Tunable Plug-In Rules with Reduced Posterior Certainty Loss in Imbalanced Datasets

Emmanouil Krasanakis* MANIOSPAS@ITI.GR
Eleftherios Spyromitros-Xioufis ESPYROMI@ITI.GR
Symeon Papadopoulos PAPADOP@ITI.GR
Yiannis Kompatsiaris IKOM@ITI.GR
Centre of Research & Technology - Hellas (CERTH), Information Technologies Institute (ITI), 6th km Xarilaou - Thermi, 57001, Thessaloniki, Greece

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

Abstract

Classifiers have difficulty recognizing under-represented minorities in imbalanced datasets, due to their focus on minimizing the overall misclassification error. This introduces predictive biases against minority classes. Post-processing plug-in rules are popular for tackling class imbalance, but they often affect the certainty of base classifier posteriors, when the latter already perform correct classification. This shortcoming makes them ill-suited to scoring tasks, where informative posterior scores are required for human interpretation. To this end, we propose the *ILoss* metric to measure the impact of imbalance-aware classifiers on the certainty of posterior distributions. We then generalize post-processing plug-in rules in an easily tunable framework and theoretically show that this framework tends to improve performance balance. Finally, we experimentally assert that appropriate usage of our framework can reduce *ILoss* while yielding similar performance, with respect to common imbalance-aware measures, to existing plug-in rules for binary problems.

Keywords: classification, class imbalance, plug-in rules, posterior certainty

1. Introduction

In class imbalance problems, where class distributions are often highly skewed, classifiers tend to favor over-represented classes. This happens due to overall performance maximization goals, which effectively place importance on classes according to their priors. However, in many real-life problems, correct identification of minority classes is at least as important as identification of majority ones. Therefore, there is a need to balance performance between majority and minority classes.

Furthermore, it is desirable to improve performance balance without considerably affecting posterior distribution certainty for correctly classified samples. This way, imbalance-aware scoring systems are prevented from shifting their posteriors towards the uniform distribution when attempting to downplay the importance of majority classes. Since shifting posteriors towards the uniform distribution can drastically impact human judgment (e.g., one can easily consider class distribution $[0.45, 0.55]$ as uninformative), avoiding such shifts would improve the trust in the respective scoring systems.

* Corresponding author. (maniospas@iti.gr)

In this work, we outline popular imbalance-aware methods and performance measures. We also present a novel metric, called $ILoss$, for measuring the certainty loss on correct posteriors. We then generalize cost-sensitive plug-in rules used to post-process classifier posteriors in a setting that only partially skews them. We theoretically show that employing a certain family of functions to do so tends to improve performance balance. Finally, we experimentally assert that certain functions from this family yield superior $ILoss$ compared to existing plug-in approaches without impacting imbalance-aware performance measures.

2. Background

Taking into account recent surveys (Fernández et al., 2017; Mundle and Chaudhari, 2017), we categorize imbalanced methods into data pre-processing (section 2.1) and learning algorithm modifications. The latter typically comprise cost-based post-processing (Section 2.2) and weighting (Section 2.3). Recent works have also used ensembling (Section 2.4) in the context of class imbalance. We briefly outline the above methods, placing more emphasis to pre-processing and cost-based post-processing, which are utilized throughout this work.

2.1. Data pre-processing

Data pre-processing aims to reduce imbalance between priors in the training set using undersampling, oversampling or hybrid approaches. Such methods enjoy widespread use, since they do not affect the classification model. Undersampling balances priors by reducing the number of majority class samples, whereas oversampling proliferates minority class samples (e.g. by duplication). Advanced methods, such as the one by Zikeba et al. (2014), try to mitigate lost or duplicate information. Hybrid approaches attempt to combine undersampling with oversampling to preserve the favorable characteristics of each approach. Pre-processing training data to obtain a uniform prior distribution is rarely optimal in imbalanced class settings (Weiss and Provost, 2003) and various class ratios have been proposed.

Notable pre-processing techniques besides random undersampling and oversampling include methods that remove majority class samples that interfere with minority classifications (Kubat and Matwin, 1997), as well as methods that perform synthetic minority oversampling (SMOTE) (Chawla et al., 2002). Seiffert et al. (2008) also propose a combination of undersampling and oversampling to preserve the total number of samples.

2.2. Cost-based post-processing

Cost-sensitive learning formulates sample misclassification costs and tries to minimize the overall cost across classes instead of the classification error. To this end, *plug-in rules* are often used to estimate misclassification costs according to class priors. To improve performance balance, such rules are designed to favor minority classes. For binary problems, this practice is equivalent to threshold adaptation (i.e. selecting a different decision threshold).

