

Document Set Expansion with Positive-Unlabeled Learning: A Density Estimation-based Approach

Haiyang Zhang^{1*}, Qiuyi Chen^{1*}, Yuanjie Zou¹, Yushan Pan¹, Jia Wang¹, Mark Stevenson²

¹Xi'an Jiaotong-Liverpool University

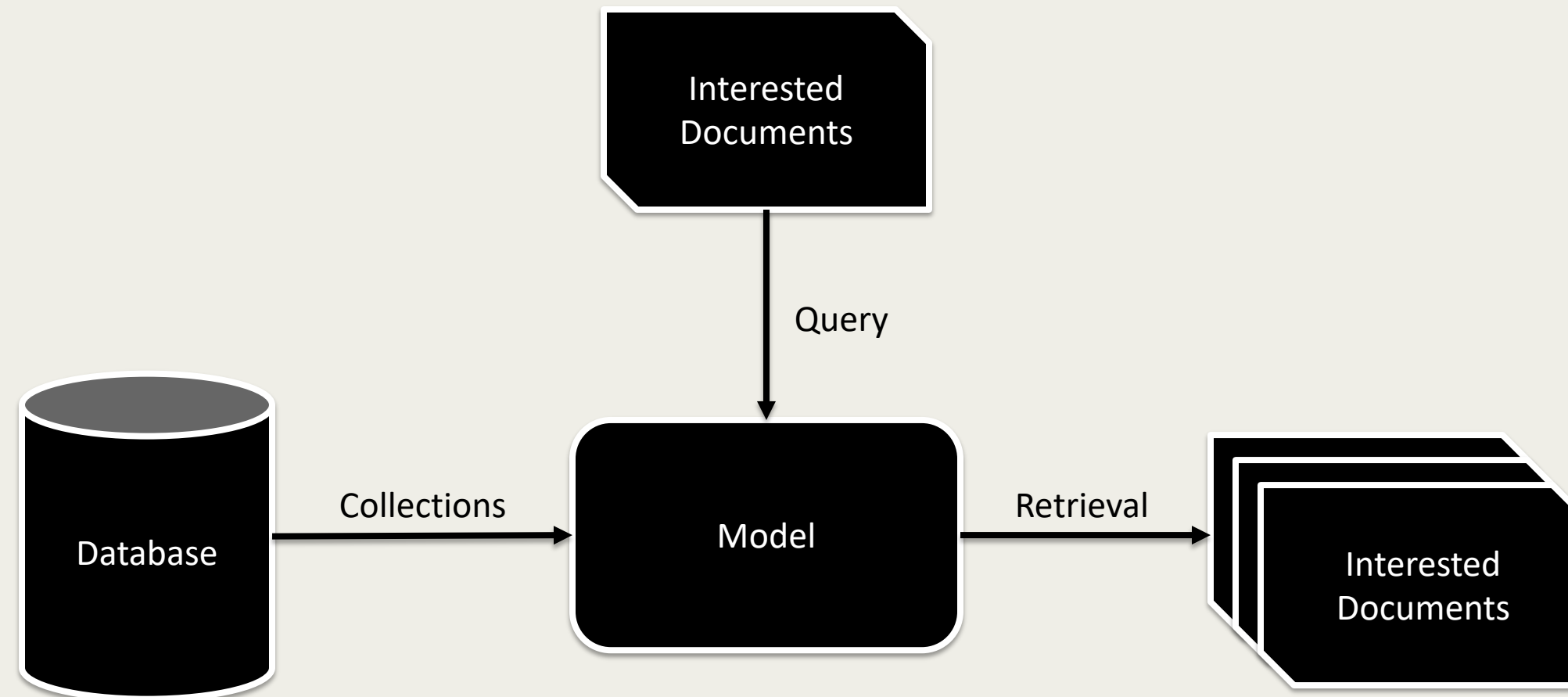
²University of Sheffield



LREC-COLING  2024

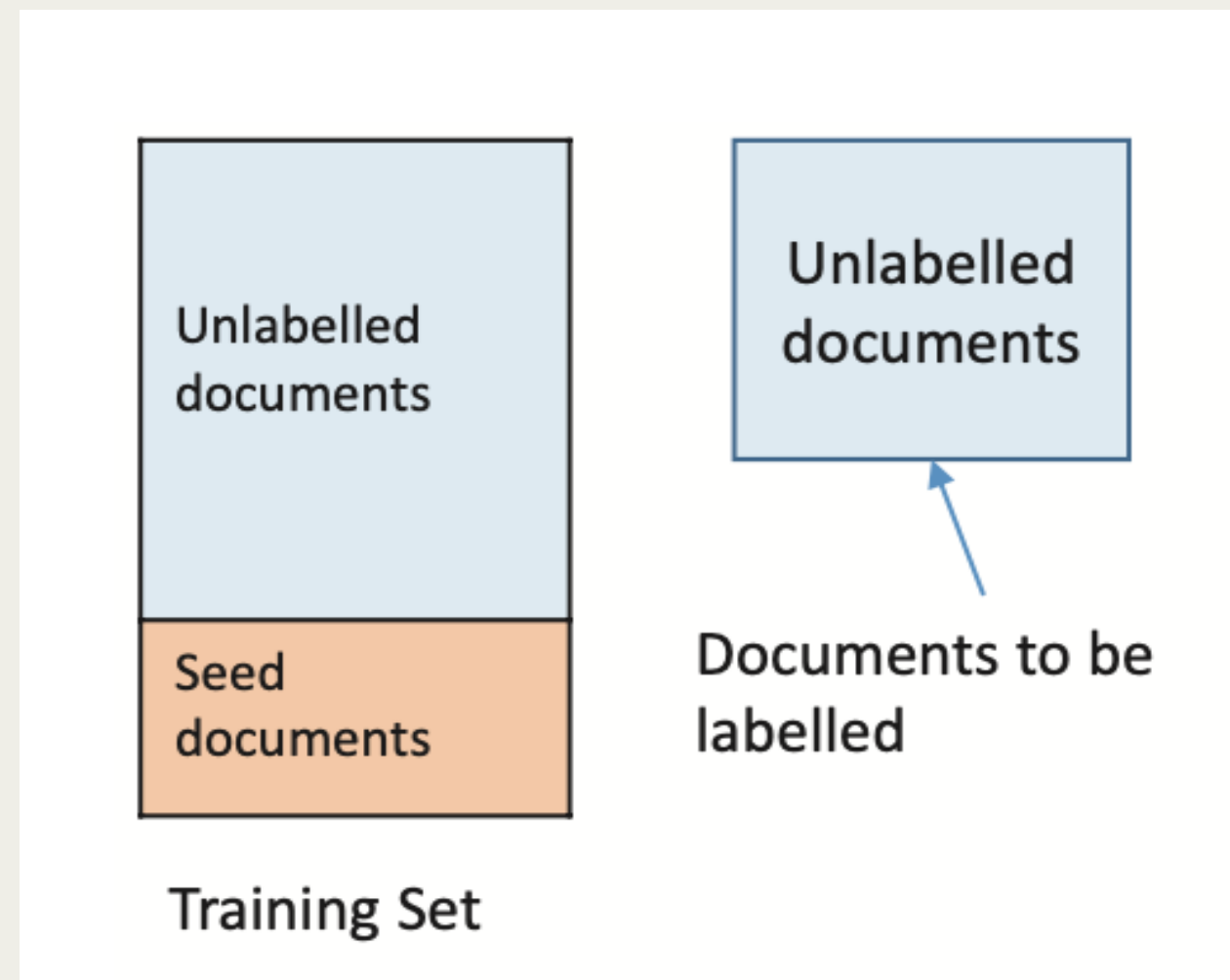
Document Set Expansion (DSE)

- We focus on the scenario where a user has access to a (possibly small) set of documents of interest and wishes to identify further similar documents within a large collections
- Query-by-document (QBD) is an approach to DSE which involves treating the set of documents as an extended query used to rank the documents in the collection.



Convert DSE into Positive-unlabeled Learning

Jacovi et al. (2021) treated the DSE task as a positive-unlabeled (PU) learning problem by learning a binary classifier from positive-unlabeled data (Kiryo et al., 2017) .



Alon Jacovi, Gang Niu, Yoav Goldberg, and Masashi Sugiyama. 2021. Scalable evaluation and improvement of document set expansion via neural positive-unlabeled learning. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 581–592, Online. Association for Computational Linguistics.

Ryuichi Kiryo, Gang Niu, Marthinus C. du Plessis, and Masashi Sugiyama. 2017. Positive-unlabeled learning with non-negative risk estimator. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 1674–1684, Red Hook, NY, USA. Curran Associates Inc.

Limitation of Existing Work

Jacovi et al. (2021) treat DSE as an inductive problem, where U is split into training and test sets, with only samples in the test set being labelled.

Existing PU(Positive-unlabeled) methods commonly utilize complex kernel machines.

