Document Set Expansion with Positive-Unlabelled Learning Using Intractable Density Estimation

Haiyang Zhang[1*], **Qiuyi Chen**[1*], **Yuanjie Zou**[1]
Yushan Pan[1], **Jia Wang**[1], **Mark Stevenson**[2]
[1]Xi'an Jiaotong Liverpool University
[2]University of Sheffield
{haiyang.zhang, yushan.pan, Jia.wang02}@xjtlu.edu.cn
{qiuyi.chen2002, yuanjie.zou}@student.xjtlu.edu.cn
mark.stevenson@sheffield.ac.uk

Abstract

The Document Set Expansion (DSE) task involves identifying relevant documents from large collections based on a limited set of example documents. Previous research has highlighted Positive and Unlabeled (PU) learning as a promising approach for this task. However, most PU methods rely on the unrealistic assumption of knowing the class prior for positive samples in the collection. To address this limitation, this paper introduces a novel PU learning framework that utilizes intractable density estimation models. Experiments conducted on PubMed and Covid datasets in a transductive setting showcase the effectiveness of the proposed method for DSE. Code is available from https://github.com/Beautifuldog01/Document-set-expansion-puDE.

Keywords: Document set expansion, PU learning, Information retrieval, Density estimation.

1. Introduction

We focus on the scenario where a user has access to a (possibly small) set of documents of interest and wishes to identify further such documents within a large collections, a problem known as Document Set Expansion (DSE) (Jacovi et al., 2021; Lee and Sun, 2018; Wang et al., 2022b). DSE is a common information seeking problem, for example when searching scientific literature for papers that are similar to a small set of relevant 'seed' publications (Wang et al., 2022a). It can also occur in the maintenance of curated databases of scientific literature where examples of relevant studies are readily available but there may not be an explicit query (Chen et al., 2021).

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 (Abolghasemi et al., 2022; Lee and Sun, 2018; Yang et al., 2009). A common QBD approach focuses on constructing an accurate query from the seed documents, using methods such as bag-of-word (Yang et al., 2009) or Monte-Carlo (MC) sampling procedure (Lechtenberg et al., 2022). However, such methods fail to capture the local or global connections between words (Jacovi et al., 2021). More recent work fine-tuned a BERT re-ranker for the QBD retrieval task (Abolghasemi et al., 2022), but requires a fully labelled dataset to train the neural models. In addition, the majority of the QBD approaches only work with a single seed document (Abolghasemi et al., 2022; Lee and Sun, 2018).

Jacovi et al. (2021) treated the DSE task as a positive and unlabelled (PU) learning problem by learning a binary classifier using only positive and unlabelled data (Bekker and Davis, 2020; Plessis et al., 2015; Kiryo et al., 2017). They introduce a new PU method based on Non-negative PU (nnPU) (Kiryo et al., 2017), and show that their methods can outperform common information retrieval (IR) solutions for the DSE task. However, some important issues remain unresolved:

- PU methods that rely on misclassification risk, such as nnPU, assumes that the class prior, $\pi = P(Y = 1)$, is known. The class prior denotes the proportion of positive data in the unlabelled data and plays an important role in PU learning. However, in practical applications, π is usually unknown and it cannot be treated as a trainable parameter (Chen et al., 2020). Several studies propose to estimate the class prior as an intermediate step for PU classification (Christoffel et al., 2016; Chang et al., 2020). However, such methods commonly utilize complex kernel machines. Moreover, inaccurate estimation may bring more errors in the PU classification (Chen et al., 2020).
- DSE is essentially a transductive problem since we aim to identify all positive documents from the unlabelled set (U). In such a case, the unlabelled set should be used for both training

* denotes equal contribution.