Cost-sensitive analysis defines a cost matrix, where each element represents a misclassification cost. Analysis then aims to minimize the overall misclassification cost. Several works (Elkan, 2001; Dembczynski et al., 2013; Hong et al., 2016) posit that misclassification costs can be estimated using class priors and plugged in the cost-sensitive function. For classes i and user-defined class weights w_i instead of costs, one can obtain the plug-in rule

for sample x and classifier C that calculates posteriors $C_i(x) = P(x \in i)$ using Eq. 1.

$$x \in \underset{i}{\operatorname{argmax}} \left[C_i(x)\, w_i\right] \tag{1}$$

2.3. Weighting

Weighting techniques assign misclassification costs on training samples. The classifier is then responsible for taking those costs into account during training. Similarly to cost-based post-processing, training weights are plugged-in using a prior-based estimation. Popular classification algorithms have been adapted to account for weighted samples. For example, weighted SVMs have been used by Silva et al. (2017). However, such analysis may prove to be difficult for new classification models, since each model has different theoretical intricacies. Seiffert et al. (2008) and Chawla et al. (2008) have found weighting practices to be slightly inferior to sampling, since treating imbalance itself can yield more robust posteriors.

2.4. Ensembling

Ensembling aims to improve performance by aggregating weak base classifiers using voting. Arguably the most popular ensembling method is AdaBoost (Rätsch et al., 2001), which trains different classifiers by adjusting the weights of training samples as it generates base classifiers. Imbalance-aware approaches have used classifier ensembles to improve base classifier performance towards minority classes while mitigating information loss. Boosting approaches either disturb training weights (Seiffert et al., 2010) or use clustering (Ofek et al., 2017) to train different base classifiers.

2.5. Notation and terminology

In this work, we use common stochastic analysis notation (Table 1). For clearer presentation, we refrain from referencing the classifier when contextually obvious.

Table 1: Notation and terminology

$E_{\mathcal{X}}[A(x)]$	Mean value of $A(x)$ over all $x \in \mathcal{X}$
$1_{condition}$	Yields 1 if *condition* holds true and 0 otherwise
$P(A)$	Probability of event A.
$P(A\|B)$	Conditional probability of event A, assuming the event B
$A(x) \propto B(x)$	Denotes that A and B are proportional: $\exists \lambda > 0 : A(x) = \lambda B(x)\, \forall x$
TPR_i	True Positive Rate for class i is defined as $\frac{TP_i}{TP_i+FN_i}$
FPR_i	False Positive Rate for class i is defined as $\frac{FP_i}{FP_i+TN_i}$
$accuracy(C_i)$	Accuracy for class i is defined as $\frac{TP_i+TN_i}{TP_i+TN_i+FP_i+FN_i}$
C_i	Posteriors for class i of the classifier C

3. Imbalance-aware Metrics

Approaches aiming to tackle class imbalance measure performance using either Area Under Curve (AUC) or performance means across classes. Mason and Graham (2002) propose AUC of the Receiver Operating Characteristics curve (ROC) to evaluate separability between different class posteriors. Oommen et al. (2011) point out that AUC is a robust measure that remains unaffected from imbalance or sampling bias. Arithmetic Mean (AM) and

Geometric Mean (GM) between True Positive Rates (TPRs) across all N classes i are also popular performance measures for imbalanced datasets, since they are easy to explain when calculating TPR trade-offs between classes. Menon et al. (2013) confirm certain favorable theoretical properties for AM, but GM focuses more on balanced classifications and is usually preferred for evaluation.

$$GM = \sqrt[N]{\prod_i TPR_i} \tag{2}$$

Unfortunately, GM is not completely imbalance-insensitive, since it still favors higher overall performance. Hence, one can also measure imbalance as average TPR disparity:

$$Imb = \frac{1}{N(N-1)} \sum_i \sum_j |TPR_i - TPR_j| \tag{3}$$

This metric completely discards performance maximization goals but can be used to compare classifiers that exhibit similar GM.

There has been little to no work towards examining the impact of imbalance-aware techniques on the certainty of posteriors in scoring systems. In such systems, it becomes important for posteriors to be easily interpreted by humans. Although AUC evaluates whether posteriors clearly separate classes, it fails to measure whether prediction certainty is preserved for originally correct classifications. For example, imbalance-aware classifiers could cause posteriors to differ from the uniform distribution by a small margin; the classification would still optimize performance goals, such as GM and AUC, but humans would interpret posterior differences as unclear. Furthermore, as pointed out by Krawczyk (2016), it is important to retain posteriors for samples correctly identified by the base classifier.

Assuming that the level of certainty for originally correct classifications is not misplaced, we propose that imbalance-aware classifiers should preserve this certainty as much as possible. To this end, we compare the certainty of an imbalance-aware method against the certainty of the base classifier. Employing $entropy(C(x)) = -\sum_i c_i(x) \log c_i(x)$ with $c_i(x) = C_i(x) / \sum_i C_i(x)$ to measure classification uncertainty (in bits), we can calculate the certainty loss between an imbalance-aware classifier $\mathcal{R}(C)$ and a base classifier C:

$$ILoss = \mathbb{E}\left[1_{y=\operatorname{argmax} C(x)} \Big(entropy\big(\mathcal{R}(C(x))\big) - entropy(C(x)) \Big)\right] \tag{4}$$

$ILoss$ increases if imbalance-aware methods yield lower certainty (i.e. posteriors are closer to the uniform distribution) than base classifiers for correct classifications.

