Evaluating the Quality of Positive Unlabeled Learning Methods if Unlabeled Instances Cannot be Validated

Praveen Kumar[1], Kristan A Schneider[1], Fariha Moomtaheen[1], Rajesh Upadhayaya[1], Scott A. Malec[1], Jeremy J. Yang[1], Cristian G. Bologa[1], Yiliang Zhu[1], Mauricio Tohen[2], Gerardo Villarreal[2,3], Douglas J. Perkins[1], Elliot M. Fielstein[4], Sharon E. Davis[4], Michael E. Matheny[4,5], Christophe G. Lambert[1]

[1]University of New Mexico, Department of Internal Medicine, Albuquerque, NM, USA
[2]University of New Mexico, Department of Psychiatry & Behavioral Sciences, Albuquerque, NM, USA
[3]VA New Mexico Healthcare System, Albuquerque, NM, USA
[4]Vanderbilt University Medical Center, Department of Biomedical Informatics, Nashville, TN, USA
[5]Tennessee Valley Healthcare System VA, Nashville, TN, USA

Background

While clinical diagnoses can confirm the presence of a medical condition, the absence of a diagnosis does not necessarily imply the absence of disease. This extends to electronic healthcare records: the absence of an International Classification of Diseases (ICD) code does not imply the absence of the underlying medical condition. When attempting to automatically classify medical conditions in the growing field of health informatics, this asymmetry in diagnostics/coding – or more generally labeling – creates a fundamental problem for traditional supervised learning, which assumes the availability of both positive and negative examples. The absence of the latter leads to the problem of positive unlabeled (PU) learning. Recently, novel methods to estimate the fraction of positive unlabeled instances were proposed[1,2,3]. However, the performance of such methods is difficult to ascertain if it is impractical or even impossible to validate unlabeled instances, as in the case of psychological conditions such as depression, bipolar disorder, and post-traumatic stress disorder (PTSD).