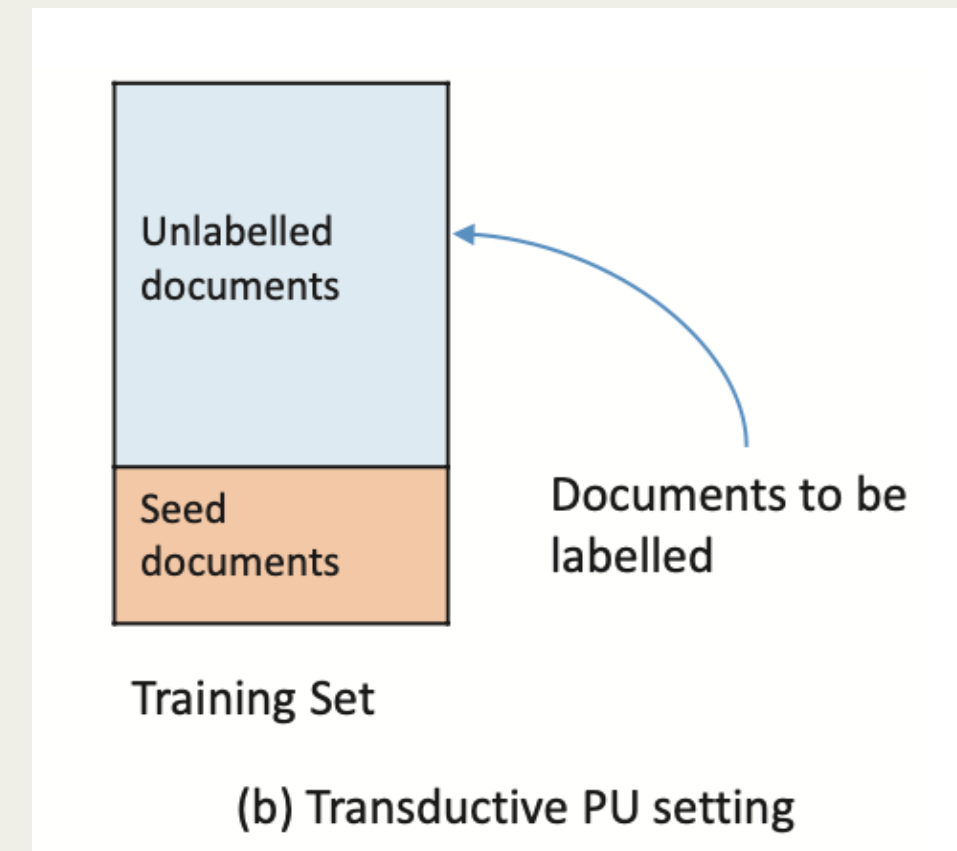
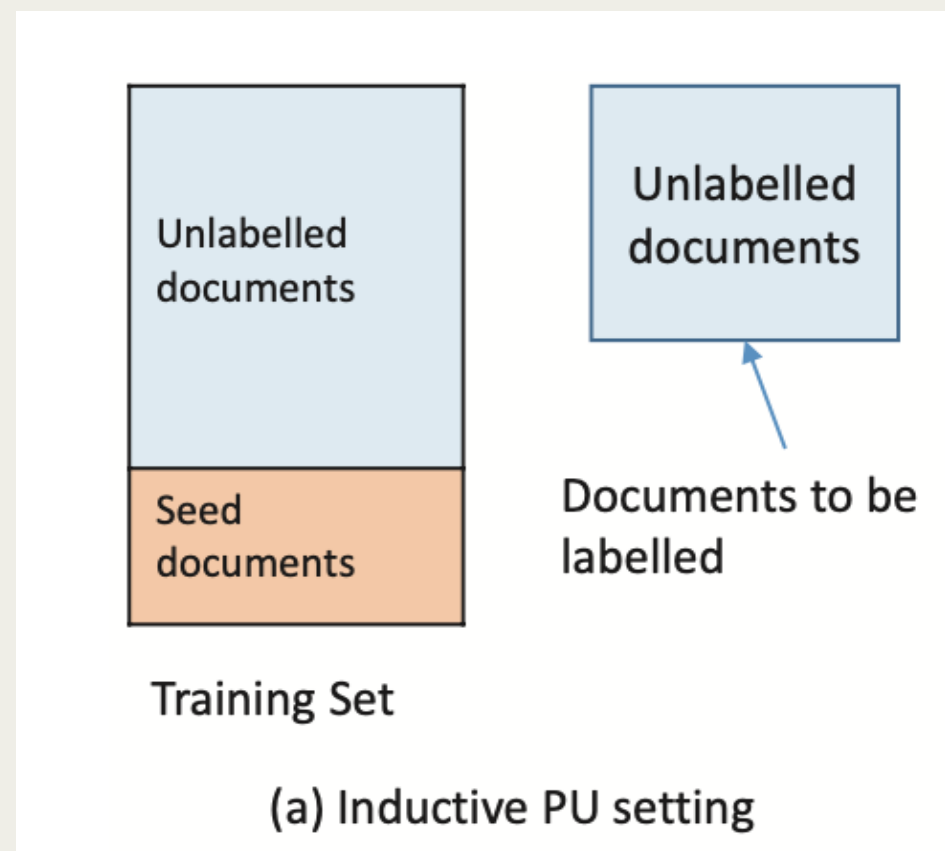
Inaccurate estimation may bring more errors in the PU classification (Chen et al., 2020).