LREC-COLING 2024, pages 5167–5173
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

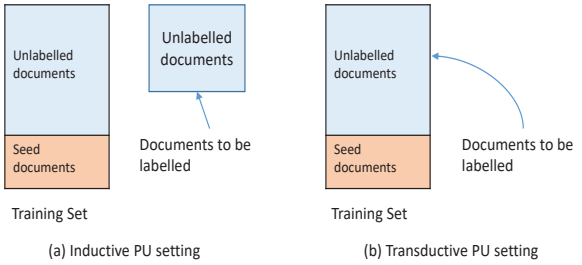

Figure 1: Inductive setting and transductive setting in PU learning.

and testing. However, 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. Such experimental settings cannot reflect the ground truth performance of the model for the DSE task. The difference between the two settings can be found in Figure 1.

To address these issues, we propose a novel PU learning framework based on intractable models, which does not require a known class prior. It aims to learn a bayesian binary classifier by merely making use of the distribution of labelled and unlabelled data, without class prior involved in training. Intractable models, i.e. Kernel Density Estimation (KDE) (Wang et al., 2023) and Energy-based model (EBM) (LeCun et al., 2006) are used to estimate the density, as they do not restrict to the tractability of the normalizing constant (Zhai et al., 2016). Consequently, it does not require assumptions on the form of data distribution to be fitted. Experiments are conducted in a transductive setting to better reflect the DSE task.

The contributions of this work are: 1) to 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.

2. Background

In the binary classification task, given $\mathbf{x} \in \mathcal{R}^d$ as the input instance and $Y \in \{1, -1\}$ as the label of $\mathbf{x}$, the goal is to learn a decision function $\Phi : X \to Y$ that can separate the positive and negative examples. In order to facilitate the training of an accurate classifier, it is assumed that the training data represents an independent and identically distributed sample of the actual underlying distribution: $\mathbb{P}(\mathbf{x}) = \pi\mathbb{P}(\mathbf{x}|Y = 1) + (1 - \pi)\mathbb{P}(\mathbf{x}|Y = -1)$ where $\pi = P(Y = 1)$ is the class prior.

The setting of PU is a special scenario within the binary classification problem, where only a small portion of positive examples are observed (Bekker and Davis, 2020). The training set is a combination of the labelled positive set X_{LP}, the unlabelled positive set X_{UP}, and the unlabelled negative set X_{UN}, such that $X = X_{LP} \cup X_{UP} \cup X_{UN}$. Let $s \in \{1, 0\}$ present the label status of y ($s = 1$ if labelled, otherwise $s = 0$), there will be:

$$X_{LP} = \{\mathbf{x}|s = 1, Y = 1\}, \quad X_{UP} = \{\mathbf{x}|s = 0, Y = 1\}$$
$$X_{UN} = \{\mathbf{x}|s = 0, Y = -1\}, \quad X_U = \{\mathbf{x}|s = 0\}$$

The label frequency can be represented as $c = P(s = 1|Y = 1)$ (Elkan and Noto, 2008). In PU scenario, X_{LP} are selected from a completed set of positive examples X_P under certain probabilistic labeling mechanism, and the probability of an example being labelled is defined as $e(\mathbf{x}) = P(s = 1|\mathbf{x}, Y = 1)$, known as propensity score (Elkan and Noto, 2008). Hence, the distribution of labelled positives $f_{LP} \triangleq P(\mathbf{x}|s = 1, Y = 1)$ can be seen as a biased version of $f_P \triangleq P(\mathbf{x}|Y = 1)$:

$$f_{LP}(\mathbf{x}) = \frac{e(\mathbf{x})}{c} f_P(\mathbf{x}), \tag{1}$$

where $c = P(s = 1|Y = 1)$ is the label frequency. The goal of PU learn is to learn a binary classifier $\Phi : X \to Y$ that can separate the positive from unlabelled examples. In this work, our objective is to estimate an optimal Bayesian classifier under the following assumption:

Assumption 1 *The positive labelled data are randomly selected from the set of positive data and are identically distributed with the positive unlabelled data: $f_{LP}(\mathbf{x}) = f_P(\mathbf{x})$, which is known as the Selected Completely At Random (SCAR) assumption (Bekker and Davis, 2020).*

