JMLR: Workshop and Conference Proceedings 25:555–570, 2012 Asian Conference on Machine Learning

Multi-view Positive and Unlabeled Learning

Joey Tianyi Zhou TZHOU1@NTU.EDU.SG
School of Computer Engineering, Nanyang Technological University, Singapore

Sinno Jialin Pan JSPAN@I2R.A-STAR.EDU.SG
Institute for Infocomm Research, Singapore

Qi Mao QMAO1@NTU.EDU.SG
School of Computer Engineering, Nanyang Technological University, Singapore

Ivor W. Tsang IVORTSANG@NTU.EDU.SG
School of Computer Engineering, Nanyang Technological University, Singapore

Editor: Steven C.H. Hoi and Wray Buntine

Abstract

Learning with Positive and Unlabeled instances (PU learning) arises widely in information retrieval applications. To address the unavailability issue of negative instances, most existing PU learning approaches require to either identify a reliable set of negative instances from the unlabeled data or estimate probability densities as an intermediate step. However, inaccurate negative-instance identification or poor density estimation may severely degrade overall performance of the final predictive model. To this end, we propose a novel PU learning method based on density ratio estimation without constructing any sets of negative instances or estimating any intermediate densities. To further boost PU learning performance, we extend our proposed learning method in a multi-view manner by utilizing multiple heterogeneous sources. Extensive experimental studies demonstrate the effectiveness of our proposed methods, especially when positive labeled data are limited.

Keywords: PU learning, density ratio estimation, multi-view learning

1. Introduction

Learning with Positive and Unlabeled instances (PU learning) has attracted a great deal of attentions in the machine learning and data mining literatures (Denis, 1998; Liu et al., 2003; Li et al., 2009; Zhao et al., 2011; Nguyen et al., 2012; Zhang and Lee, 2008). In PU learning, the training data are composed of a set of positive data and a large amount of unlabeled data which can be positive or negative. This makes PU learning different from supervised learning and semi-supervised learning where both positive and negative data are required as inputs, and also different from unsupervised learning where only unlabeled data are available. Furthermore, in contrast to one-class classification (Schölkopf et al., 2001) where only positive data are used for training, PU learning assumes that a large amount of unlabeled data are available, and aims to fully exploit the unlabeled data together with the limited positive data to learn more precise predictive models.

Many real-world applications can be considered as PU learning tasks. For instance, in a book recommender system, each user can bookmark a set of items which can be regarded as *positive* instances. The items which a user does not bookmark are referred to as *unlabeled* instances. However, the unlabeled instances may be *implicit positive* instances which are of

users' interest but have not been read, or *negative* instances which have been read but are not of users' interest (Pan et al., 2008). Another example is information retrieval based on clickthrough data of search engines. For a search engine, a user may submit a query and click some webpages returned by the search engine. The webpages clicked by the user can be regarded as *positive* instances with respect to the query. However, for those webpages the user does not click, it is hard to decide whether they are irrelevant (i.e., *negative* instances) or relevant but are not noticed by the user (i.e., *implicit positive* instances). In this case, the explicit positive data may be extremely sparse and how to leverage the unlabeled data to improve learning performance is crucial to PU learning.

One prominent solution of PU learning is a two-step approach, proposed by Liu et al. (2002), which first identifies a certain reliable negative instances from the unlabeled data. After that traditional classification methods, such as Naive Bayes (NB) (Mitchell, 1997) or Support Vector Machines (SVMs) (Cristianini and Shawe-Taylor, 2000), can be directly applied on the positive and identified negative instances to train predictive models. To the extreme, all the unlabeled instances are treated as negative instances with different misclassification cost of the positive and negative instances respectively (Liu et al., 2003). Although this approach can efficiently reuse existing classification methods for PU learning, its performance highly depends on the quality of the identified negative instances. Apart from this approach, another family of PU learning methods is to estimate the conditional probability of the positive class given inputs (i.e., feature vectors) directly, which is further used for making predictions (Elkan and Noto, 2008; Zhang, 2005). In these methods, some probability density functions need to be estimated as an intermediate step. For instance, in (Elkan and Noto, 2008), both the marginal probability of the positive class and the conditional probability of the labeled positive instance need to be estimated. In (Zhang, 2005), the probabilities of an instance being labeled or unlabeled, and a positive instance being labeled are required to be estimated.

However, estimation of conditional probability densities is still very challenging, especially with only limited positive labeled data. In this work, based on density ratio estimation (Sugiyama et al., 2012), we propose a new PU learning method named *Density-Ratio-based PU learning (DRPU)*, which avoids estimating densities separately. More specifically, we first model the problem of PU learning as estimation of a density ratio between the positive instances and all training instances (i.e., including both positive and unlabeled data). After that, the estimated density ratio can be considered as a ranking function to make predictions on unseen data. Compared to previous PU learning methods, our proposed method can handle both discrete and continuous feature values naturally, and can be applied to high dimensional data effectively.

