Analysis of Imbalanced Datasets

Miguel Jiménez Aparicio
Atlanta, USA
maparicio6@gatech.edu

Belén Martín Urcelay
San Sebastián, Spain
burcelay3@gatech.edu

Cristian Gómez Peces
Atlanta, USA
cpeces3@gatech.edu

***Abstract*—This document analyzes the impact of imbalanced datasets on machine learning classifiers. It discusses several techniques to deal with the imbalance. The results are applied to train a Twitter fake account detection algorithm where the performance of these techniques is compared.**

***Index Terms*—imbalance**

I. Introduction

A. Motivation

Many canonical machine learning algorithms used for classification assume that the number of objects in the respective classes is roughly the same. However, in reality, classes are rarely represented equally. In fact, class distribution skews are not only common, but many times expected [**?**], especially in decision systems aiming to detect rare but important cases. For instance, Covid-19 testing at Georgia Institute of Technology showed that less than 1% of the samples contained the virus. This means that a naive classifier could achieve a 99% accuracy just by labeling all samples as negative for Covid-19.

Imbalanced datasets significantly compromise the performance of many traditional learning algorithms. The disparity of classes in the training dataset may lead the algorithm to bias the classification towards the class with more instances, or even to ignore the minority class altogether. Therefore, it is vital to find efficient ways of dealing with data imbalances.

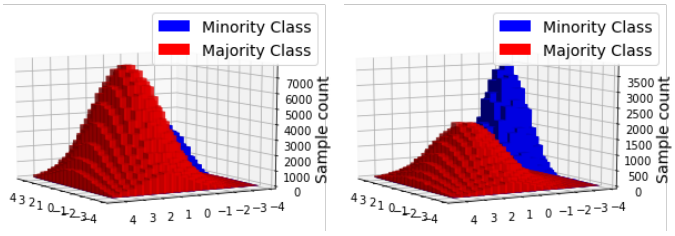
The overall goal of our project is to provide an overview of the state-of-the-art approaches to solve the issues introduced by imbalanced datasets. Including, a performance comparison of the various techniques. We also aim to implement an efficient scheme that is able to deal with highly complex and imbalanced datasets.

B. Methodology

Firstly, we study a synthetic dataset characterized by its simplicity. It is made up of two classes following $N(\boldsymbol{\mu}_0, \Sigma_0)$ and $N(\boldsymbol{\mu}_1, \Sigma_1)$), where

$$\boldsymbol{\mu}_0 = \begin{pmatrix} -1 \\ -0.5 \end{pmatrix}, \ \boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \ \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \ \Sigma_1 = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$$

and where the minority class only accounts for 15% of the samples. This simple dataset is especially useful to analyze the imbalance-compensating techniques from a mathematical perspective. Not only do we study the concepts learnt in class at a theoretical level, but we also use plugin machine learning models to illustrate how they affect density distributions.



(a) Original. (b) Undersampled. $\beta = 0.25$.

Fig. 1: Gaussian dataset.

Secondly, we target a more complex dataset. TODO

The performance of the classification will be evaluated using the F_1 score $\in [0,1]$, where the best possible score is 1. This metric is computed as

$$F_1 = 2\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}},$$

where the precision is the ratio between correctly identified minority samples and the total number of minority samples, while the recall is given by the fraction of correctly identified minority samples over all samples.

C. Accomplishments

II. Overview of the techniques

A. Undersampling

Undersampling is frequently employed to balance datasets before any machine learning algorithm is applied. Undersampling involves randomly removing entries from the majority class. Figure 1 shows the effects of undersampling on the Gaussian training dataset. The class imbalance is somewhat countered. However, the algorithm learnt from this undersampled dataset will be affected. Namely, its ability to generalize and its posterior distribution.

1) Generalization Ability: Induction algorithms require a sufficient amount of data to learn a model that generalizes well. If the training set is not large, a classifier may just memorize the characteristics of the training data. Moreover, undersampling has the potential of eliminating valuable samples from consideration of the classifier entirely [**?**], so it may exacerbate this problem of lack of data. The obtained training set may vary greatly from one undersampling to another, this leads to a high variance of the learned model.

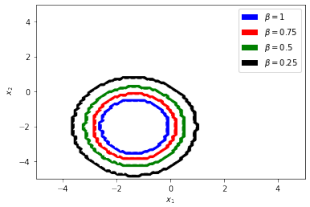


Fig. 2: Influence of undersampling on the classification region of a naive Bayes classifier trained with the Gaussian dataset. The area within each circle corresponds to the cluster of points that are classified as the minority class for a given undersampling rate.

Hence, the achievable complexity of the hypothesis set must be reduced to ensure a good generalization.