4. Tunable Plug-in Rules

In this paper, we generalize post-processing plug-in rules in a framework that can use a wide range of posterior distribution editing mechanisms. In that way, it is possible to improve balance while exploring different heuristics that intertwine priors with posterior rebalancing[1]. Furthermore, following previous concerns against heuristic costs (Krawczyk

1. Researchers often use the term *rebalancing* to indicate a process that yields performance balance between majority and minority classes. For example, this terminology is used by Zhou and Liu (2010).

and Woźniak, 2015; Yu et al., 2016), our framework allows non-heuristic tuning towards the set goal. For our analysis, we examine cases where there exists a monotonic relation between priors and classifier performance (which is always true for binary problems).

Definition 1 *A classifier C will be called TPR-imbalanced if and only if, for every class pair i, j with priors $f_i \geq f_j$, $TPR_i(C) \geq TPR_j(C)$ and more samples are classified in class i.*

A simple way to rebalance posteriors would be skewing their distribution in the opposite way imbalance does. Intuitively, if a class produces, on average, higher posterior scores than other classes, we expect more classifications to be attributed to it, thus increasing TPR and FPR. Of course, new assignments may be only erroneous or only correct, but the general *tendency* is for an even distribution between TP and FP. Since our work aims to retain posterior certainties, we also theorize that the posterior distribution of the base classifier should preferably be only partially edited. Hence, we propose a weighted voting mechanism between the base and the plug-in classifier posteriors. This formulation shares certain similarities to the work of Weng and Poon (2008), who employ asymmetric ROC curves to skew only a portion of the imbalanced posteriors.

Definition 2 *For a probabilistic classifier $C : \mathcal{S} \to \mathcal{W}^N \subseteq \mathbf{R}^N$ in feature space $\mathcal{S}$, we may use a suitable bounded function $t : [0,1] \times \mathcal{W} \to \mathbf{R}$ and $a : \mathcal{W}^N \to (0,1]$ to obtain the rebalanced classifier $R(C)$ with:*

$$\mathcal{R}(C)_i(x) = \big(1 - a(x)\big)C_i(x) + a(x)t\big(f_i, C_i(x)\big) \tag{5}$$

The above process yields the classification rule $x \in \operatorname{argmax}_i \mathcal{R}(C)_i(x)$, which generalizes Eq. 1. For example, it can be shown that using constant a and a posterior-independent function t yields classification decisions equivalent to thresholding. We will later show that, for an appropriate selection of $t(f, w)$ and constant a, this rule tends to reduce imbalance. In particular, in an imbalanced setting where classification performance and class frequencies are monotonically related, rebalanced posteriors need to be staged in favor of minority classes. Therefore, the function $t(f, w)$ must be decreasing for class priors f. To preserve same-class posterior orderings, this function should also be increasing for w.

Definition 3 *A function $t : [0,1] \times \mathcal{W} \to \mathbf{R}$ differentiable on $\mathcal{W}$ will be called a rebalance function if:*

$$\begin{aligned} & f_i \geq f_j \Leftrightarrow t(f_i, w) \leq t(f_j, w) \\ & \frac{\partial t(f,w)}{\partial w} \geq 0, \quad \forall w \neq 0 \\ & s.t. \quad t(f,0) = 0 \quad \textit{and} \quad t(1,w) \geq w \end{aligned} \tag{6}$$

Theorem 4 *For a binary imbalanced classifier, thresholding is equivalent to rebalancing with rebalance function $\frac{\partial t(f,w)}{\partial w} = 0$ and $a \in [0,1)$.*

Proof For a two-class problem with priors $f_1 \geq f_2$, rebalance function $t(f, w)$ and posteriors w_1, w_2, the classification rule can be reduced to calculating the sign of quantity

$$\begin{aligned}
&\text{sign}((1-a)w_1 + at(f_1, w_1) - ((1-a)w_2 + at(f_2, w_2))) \\
&\quad = \text{sign}((1-a)(w_1 - w_2) - a(t(f_1, w_1) - t(f_2, w_2))) \\
&\quad = \text{sign}(w_1 - w_2 + \theta)
\end{aligned}$$

where $\theta = -\frac{a}{1-a}(t(f_1, w_1) - t(f_2, w_2))$. However, if $\frac{\partial t(f,w)}{\partial w} = 0$ then $t(f_1, w_1) - t(f_2, w_2) = const \geq 0$. For $a \in [0, 1)$ we obtain $\theta = const \in [0, \infty)$, reducing the classification rule to binary thresholding. ■

Theorem 5 *For a TPR-imbalanced classifier C and rebalance function $t : [0, 1] \times \mathcal{W} \to \mathbf{R}$, there exists $a_{max} \leq 2$ such that, $\forall a \in (0, a_{max}]$, rebalancing with Eq. 5 tends to induce:*

$$Imb(C) > Imb(\mathcal{R}(C)) \tag{7}$$

Proof Defining Mean Class Scores (MCS) for a class i as $MCS_i = E_x[C_i(x)]$, for $C_i(x) \propto P(C(x) \in i)$ we obtain:

$$\begin{aligned}
MCS_i &\propto E_x[P(C(x) \in i)] = P(C(x) \in i) \\
&= P(C(x) \in i | x \in i)P(x \in i) + P(C(x) \in i | x \notin i)P(x \notin i) \\
&= TPR_i f_i + FPR_i(1 - f_i)
\end{aligned}$$