Besides leveraging the unlabeled data, there exists other auxiliary information that can be exploited to address the data sparsity issue in PU learning, and thus further improve the prediction performance of DRPU. Here, we also use information retrieval as a motivating example. In general, a webpage can be represented by its content in text. Alternatively, it can also be represented by its hyperlinks, or its content in images or videos. In the machine learning community, incorporating multiple *view* information to improve learning performance is called *multi-view learning*. To incorporate multi-view information into PU learning, we extend DRPU by adding a co-regularizer on multiple views, such that the results of the predictive functions learned from different views on the same instance tend

to be the same. To the best of our knowledge, there is only one method on multi-view PU learning, namely PNCT (Denis et al., 2003). Its basic idea is to combine co-training (Blum and Mitchell, 1998) with a NB-based PU learning classifier for multi-view PU learning. However, their proposed method requires prior knowledge on the probability of the positive class, and also requires a feature discretization preprocessing to deal with continuous data, which may potentially discard some discriminative information.

We conduct extensive experiments on both toy and real-world datasets to verify the effectiveness of our proposed DRPU and its multi-view extension. Experimental results show that, when positive label ratio is small, DRPU outperforms state-of-the-art PU learning methods considered in the present paper. The two extended versions of DRPU in a multi-view manner further improve the prediction performance, and performs much better than PNCT. In summary, the main contributions of this paper are listed as follows:

1. We propose a new method for PU learning based on density ratio estimation. Compared to previous methods, our proposed method does not require to identify negative instances from the unlabeled data iteratively, and avoids density estimation as an intermediate step. Furthermore, our proposed method is flexible to be applied to both discrete and continuous data, and works effectively on high-dimensional feature space.

2. We integrate the density ratio estimation techniques into a co-regularization framework for multi-view PU learning. Compared to single-view PU learning methods, our proposed multi-view PU learning methods can fully exploit the multi-view information through a co-regularization term to boost the PU learning performance.

The rest of this paper is organized as follows. We begin by reviewing some related works in Section 2. We present our proposed PU learning method based on density ratio estimation in Section 3. After that, we extend our PU learning method in a multi-view manner in Section 4. Extensive experiments are conducted in Section 5. Finally, we conclude this work in Section 6.

2. Related Work

In the past decade, PU learning has been explored widely in the literature (Denis, 1998; Denis et al., 2002; Liu et al., 2002, 2003; Li and Liu, 2003; Zhang, 2005; Elkan and Noto, 2008). PU learning methods can be categorized into two approaches: negative set construction and density-estimation-based approaches. For the former category, the first step is to identify a set of reliable negative instances from the unlabeled instances by using various techniques, such as the Expectation Maximization (EM) technique (Liu et al., 2002), the Rocchio algorithm (Li and Liu, 2003), and NB (Liu et al., 2003). In the second step, various classifiers are built from the positive and the identified negative instances, such as a NB classifier with the EM algorithm (Liu et al., 2002) and SVMs (Li and Liu, 2003). One drawback of this category is that if the identification step of negative instances is inaccurate, the error will propagate towards the further classification step.

For the latter category, the goal is to estimate the conditional probability of the positive class given inputs to make predictions. Most existing methods are based on estimation of some probability distributions to obtain the target conditional probability. The method proposed by Elkan and Noto (2008) requires to estimate the marginal probability of the positive class and the conditional probability that a positive instance is labeled respectively.

Similarly, Zhang (2005) proposed to transform the problem of estimating the conditional probabilities of the positive class and the negative class given inputs to density estimations of an instance being labeled or unlabeled, and a positive instance being labeled, respectively. One drawback of the density-estimation-based approaches is that in general density estimation is a difficult task, especially when the input feature space is high dimensional or there are only a few positive labeled data.

Among these methods, Biased SVM (B-SVM) (Liu et al., 2003), which assigns different costs to the identified negative instances and positive instances respectively, is one of the state of the art. Though B-SVM has shown promising performance on PU learning, when the number of positive labeled instances is relatively small, the results of B-SVM become sensitive, which will be shown in the experimental section.

Multi-view learning has been also studied widely in machine learning. Blum and Mitchell (1998) proposed a co-training framework for multi-view learning, which first learns a separate classifier on each view using the labeled data. The instances with the highest confidence from the unlabeled data are then used to iteratively construct additional labeled training data in the next round. Recently, Sindhwani and Niyogi (2005) and Sindhwani and Rosenberg (2008) proposed a co-regularization framework, which aims to optimize the agreement on different views and the smoothness of labeled and unlabeled data in a unified regularization framework. Denis et al. (2003) borrowed the idea from co-training, and proposed an algorithm named PNCT to combine co-training with a existing PU learning method PNB (Denis et al., 2002) for multi-view PU learning. However, PNCT has two major drawbacks: 1) it requires the marginal distribution of the positive class as the prior. This limitation renders this method unpractical in many real-world applications where this prior is hard to obtain. 2) PNCT is proposed for discrete data. To extend it for continuous data, a feature discretization preprocessing is needed, which may potentially discard some discriminative information. In contrast, our proposed multi-view PU learning framework is flexible for both discrete and continuous data.

