# Positive Unlabeled Gradient Boosting

Caitlin Timmons*
Department of Statistical and Data Sciences
Smith College
Northampton, MA, USA
ctimmons@smith.edu

Andrea Boskovic*
Department of Mathematics and Statistics
Amherst College
Amherst, MA, USA
aboskovic21@amherst.edu

Sreeharsha Lakamsani*
Department of Computer Science
Arizona State University
Tempe, AZ, USA
hlakamsani@asu.edu

Walter Gerych
Department of Data Science
Worcester Polytechnic Institute
Worcester, MA, USA
wgerych@wpi.edu

Luke Buquicchio
Department of Data Science
Worcester Polytechnic Institute
Worcester, MA, USA
ljbuquicchio@wpi.edu

Elke Rundensteiner
Department of Data Science
Worcester Polytechnic Institute
Worcester, MA, USA
rundenst@wpi.edu

***Abstract*—Classification applied to medical datasets from diagnosis or survey application data often must address the challenge that such data is weakly labeled, with only some positive labels and the rest of the data instances mostly being unlabeled. Standard classifiers struggle to learn the correct class for these positive but unlabeled instances, particularly within imbalanced datasets. The standard Gradient Boosting Classifier is one algorithm that works well on balanced data with completely labeled examples but performs poorly otherwise. In order to improve upon this state-of-the-art method, we propose a modification to its loss function that empowers it to learn a decision boundary more reflective of the data's true distribution. We call this the novel gradient boosting classifier. Our experimental study demonstrates that our proposed new classifier outperforms the state-of-the-art by 8.3% on average across several public medical data sets. This classifier can be applied to healthcare settings, where imbalanced and positive unlabeled data sets are common.**

***Index Terms*—gradient boosting, machine learning, neural network, positive unlabeled**

## I. Introduction

In health fields, it is especially crucial that we develop highly accurate machine learning algorithms, as they may determine whether or not a patient receives appropriate treatment. However, some health-related data, particularly surveys or diagnoses, can be described as positive unlabeled: datasets in which the only labeled instances are positive [7]. The remaining examples are unlabeled, but may actually belong to either class. Positive unlabeled health data may occur for multiple reasons, such as stigma around reporting mental health symptoms, incorrect patient reports, infrequent medical visits, or errors in data labeling. This data type makes classification tasks much more difficult [15]. Standard classifiers typically expect completely labeled training sets in order to make predictions, which forces the assumption that all unlabeled examples belong to the negative class and typically results in an erroneous decision boundary, as described in Figure 1.

### A. Problem Statement

Standard machine learning techniques are ill-suited for learning from positive unlabeled data. In positive unlabeled datasets, only a subset of the positive class is labeled, often assumed to be a representative sample of the entire positive distribution from which the data are drawn [1]. The remaining data from both classes are unlabeled. Relatively few learning techniques explicitly designed for positive unlabeled data exist. Of those techniques, most perform poorly on small datasets. However, in health fields, where positive unlabeled data are more common, datasets are often small due to the difficulty of collecting data or accessing patients.

### B. State-of-the-Art

One popular machine learning classifier is the standard Gradient Boosting Classifier [4]. Gradient boosting has gained popularity due to its success with many classification problems. Its ensembling method creates a model based on a combination of weak learners in a successive manner, and it achieves optimal results through loss minimization. Gradient Boost can work well for small datasets, though is usually ill-suited for positive unlabeled data, as it incorporates loss and residual minimization steps requiring data from both classes [1]. Other approaches to positive unlabeled learning include method modification, where existing techniques are modified for positive unlabeled data, and preprocessing techniques like deep learning and Support Vector Machines [3], [6], [8]. However, these methods typically require large datasets, which are not always readily available, in order to perform well.

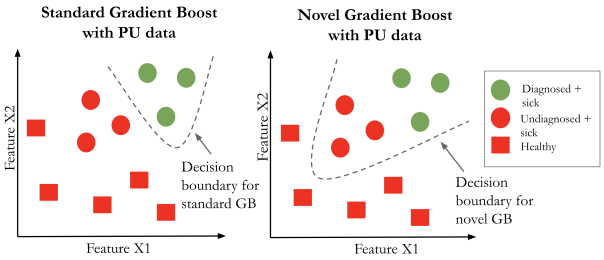
### C. Limitations of State-of-the-Art

The main shortcoming of the Gradient Boosting Classifier and other standard machine learning techniques is that they expect completely, and correctly, labeled data as input. Using these techniques on positive unlabeled data forces the

*Work on this project was done while the authors worked remotely through Worcester Polytechnic Institute.