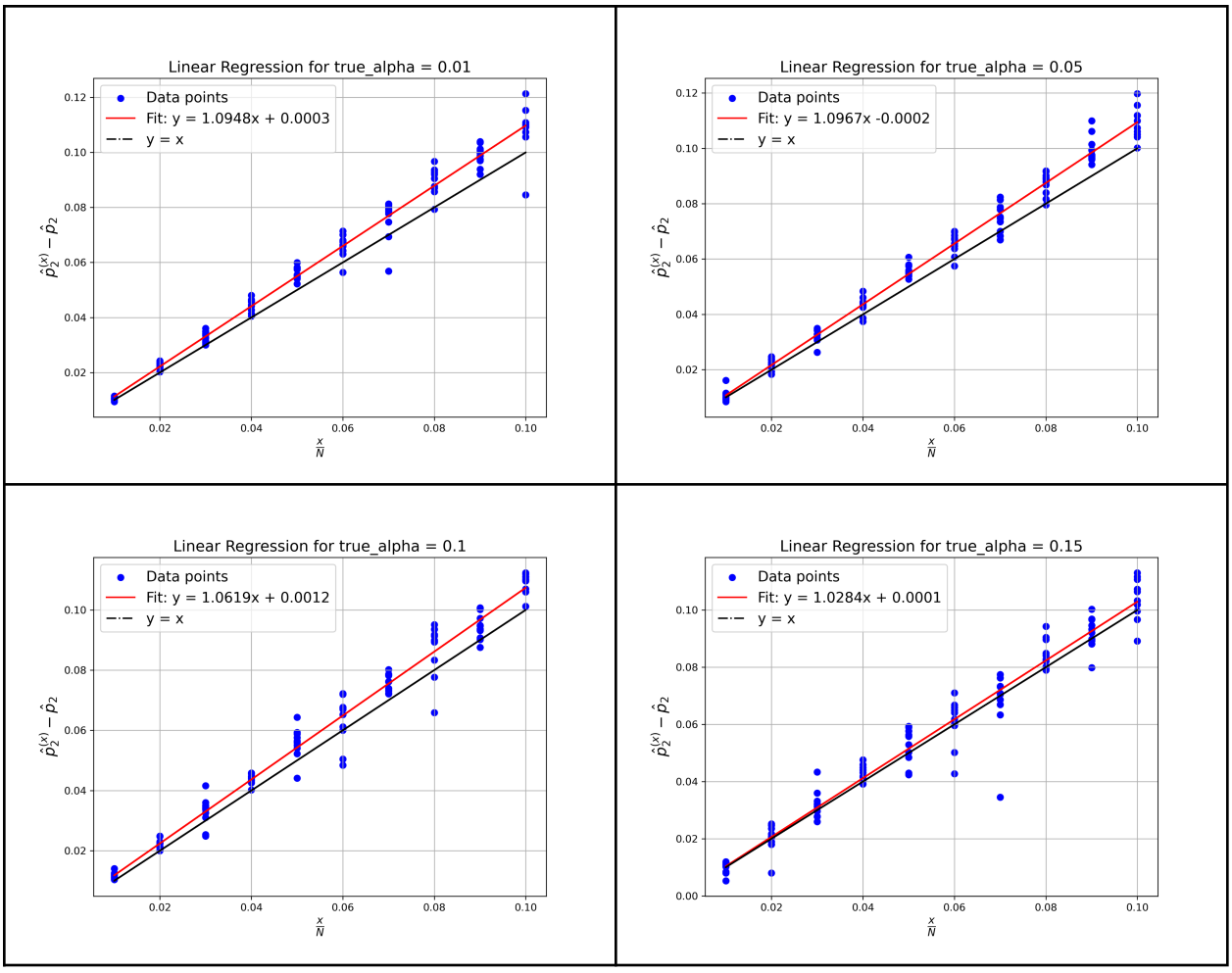
Materials and Methods

As such, we propose a method to validate the performance of PU learning algorithms by manipulating the set of labeled instances. Specifically, when removing its label, an instance becomes an unlabeled positive. Thus, the percentage of unlabeled cases increases by exactly the amount by which the percentage of labeled instances decreases. Formally, consider a sample of size N, with N_1 labeled positive, N_2 unlabeled positive, and N_3 unlabeled negative instances ($N=N_1+N_2+N_3$ – the sum N_2+N_3 is known, while N_2 and N_3 are unknown). When removing the labels of x ($0 < x < N_1$) positive instances, the modified data set contains N_1-x positive labeled and N_2+x positive unlabeled instances. Hence, the percentage of positive unlabeled instances increases by x/N. This relation can be utilized to validate the PU learning method using a linear regression approach. The slope of a straight line regression with the fraction of removed labels (x/N) as independent and the estimated percentage of PU cases as the dependent variable should theoretically be equal to one, if the PU learning algorithm is unbiased. The intercept of the straight line regression corresponds to the true fraction of PU instances.

We use synthetic data and three benchmark datasets (KDD cup 2004 particle physics[4], Magic gamma telescope[5], Mice protein expression[6]) to validate the four PU learning algorithms (PULSCAR[1], KM1[2], KM2[2], and TIcE[3]).

Results

The results show that the linear regression approach is well-suited to ascertain the bias of four different PU learning algorithms. Particularly, the method exposes the bias of the different algorithms. As expected, the novel method PULSCAR overestimates the true fraction of PU instances, because it estimates an upper bound rather than the true percentage. However, the slope of the regression line can be used to bias-correct the estimates. In accordance with previous results PULSCAR is superior to the other three methods (KM1, KM2, TIcE), as reflected by larger deviations of the respective regression slopes from 1.