For $f_i \geq f_j$, $TP_i + FP_i \geq TP_j + FP_j \Leftrightarrow MCS_i \geq MCS_j$ and $t(f, w) - w$ decreasing for w:

$$E_x\big[t(f_i, C_i(x)) - C_i(x)\big] \leq E_x\big[t(f_j, C_j(x)) - C_j(x)\big] \Leftrightarrow \Delta_{\mathcal{R}} MCS_i \leq \Delta_{\mathcal{R}} MCS_j$$

Otherwise, for $f_i \geq f_j$, $MCS_i \geq MCS_j$, finite many samples and set $\mathbb{K} = \big[0, 1/\inf \frac{\partial t(f,w)}{\partial w}\big]$, any $K \in \mathbb{K}$ yields $\frac{\partial (Kt(f,w) - w)}{\partial w} < 0 \Rightarrow Kt(f, w) - w$ is decreasing. For $T_{a,K}(w) = w\frac{1-a}{1-\frac{a}{K}}$, rebalancing $(1-a)w + at(f, w) = (1 - \frac{a}{K})T_{a,K}(w) + \frac{a}{K}Kt(f, w)$ is equivalent to rebalancing with $a' = \frac{a}{K}$ and $t'(f, w) = t\big(f, T^{-1}_{a,K}(w)\big)$ on classifier $T_{a,K}(C)$. Trionym analysis then shows that, $\forall a \leq 1.5 + 0.5/\inf \frac{\partial t(f,w)}{\partial w} \Rightarrow a \leq a_{\max}, a_{\max} \leq 1.5 + 0.5/t(1, 1) \leq 2\ \exists K \in \mathbb{K}$:

$$\begin{aligned}
&(1-a)K^2 - K + a \geq 0 \Leftrightarrow \tfrac{1}{K}\tfrac{1-\frac{a}{K}}{1-a} - 1 \leq 0 \\
&\quad \Rightarrow \tfrac{\partial t}{\partial w}(f, T^{-1}_{a,K}(w))\tfrac{\partial T^{-1}}{\partial w}(w) - 1 \leq 0 \Leftrightarrow \tfrac{\partial t'(f,w)}{\partial w} - 1 \leq 0 \\
&\quad \Leftrightarrow t'(f, w) - w \text{ is decreasing} \\
&\quad \Rightarrow \Delta_{\mathcal{R}(T_{a,K})} MCS_i \leq \Delta_{\mathcal{R}(T_{a,K})} MCS_j \\
&\quad \Rightarrow \Delta_{\mathcal{R}} MCS_i \leq \Delta_{\mathcal{R}} MCS_j
\end{aligned}$$

We can also obtain a linear approximation to $\Delta_{\mathcal{R}} TPR_i$ for small MCS fluctuations:

$$\Delta_{\mathcal{R}} MCS_i \approx \frac{\partial MCS_i}{\partial TPR_i} \Delta_{\mathcal{R}} TPR_i \propto f_i \Delta_{\mathcal{R}} TPR_i$$

Therefore, since $t(f,w) \geq w \Rightarrow \Delta_{\mathcal{R}} MCI_j \geq 0$, for $f_i \geq f_j$:

$$\begin{aligned} &\Delta_{\mathcal{R}} MCS_i - \Delta_{\mathcal{R}} MCS_j \leq 0 \\ &\quad \Leftrightarrow f_i(\Delta_{\mathcal{R}} TPR_i - \Delta_{\mathcal{R}} TPR_j) + \tfrac{f_i - f_j}{f_j} \Delta_{\mathcal{R}} MCS_j \leq 0 \\ &\quad \Rightarrow \Delta_{\mathcal{R}} TPR_i - \Delta_{\mathcal{R}} TPR_j \leq 0 \end{aligned}$$

Hence, $\Delta_{\mathcal{R}} |TPR_i - TPR_j| \leq 0$. ■

Finally, we recognize the case of inverse imbalance; a classifier could yield improved performance for minority classes, e.g. due to class cost over-skewing.

Definition 6 *A classifier C will be called inverse TPR-imbalanced if and only if $f_i \geq f_j \Leftrightarrow TPR_i(C) \leq TPR_j(C)$ and less samples are classified in class i for class priors f_i, f_j.*

Theorem 7 *For an inverse TPR-imbalanced classifier C and any rebalance function $t_{pos} : [0,1] \times \mathcal{W} \to \mathbf{R}$, there exists $a_{max} \leq 2$ such that, $\forall a \in (0, a_{max}]$, rebalancing with Equation 5 and $t(f,w) = t_{pos}(1-f,w)$ tends to induce:*

$$Imb(C) > Imb(\mathcal{R}(C))$$

Proof Proof is the dual of Theorem 5. ■

5. Experiments

5.1. Experimental Setup

To showcase the effectiveness of the presented rebalancing framework, we propose a post-processing scheme based on Eq. 5, which we compare with other imbalance-aware schemes on various imbalanced datasets. In all cases, we employ as base classifier Weka's (Hall et al., 2009) implementation of Logistic Regression (LR) with default parameters. The source code for our experiments is available online.[2]

Our proposed scheme is called *Tuned Inverse function Rebalance* (TIR) and utilizes the rebalance function w/f and parameter $a \in [0,2]$ tuned over the training set[3] towards maximizing $GM - Imb/2$ by Algorithm 1. This objective reflects that GM is more important than imbalance and was empirically found to provide a good performance trade-off.