Alon Jacovi, Gang Niu, Yoav Goldberg, and Masashi Sugiyama. 2021. Scalable evaluation and improvement of document set expansion via neural positive-unlabeled learning. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 581–592, Online. Association for Computational Linguistics.

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

Our Main Contributions

1. identify the limitations of previous for the DSE task (Jacovi et al., 2021);
2. propose **puDE**, a new PU learning framework by using intractable models for density estimation that does not require any knowledge of class prior ;
3. demonstrate that **puDE** outperforms state-of-the-art PU methods for the DSE task on real-world datasets.



Notation

- Random Variable: Input $X \in \mathbb{R}^d$
 Output $Y \in \pm 1$
- Density: $p_p(x) = p(x \mid Y = +1)$
 $p_n(x) = p(x \mid Y = -1)$
- Class-prior probability: $\pi = \pi_p = p(Y = +1)$
 $\pi_n = p(Y = -1)$
- Expectation: $\mathbb{E}_p[\cdot] = \mathbb{E}_{X \sim p_p}[\cdot]$
 $\mathbb{E}_n[\cdot] = \mathbb{E}_{X \sim p_n}[\cdot]$
- Dataset: $x_p = \{x_i^p\}_{i=1}^{n_p} \stackrel{\text{i.i.d.}}{\sim} p_p(x)$
 $x_n = \{x_i^n\}_{i=1}^{n_n} \stackrel{\text{i.i.d.}}{\sim} p_n(x)$
 $x_u = \{x_i^u\}_{i=1}^{n_u} \stackrel{\text{i.i.d.}}{\sim} p(x)$

Unbiased risk estimator

Let g be a decision function ℓ be a non-convex **0-1 loss function** such that

$$\ell(Y, g(X)) = \begin{cases} 0 & Y * g(X) > 0 \\ 1 & Y * g(X) < 0 \end{cases}$$

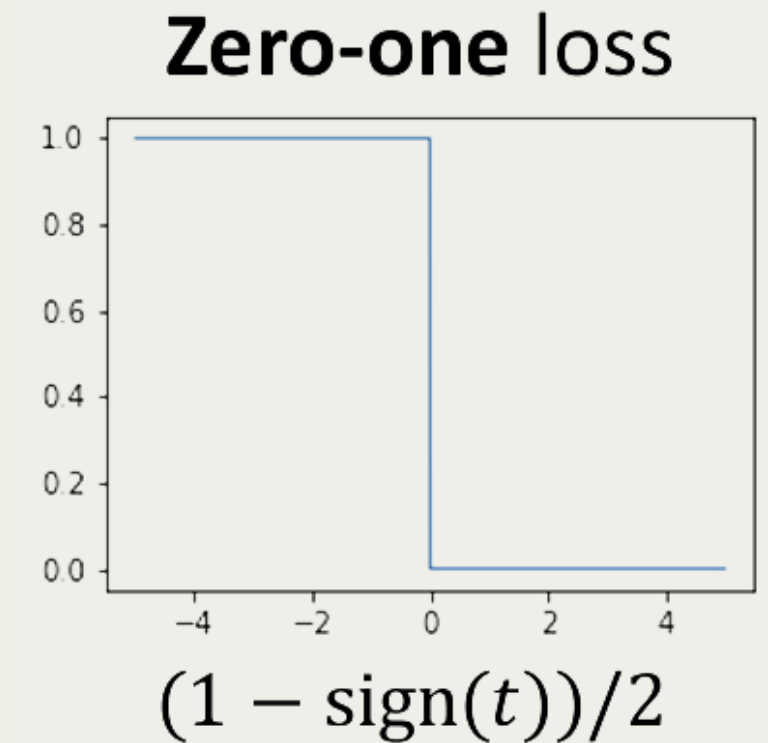
The risk of g is

$$\begin{aligned} R(g) &= \mathbb{E}_{(X,Y) \sim p(x,y)} [\ell(Y, g(X))] \\ &= \pi_p \mathbb{E}_p[\ell(g(X))] + \pi_n \mathbb{E}_n[\ell(-g(X))] \end{aligned}$$

where $\pi_n = 1 - \pi_p$. The risk can be approximated directly by

$$\hat{R}_{pn}(g) = \frac{\pi_p}{n_p} \sum_{x \in \mathcal{X}_p} \ell(g(x)) + \frac{\pi_n}{n_n} \sum_{x \in \mathcal{X}_n} \ell(-g(x))$$