3. PU Learning with Intractable Models

We consider the following task: we have a set of labelled positive documents X_{LP} on a fine-grained topic and want to find more documents about that topic from a large unlabelled collection X_U. Given X_{LP} and X_U, the objective of our method is to learn a Bayesian classifier Φ to approximate $\mathbb{P}(Y = 1|\mathbf{x})$. According to the Bayesian rule, we have:

$$\mathbb{P}(Y = 1|\mathbf{x}) = \frac{\mathbb{P}(\mathbf{x}|Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(\mathbf{x})} = \frac{f_p(\mathbf{x})}{f(\mathbf{x})}\pi \tag{2}$$

where $f_p(\mathbf{x})$ is the positive data distribution and $f(\mathbf{x})$ is the distribution of the whole dataset. It is intuitive to estimate the probability density of $f_p(\mathbf{x})$ and $f(\mathbf{x})$ respectively, so that π can be treated as a constant for each $\mathbf{x}$ and not involved in the training process. In such a way, we can learn a classifier

without the need for class prior estimation which is an intermediate step for the PU classification task (Chang et al., 2020).

Let $p_\theta(\mathbf{x}) : \mathcal{R}^d \to [0,1]$ and $q_\theta(\mathbf{x}) : \mathcal{R}^d \to [0,1]$ be the two models to estimate $f_p(\mathbf{x})$ and $f(\mathbf{x})$, $\mathbb{P}(Y = 1|\mathbf{x})$ can be then approximated by:

$$\mathbb{P}(Y = 1|\mathbf{x}) \approx \Phi(\mathbf{x}) = \frac{p_\theta(\mathbf{x})}{q_\theta(\mathbf{x})}\pi \tag{3}$$

Under Assumption 1, i.e. $f_{LP}(\mathbf{x}) = f_P(\mathbf{x})$, we can estimate f_P using samples from X_{LP}. In this paper, we try to make less restriction on the underlying distribution on the data we fit. Therefore, intractable density estimation methods in both nonparametric and parametric forms are adopted.

3.1. Nonparametric Density Estimation

Kernel Density Estimation (KDE) is a nonparametric density estimation technique, which has been applied in recommender systems and information retrieval (Silverman, 2018; Chakraborty et al., 2022). For a given dataset $\{\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_n\}$, the estimated density $\hat{f}$ at $\mathbf{x}$ using KDE is defined as:

$$\widehat{f}_{kde}(\mathbf{x}) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \tag{4}$$

where h is the bandwidth hyperparameter, and K is a non-negative kernel function. In the DSE task, given a set of labelled documents X_{LP}, $\Phi(\mathbf{x})$ estimated with KDE is represented as follows:

$$\Phi(\mathbf{x}) = \frac{p_\theta(\mathbf{x})}{q_\theta(\mathbf{x})}\pi = \frac{\widehat{f}_{p,kde}(\mathbf{x})}{\widehat{f}_{kde}(\mathbf{x})}\pi \tag{5}$$

where $\widehat{f}_{p,kde}$ is the estimated density of positive data which can be estimated by samples from X_{LP}, and $\widehat{f}_{kde}$ is the estimated density of the whole data X. Gaussian density function $K(\mathbf{x}) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\mathbf{x}^2}$ is used as the kernel function.

3.2. Parametric Density Estimation

The parametric approach used to estimate the density is the energy-based model (EBM) (LeCun et al., 2006), which is a powerful tool for representing complex high-dimensional data distributions. It aims to learn an energy function that assigns a low energy value to observed data and a high energy value to different values. Compared with other parametric density estimation methods, such as VAE (Kingma and Welling, 2013) and Masked Autoregressive Density Estimators (MADE) (Papamakarios et al., 2017), EBM does not make any assumption on the form of data distribution they fit. An EBM parameterizes any probability density for $\mathbf{x} \in \mathbb{R}^d$ as:

$$f_{EBM,\theta}(\mathbf{x}) = \frac{e^{-E_\theta(\mathbf{x})}}{Z_\theta} \qquad Z_\theta = \int e^{-E_\theta(\mathbf{x})}dx \tag{6}$$