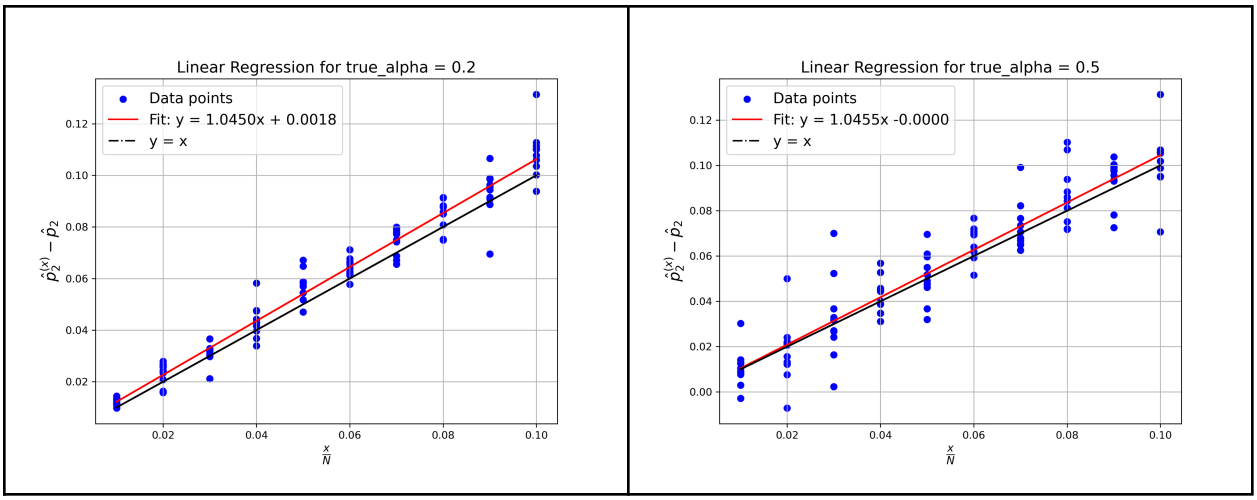


Figure 1. Linear regression plots based on synthetic datasets for six different proportions of positive examples within the unlabeled set: 1%, 5%, 10%, 15%, 20%, and 50%. For each proportion value, 10 distinct synthetic datasets were generated, varying in the number of labeled positives and unlabeled examples. The counts of labeled positives and unlabeled examples were (1000, 2000), (2000, 4000), …, (10,000, 20,000), where the first number indicates the number of labeled positives and the second the number of unlabeled examples. The parameters in the plot are defined as follows: $p_1 = \frac{N_1}{N}$, $p_2 = \hat{\alpha}(1 - p_1)$, and $\hat{p}_2^{(x)} = \hat{\alpha}^{(x)}(1 - p_1 + \frac{x}{N})$. Here $\hat{\alpha}$ is the estimated proportion of positive examples within the unlabeled set using PULSCAR, and $\hat{\alpha}^{(x)}$ is the estimated proportion of positive examples within the unlabeled set using PULSCAR when labels of x positives were changed from 1 to 0.

Discussion and Conclusion

The regression approach adapted here performs well on synthetic data and benchmark datasets, for which the true labels of all instances are known. The agreement between theoretical predictions and observations on these selected data sets, suggest that the proposed method is adequate for ascertaining the performance of PU learning algorithms, when applied to empirical data sets for which the status of unlabeled instances cannot be assessed, such as in the case of depression, bipolar disorder, or PTSD.

Acknowledgment

This research was supported by funding from the US National Institutes of Health, specifically, the National Institute of Mental Health grant R01MH129764, the National Library of Medicine grant R00LM013367, and infrastructure support from the National Center for Advancing Translational Sciences grant UL1TR001449.

References

1. Kumar P, Lambert CG. 2024. Positive Unlabeled Learning Selected Not At Random (PULSNAR): class proportion estimation without the selected completely at random assumption. PeerJ Computer Science 10:e2451 https://doi.org/10.7717/peerj-cs.2451

2. Ramaswamy H, Scott C, Tewari A. Mixture proportion estimation via kernel embeddings of distributions. InInternational conference on machine learning 2016 Jun 11 (pp. 2052-2060). PMLR.
3. Bekker, J., & Davis, J. (2018). Estimating the Class Prior in Positive and Unlabeled Data Through Decision Tree Induction. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11715
4. Caruana, R., Joachims, T., and Backstrom, L. (2004). KDD-Cup 2004: results and analysis. SIGKDD529 Explor. Newsl., 6(2):95–108.530
5. Bock, R. (2007). MAGIC Gamma Telescope. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52C8B.
6. Higuera, C., Gardiner, K., and Cios, K. (2015). Mice Protein Expression. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C50S3Z.570