POSITIVE-UNLABELED DATA CLASSIFICATION WITH USE OF DEEP LEARNING METHODS

Paweł Karol Koźmiński[1], Jan Mielniczuk[1,2]

[1] Faculty of Mathematics and Information Science, Warsaw University of Technology
[2] Institute of Computer Science, Polish Academy of Sciences
pkozminski99@gmail.com, jan.mielniczuk@pw.edu.pl

Positive-Unlabeled Data Classification

Positive-Unlabeled data classification is a special type of binary classification task, where the model has a limited access to the observations' labels: most of them are unknown, except for a part of positively marked ones. The problem naturally emerges in a vast number of applications, such as:

- disease diagnosis – besides patients who are diagnosed with a disease, there are those who are not diagnosed at all (and thus are either healthy or ill),
- surveys with answers possibly stigmatized by the society – when the people are likely to respond with a false negative. [1]

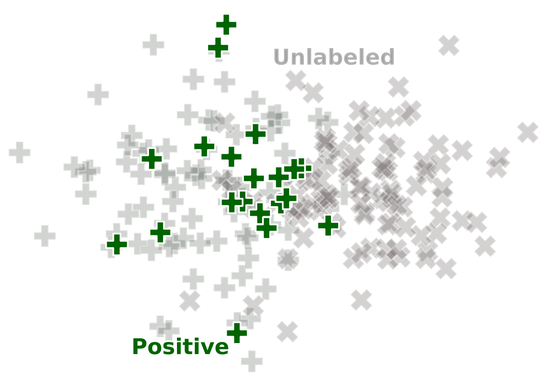

Fig. 1: PU data illustration

We consider **case-control** scenario for PU data, corresponding to the first example. Available samples: sample U (Unlabeled) pertaining to a general population governed by density f and sample L (Labeled) from positive class with density f_P.

The Variational Approach

Class prior π is a valuable information in PU learning methods design. However, its acquisition is difficult and based only on PU case-control data - infeasible. Thus, methods operating with no class prior provided are of a great importance. One of them, **VPU** [2], involves neural networks with a novel loss function to solve the problem:

$$\mathcal{L}(\Phi) = logE_f[\Phi(x)] - E_{f_P}[log\Phi(x)] + \\ + \lambda \left(E_{\tilde{\Phi},\tilde{x}} \left[\left(log\tilde{\Phi} - log\Phi(\tilde{x}) \right)^2 \right] \right)$$

where:

- $\tilde{x}$ are artificial examples, prepared as convex combinations of random positive and unlabeled examples
- Φ is an analogous combination of model's expected output (i.e. 1 for positive and $\Phi(x)$ for unlabeled input)

The loss is a weighted sum of variational and *MixUp* regularization components. The ideal Bayesian classifier related to $P(Y = 1 \mid x)$ is approximated with a parametric model Φ^*, s.t. $\Phi^* = \min_\Phi \mathcal{L}(\Phi)$.

Method Analysis

The variational part of the proposed loss function is related to the *Donsker-Varadhan representation* of the **Kullback-Leibler divergence**, known from information theory:

$$KL(P \,||\, Q) = \sup_{T \in \mathcal{F}} \left(E_P[\log T] - \log(E_Q[T]) \right)$$

In VPU maximizers are proportional to posterior probability $P(Y = 1 \mid x)$ and penalization terms are introduced to constrain them to be equal to it.

Modifications

Kullback-Leibler divergence can be bounded with use of different variational representations. Along with the original VPU, we adapt **other representations** to address PU learning problem and denote:

- R^I [3]

$$KL(P \,||\, Q) = \sup_{T \in \mathcal{F}} \left(E_P[\log T] - \log(E_Q[T]) \right)$$

- R^{II} [4]

$$KL(P \,||\, Q) = \sup_{T \in \mathcal{F}} \left(E_P[\log T] - E_Q[T/e] \right)$$

- R^{III}

$$KL(P \,||\, Q) = \sup_{T \in \mathcal{F}} \left(E_P[\log T] - E_Q[T] + 1 \right)$$

The above formulas are applied to P corresponding to f_P and Q corresponding to f.
R^I is the only method which yields posterior probability estimator without knowledge of π. Given a $T \in \mathcal{F}$, it obtains greater values than others. Nonetheless, the effectiveness of optimization by deep learning models is the subject of experiments. Moreover, for the original version of variational loss, we also propose and employ other **regularization terms** to the R^I representation:

- MixUp – original, with use of augmented synthetic examples:

$$\mathcal{L}_{\text{MixUp}}(\Phi) = E_{\tilde{\Phi},\tilde{x}} \left[\left(log\tilde{\Phi} - log\Phi(\tilde{x}) \right)^2 \right]$$

- π-based – tightening the average prediction over unlabeled prediction to π:

$$\mathcal{L}_\pi(\Phi) = \left(\log \bar{\Phi}(x_u) - \log \pi \right)^2$$

- Mixed – a combination of two previous functions:

$$\mathcal{L}_{\text{Mixed}}(\Phi) = \mathcal{L}_{\text{MixUp}}(\Phi) + \lambda_1 \mathcal{L}_\pi(\Phi)$$

Finally, KL divergence determination can also be represented as a dual problem:

$$KL(P \,||\, Q) = \sup_{T \in \mathcal{F}} E_P[s] \\ \text{w.r.t.} E_Q[e^s] = 1$$

Thus, we propose a new loss function with components responsible for seeking the supremum and penalizing deviations from the constraint.