978-1-7281-7571-3/20/$31.00 ©2020 IEEE

Fig. 1: Unlike standard classifiers, the novel gradient boosting classifier creates a new decision boundary to capture positive unlabeled instances in data.




assumption that all unlabeled data simply must belong to the negative class, which is not always true. Method modification techniques, while more sophisticated, also require that at least some unlabeled data are designated as negative—either at random or using prior knowledge of the dataset. Preprocessing approaches like deep learning do not always require negative examples, but often work poorly on small datasets. Small data is common in health areas, as patients may be scarce or data difficult or expensive to collect. All of these techniques share a common pitfall, they are not attuned to severe class imbalance and struggle to make accurate predictions for these datasets.

### D. Proposed Solution

In order to more accurately classify positive unlabeled examples, especially in small, imbalanced datasets, we present a novel gradient boosting classifier designed to handle positive unlabeled data. We modify the loss minimization steps in the gradient boosting algorithm, such that they requires only positive and unlabeled examples. This eliminates the need for a negative data class, thereby modifying the classifier's decision boundary to correctly capture positive unlabeled examples, as shown in Figure 1.

## II. Background

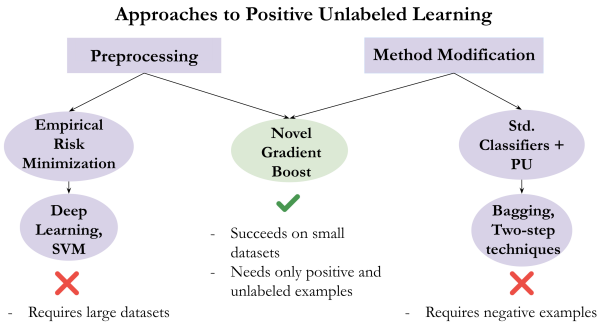
### A. Gradient Boosting and Class Imbalance

The gradient boosting technique ensembles together weak learners (decision trees) to form an extensible classifier [4]. Gradient boosting can succeed on small datasets, but like most classifiers, performs poorly on imbalanced datasets. With a skewed class distribution, it predicts almost exclusively the majority class. The minority class in health data often represents an abnormal case, such as an ill patient. As such, the cost of misclassifying minority instances is high, and imbalanced classification becomes even more difficult. Both issues' effects are compounded in the case of small datasets. Modifications such as resampling can mitigate class imbalance, although gradient boosting may be particularly prone to overfitting on such datasets, as it focuses on individual learners' errors.

### B. Empirical Risk Minimization

One method for handling positive unlabeled data is preprocessing, which includes empirical risk minimization. Empirical risk is a reflection of model error, and thus an ideal model minimizes its value. We define empirical risk below in

Fig. 2: Our proposed novel gradient boosting classifier improves upon the shortcomings of other approaches to positive unlabeled learning, allowing us to work with smaller datasets and datasets lacking negative examples.




Equation (1). We assume an empirical probability distribution $D$ based on our dataset, from which we can draw $(x, y)$ pairs, and a loss function $L$ measuring the difference between our model's prediction $\hat{y}$ and the true value $y$ [12].

$$R(f, D) = \mathbb{E}_{(x,y)\sim D}[L(f(x), y)] \tag{1}$$

For simplicity, we redefine empirical risk as follows:

$$R(f) = \mathbb{E}L(f(x), y).$$

## III. Proposed Method: Gradient Boosting Algorithm for Positive Unlabeled Data

We have developed a novel gradient boosting algorithm designed to perform classification tasks on positive and unlabeled data. Loss minimization is an integral step in gradient boosting, as loss estimates the model's predictive ability, yet standard differentiable loss functions require completely labeled data as inputs. Instead, the novel gradient boosting algorithm minimizes an empirical risk function that requires only positive and unlabeled data as inputs.

