. On the Deft pump and alphabets datasets, which have strongest evidence multi-modality, BCs start out strong, but their performance erodes sharply. Conversely, the diabetes and sonar datasets have less pronounced modality, but are small with overlap, thus being difficult for the OCCs and the BCs from the outset. Thus, the two scenarios are *a*) a big gap with sharp erosion, and *b*) a small gap with moderate erosion in binary performance. In all of these cases, sampling is able to significantly extend the lifeline of the best binary classifier.

In Biomed, the dataset with the highest *OCC*, we find the opposite properties in the data. This is shown in Figure 9. Here, the target class that the OCC is trained on is clearly unimodal. Furthermore, there is little class overlap, making this a relatively easy task for both the OCC and the unsampled BC. The degradation in the g-mean is very slight in comparison to that over the more complex domains. Consequently, the improvement by sampling is relatively small compared to the more complex datasets.

Let us examine the results over the Arrhythmia dataset, which has the smallest lifeline. As shown in Figure 10, the classification task is challenging due to exceptional overlap. The

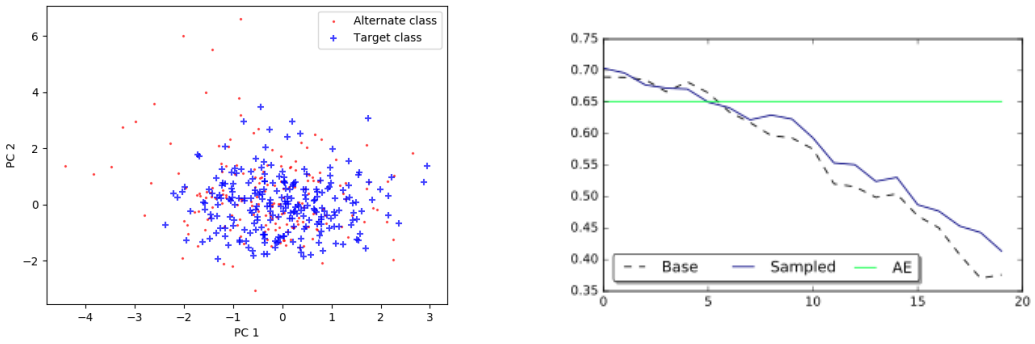


Figure 10: PCA plot and results curves for the arrhythmia dataset.

target class, however, is unimodal, and thus, the OCC can model it nearly as well as the baseline BC. Even at moderate levels, sampling is not superior to the best OCC. Indeed, all of the datasets with the shorter lifelines have the common property of being unimodal, and, with the exception of the cancer wpbc, having good OCC results and relatively robust baseline BCs.

The results presented in this section indeed validate our initial hypotheses. We see: *1*) sampling provides a significant boost in the life of a BC on complex, imbalanced datasets, *2*) the boost naturally leads to greater robustness to imbalance, *3*) the performance of OCCs is negatively related to the modality of the domain. As a result, on complex domains sampling is able to maintain a higher g-mean than OCC even in extreme imbalance. Alternatively, in simpler domains, OCC is more likely to dominate over sampled BCs starting at moderate levels of imbalance. To summarize, we find that sampled BCs have an edge over OCCs in multi-modal domains even in extreme imbalance, whereas OCCs are a good option on unimodal domains. Class overlap has a more uniform impact across methods.

6. Concluding Remarks and Future Work

The importance of class imbalance is well established in the machine learning literature. Indeed, much work has focused on new algorithms and understanding the root causes of the problem. However, it remains difficult to extrapolate from research to practice. In this paper, we build a framework from the wealth of research into the root causes of the class imbalance problem to guide the most practical of decisions. Specifically, given a dataset with some level of imbalance and complexity, which classification strategy should I apply? We explore this question over a wide range of imbalance levels, and data complexities, with the latest sampling, binary and one-class classification methods.

Our results show that modality has the greatest impact on the choice of classification strategy. In particular, sampled binary classifiers have improved robustness, and outperform one-class classifiers on multi-modal domains even in extreme imbalance. On the other hand, one-class classifier induce models on unimodal domains that are superior at moderate to high levels of imbalance.

For future work will expand the study beyond the impact of modality and class overlap over to other domain properties, such as dimensionality, variance and feature correlations *etc.* We are interested in meta learning as a way to better understand how these factors combine to influence the choice of classification method. A very important aspect of the future work will be to apply this study in the context of multi-class classification.

References

Nafees Anwar, Geoff Jones, and Siva Ganesh. Measurement of data complexity for classification problems with unbalanced data. *Statistical Analysis and Data Mining*, 7(3), 2014.

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

C. Bellinger, S. Sharma, and N. Japkowicz. One-class versus binary classification: Which and when? In *11th ICMLA*, volume 2, pages 102–106. IEEE, 2012.

C. Bellinger, C. Drummond, and N. Japkowicz. Beyond the boundaries of smote-a framework for manifold-based synthetically oversampling. In *ECML/PKDD*, pages 248–263, 2016.

José-Ramón Cano. Analysis of data complexity measures for classification. *Expert Systems with Applications*, 40(12):4820–4831, 2013.

N.V. Chawla, K.W. Bowyer, L.O. Hall, and Kegelmeyer W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artificial Intelligence Research*, 16:321–357, 2002.

M. Denil and T. P. Trappenberg. Overlap versus imbalance. In *Canadian Conference on AI*, pages 220–231. Springer, 2010.

P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *5th ACM SIGKDD*, pages 155–164. ACM, 1999.

C. Elkan. The foundations of cost-sensitive learning. In *IJCAI*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

V. García, J. Sánchez, and R. Mollineda. An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. *Progress in Pattern Recognition, Image Analysis and Applications*, pages 397–406, 2007.

H. Han, W. Y. Wang, and B. H. Mao. Borderline-SMOTE : A New Over-Sampling Method in Imbalanced Data Sets Learning. *Advances in intelligent computing*, pages 878–887, 2005.

H. He, Y. Bai, E. A. Garcia, and S. Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE IJCNN*, 3:1322–1328, jun 2008.

N. Japkowicz. Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68, pages 10–15. Menlo Park, CA, 2000.

N. Japkowicz and S. Stephen. The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis*, 6:429–450, 2002.

M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *14th ICML*, pages 179–186. Morgan Kaufmann, 1997.

H. Lee and S. Cho. The novelty detection approach for different degrees of class imbalance. In *NIPS*, pages 21–30. Springer, 2006.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

R.C. Prati, E.A.P.A. Batista, and M. C. Monard. Class imbalances versus class overlapping : an analysis of a learning system behavior. In *MICAI 2004*, pages 312–321. Springer Berlin Heidelberg, 2004.

S. Sharma. *Learning the Sub-Conceptual Layer: A Framework for One-Class Classification.* PhD thesis, Université d'Ottawa/University of Ottawa, 2016.

B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos. Class Imbalance, Redux. In *IEEE 11th ICDM*, pages 754–763. Ieee, dec 2011.

S. Wu, K. Lin, C. Chen, and M. Chen. Asymmetric support vector machines: low false-positive learning under the user tolerance. In *14th ACM SIGKDD*, pages 749–757. ACM, 2008.