| | COVID-19 | | | | Virology | | | | Immunotherapy | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1-S | F1-M | F1-P | LRAP | F1-S | F1-M | F1-P | LRAP | F1-S | F1-M | F1-P | LRAP |
| penL1-NNPU | 0.71 | 0.70 | 1.35 | 0.45 | 0.61 | 0.71 | 1.83 | 0.63 | 0.53 | 0.63 | 1.62 | 0.73 |
| TRP-NNPU | 0.80 | 0.82 | 1.68 | 0.89 | 0.73 | 0.82 | 2.38 | 0.83 | 0.67 | 0.78 | 2.18 | 0.76 |
| penL1-UPU | 0.85 | **0.88** | 1.86 | 0.89 | **0.74** | **0.83** | **2.45** | **0.81** | 0.70 | 0.81 | 2.33 | 0.72 |
| TRP-UPU | **0.86** | **0.88** | **1.88** | **0.91** | **0.74** | **0.83** | 2.38 | **0.81** | **0.71** | **0.82** | **2.35** | **0.77** |

We would like to thank the reviewers for the recognition of the novelty of our studied problem and our proposed solution, the attention to our evaluation including a Covid-19 dataset, as well as their valuable suggestions and comments.

**Organization and Writing (ALL)** We have gone through the paper multiple times to correct the grammatical and structural mistakes in the paper. We also address the issues raised by the reviewers, which are: 1) explain all notations before usage, 2) clearly denote the cost function $L^E$ in Eq(7), and 3) spell out all acronyms. Given that the page limit for review submission is eight pages, the images were reduced to fit the limit. We will increase the fonts and image sizes of the figures in the camera-ready version, which allows for an additional ninth page.

**Related works Citation and Comparison (R1)** We have cited the suggested papers [**I-III**], and also compared with the *pen-L1* method proposed in [**I**], for further evaluating the positive prior estimation in our TRP method. The results presented in the table above show that our TRP outperforms or is comparable to the *pen-L1* method in both UPU and NNPU setting. Thus, for positive prior estimation, it is more effective to use KL-divergence to estimate the parameters of class mixture distribution, than to assign higher penalties by penalized L1 divergences (*pen-L1*) to class priors that scale the positive distribution more than the total distribution. We will add this comparison in our paper.

**Prior Estimation Assumptions (R2,R4)** We follow a basic and the most generic SCAR assumption in PU learning: the labeled positive examples are chosen completely randomly from all positive examples. This means, for all positive $x$, they have the same probability to be labeled (sampled as observed positive instances). This is the positive prior probability $\pi_P = p(y = 1)$ to estimate, since there exist unobserved positive instances in the unlabeled set. Then we follow the assumption of *positive subdomain/anchor set* in [5] and [**I**], which realistically assumes the distribution of positive and negative may overlap, but the labeled positive form an anchor set in the positive subdomain. We thus use a Gaussian mixture model (GMM) to fit the data distribution, which is made up of two components: positive and negative. The estimated prior $\pi_P$ is the weight of the positive component. Thus, GMM is a suitable model under the *positive subdomain/anchor set* assumption for the positive prior estimation.

**Exp Result and Discussion (ALL)** (**R1**) In our experiments, we compared against recent SOTA methods that jointly learn a PU classifier while estimating the class prior or propensity score. They are: Elkan [14] proposed in 2008, SAR-EM and SCAR-C [7] proposed in 2019, SCAR-KM2 [35] proposed in 2016, and SCAR-TIcE [5] proposed in 2018. We now include *pen-L1* method proposed in 2015.
(**R2**) An increase in the number of positive samples improves the effectiveness of the model, as shown in Figure 3, where the number of training instances increases with the introduction of labeled pairs observed in incremental years.
(**R2,R3**) It is hard to provide a detailed comparison of the prior estimates given our dataset since the true prior is unknown in our real-world setting. Hence we evaluate the performance of the model as a whole based on the classification performance metrics. The results presented in Table 2 in the paper and in the above table validate the effectiveness of our prior estimation method, comparing to other prior estimation solutions.
(**R2**) We also found interesting that NNPU was worse than UPU in our experimental results. However, it is not uncommon for UPU to outperform NNPU in evaluation with real-world datasets. Similar observations were found in the results in [**I**] and [**II**]. In our case, we attribute this observation to the joint optimization of the loss from the classifier and the prior estimation. In the loss of UPU (Eq. (3)), $\pi_P$ played in weighting both $\hat{\mathcal{R}}_P^+(f)$ and $\hat{\mathcal{R}}_P^-(f)$. However, in the loss of NNPU (Eq. (4)), $\pi_P$ only weighted $\hat{\mathcal{R}}_P^+(f)$ when $\hat{\mathcal{R}}_U^- - \pi_P\hat{\mathcal{R}}_P^-(f)$ is negative. In real-world applications, especially when the true prior is unknown, the loss selection affects the estimation of $\pi_P$, and thus the final classification results.
(**R1,R3**) The full **code** and **datasets** will be provided upon acceptance.
(**R4**) We have included confidence intervals of TRP performance in Table 2.

**Usefulness of HG (R4)** Hypothesis generation can provide domain experts clues about the relevant concepts to explore. Following these clues, domain experts can retrieve the documents, including the relevant concepts, and then investigate the hypothesis's usefulness. For instance, *Schizophrenia* and *Calcium-Independent Phospholipase A2* were independently studied in different papers in 1997 and 1995. They were connected in 1998 [**IV**] because they had *oxidative stress* as a common factor. The *Five Golden Test Cases* are also often used in the qualitative evaluation of hypothesis generation papers [41,25,45]. Our next target is to develop a system that can generate a sentence for the generated hypothesis, including the context explaining why and how they are relevant. This requires to enrich the currently studied graph with the contextual information for the nodes and edges. The system will help to new scientific discoveries in a more explainable way.

## Reference


[I] Christoffel et al., "Class-prior estimation for learning from positive and unlabeled data." ACML. 2016.
[II] Gong et al., "Loss decomposition and centroid estimation for positive and unlabeled learning." PAMI. 2019.
[III] Shi et al., "Positive and Unlabeled Learning via Loss Decomposition and Centroid Estimation." IJCAI. 2018.
[IV] Smalheiser and Swanson, "Calcium-independent phospholipase A2 and schizophrenia." Archives of General Psychiatry. 1998.