2) Posterior Bias: One goal of undersampling is to change the priori probabilities of the classes to make them more balanced. The classifier assumes that the features it encounters at testing follow the same distribution as the training set. This mismatch introduced by design is known as sampling selection bias [**?**] on posterior distribution.

Let $(\mathscr{X},\mathscr{Y})$ denote the pairs of feature vectors, $\mathbf{x}\in\mathbb{R}^n$, and binary labels, $y\in\{0,1\}$, contained in our original dataset. We assume that the number of samples labeled as zero is small compared with the number of samples in class one. Undersampling randomly removes points from the majority class, we describe this sampling with the binary random variable S, which takes the value 1 if a sample is selected.

It is reasonable to assume that the selection is independent of the features given the class. Then, applying Bayes rule, the law of total probability and noting that the samples from the minority class are always selected we obtain

$$\begin{aligned} p' &= P(y=0|\mathbf{x},s=1) = \frac{P(s=1|y=0,\mathbf{x})P(y=0|\mathbf{x})}{P(s=1|\mathbf{x})} = \\ &= \frac{P(y=0|\mathbf{x})}{P(y=0|\mathbf{x})+P(s=1|y=1)P(y=1|\mathbf{x})} = \\ &= \frac{p}{p+\beta(1-p)}, \end{aligned}$$

where p and p' denote the posterior probability of encountering a sample from the minority class when employing the original and the undersampled dataset respectively. Whereas β denotes the probability of keeping a sample from the majority class.

The posterior is highly affected by the rate of the sampling. As more samples are removed, the classification is more biased towards the minority class. Figure 2 shows the decision region of a naive Bayes classifier. As the training set is undersampled the region of points that are labeled as the minority class grows. The rate of undersampling not only influences the posterior bias, but also the algorithm's ability to generalize. Thus, β should be chosen with care. Figure 3 presents the average F1-score over different training sets. We observe that the score is concave and in this case the optimum occurs with $\beta = 0.82$.

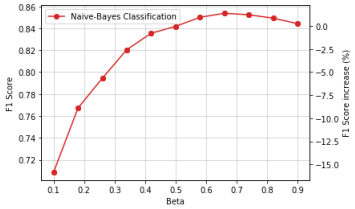
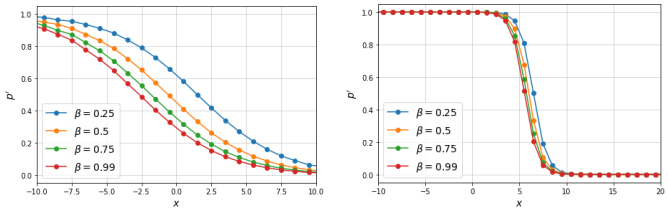


Fig. 3: F1-score vs. undersampling rate.



(a) $|\mu_0-\mu_1|=3$. (b) $|\mu_0-\mu_1|=13$.

Fig. 4: Influence of undersampling on posterior probability of the minority class.

Another factor that strongly influences the posterior bias is class separability. The bias is higher when conditional distributions are similar across the classes [**?**]. To analyze this behaviour we reduced the problem to a one-dimensional setting, the results are depicted in Figure 4. We confirm that undersampling shifts the posterior distribution in favor of the minority class. Nevertheless, the shift caused by β is lower under the configuration with lower overlap.

B. Oversampling

C. Cost-Sensitive Techniques

Blabla

III. Classification impact on Real Data

Ensemble of classifiers are known to improve the performance of single classifiers in imbalanced datasets [A Review on Ensembles for the Class Imbalance] by combining separate weak learners into a composite whole. Both bagging and boosting algorithms have been implemented from scratch to explore their advantages and compare them with highly-effective classifiers such as XGBoost.

A. Bagging

Important to mention that these algorithms are purely 'variance' reducing procedures intended to mitigate instability (espcially associated to decision trees).

B. AdaBoost

Unlike bagging, boosting fits the weak learners in an adaptive way: a different significance is assigned to each of the base learners based on the samples that were misclassified in the previous iteration. It was first introduced by Schapire in 1990, who proved in [The strength of weak learnability]that a weak learner can be turned into a strong learner in the sense of probably approximately correct (PAC) learning framework. In particular, the AdaBoost algorithm stands out in the field of ensemble learning [Cost-sensitive boosting algorithms: Do we really need them?] and has been appointed as one of the top ten data mining algorithms [Top 10 algorithms in data mining]. One its main advantages is the versatility to incorporate cost-sensitive techniques. We implemented a custom AdaBoost classifier that enables us to gain control over the algorithm at a lower level, make adjustments when necessary and create other Adaboost-based classifiers such as AdaCost or Boosted SVM. The algorithm is fully detailed below.