Algorithm 1 Tuning Mechanism

1 Detect whether problem is TPR-imbalanced or negative TPR-imbalanced
2 Set $a \leftarrow 1, range \leftarrow 1, step \leftarrow 0.1$
3 Select $aBest \in a - range : step : a + range$ maximizing the objective over the training set
4 $a \leftarrow aBest, range \leftarrow step, step \leftarrow 0.1 \times range$
5 If training objective value changed more than 0.1%, go to 3

2. https://github.com/MKLab-ITI/Posterior-Rebalancing
3. Further splitting the training set to tune the parameter was found to yield inferior results.

TIR is compared with the base classifier, tuned plug-in rules (employing the same tuning mechanism) and a *ROC-based Moving Threshold* (RMT) approach. The latter bears a similarity to our work, since it sets minority class weights equal to class ratio and tunes majority class weights in the range $[0, 200]$ for Eq. 1. We also compare post-processing mechanisms when resampling is employed to pre-process datasets.[4] Table 2 lists the compared schemes.

Table 2: Evaluated Schemes

Abbreviation	Source	Method
Base (LR)	Weka (Hall et al., 2009)	No rebalancing
TIR	This paper	Rebalancing with tuned $a \in [0, 2]$ and w/f
TPlugIn	Yu et al. (2016)	Tuned plug-in rule for adaptive thresholding
RMT	Krawczyk and Woźniak (2015)	ROC-based moving threshold
Resampling	Seiffert et al. (2008)	Combination of under/oversampling
Resampling+TIR	This paper	Post-process resampling using TIR
Resampling+TPlugIn	Thai-Nghe et al. (2010)	Post-process resampling using TPlugIn
Resampling+RMT	This paper	Post-process resampling using RMT

As per Theorem 4, tuned plug-in rules make the same classification decisions as binary adaptive thresholding that shifts the decision threshold. However, thresholding does not alter posteriors and is thus unsuitable for scoring mechanisms.

Experiments are carried out on a collection of 28 imbalanced datasets commonly used in the literature for imbalanced classification experiments. Of those, 19 come from the publicly available *HDDT* algorithm validation (Cieslak and Chawla, 2008), whereas the rest were extracted from the public *UCI* repository (Lichman, 2013). Although Eq. 5 supports multi-class problems, in practice priors and performance are not always correlated in multi-class datasets among multiple minority classes. Thus, in case of multi-class labels, we focus on recognizing the smallest minority class. As seen in Table 3, our experiments encompass datasets with various numbers of instances, features and, importantly, degrees of imbalance.

Evaluation uses stratified 10-fold cross-validation and applies the same random splits each time. To measure the effectiveness of various schemes in tackling class imbalance, we employ the GM and Imb metrics described in Section 3, whereas AUC is used to measure the impact on base classifier posterior ranks, and confidence loss ($ILoss$) is used to measure the impact on rank descriptiveness.

To test for statistically significant differences between schemes across datasets, we follow the methodology suggested by Demšar (2006). In particular, we first use the Friedman test as a non-parametric alternative of repeated ANOVA measures. The Friedman test operates on the average ranks of the methods and checks the validity of the (null) hypothesis that all methods are equivalent. When the null hypothesis of the Friedman test is rejected ($p < 0.05$), we proceed with the Nemenyi post-hoc test, which compares schemes to each other across datasets and finds the statistical significance of differences between their average performance ranks. Lower ranks indicate superior performance but only differences over a

4. The selected resampling methodology of Seiffert et al. (2008) performs sampling (with replacement) to create a new dataset of the same size with specified class frequencies. This process is equivalent to simultaneously performing both undersampling and oversampling. For class ratios less than 10, we set sampling to produce a $50:50$ split, but for higher imbalances we set a $65:35$ resampled dataset split, which Khoshgoftaar et al. (2007); Anand et al. (2010) have found to yield better results.

Table 3: Details of binary evaluation datasets. Class ratio is calculated as f_{max}/f_{min}.

Name	Source	Instances	Features	Class Ratio
Boundary	HDDT	3,505	176	27.5
Breast-Y	HDDT	286	10	2.4
Cam	HDDT	18,916	133	19.1
CompuStat	HDDT	13,657	21	25.3
CovType	HDDT	38,500	11	13.0
Credit-G	HDDT	1,000	21	2.3
Estate	HDDT	5,322	13	7.4
Heart-V	HDDT	200	14	2.9
Hypo	HDDT	3,163	26	19.9
ISM	HDDT	11,180	7	42.0
Letter	HDDT	20,000	17	24.3
Oil	HDDT	937	50	21.9
Page	HDDT	5,473	11	8.8
PenDigits	HDDT	10,992	17	8.6
Phoneme	HDDT	5,404	6	2.4
PhosS	HDDT	11,411	481	17.6
SatImage	HDDT	6,430	37	9.3
Segment	HDDT	2,310	20	6.0
Sick	HDDT	3,772	30	15.3
Adult	UCI	32,561	15	3.2
Car	UCI	214	7	25.6
Contraception	UCI	1,473	10	3.4
Glass	UCI	214	11	22.8
Lung Cancer	UCI	32	56	2.6
Page Blocks	UCI	1,728	7	194.5
Sick Euthyroid	UCI	3,163	26	9.8
Thoraric Surgery	UCI	470	16	5.7
Yeast	UCI	1,484	10	295.8

certain Critical Difference (CD) are considered statistically significant. Since this test is considerably strict, we use a slightly lower confidence ($p < 0.1$) to calculate CD.