This doesn't work for PU Learning



$$(l_n = -(y_n * \log(z_n) + (1 - y_n) * \log(1 - z_n)))$$

Unbiased risk estimator

- Key observation:

- $\pi_n p_n(x) = p(x) - \pi_p p_p(x)$ --- *Total Probability Formula*

- $\pi_n \mathbb{E}_n[\ell(-g(X))] = \mathbb{E}_X[\ell(-g(X))] - \pi_p \mathbb{E}_p[\ell(-g(X))]$

- Thus, the risk can be expressed as

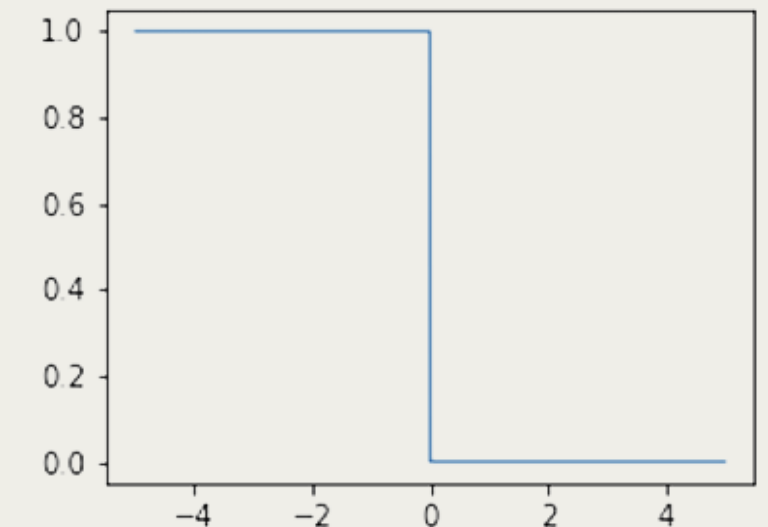
$$R(g) = \pi_p \mathbb{E}_p[\ell(g(X)) - \ell(-g(X))] + \mathbb{E}_X[\ell(-g(X))]$$

- By the property of non-convex 0-1 function $\ell(t) + \ell(-t) = 1$, we have:

$$R(g) = 2\pi_p \mathbb{E}_p[\ell(g(X))] + \mathbb{E}_X[\ell(-g(X))] - \pi_p$$

$$\begin{aligned} R(g) &= \mathbb{E}_{(X,Y) \sim p(x,y)}[\ell(Y, g(X))] \\ &= \pi_p \mathbb{E}_p[\ell(g(X))] + \pi_n \mathbb{E}_n[\ell(-g(X))] \end{aligned}$$

Zero-one loss



$$(1 - \text{sign}(t))/2$$

PU Learning with Density Estimation

By Bayesian rule, our desire distribution $\mathbb{P}(Y = 1 | \mathbf{x})$ could be express as:

$$\mathbb{P}(Y = 1 | X) = \frac{\mathbb{P}(X | Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(X)} = \frac{f_p(\mathbf{x})}{f(\mathbf{x})}\pi$$

If we can estimate the probability density ratio of $f_p(\mathbf{x})$ and $f(\mathbf{x})$, π will be a constant for each \mathbf{x} and can be ignored in training.

So, we train two separately model for predicting label:

$$g(\mathbf{x}) = \frac{p(\mathbf{x})}{q(\mathbf{x})}\pi \approx \mathbb{P}(Y = 1 | X)$$

Nonparametric Density Estimation

- We implement Kernel Density Estimation (KDE) as the model for density estimation
- One major advantage of KDE is that there is no need to make assumptions about the data distribution.

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

where h is a hyperparameter called bandwidth, and K is a non-negative kernel function.

