Viraj Biharie
(V.A.A.Biharie@tudelft.nl)
EEMCS, Delft University of Technology

Assessing Machine Learning Robustness to Sample Selection Bias

Evaluating the effectiveness of semi-supervised learning techniques

1. Introduction

In many real-world scenarios the data used to train a machine learning model is not representative of the entire population → sample selection bias

Limited availability of labeled data is a common challenge in practice. To address this, semi-supervised learning [1] utilizes a small labeled dataset along with a larger amount of unlabeled data. These techniques will be used to mitigate sample selection bias in this study.

This study will investigate the following research questions:

- How effectively do the synthetic datasets generated in this research manifest the intended bias?
- How does the performance of self-training and co-training adapt to varying levels of bias in the data?
- How does the performance of self-training and co-training vary with different quantities of labeled data?

2. Methodology

1. Synthesize datasets

2. Induce bias in the synthesized datasets

3. Train following semi-supervised learning techniques on the synthesized datasets:
   - Self-Training [2]
   - Co-Training [3]

4. Evaluate performance using accuracy, F1 score [4] and AUC-ROC score [5] on different dataset instances and iterations while varying bias weight and the unlabeled set size

3. Results

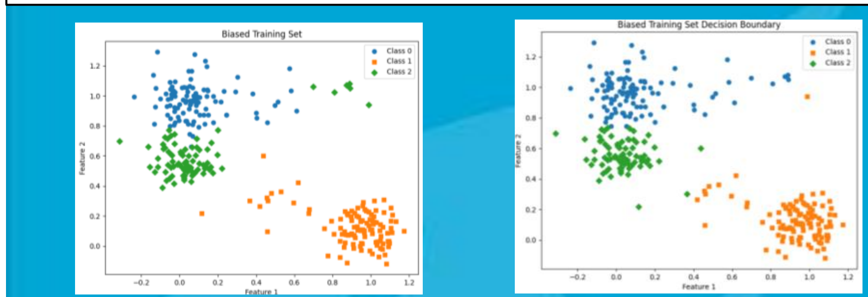

Figure 1. This figure illustrates the training set after the partitioning into the training set, unlabeled set, and test set. The figure on the left displays the training set after it has been biased. The figure on the right depicts the data after applying the decision boundary obtained through logistic regression [6]. The image on the left displays the data points with their true labels with a unique color representing the class. The image on the right displays the data points with their assigned labels based on the decision boundary created through logistic regression.

From figure 1 it can be observed that the upper right cluster is misclassified as class 0 (mostly blue in the right figure), while originally belonging to class 2 (green in left figure figure). These points have now a different label than before, thus the decision boundary has been changed due to the biasing as was intended.

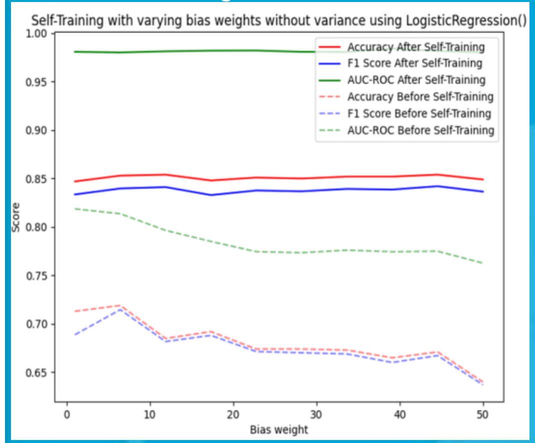

Figure 2. This figure showcases the average scores of ten dataset instances trained using ten iterations of the self-training algorithm.

From figure 2 it can be observed that the bias weight does not influence the performance of the self-training model. Furthermore, the figure demonstrates that the self-training algorithm (displayed with solid lines) surpasses the baseline approach (displayed with dashed lines).

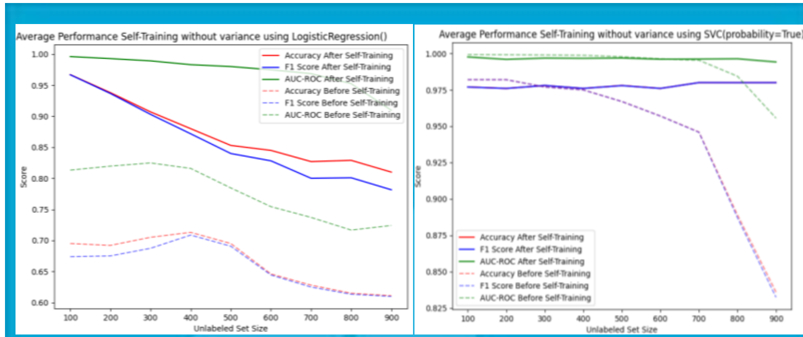

Figure 3. These figures present the mean scores obtained by splitting the entire dataset into different proportions of labeled and unlabeled data. The unlabeled data is utilized in conjunction with the labeled biased training set for self-training. The size of the test set remains constant at 100 samples throughout the experiment. The figure on the left uses logistic regression, whereas the figure on the right uses support vector classification (SVC) [7].

From figure 3 it can be observed that the performance of the baseline and self-training using logistic regression decreases as the unlabeled set size increases. Furthermore, self-training outperforms the baseline in the figure on the left and matches and outperforms it in the figure on the right. Comparing both figures it can be concluded (based on the percentages on the y-axis) that SVC has overall a better performance than logistic regression.

4. Conclusion

- Semi-supervised models outperform baseline
- Self-training outperforms co-training
- SVC outperforms Logistic Regression

Responsible Professor: Joana de Pinho Gonçalves
Supervisor: Yasin Tepeli
Examiner: Julian Urbano Merino

References
[1] Kouw, W. M., & Loog, M. (n.d.). A review of domain adaptation without target labels. Retrieved from https://arxiv.org/abs/1901.05335
[2] . C. A. P. Reddy, P. Viswanath, and B. E. Reddy. Semi-supervised learning: A brief review. Int. J. Eng. Technol, 7(1.8):81, 2018.
[3] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT '98, pages 92–100, New York, NY, USA, 1998. Association for Computing Machinery.
[4] Zachary Chase Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. Thresholding classifiers to maximize f1 score, 2014.
[5] Sarang Narkhede. Understanding auc-roc curve. Towards Data Science, 26(1):220–227, 2018.
[6] Haoran Chai, Yiqing Liang, Siyu Wang, and et al. A novel logistic regression model combining semisupervised learning and active learning for disease classification. Scientific Reports, 8:13009, August 2018.
[7] Yuanqing Li, Huiqi Li, Cuntai Guan, and Zhengyang Chin. A self-training semi-supervised support vector machine algorithm and its applications in brain computer interface. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, volume 1, pages I–385–I–388, 2007.