5.2. Experimental Results

Table 4 indicates that all three posterior editing methods yield similar GM and AUC values, i.e. the Friedman null hypothesis is not rejected. The Nemenyi post-hoc test shows that TIR yields significantly better ranks for Imb compared to RMT, whereas TIR and RMT yield significantly better $ILoss$ ranks compared to TPlugIn. These results indicate that TIR is superior to the other two methods, since it exhibits desirable behavior both for obtaining more balanced classification and for retaining posterior certainty. It must be noted that all three posterior editing schemes dominate the GM and Imb of the base classifier.

In agreement with previous works, the findings in Table 5 provide evidence in favour of TIR and TPlugIn to improve balance on base classifiers trained under resampling. Although the Friedman test reveals statistically significant differences only for AUC and $ILoss$, where Resampling outperforms its combination with post-processing techniques, TPlugIn and TIR are able to improve GM and Imb for a significant portion of datasets. Instead, RMT fails to produce similar improvements in this setting. These findings indicate that, for certain problems, there is merit in using TIR and TPlugIn to improve resampling GM and Imb.

6. Conclusions and Future Work

In this paper, we explore the concept of partially editing classifier posteriors as a generalization to plug-in rules. We show the relation of such practices to thresholding for binary problems and prove that they tend to reduce imbalance. Our analysis outlines the required characteristics for rebalance functions. Experiments show that there is merit to applying Equation 5 on binary imbalance problems, as suitable rebalance functions are superior to

Table 4: Posterior Editing LR Experiments. Arrows indicate whether metrics need to be maximized (↑) or minimized (↓). For all metrics, lower Average Ranks indicate superior results. Nemenyi CD is 0.55. Friedman test is rejected for Imb and $ILoss$.

	Base			TIR				TPlugIn				RMT			
Dataset	**GM↑**	**Imb↓**	**AUC↑**	**GM↑**	**Imb↓**	**AUC↑**	**ILoss↓**	**GM↑**	**Imb↓**	**AUC↑**	**ILoss↓**	**GM↑**	**Imb↓**	**AUC↑**	**ILoss↓**
Boundary	.39	.83	.76	.56	.22	.60	.64	.56	.22	.60	.85	.56	.22	.60	.64
Breast-Y	.50	.61	.66	.62	.08	.67	.10	.61	.09	.68	.18	.62	.10	.68	.10
Cam	.32	.89	.84	.49	.02	.61	.58	.50	.02	.61	.76	.50	.02	.61	.58
CompuStat	.15	.98	.80	.58	.01	.58	.58	.58	.01	.58	.77	.58	.01	.58	.58
CovType	.51	.72	.90	.60	0	.72	.55	.61	.01	.72	.72	.61	.02	.72	.55
Credit-G	.60	.49	.76	.71	0	.75	.09	.70	0	.75	.21	.70	.01	.75	.10
Estate	.16	.97	.63	.61	0	.63	.42	.61	.01	.63	.44	.60	.05	.63	.42
Heart-v	.55	.58	.71	.65	.08	.71	.13	.65	.08	.71	.27	.64	.06	.71	.13
Hypo	.86	.25	.97	.87	.03	.90	.77	.87	.03	.90	.89	.87	.03	.90	.77
ISM	.63	.59	.92	.69	.20	.69	.80	.69	.19	.69	.90	.69	.20	.69	.81
Letter	.92	.16	.99	.90	.01	.92	.85	.90	.01	.92	.92	.90	.01	.92	.85
Oil	.74	.41	.82	.77	.25	.80	.40	.77	.25	.77	.88	.77	.25	.80	.40
Page	.80	.34	.94	.77	.01	.86	.49	.78	.01	.86	.72	.79	.05	.86	.49
PenDigits	.94	.10	.99	.93	0	.95	.76	.93	.01	.95	.86	.93	.01	.95	.75
Phoneme	.64	.40	.81	.75	0	.81	.08	.75	0	.81	.25	.75	0	.81	.08
PhosS	.33	.88	.76	.49	.07	.55	.58	.49	.07	.55	.75	.49	.07	.56	.58
SatImage	.15	.97	.77	.55	.02	.56	.47	.55	.01	.56	.58	.55	.01	.56	.47
Segment	.99	.01	1	.99	.01	1	.10	.99	.01	1	.86	.99	.01	1	.10
Sick	.78	.37	.94	.83	.04	.87	.69	.84	.04	.87	.82	.83	.05	.87	.69
Adult	.65	.50	.85	.75	0	.84	.18	.75	0	.84	.37	.76	.06	.84	.17
Car	.66	.54	.98	.80	.04	.91	.48	.80	.04	.90	.88	.80	.03	.91	.48
Contraception	.30	.88	.72	.67	0	.72	.16	.67	0	.72	.27	.67	.01	.72	.16
Glass	1	0	1	.99	.02	1	.07	.99	.02	1	.95	.82	.33	1	.04
Lung Cancer	.74	.16	.72	.72	.12	.81	.49	.69	.17	.75	.65	.56	.62	.71	.14
Page Blocks	.71	.50	.99	.82	0	.87	.73	.81	.03	.87	.97	.80	.04	.87	.73
Sick Euthyroid	.82	.30	.96	.90	.01	.95	.55	.90	.01	.95	.76	.91	.02	.95	.54
Thoracic Surgery	.23	.91	.65	.62	.13	.66	.33	.62	.13	.65	.42	.62	.10	.66	.33
Yeast	1	0	1	1	0	1	.02	1	0	1	.98	1	0	1	.01
Average Ranks				1.98	1.77	1.96	1.57	1.96	1.89	2.11	3.00	2.05	2.34	1.93	1.43