Parametric Density Estimation

- **Energy-based models** have gained significant attention and have been demonstrated to perform well in diffusion models doing density estimation task
- Requirement of probability density function
 - $f(x)$ must be nonnegative for each value of the random variable
 - the integral over all values of the random variable must equal one

$$p_{\theta}(\mathbf{x}) = \frac{\exp(-E_{\theta}(\mathbf{x}))}{Z_{\theta}}$$

$$Z_{\theta} = \int \exp(-E_{\theta}(\mathbf{x}))d\mathbf{x}$$

Parametric Density Estimation

PU Learning with Density Estimation (puDE-EM)

Base on the Bayesian rule $g(\mathbf{x}) = \frac{p(\mathbf{x})}{q(\mathbf{x})} \pi \approx \mathbb{P}(Y = 1 | X)$

$$p_{(\mathbf{x})} \approx p_{\theta}(\mathbf{x}) = \frac{e^{-g_{p\theta}(\mathbf{x})}}{Z_{p\theta}}, \quad q_{(\mathbf{x})} \approx q_{\theta}(\mathbf{x}) = \frac{e^{-g_{q\theta}(\mathbf{x})}}{Z_{q\theta}}$$

where $g_{p\theta}$ and $g_{q\theta}$ represent two parameterized neural networks configured with optimal θ respectively, and $Z_{p\theta}$, $Z_{q\theta}$ represent the partition functions:

$$Z_{p\theta} = \int e^{-g_{p\theta}(\mathbf{x})} d\mathbf{x}$$

$$Z_{q\theta} = \int e^{-g_{q\theta}(\mathbf{x})} d\mathbf{x}$$

The classifier is then rewritten as:

$$f(\mathbf{x}) = \frac{e^{-g_{p\theta}(\mathbf{x})}}{Z_{p\theta}} / \frac{e^{-g_{q\theta}(\mathbf{x})}}{Z_{q\theta}} \pi = e^{(g_{q\theta}(\mathbf{x}) - g_{p\theta}(\mathbf{x}))} \left(\frac{Z_{q\theta}}{Z_{p\theta}} \pi \right)$$

Parametric Density Estimation

$$f(\mathbf{x}) = \frac{e^{-g_{p_\theta}(\mathbf{x})}}{Z_{p_\theta}} / \frac{e^{-g_{q_\theta}(\mathbf{x})}}{Z_{q_\theta}} \pi = e^{(g_{q_\theta}(\mathbf{x}) - g_{p_\theta}(\mathbf{x}))} \left(\frac{Z_{q_\theta}}{Z_{p_\theta}} \pi \right)$$

The term $\left(\frac{Z_{q_\theta}}{Z_{p_\theta}} \pi \right)$ in the above is a constant for each \mathbf{x} and can be ignored in practice. Hence, the classifier can be approximated by the exponent:

$$f(\mathbf{x}) := g_{q_\theta}(\mathbf{x}) - g_{p_\theta}(\mathbf{x})$$

The standard maximum likelihood training algorithm with Markov Chain Monte Carlo (MCMC) sampling is employed to train the neural networks g_{p_θ} and g_{q_θ} :

- $\nabla_\theta \log p_\theta(\mathbf{x}) = -\nabla_\theta g_p(\mathbf{x}) - \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x})} [-\nabla_\theta g_p(\mathbf{x})]$
- $\mathbf{x}_{t+1} = \mathbf{x}_t + \frac{\varepsilon}{2} \nabla \log p_\theta(\mathbf{x}_t) + \mathcal{N}(0, \varepsilon)$

Parametric Density Estimation

- It should be noted that Langevin dynamics can be **unreliable in high-intensity areas** for high-dimensional datasets, which leads to low-performance models.
- To address this concern, we add a normal PU loss component in the early stages of training. The total loss function is defined as:

$$\alpha(\nabla_{\theta} \log p_{\theta}(\mathbf{x})) + \beta(\nabla_{\theta} \log q_{\theta}(\mathbf{x})) + \gamma(R_{\ell_{0-1}}(f(\mathbf{x})))$$

where α , β and γ are coefficients. The value of γ decreases as training progresses.

Experiment - Setting

Dataset

- PubMed datasets with 3 fine-grained topics, generated by Jacovi et al. (2021) for the DSE task
- A single dataset used for Covid-19 study classification Shemilt et al.
- Training data = Labeled data + Unlabeled data
- Validation and test data = Unlabeled data

dataset	LP	N_U	N_{UP}	N_{UN}
Pubmed-topic1	20	10012	1844	8168
	50	10027	2568	7459
Pubmed-topic2	20	10012	2881	7131
	50	10027	3001	7026
Pubmed-topic3	20	7198	1201	5997
	50	10025	1916	8109
Covid	{47..4722}	4722	2310	2412

Alon Jacovi, Gang Niu, Yoav Goldberg, and Masashi Sugiyama. 2021. Scalable Evaluation and Improvement of Document Set Expansion via Neural Positive-Unlabeled Learning. *In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 581–592, Online. Association for Computational Linguistics.