3. Density Ratio Estimation for PU Learning

Denote $\mathbf{x}$ an input instance and $y \in \{-1, 1\}$ a binary label. Following the notation in (Elkan and Noto, 2008), we introduce an additional random variable s on each instance, where $s = 1$ if the instance $\mathbf{x}$ is labeled, and $s = 0$ if the instance $\mathbf{x}$ is unlabeled. In PU learning, all the labeled instances are positive (i.e., $p(s = 1|\mathbf{x}, y = -1) = 0$), and the unlabeled data can be either positive or negative. Suppose we are given a set of positive instances $\{\mathbf{x}_i, y_i\}_{i=1}^{n_1}$, where $y_i = 1$ for $i = 1, ..., n_1$, and a large amount of unlabeled data $\{\mathbf{x}_j\}_{j=n_1+1}^{n}$, where $n_1 \mathcal{L} n$. Our goal is to learn a predictive function $f(x)$ from $\{\mathbf{x}_i, y_i\}_{i=1}^{n_1}$ and $\{\mathbf{x}_j\}_{j=n_1+1}^{n}$ such that

$$f(\mathbf{x}) \propto p(y = 1|\mathbf{x}).$$

However, since

$$p(y = 1|\mathbf{x}) = p(y = 1, s = 1|\mathbf{x}) + p(y = 1, s = 0|\mathbf{x}),$$

where $p(y = 1, s = 0|\mathbf{x})$ is unknown, it is hard to estimate $p(y = 1|\mathbf{x})$ directly. Fortunately, the problem of estimating $p(y = 1|\mathbf{x})$ can be simplified to the problem of estimating $p(s = 1|\mathbf{x})$ based on the following Lemma introduced in (Elkan and Noto, 2008),

Lemma 1 *Assume* [1] $p(s=1|y=1,\mathbf{x}) = p(s=1|y=1)$, *then* $p(y=1|\mathbf{x}) = p(s=1|\mathbf{x})/c$, *where* c *is a positive constant.*

We only concern the ranking order rather than the classification in information retrieval applications, so the constant c does not affect the ranking order. By using the Bayes rule, $p(s=1|\mathbf{x}) = \frac{p(s=1,\mathbf{x})}{p(\mathbf{x})}$. One solution to estimate $p(s=1|\mathbf{x})$ is to estimate the densities $p(s=1,\mathbf{x})$ and $p(\mathbf{x})$ separately. However, this solution may suffer from the difficulty in density estimation, especially in high-dimensional feature space. Instead, we propose to estimate the density ratio $\frac{p(s=1,\mathbf{x})}{p(\mathbf{x})}$ directly by using the unconstrained Least-Square Importance Fitting (uLSIF) method (Kanamori et al., 2009). Therefore, our objective can be written as the following minimization problem,

$$\min_{r(\mathbf{x})} \frac{1}{2} \int \left(r(\mathbf{x}) - \frac{p(s=1,\mathbf{x})}{p(\mathbf{x})} \right)^2 p(\mathbf{x}) d\mathbf{x}, \tag{1}$$

where $r(\mathbf{x})$ is the density ratio function, which can be regarded as the predictive function $f(\mathbf{x})$ for unseen test data. By expanding (1), we obtain

$$\min_{r(\mathbf{x})} \frac{1}{2} \int r(\mathbf{x})^2 p(\mathbf{x}) d\mathbf{x} - \int r(\mathbf{x}) p(s=1,\mathbf{x}) d\mathbf{x} + M, \tag{2}$$

where $M = \frac{1}{2} \int \frac{p(s=1,\mathbf{x})^2}{p(\mathbf{x})} d\mathbf{x}$ is a constant term that is irrelevant to $r(\mathbf{x})$. Given the positive instances $\{\mathbf{x}_i, y_i\}_{i=1}^{n_1}$ and the unlabeled instances $\{\mathbf{x}_j\}_{j=n_1+1}^{n}$, where $\{\mathbf{x}_i\}_{i=1}^{n_1}$ are drawn i.i.d. from $p(s=1,\mathbf{x})$, and $\{\mathbf{x}_k\}_{k=1}^{n} = \{\mathbf{x}_i\}_{i=1}^{n_1} \bigcup \{\mathbf{x}_j\}_{j=n_1+1}^{n}$ are drawn i.i.d from $p(\mathbf{x})$. By dropping the constant term and adding a regularization term $R(r)$ on $r(\mathbf{x})$ to avoid overfitting, the empirical approximation of (2) can be written as follows,