Table 5: Resampling LR experiments. Arrows indicate whether metrics need to be maximized (↑) or minimized (↓). For all metrics, lower Average Ranks indicate superior results. Nemenyi CD is 0.79. Friedman test is rejected for AUC and $ILoss$.

	Resampling				Resampling+TIR				Resampling+TPlugIn				Resampling+RMT			
Dataset	**GM↑**	**Imb↓**	**AUC↑**	**ILoss↓**	**GM↑**	**Imb↓**	**AUC↑**	**ILoss↓**	**GM↑**	**Imb↓**	**AUC↑**	**ILoss↓**	**GM↑**	**Imb↓**	**AUC↑**	**ILoss↓**
Boundary	.63	.43	.75	.01	.62	.46	.73	.26	.62	.48	.66	-.01	.62	.48	.73	.19
Breast-Y	.65	.05	.68	.07	.66	.06	.68	.08	.66	.06	.68	.06	.67	.04	.68	.08
Cam	.76	.04	.84	.42	.76	.05	.83	.45	.76	.05	.83	.42	.73	.11	.81	.47
CompuStat	.75	.04	.81	.55	.75	.01	.81	.56	.75	.02	.81	.58	.45	.78	.82	.56
CovType	.85	.02	.91	.28	.85	0	.90	.35	.85	0	.90	.26	.80	.23	.90	.46
Credit-G	.69	.03	.74	.07	.69	.02	.74	.07	.69	.02	.74	0	.69	.01	.74	.07
Estate	.60	.03	.63	.41	.61	0	.63	.41	.61	0	.63	.41	.35	.81	.63	.39
Heart-v	.61	.10	.64	.02	.62	.09	.66	.07	.64	.05	.67	-.12	.65	.04	.66	.07
Hypo	.95	.02	.96	.04	.94	.05	.96	.41	.94	.05	.96	.11	.90	.06	.95	.46
ISM	.87	.05	.92	.46	.87	.01	.91	.54	.87	.01	.91	.66	.87	.01	.91	.54
Letter	.95	.01	.99	.13	.95	0	.97	.62	.95	0	.97	.69	.82	.27	.96	.70
Oil	.80	.25	.83	-.01	.80	.25	.84	.02	.80	.25	.79	.65	.79	.28	.82	.01
Page	.91	.02	.96	.20	.91	.01	.95	.36	.91	.01	.95	.45	.91	.02	.95	.40
PenDigits	.95	0	.99	.08	.95	.01	.97	.54	.95	.01	.97	.52	.95	.03	.97	.66
Phoneme	.74	.10	.81	.06	.74	.02	.81	.05	.74	0	.81	-.14	.74	0	.81	.05
PhosS	.66	.20	.74	.26	.66	.25	.72	.36	.66	.25	.72	.22	.66	.25	.72	.36
SatImage	.70	.23	.76	.22	.68	.01	.76	.27	.67	.04	.76	.10	.67	.04	.76	.27
Segment	.99	0	1	0	.99	0	1	.02	.99	0	.99	.76	.99	0	1	.02
Sick	.89	0	.94	.20	.89	.03	.93	.36	.89	.03	.93	.18	.89	.03	.93	.36
Adult	.77	.02	.85	.13	.77	0	.84	.16	.77	0	.84	.18	.77	0	.84	.16
Car	.94	.03	.98	.03	.93	.01	.97	.19	.92	.03	.97	0	.92	.03	.97	.16
Contraception	.66	.13	.72	.15	.66	.01	.72	.14	.67	.02	.72	.05	.67	.02	.72	.14
Glass	.98	.04	.98	0	.98	.04	.98	.13	.98	.04	.98	.79	.93	.08	.98	.07
Lung Cancer	.60	.10	.68	-.06	.60	.10	.64	.09	.60	.10	.65	.60	.62	.43	.68	.04
Page Blocks	.93	.08	.98	.05	.93	.09	.98	.23	.94	.09	.94	.01	.86	.06	.95	.26
Sick Euthyroid	.90	.05	.95	.15	.91	.02	.94	.32	.92	0	.94	0	.92	0	.94	.29
Thoracic Surg.	.60	.15	.61	.16	.60	.09	.61	.22	.60	.09	.60	.26	.60	.06	.61	.22
Yeast	1	0	1	0	1	0	1	.01	1	0	1	.95	1	0	1	0
Average Ranks	2.41	2.70	1.79	1.75	2.43	2.20	2.59	3.00	2.32	2.32	2.89	2.30	2.84	2.79	2.73	2.95

previous plug-in rule approaches in reducing imbalance while retaining classification certainty. Hence, the proposed framework can be used to obtain better scoring mechanisms. Furthermore, we confirm that it is possible for posterior editing approaches to improve sampling GM and balance for certain problems.