Algorithm 1: AdaBoost Algorithm

Input: Training set $D = \{x_i, y_i\}, i = 1,\dots,N$; and $y_i \in \{-1,+1\}$; T: Number of iterations; I: Weak learner

Output: Boosted Classifier: $H(x) = sign\left(\sum_{t=1}^{T} \alpha_i h_t(x)\right)$ where h_t, α_t are the induced classifiers and their significance, respectively

$W_1(i) \leftarrow 1/N$ for $i = 1,\dots,N$

/* Create a weak learner in each iteration */

for *t=1 to T* **do**

  $h_t \leftarrow I(D, W)$

  $\epsilon_t \leftarrow \sum_{i=1}^{N} W_t(i)[h_t(x_i) \neq y_i]$

  if $\epsilon_t > 0.5$ **then**

    $T \leftarrow t - 1$

    return

  $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$

  /* Update Weights */

  $W_{t+1}(i) = W_t(i) e^{(-\alpha_t h_t(x_i) y_i)}$ for $i = 1,\dots,N$

  Normalize W_{t+1} such that $\sum_{i=1}^{N} W_{t+1}(i) = 1$

This custom implementation uses single decision trees with 2 leaf nodes. It is given the number of maximum iterations (weak learners) that will be fitted, and a dataset D formed by N training samples, each constituted by a feature vector $\boldsymbol{x}_i$ and an associated binary label y_i. For each iteration, a weak classifier is created using the weighted training samples. For the first iteration, we allocate the same weight for all the samples $(1/N)$. Once the decision tree has been created, we predict the class of the training samples and compute the weighted misclassification error ϵ_t. Then, the significance α_t is calculated based on that error. Finally, the weights of each training samples are updated depending on the significance and whether the sample was correctly classified or not. Note that the weights are increased by e^{α} if the sample is misclassified or decreased by $e^{-\alpha}$ if correctly classified, with $\alpha_t \in [0,1]$ because $\epsilon_t \in [0, 0.5]$.

AdaBoost uses the whole training samples to create each weak learner serially –in contrast to *Bagging*, where *bootstrap aggregating* is used to construct ensembles– giving more focus to challenging instances and turn the incorrect classifications into good predictions in the next iteration. Hence, AdaBoost can be considered as a "bias" reducing procedure, intended to increase the flexibitily of stable (highly biased) weak learners when incorporated properly in a serial additive expansion. That is the reason behind why weak learners with low variance but high bias –such as decision trees– are well adapted for boosting. Lastly the overall literature suggest AdaBoost algorithms are generally resistant to overfitting regardless the number of weak estimators we use, and there are very few cases reported to overfit the training data []. This fascinating issue is explained due to 1) as the iterations proceed the weak learners tend to have less significance and 2) similarly to SVM, ensemble classifiers maximize the margin which allows promoting good generalization [The dynamics of AdaBoost: Cyclic behavior and convergence of margin]. The impact of its complexity on overfitting is covered in the result section.

C. AdaCost

Blabla

D. Boosting SVM

Algorithm 2: Boosted SVM Algorithm

Input: Training set $D = \{x_i, y_i\}, i = 1,\dots,N$; and $y_i \in \{-1,+1\}$; T: Maximum number of iterations; The initial $\sigma = \sigma_{ini}, \sigma_{min}, \sigma_{step}$

Output: Boosted Classifier: $H(x) = sign\left(\sum_{t=1}^{T} \alpha_i h_t(x)\right)$ where h_t, α_t are the induced classifiers and their significance, respectively

$W_1(i) \leftarrow 1/N$ for $i = 1,\dots,N$

/* Create a weak learner in each iteration */

while $\sigma > \sigma_{min}$ *and* $t < T$ **do**

  $t \leftarrow t + 1$

  $h_t \leftarrow RBFSVM(D, W, \sigma)$; // Train RBFSVM

  $\epsilon_t \leftarrow \sum_{i=1}^{N} W_t(i)[h_t(x_i) \neq y_i]$

  if $\epsilon_t > 0.5$ **then**

    $\sigma \leftarrow \sigma - \sigma_{step}$

  else

    $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$

    /* Update Weights */

    $W_{t+1}(i) = W_t(i) e^{(-\alpha_t h_t(x_i) y_i)}$ for $i = 1,\dots,N$

    Normalize W_{t+1} such that $\sum_{i=1}^{N} W_{t+1}(i) = 1$

E. AdaMEC

Blabla

F. XGBoost

Blabla

IV. Results and Discussion

Finally, we have combined them with downsampling to explore the effectiveness of hybrid algorithms. Blabla

V. Conclusions

Blabla

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BIBTEX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BIBTEX to produce a bibliography you must send the .bib files.

LATEX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LATEX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [**?**].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 5", even at the beginning of a sentence.

TABLE I: Table Type Styles

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not

Fig. 5: Example of a figure caption.

label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks …”. Instead, try “R. B. G. thanks…”. Put sponsor acknowledgments in the unnumbered footnote on the first page.