Team Members

Liu Boyu A0177847J, Liu Yu A0177906R, Wang Zihong A0230339L, Wu Ruichi A0177880N, Zhu Xinji A0177866H

Paper Chosen and Project Type

We choose the paper "A Variational Approach for Learning from Positive and Unlabeled Data" written by Chen et al. [1] and our project focuses on the application.

Introduction to Paper

In real-world projects in the fields including web text classification, disease gene identification and fraud detection, we may face the problems of having a small proportion of positive labels and massive amounts of unlabelled data. Therefore, it is important to develop a proper way to learn information from such datasets. In our background reading 2, a Positive-unlabeled (PU) learning method called non-negative PU (nnPU) learning was developed. However, in this method, an inaccurate prior estimation was introduced, which may limit its application. To overcome the limitations, a novel method named Variational PU (VPU) algorithm is proposed in another paper which is the one we are trying to reproduce. VPU can learn information from positive and unlabeled data bypassing the prior estimation.

Background Reading

1. Learning From Positive and Unlabeled Data: A Survey (2018) [2]
This report analyzed the current state of the art in PU learning. It proposed seven key research questions that frequently arise in this field, as well as a broad overview of how the field has attempted to address them. It provides preliminary knowledge on PU Learning, including the key concepts, models, measurement methods and how PU transfer to real-world applications. With reference to this paper, we would be able to establish a complicated understanding of PU learning.
2. Positive-Unlabeled Learning with Non-Negative Risk Estimator (2017) [3]
This paper reviewed the drawback of unbiased PU (uPU) learning and proposed nnPU learning. In the experiments, PN, uPU and nnPU were implemented using four datasets and the performance of the models are compared. We will refer to this paper to learn the theory and implementation of nnPU.

Reproduction and Extension

We plan to reproduce the VPU algorithm based on the methodology introduced in the paper. We will test the effectiveness of the VPU on the same datasets used in the paper. The performance of VPU and other benchmark models such as uPU and nnPU will also be compared and investigated. The goal of the reproduction is to provide a thorough elaboration of the proof of the main results. For the extension, we plan to generalize the results by using different datasets.

Evaluation

The paper compared the classification accuracies of various models on different datasets. In our study, we also use accuracy as the metric to evaluate the models on the same datasets, allowing direct comparison with the previous work. Similar to the paper, we will test the accuracies of different PU methods, and compare them with the VPU method.

Planned Division of the Work

Background (literature review): Liu Yu
Data Preprocessing: Wu Ruichi
Model Construction: Liu Boyu, Wang Zihong and Zhu Xinji
Training and Testing: Liu Boyu, Wang Zihong and Zhu Xinji
Results Evaluation: Liu Yu and Wu Ruichi
Report Drafting: All team members
The distribution of work might be changed based on the progress of the project.

References

[1] Hui Chen, Fangqing Liu, Yin Wang, Liyue Zhao, and Hao Wu. A variational approach for learning from positive and unlabeled data. *arXiv preprint arXiv:1906.00642*, 2019.

[2] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109:719–760, 2020.

[3] Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.