where $E_\theta(\mathbf{x})$ is the energy function, which is a nonlinear regression function configured with optimal θ, and Z_θ is the partition function, which is a function of θ but is a constant with respect to $\mathbf{x}$. For the DSE task, we use two neural networks (g_{p_θ} and g_{q_θ}) as the energy function to estimate p_θ and q_θ respectively. Thus, $\Phi(\mathbf{x})$ is rewritten as:

$$\begin{aligned}\Phi(\mathbf{x}) = \frac{p_\theta(\mathbf{x})}{q_\theta(\mathbf{x})}\pi &= \frac{e^{-g_{p_\theta}(\mathbf{x})}}{Z_{p_\theta}} \Big/ \frac{e^{-g_{q_\theta}(\mathbf{x})}}{Z_{q_\theta}}\pi \\ &= e^{(g_{q_\theta}(\mathbf{x}) - g_{p_\theta}(\mathbf{x}))}\left(\frac{Z_{q_\theta}}{Z_{p_\theta}}\pi\right)\end{aligned} \tag{7}$$

where $\frac{Z_{q_\theta}}{Z_{p_\theta}}\pi$ is a constant for each $\mathbf{x}$ and can be ignored in practise. Hence, $\Phi(\mathbf{x})$ can be approximated by the exponent: $\Phi(\mathbf{x}) := g_{q_\theta}(\mathbf{x}) - g_{p_\theta}(\mathbf{x})$.

Model Training We employ the maximum likelihood training with Markov Chain Monte Carlo (MCMC) sampling to train the energy models, such that there will be no need to calculate the constant term $\frac{Z_{q_\theta}}{Z_{q_\theta}}\pi$ during the training process. With maximum likelihood estimation (MLE), we can fit p_θ to $f_{LP}(\mathbf{x})$ and q_θ to $f(\mathbf{x})$ by maximizing the following expected log-likelihood:

$$\mathrm{E}_{X_{LP}}\left[\log \mathrm{p}_\theta(\mathbf{x})\right] \qquad \mathrm{E}_X\left[\log \mathrm{q}_\theta(\mathbf{x})\right]$$

which are equivalent to minimizing the following KL divergence:

$$\arg\min_\theta \mathrm{KL}(f_{LP}(\mathbf{x})\|\mathrm{p}_\theta) \qquad \arg\min_\theta \mathrm{KL}(f(\mathbf{x})\|\mathrm{q}_\theta)$$

where $f_{LP}(\mathbf{x}) = f_P(\mathbf{x})$, and $f(\mathbf{x})$ is the real distribution of positive data and the whole data, respectively. The loss function to minimize is defined as:

$$\alpha\left(-\mathrm{E}_{X_{LP}}\left[\log \mathrm{p}_\theta(\mathbf{x})\right]\right) + \beta\left(-\mathrm{E}_X\left[\log \mathrm{q}_\theta(\mathbf{x})\right]\right) \tag{8}$$

where α and β are coefficients. By using the MCMC sampling approach, the gradient of the log-likelihood of $p_\theta(\mathbf{x})$ and $q_\theta(\mathbf{x})$ are defined as:

$$\begin{aligned}\nabla_\theta \log p_\theta(\mathbf{x}) &= -\nabla_\theta g_{p_\theta}(\mathbf{x}) - \mathrm{E}_{\mathbf{x}\sim p_\theta(\mathbf{x})}\left[-\nabla_\theta g_{p_\theta}(\mathbf{x})\right] \\ \nabla_\theta \log q_\theta(\mathbf{x}) &= -\nabla_\theta g_{q_\theta}(\mathbf{x}) - \mathrm{E}_{\mathbf{x}\sim q_\theta(\mathbf{x})}\left[-\nabla_\theta g_{q_\theta}(\mathbf{x})\right]\end{aligned}$$

The first terms in both equations above are straightforward to obtain. To approximating the second terms, Langevin Dynamics is used to sample from $p_\theta(\mathbf{x})$ and $q_\theta(\mathbf{x})$:

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{x}_t + \frac{\varepsilon}{2}\nabla \log p_\theta(\mathbf{x}_t) + \mathcal{N}(0, \varepsilon) \\ \mathbf{x}_{t+1} &= \mathbf{x}_t + \frac{\varepsilon}{2}\nabla \log q_\theta(\mathbf{x}_t) + \mathcal{N}(0, \varepsilon)\end{aligned}$$

where t denotes the iteration step, $\mathcal{N}$ is the normal distribution. Since Langevin dynamics can be unreliable in high-intensity areas for high-dimensional datasets, which will effect the model performance.

To address this issue, we add a risk estimator in the loss function:

$$\alpha\left(-\mathrm{E}_{X_{LP}}\left[\log \mathrm{p}_\theta(\mathbf{x})\right]\right)+\beta\left(-\mathrm{E}_{X}\left[\log \mathrm{q}_\theta(\mathbf{x})\right]\right) +\gamma\left(R_{\ell_{0-1}}(\Phi(\mathbf{x}), s)\right) \quad (9)$$

where γ is a coefficient and it decreases as training progresses, $R_{\ell_{0-1}}(\Phi(\mathbf{x}), s)$ represents the loss generated by binary classification using 's' as the label.

4. Experiment

4.1. Dataset

Experiments use PubMed datasets on three fine-grained topics derived from Jacovi et al. (2021). Additionally, we use the Covid-19 dataset that is used for Covid-19 literature classification (Shemilt et al., 2022) to simulate real-world literature curation. All datasets were originally designed for inductive classification, where each dataset is split into training, validation, and testing sets. In our experiments, to simulate real-world DSE (transductive case), we treat the test set in original data slipt settings as X_U and use X_U for both training and testing (X_{LP} and X_U for training and X_U for testing). Following Jacovi et al. (2021), the number of labelled positives |LP| is set to $\{20, 50\}$ on Pubmed datasets. For the Covid-19 dataset, the labelled positives are randomly sampled from their original positive training set, and the number of $|LP|$ is set with respect to the ratio of X_{LP} over X_U, ranging from 0.01 to 0.1 with step of 0.01, and from 0.1 to 1 with step 0.1. The statistics of each set is summarized in Table 1.

4.2. Comparison Methods

The performance of puDE-*kde* and puDE-*em* are compared with the following methods:

- **BM25** BM25 (Robertson et al., 1995) serves as a strong baseline in various IR tasks. In our paper, following the method in Jacovi et al. (2021), we vary the number of top documents (K) to be considered as positive examples, $K \in \{i\}_{|LP|}^{5000}$, and report the F1 mean and standard deviation across the $5000 - |LP|$ cases.
- **nnPU** nnPU (Kiryo et al., 2017) is a recent PU method that are based on unbiased risk estimators. It is used as the baseline in various PU studies, and is the first method being used for DSE task (Jacovi et al., 2021).
- **VPU** VPU (Chen et al., 2020) is the a state-of-the-art PU method that do not require knowledge of class prior. It uses a variational principle to modeling the error of the Bayesian classifier directly from the provided data.

PU classifiers, i.e. nnPU and VPU, were implement them in transductive fashion to complement the DSE task.

4.3. Settings

We use puDE-*kde* and puDE-*em* to denote our proposed PU models that is based on KDE and EBM, respectively. For puDE-*kde*, the bandwidth is set to 1.9 for both $\widehat{f}_{p,kde}$ and $\widehat{f}_{kde}$, and Gaussian function is used as the kernel. Since KDE suffers from the curse of dimensionality, we use Variational Autoencoders (VAE) (Girin et al., 2020) with 50 latent dimensions to reduce the high text dimension in this work. For puDE-*em*, we use 512D 4-layer fully connected neural network as the energy function for g_{p_θ} and g_{q_θ}. The weights for the total loss function are set as $\alpha = 1$, $\beta = 1$, and $\gamma = 1$.

nnPU is implemented using the tricks from Jacovi et al. (2021) but in transductive version. For both nnPU and VPU, the classifiers are modeled by 512D 4-layer fully connected neural network, with Batch normalization (Ioffe and Szegedy, 2015) and leaky ReLU (Maas et al., 2013) applied. Adam optimizer with a learning rate of 1e-3 is employed. For all methods, SciBERT is used as the pre-trained embedding. Following Jacovi et al. (2021), F1 score is used as the evaluation metric.