Shemilt, Ian, et al. "Machine learning reduced workload for the Cochrane COVID-19 Study Register: development and evaluation of the Cochrane COVID-19 Study Classifier." *Systematic Reviews* 11.1 (2022): 15.

Experiment

Baseline

- nnPU (5-layer neural network with layers sized 768, 512, 256, 128, 64, and an output layer with 1)
- vPU (5-layer neural network with layers sized 768, 512, 256, 128, 64, and an output layer with 1)
- BM25
- puDE-kde (VAE with 256 hidden dimensions and 50 latent dimensions)
- puDE-em (5-layer neural network with layers sized 768, 512, 256, 128, 64, and an output layer with 1)

Experiment – Result On PubMed datasets

Performance of nnPU is much worse than that reported by [Jacovi et al. \(2021\)](#) and is similar to BM25, which indicate that the PU solutions proposed in ([Jacovi et al., 2021](#)) is not as effective as they stated for the DSE task in transductive setting.

Both puDE methods outperform other methods, with one exception where BM25 get the best result on the last topic. It should be noticed that result reported for BM25 is the average across 5000-|LP| cases, which is not the direct classification result, and it serves as references to the state-of-the-art ([Jacovi et al., 2021](#))

Topic	LP	BM25	nnPU	VPU	puDE- <i>kde</i>	puDE- <i>em</i>
Animals+Brain+Rats	20	32.25 ± 11.6	33.03	25.62	37.31	40.59
	50	32.80 ± 10.9	38.76	29.32	44.65	44.91
Adult+Middle Aged +HIV infections	20	26.75 ± 7.22	31.30	29.77	36.18	39.67
	50	31.85 ± 10.7	34.16	31.42	44.03	46.22
Renal Dialysis + Chronic Kidney Failure+ Middle Aged	20	41.23 ± 8.95	27.76	21.59	36.63	35.59
	50	35.78 ± 9.13	32.84	19.42	36.63	36.57

Table 2: F1 comparison against baseline and state-of-the-art DES methods with transductive setting.

Experiment – Result on Covid Dataset

- nnPU and vPU get stable results only when more than 20% of labelled data is available
- puDE methods perform well with less data (<10%) and consistently shown significant improvements over other methods with the increase of labelled data
- When the number of labelled positives is small, the performance of vPU is poor since its training strategy needs equal batch size of unlabeled (U) and labelled (LP) samples feeding into the model to calculate the variational loss