Experiments

The proposed modifications were evaluated with comprehensive experiments, following these principles:

- the models' architectures were consistent with the ones proposed by VPU authors;
- we used **6 datasets** (3 tabular and 3 image) and measured accuracy on the test set;
- every experiment was performed with a set of **hyperparameters optimized** with Optuna;
- final result is the average accuracy over at least **10 runs**, reported with standard error;
- the methods are compared with **3 baselines**, considered as state-of-the-art in case-control PU learning area: unbiased PU (uPU), non-negative PU (nnPU) [5] and the original VPU;
- the starting ratio of labeled positive examples was $c = 0.4$. As an **ablation study** we carried out experiments with different values of c.

Results

Fig. 2 presents average accuracy of each method on all datasets. In general, R^I with mixed regularization and R^{III} consistently obtain relatively high values, in most cases higher than the baselines. On contrary, R^I with π-based regularization and dual-problem-based method perform poorly with unstable results.

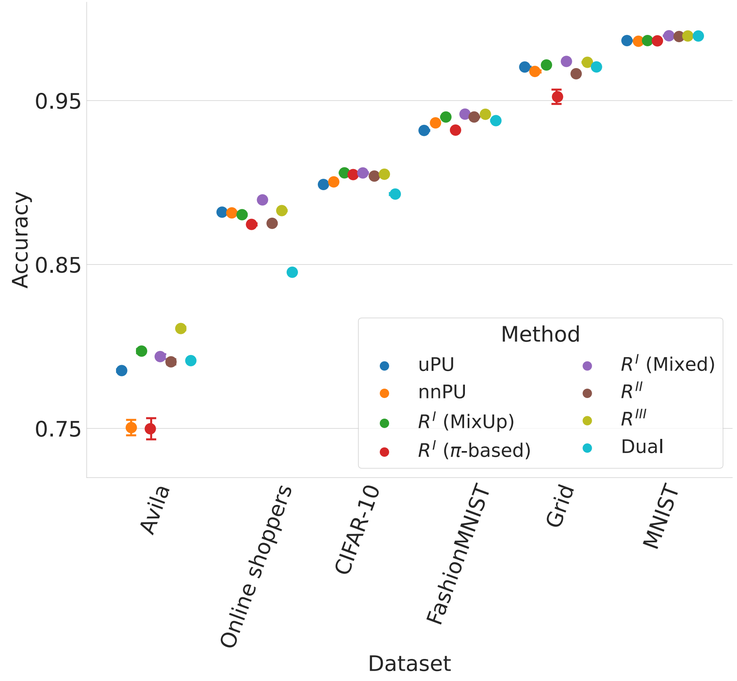

Fig. 2: Results for various datasets with positive-labeled ratio: 0.4

Fig. 3 presents the results of ablation study - multiple experiments were run on Avila dataset. As expected, the higher is the level of labeled positive observations, the more accurate are the methods. Again, R^{III} appears to perform the best as it obtains the greatest accuracy for 6 out of 9 c levels.

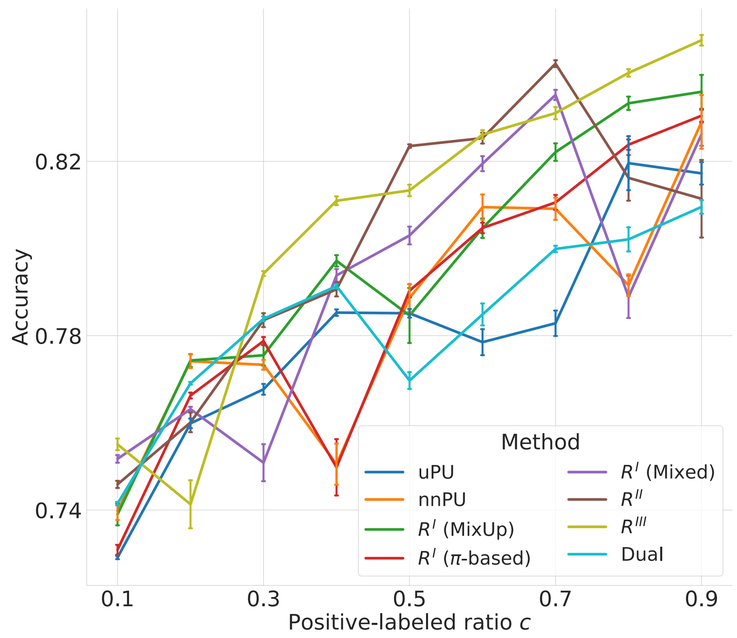

Fig. 3: Ablation study: Results on Avila dataset vs. positive-labeled ratio

Conclusions

In this work we exhaustively analyzed and further developed VPU method for PU learning method. Part of the proposed modifications are proved to be successful by broad experiments on various data. Though not consistently, R^I (Mixed), R^{II} and R^{III} usually obtain higher accuracy than other methods and can be considered in real applications.

References

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

[2] Hui Chen, Fangqing Liu, Yin Wang, Liyue Zhao, and Hao Wu. A variational approach for learning from positive and unlabeled data. *Advances in Neural Information Processing Systems*, 33:14844–14854, 2020.

[3] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212, 1983.

[4] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[5] Ryuichi Kiryo, Gang Niu, Marthinus C Du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems*, 30, 2017.

Faculty of Mathematics and Information Science
WARSAW UNIVERSITY OF TECHNOLOGY