4.4. Results

The F1 results across all methods are reported in Table 2, where the best performance are shown in bold font. 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). Both puDE methods show significant improvement over nnPU and VPU, demonstrating that the proposed PU framework based on density estimation is a better alternative for the DSE task.

Figure 1 demonstrates the F1 score for all methods on Covid-19 dataset, with the ratio of |LP| over |U| ranging from 0.01 to 1. It can be seen that nnpu and VPU get stable results only when more than 20% of labelled data is available. Both puDE methods perform well with less data (<10%) and consistently shown significant improvements over other methods with the increase of labelled data. As a state-of-the-art PU method, VPU can get similar performance with our models when the label

dataset	|LP|	N_U	N_{UP}	N_{UN}
Animals+Brain+Rats	20	10012	1844	8168
	50	10027	2568	7459
Adult+Middle Aged	20	10012	2881	7131
+HIV infections	50	10027	3001	7026
Renal Dialysis + Chronic	20	7198	1201	5997
Kidney Failure+ Middle Aged	50	10025	1916	8109
Covid	$\{47..4722\}$	4722	2310	2412

Table 1: Statistics of X_U for each set, where N_U, N_{UP} and N_{UN}, the total number of unlabelled samples, the number of true positive samples and true negatives in the training set.

Topic	|LP|	BM25	nnPU	VPU	puDE-*kde*	puDE-*em*
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	20	26.75 ± 7.22	31.30	29.77	36.18	**39.67**
+HIV infections	50	31.85 ± 10.7	34.16	31.42	44.03	**46.22**
Renal Dialysis + Chronic	20	$\mathbf{41.23 \pm 8.95}$	27.76	21.59	36.63	35.59
Kidney Failure+ Middle Aged	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.

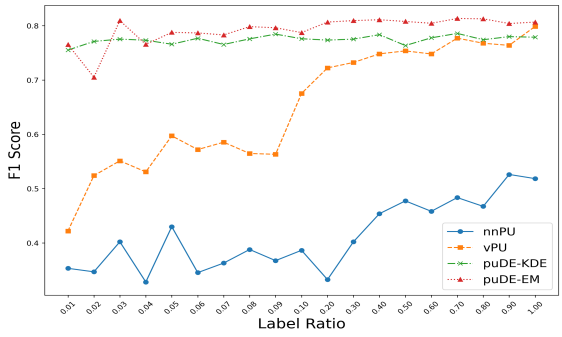

Figure 2: F1 comparison on covid dataset with respect to the ratio of |LP| over |U| ranging from 0.01 to 0.1 with step of 0.01 and from 0.1 to 1 with step of 0.1.

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.

ratio goes over 0.5. However, when the the number of labelled positives is small, the performance is poor. This is due to its training strategy, where equal batch size of unlabelled (U) and labelled (LP) samples are fed into the model to calculate the variational loss. When $|LP| \ll |U|$, the distribution of LP may be different from that of U. Replicating LP, until the size of LP equals the size of U, can lead to instability in the model, making it difficult for the model to converge during training and resulting in poor prediction performance.

We further conduct experiments for ranking task on Covid dataset to simulate screening process in literature curation. The experiment settings are the same as previous ones except that ranking-based evaluation metrics for systematic reviews (Kanoulas et al., 2019; Wang et al., 2023) are adopted. Precision at top k% documents in U (p@k%), and recall at top k% documents in U (r@k%) are reported. Table 3 shows the ranking effectiveness of all methods with the number of labelled documents equals 50. It can be seen that our methods produces the best overall performance.