Empirical risk is the average expectation of the loss over the entire data distribution. This definition can be decomposed into the positive expectation of the positive loss multiplied by the class prior, and the negative expectation of the negative loss, multiplied by the class prior.

$$\begin{aligned}\hat{R}(f) &= \mathbb{E}L(f(x), y) \\ &= P(y=1)\mathbb{E}_{D_+}L^+(f(x)) + P(y=0)\mathbb{E}_{D_-}L^-(f(x)) \\ &= \pi\mathbb{E}_{D_+}L^+(f(x)) + (1-\pi)\mathbb{E}_{D_-}L^-(f(x))\end{aligned}$$

However, we cannot use this exact definition to calculate empirical risk for positive unlabeled datasets, which lack labeled negative examples. In order to eliminate the need for labeled negative examples, we redefine the negative expectation of the negative loss in terms of the positive expectation of the negative loss, multiplied by the class prior, and the expectation of the negative loss over the entire data distribution [11].

$$\begin{aligned}&\mathbb{E}_D L^-(f(x)) = \pi\mathbb{E}_{D_+}L^-(f(x)) + (1-\pi)\mathbb{E}_{D_-}L^-(f(x)) \\ &\qquad\qquad \Longleftrightarrow \\ &\mathbb{E}_D L^-(f(x)) - \pi\mathbb{E}_{D_+}L^-(f(x)) = (1-\pi)\mathbb{E}_{D_-}L^-(f(x))\end{aligned}$$

Therefore, we may now determine empirical risk by calculating the positive expectation of the positive loss multiplied by the class prior, subtracting the positive expectation of the

negative loss multiplied by the class prior, and adding the expectation of the negative loss over the entire distribution. This method eliminates the need for labeled negative examples in the loss minimization step of a gradient boosting algorithm, making it ideal for work on positive unlabeled datasets.

$$\begin{aligned}\hat{R}(f) &= \mathbb{E}L(f(x), y) \\ &= P(y=1)\mathbb{E}_{D+}L^{+}(f(x)) + P(y=0)\mathbb{E}_{D-}L^{-}(f(x)) \\ &= \pi\mathbb{E}_{D+}L^{+}(f(x)) + (1-\pi)\mathbb{E}_{D-}L^{-}(f(x)) \\ &= \pi\mathbb{E}_{D+}L^{+}(f(x)) + \big(E_{D}L^{-}(f(x)) - \pi E_{D+}L^{-}(f(x))\big) \\ &= \pi\mathbb{E}_{D+}L^{+}(f(x)) - \pi E_{D+}L^{-}(f(x)) + E_{D}L^{-}(f(x))\end{aligned} \tag{2}$$

We implement our novel classifier in PyTorch [10]. The PU empirical risk function defined in Equation (2) replaces a standard differentiable function at gradient boost's loss and residual minimization steps. We compute its first term as the binary cross entropy loss over the predictions for positive examples and positive labels. We calculate its second term as the binary cross entropy loss over the predictions for positive examples and the same number of synthetic negative labels. We use synthetic negatives as we assume the data contains only positive and unlabeled examples. Lastly, we calculate its third term as the binary cross entropy loss over the predictions for all examples and the same number of synthetic negative labels. We average each function term before taking the sum.

While we specifically incorporate the empirical risk function for positive unlabeled data (Equation (2)) within a Gradient Boosting classifier, our approach is flexible enough to be implemented into multiple algorithms. Though not discussed in this paper, this method of loss function modification may be used to equip other classifiers for positive unlabeled learning, such as XGBoost, SVM, or Stochastic Gradient Descent.

## IV. Experimental Study Evaluation

TABLE I: Overview of the datasets used to evaluate the novel gradient boosting classifier.

| Dataset | Num. Instances | Prop. Pos. Instances |
|---|---|---|
| HBC [2] | 306 | 26% |
| PID [2] | 768 | 35% |
| CHD [2] | 303 | 55% |
| RC: Nasal [7] | 7167 | 5.2% |
| RC: Cough [7] | 7167 | 3.8% |
| RC: Stressed [7] | 7167 | 3.9% |
| RC: Sad [7] | 7167 | 1.7% |
| RC: Nausea [7] | 7167 | 1.0% |
| RC: Any [7] | 7167 | 9.7% |

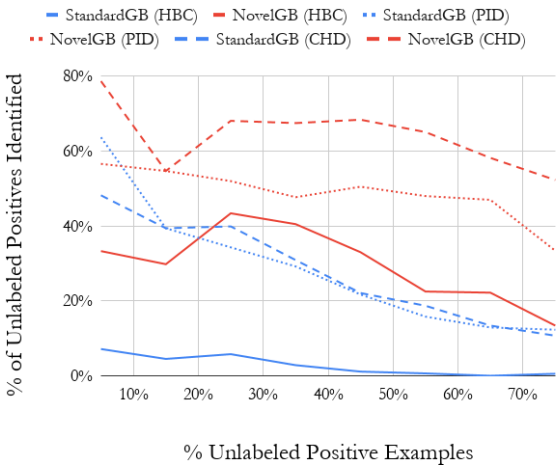
### A. Novel Gradient Boost on Three Standard Health Data Sets