In the future, we are interested in conducting experiments on more base classifiers, datasets and rebalance schemes. Since performance may not directly correlate to minority priors, if there are multiple minority classes, we also propose extending the multiclass aspect of our analysis to using an increasing function of the imbalanced metric instead of priors.

Acknowledgements

This work has been supported by the USEMP and STEP projects, partially funded by the EC under contract numbers FP7-611596 and H2020-649493 respectively.

References

Ashish Anand, Ganesan Pugalenthi, Gary B Fogel, and PN Suganthan. An approach for classification of highly imbalanced data using weighting and undersampling. *Amino Acids*, 39(5):1385–1391, 2010.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *JAIR*, 16:321–357, 2002.

Nitesh V Chawla, David A Cieslak, Lawrence O Hall, and Ajay Joshi. Automatically countering imbalance and its empirical relationship to cost. *DMKD*, 17(2):225–252, 2008.

David A Cieslak and Nitesh V Chawla. Learning decision trees for unbalanced data. *ECML PKDD*, pages 241–256, 2008.

Krzysztof Dembczynski, Arkadiusz Jachnik, Wojciech Kotlowski, Willem Waegeman, and Eyke Hüllermeier. Optimizing the f-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. *ICML (3)*, 28:1130–1138, 2013.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Charles Elkan. The foundations of cost-sensitive learning. *IJCAI*, 17(1):973–978, 2001.

Alberto Fernández, Sara del Río, Nitesh V Chawla, and Francisco Herrera. An insight into imbalanced big data classification: outcomes and challenges. *Complex & Intelligent Systems*, pages 1–16, 2017.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *SIGKDD*, 11(1):10–18, 2009.

Charmgil Hong, Rumi Ghosh, and Soundar Srinivasan. Dealing with class imbalance using thresholding. *arXiv preprint arXiv:1607.02705*, 2016.

Taghi M Khoshgoftaar, Chris Seiffert, Jason Van Hulse, Amri Napolitano, and Andres Folleco. Learning with limited minority class data. *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on*, pages 348–353, 2007.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

Bartosz Krawczyk and Michał Woźniak. Cost-sensitive neural network with roc-based moving threshold for imbalanced classification. *ICIDEAL*, pages 45–52, 2015.

Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. *ICML*, pages 179–186, 1997.

M. Lichman. UCI machine learning repository, 2013. URL archive.ics.uci.edu/ml.

Simon J Mason and Nicholas E Graham. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *JRMC*, 128(584):2145–2166, 2002.

Aditya Krishna Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. *ICML*, (3):603–611, 2013.

RV Mundle and MS Chaudhari. Survey on class imabalance problem using data mining techniques. *IJCMS*, 2017.

Nir Ofek, Lior Rokach, Roni Stern, and Asaf Shabtai. Fast-cbus: A fast clustering-based undersampling method for addressing the class imbalance problem. *Neurocomputing*, 243: 88–102, 2017.

Thomas Oommen, Laurie G Baise, and Richard M Vogel. Sampling bias and class imbalance in maximum-likelihood logistic regression. *Mathematical Geosciences*, 43(1):99–120, 2011.

Gunnar Rätsch, Takashi Onoda, and K-R Müller. Soft margins for AdaBoost. *Machine Learning*, 42(3):287–320, 2001.

Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Resampling or reweighting: a comparison of boosting implementations. *ICTAI*, 1:445–451, 2008.

Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1):185–197, 2010.

Joel Silva, Fernando Bacao, Maguette Dieng, Giles M Foody, and Mario Caetano. Improving specific class mapping from remotely sensed data by cost-sensitive learning. *IJRS*, 38(11): 3294–3316, 2017.

Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. Cost-sensitive learning methods for imbalanced data. *IJCNN*, pages 1–8, 2010.

Gary M Weiss and Foster Provost. Learning when training data are costly: the effect of class distribution on tree induction. *JAIR*, 19:315–354, 2003.

Cheng G Weng and Josiah Poon. A new evaluation measure for imbalanced datasets. *AusDM*, 87:27–32, 2008.

Hualong Yu, Changyin Sun, Xibei Yang, Wankou Yang, Jifeng Shen, and Yunsong Qi. Odoc-elm: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data. *KBS*, 92:55–70, 2016.

Zhi-Hua Zhou and Xu-Ying Liu. On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257, 2010.

Maciej Zikeba, Jakub M Tomczak, Marek Lubicz, and Jerzy Swikatek. Boosted svm for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *WFSC*, 14:99–108, 2014.