5. Conclusion

This paper addresses the limitations of previous Positive-Unlabeled (PU)-based approaches in solving the Document Set Expansion (DSE) task (Jacovi et al., 2021). It demonstrates that experimental results obtained from an inductive setting cannot be directly transferred to a real-world transductive DSE scenario. To overcome these challenges, we propose a novel PU learning framework based on intractable density estimation methods. A key advantage of our approach is that it does not rely on prior knowledge of class proportions. Experimental results validate the effectiveness of our proposed methods. In conclusion, we assert that our approach represents a superior solution for the DSE task compared to existing methods.

Acknowledgements

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.

References

Amin Abolghasemi, Suzan Verberne, and Leif Azzopardi. 2022. Improving bert-based query-by-document retrieval with multi-task optimization. In *European Conference on Information Retrieval*, pages 3–12. Springer.

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

Anirban Chakraborty, Debasis Ganguly, Annalina Caputo, and Gareth JF Jones. 2022. Kernel density estimation based factored relevance model for multi-contextual point-of-interest recommendation. *Information Retrieval Journal*, 25(1):44–90.

Shizhen Chang, Bo Du, and Liangpei Zhang. 2020. Positive unlabeled learning with class-prior approximation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2014–2021. International Joint Conferences on Artificial Intelligence Organization. Main track.

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.

Qingyu Chen, Alexis Allot, and Zhiyong Lu. 2021. Litcovid: an open database of covid-19 literature. *Nucleic acids research*, 49(D1):D1534–D1540.

Marthinus Christoffel, Gang Niu, and Masashi Sugiyama. 2016. Class-prior estimation for learning from positive and unlabeled data. In *Asian Conference on Machine Learning*, pages 221–236. PMLR.

Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 213–220, New York, NY, USA. Association for Computing Machinery.

Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. 2020. Dynamical variational autoencoders: A comprehensive review. *arXiv preprint arXiv:2008.12595*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr.

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.

Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2019. Clef 2019 technology assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, volume 2380, page 250.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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.

Fabian Lechtenberg, Javier Farreres, Aldwin-Lois Galvan-Cara, Ana Somoza-Tornos, Antonio Espuña, and Moisès Graells. 2022. Information retrieval from scientific abstract and citation databases: A query-by-documents approach based on monte-carlo sampling. *Expert Systems with Applications*, 199:116967.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Grace E. Lee and Aixin Sun. 2018. Seed-driven document ranking for systematic reviews in evidence-based medicine. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 455–464, New York, NY, USA. Association for Computing Machinery.

Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, Georgia, USA.

George Papamakarios, Theo Pavlakou, and Iain Murray. 2017. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30.

Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. 2015. Convex formulation for learning from positive and unlabeled data. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1386–1394, Lille, France. PMLR.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Ian Shemilt, Anna Noel-Storr, James Thomas, Robin Featherstone, and Chris Mavergames. 2022. 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):1–8.

B.W. Silverman. 2018. *Density Estimation for Statistics and Data Analysis*. Routledge.

Shuai Wang, Harrisen Scells, Justin Clark, Bevan Koopman, and Guido Zuccon. 2022a. From little things big things grow: A collection with seed studies for medical systematic review literature search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3176–3186, New York, NY, USA. Association for Computing Machinery.

Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2023. Neural rankers for effective screening prioritisation in medical systematic review literature search. In *Proceedings of the 26th Australasian Document Computing Symposium*, ADCS '22, New York, NY, USA. Association for Computing Machinery.

Shuai Wang, Harrisen Scells, Ahmed Mourad, and Guido Zuccon. 2022b. Seed-driven document ranking for systematic reviews: A reproducibility study. In *Advances in Information Retrieval*, pages 686–700, Cham. Springer International Publishing.

Yin Yang, Nilesh Bansal, Wisam Dakka, Panagiotis Ipeirotis, Nick Koudas, and Dimitris Papadias. 2009. Query by document. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 34–43.

Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. 2016. Deep structured energy based models for anomaly detection. In *International conference on machine learning*, pages 1100–1109. PMLR.