We first compare the overall performance with respect to positive unlabeled instances of our novel gradient boosting classifier (NovelGB) to the state-of-the-art Gradient Boosting Classifier from scikit-learn (StandardGB) [10] using three standard health datasets: *Haberman Breast Cancer* (HBC), *Pima Indians Diabetes* (PID), and *Cleveland Heart Disease* (CHD) datasets, which are detailed in Table I. These datasets likely do not contain many positive unlabeled examples, as their target variables are measured rather than self-reported. We manually increase the proportion of positive unlabeled examples in each dataset to learn about the impact of such unlabeling on the methods. We use a 70:30 train test split and perform 10-fold cross validation at each proportion.

TABLE II: At a high proportion of positive unlabeled examples (75%), the NovelGB outperforms the StandardGB for each standard health dataset.

| Dataset | Method | % PU Examples | Balanced Accuracy | Recall |
|---|---|---|---|---|
| HBC | Novel | 75% | **0.508** | **0.113** |
| HBC | Standard | 75% | 0.497 | 0.004 |
| PID | Novel | 75% | **0.570** | **0.311** |
| PID | Standard | 75% | 0.537 | 0.117 |
| CHD | Novel | 75% | **0.564** | **0.520** |
| CHD | Standard | 75% | 0.534 | 0.004 |

Fig. 3: The NovelGB predicts the correct class for more positive unlabeled examples than the StandardGB for the HBC, PID and CHD datasets. Each result shown represents an average over 10 experiments.




NovelGB outperforms StandardGB on each dataset as the proportion of positive unlabeled unlabeled instances, exemplified in Table II. It achieves a higher average balanced accuracy in 13 of 24 experiments, 8 of which occur when more than half the positive examples in a dataset were unlabeled. NovelGB obtains a superior recall in every experiment, ranging from 7% to 51% higher than StandardGB. Importantly, the NovelGB predicts the correct label for substantially more positive unlabeled examples than the StandardGB in every dataset, as shown in Figure 3.

### B. Novel Gradient Boost on Reality Commons Dataset

Reality Commons is unique among the datasets used in that it likely contains many positive unlabeled examples. Its target variables—which indicate if an individual experienced the given flu symptom on that day—are all self-reported. We compare the NovelGB's performance against two other methods: the StandardGB and a positive unlabeled bagging technique (PU-SVM), [8], which is a method modification

TABLE III: Comparing performance of NovelGB (dark grey), StandardGB (white) [15], and PU-SVM (light grey) [8] on Reality Commons. NovelGB achieves the highest overall performance.

| Method | Target | Balanced Accuracy | Recall |
|---|---|---|---|
| NovelGB | Nasal | **0.511** | 0.279 |
| StandardGB | Nasal | 0.413 | **0.313** |
| PU-SVM | Nasal | 0.497 | 0.012 |
| NovelGB | Cough | **0.541** | **0.230** |
| StandardGB | Cough | 0.453 | 0.106 |
| PU-SVM | Cough | 0.494 | 0.006 |
| NovelGB | Stress | **0.532** | 0.328 |
| StandardGB | Stress | 0.480 | **0.759** |
| PU-SVM | Stress | 0.5 | 0.017 |
| NovelGB | Sad | **0.508** | **0.295** |
| StandardGB | Sad | 0.483 | 0.016 |
| PU-SVM | Sad | 0.506 | 0.016 |
| NovelGB | Nausea | **0.648** | **0.364** |
| StandardGB | Nausea | 0.486 | 0 |
| PU-SVM | Nausea | 0.491 | 0.017 |
| NovelGB | Any | **0.544** | 0.230 |
| StandardGB | Any | 0.472 | **0.745** |
| PU-SVM | Any | 0.497 | 0.011 |

approach to PU learning. Like the StandardGB, this method requires that we treat unlabeled data points as negative. We implement the PU-SVM as detailed in [13] using SVM as the base classifier.

We use a 60:40 train test split for all experiments, and apply SMOTETomek resampling to our training data for all experiments to handle class imbalance. We tune hyperparameters in each algorithm using grid search.