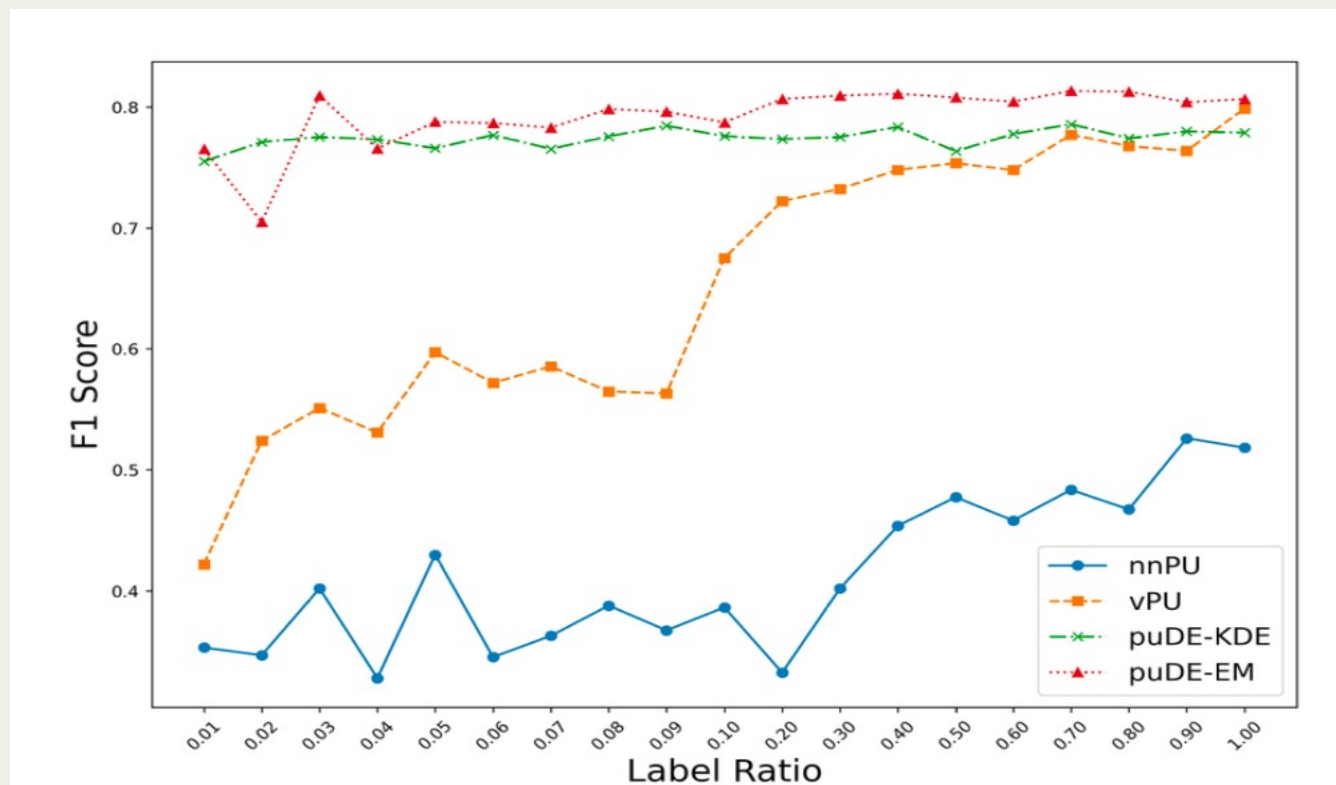


Figure 2: F1 comparison on covid dataset with respect to the ratio of $|LP|$ over $|U|$ ranging from 0.01 to 0.1 with step of 0.01 and from 0.1 to 1 with step of 0.1.

method	P@10%	P@20%	R@10%	R@20%
BM25	54.66	52.64	11.16	21.51
nnPU	52.54	67.16	10.74	27.45
VPU	56.77	57.30	11.90	23.41
puDE-kde	70.26	72.88	16.91	28.67
puDE-em	76.91	75.11	15.71	30.69

Table 3: Performance comparison for ranking task on Covid dataset with $|LP| = 50$.

Conclusion and Acknowledgements

- The transductive setting addresses the limitations of previous Positive-Unlabeled (PU)-based approaches in solving the Document Set Expansion (DSE) task (Jacovi et al., 2021)
- We propose a novel PU learning framework based on intractable density estimation methods
- Experimental results validate the effectiveness of our proposed methods

We would like to express our gratitude for the support provided by the XJTLU AI University Research Centre and the Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU. Additionally, we acknowledge the support from the SIP AI Innovation Platform (YZCXPT2022103) and the Research Development Funding (RDF) at Xi'an Jiaotong-Liverpool University, under contract numbers RDF-21-02-044 and RDF-21-02-008.

Reference

- Chen, H., Liu, F., Wang, Y., Zhao, L. and Wu, H., 2020. A variational approach for learning from positive and unlabeled data. *Advances in Neural Information Processing Systems*, 33, pp.14844-14854.
- Kato, M., Teshima, T. and Honda, J., 2018, September. Learning from positive and unlabeled data with a selection bias. In *International conference on learning representations*.
- Nakajima, S. and Sugiyama, M., 2023. Positive-unlabeled classification under class-prior shift: a prior-invariant approach based on density ratio estimation. *Machine Learning*, 112(3), pp.889-919.
- Jiang, L., Li, D., Wang, Q., Wang, S. and Wang, S., 2020. Improving positive unlabeled learning: Practical aul estimation and new training method for extremely imbalanced data sets. *arXiv preprint arXiv:2004.09820*.
- Du Plessis, M.C., Niu, G. and Sugiyama, M., 2014. Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems*, 27.
- Du Plessis, M., Niu, G. and Sugiyama, M., 2015, June. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning* (pp. 1386-1394). PMLR.
- Bekker, J. and Davis, J., 2020. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109, pp.719-760.
- Alon Jacovi, Gang Niu, Yoav Goldberg, and Masashi Sugiyama. 2021. Scalable Evaluation and Improvement of Document Set Expansion via Neural Positive-Unlabeled Learning. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 581–592, Online. Association for Computational Linguistics.
- Shemilt, Ian, et al. "Machine learning reduced workload for the Cochrane COVID-19 Study Register: development and evaluation of the Cochrane COVID-19 Study Classifier." *Systematic Reviews* 11.1 (2022): 15.
- Geng, C., Wang, J., Gao, Z., Frellsen, J. and Hauberg, S., 2021. Bounds all around: training energy-based models with bidirectional bounds. *Advances in Neural Information Processing Systems*, 34, pp.19808-19821.
- Song, Y. and Kingma, D.P., 2021. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*.

Thank you!