We find that the NovelGB outperforms the StandardGB for every target variable, as shown in Table III. It achieves the highest balanced accuracy in every case, and the highest recall for half of the targets. Interestingly, the NovelGB achieves the greatest improvement over the other two methods for the nausea target, which is the most imbalanced with positive instances representing 1% of all instances.

### C. Preprocessing Techniques for Positive Unlabeled Data on Reality Commons Dataset

In order to compare the NovelGB to preprocessing techniques for positive unlabeled data, we test a feed-forward neural network (PU-FFNN) equipped with the PU empirical risk function defined in Equation (2). Its architecture consists of three linear layers with hidden dimension size ten, and a dropout layer with probability 0.3. We use the Adam optimizer. Batch size and learning rate are kept at 256 and $10^{-5}$. Training epochs vary between 750-1000 depending on the experiment.

We find that the NovelGB outperforms the PU-FFNN on the three most imbalanced targets. It always obtains a false positive rate between 21% and 61% lower than the PU-FFNN, depending on the symptom. A low false positive rate is desirable in a medical context, as it indicates that fewer healthy patients are falsely identified as ill.

## V. Conclusion

Through experimentation on standard health datasets and the Reality Commons dataset, we demonstrate that our NovelGB outperforms the state-of-the-art on positive unlabeled data. The StandardGB expects completely labeled data and thus likely misclassifies many unlabeled positives. The PU-SVM requires that we assume a subset of unlabeled examples are negative, which may account for its poor performance. The PU-FFNN's relative failure may be attributed to Reality Commons' smaller size, as deep learning techniques are typically better suited for very large datasets [14].

We have improved upon a popular classification algorithm, making it applicable to fields where a dearth of machine learning methods for such data exist. Potential applications include the battle against COVID-19, where collected data would likely be positive and unlabeled due to sick individuals who go untested. Future directions include also adapting the XGBoost algorithm for positive unlabeled learning, because its regularization step may help avoid overfitting tendencies that could affect Gradient Boost.

## VI. Acknowledgements

We would like to thank the Worcester Polytechnic Institute Data Science department, the NSF REU site: Data Science Research for Healthy Communities in the Digital Age, the NSF IIS grant #1815866 and the DARPA WASH program HR001117S0032.

## References

[1] J. Bekker and J. Davis. 2020. Learning from positive and unlabeled data: a survey. Machine Learning 109:719-760.

[2] D. Dua and C. Graff. 2019. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] C. Elkan and K. Noto. 2008. Learning classifiers from only positive and unlabeled data. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 213-220.

[4] J. Friedman. 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics 29(5):1189-1232.

[5] S. Jain, M. White, and P. Radivojac. 2017. Recovering true classifier performance in positive-unlabeled learning. arXiv preprint arXiv:1702.00518.

[6] A. Kaboutari, J. Bagherzadeh, F. Kheradmand. 2014. An Evaluation of Two-Step Techniques for Positive-Unlabeled Learning in Text Classification. International Journal of Computer Applications Technology and Research 3(9):592-594.

[7] A. Madan, M. Cebrian, S. Moturu, K. Farrahi, A. Pentland. 2012. Sensing the 'Health State' of a Community. Pervasive Computing 11(4):36-45.

[8] F. Mordelet and J.P Vert. 2014. A bagging SVM to learn from positive and unlabeled examples. Pattern Recognition Letters 37:201-209.

[9] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer. 2017. Automatic differentiation in Pytorch.

[10] F. Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12:2825-2830.

[11] M. du Plessis, G. Niu, M. Sugiyama. 2014. Analysis of Learning from Positive and Unlabeled Data. In Proceedings of the 27th International Conference on Neural Information Processing Systems. 703-711.

[12] V. Vapnik. 1991. Principles of risk minimization for learning theory. Proceedings of the 4th International Conference on Neural Information Processing Systems. 831-838.

[13] R. Wright. 2017. Bagging Meta-Estimator for PU Learning. [https://roywrightme.wordpress.com/2017/11/16/positive-unlabeled-learning/].

[14] M. Zahangir Alom et al., 2019. A State-of-the-art Survey on Deep Learning Theory and Architectures. Electronics 8(3):292.

[15] Ayyadevara V.K. 2018. Gradient Boosting Machine. In: Pro Machine Learning Algorithms. https://doi.org/10.1007/978-1-4842-3564-5_6.