$$\min_{r(\mathbf{x})} \frac{1}{2} \sum_{i=1}^{n} r(\mathbf{x}_i)^2 - \frac{1}{n_1} \sum_{j=1}^{n_1} r(\mathbf{x}_j) + \lambda_1 R(r), \tag{3}$$

where $\lambda_1 > 0$ is a tradeoff parameter. Assume that the density ratio function $r(\mathbf{x})$ is represented by the following parametric form,

$$r(\mathbf{x}) = \sum_{l=1}^{b} \theta_l \psi_l(\mathbf{x}) = \boldsymbol{\psi}(\mathbf{x})^T \boldsymbol{\theta}, \tag{4}$$

where $\{\theta_l\}$'s are parameters to be learned, and $\{\psi_l(\mathbf{x})\}$'s are nonnegative basis functions, which can be linear or nonlinear. Suppose $\psi_l \in \mathcal{H}$, for $l = 1, ..., b$, where $\mathcal{H}$ is a reproducing Kernel Hilbert Space (RKHS), we can define the basic functions by using kernel functions (Scholkopf and Smola, 2001) as $\psi_l(\mathbf{x}) = k(\mathbf{x}, \mathbf{c}_l)$ where $\mathbf{c}_l$ is the l^{th} center point in the kernel space defined by the kernel function $k(\cdot,\cdot)$. Then (4) can be written as

$$r(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T \boldsymbol{\theta}, \tag{5}$$

where $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x}, \mathbf{c}_1), \cdots, k(\mathbf{x}, \mathbf{c}_b))^T$. By substituting (5) back to (3), and setting $R(r) = \boldsymbol{\theta}^T \boldsymbol{\theta}$, the objective can be rewritten as,

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \boldsymbol{\theta}^T \mathbf{H} \boldsymbol{\theta} - \mathbf{h}^T \boldsymbol{\theta} + \lambda_1 \boldsymbol{\theta}^T \boldsymbol{\theta}, \tag{6}$$

1. This assumption implies that the labeled positive instances are chosen totally random from all positive instances. This is a common assumption of previous probabilistic approaches to PU learning.

where $\mathbf{H} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{k}(\mathbf{x}_i)\mathbf{k}(\mathbf{x}_i)^T$, and $\mathbf{h} = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{k}(\mathbf{x}_j)$.

Note, the density ratio is always non-negative by definition $p(s=1|\mathbf{x})$. However the solution obtained from the above optimization problem cannot guarantee non-negativity. To tackle this problem, we follow (Kanamori et al., 2009) to modify the solution as $\boldsymbol{\theta}^* = \max\{\boldsymbol{\theta}^*, \mathbf{0}_b\}$.

As shown in (Kanamori et al., 2012), the leave-one-out cross-validation (LOOCV) score for the objective (6) can be obtained analytically as well as the coefficient parameters of the kernel function. Thus the leave-one-out solution can be computed efficiently by the use of the Sherman-Woodbury-Morrison formula (Golub and Van Loan, 1996). Thus in this work, we can use LOOCV for model selection. LOOCV is defined as

$$\text{LOOCV} = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\frac{1}{2}(\widehat{r}_{(j)}(\mathbf{x}_j))^2 - \widehat{r}_{(j)}(\mathbf{x}_j)\right),$$

where $\widehat{r}_{(j)}$ is the estimator based on training samples except $\mathbf{x}_j$. The hyper-parameters achieving the minimum value of LOOCV are chosen.

In the sequel, we name the proposed PU learning method *Density-Ratio-based PU learning (DRPU)*. Note that density ratio estimation techniques have been applied to various machine learning problems, such as covariate shift adaptation (Sugiyama et al., 2008), clustering (Sugiyama et al., 2011), and outlier detection (Hido et al., 2008). To our knowledge, this is the first work to explore density ratio estimation techniques for PU learning.

4. Multi-View PU Learning

In many real-world applications, data may be represented by multiple "views". A simple way to use multi-view information is to concatenate different views to generate unified features. However, this may not be an optimal solution because of the redundancy and noise issues of different views. In this section, we extend the proposed DRPU in a multi-view manner by using a co-regularization framework (Sindhwani and Rosenberg, 2008).

4.1. Co-regularization for Multi-View PU Learning

Suppose an instance $\mathbf{x}$ can be represented by two views $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$. Given a set of positive instances with two views $\{\mathbf{x}_i^{(1)}, y_i\}_{i=1}^{n_1}$ and $\{\mathbf{x}_i^{(2)}, y_i\}_{i=1}^{n_1}$, where $y_i = 1$ for $i = 1, ..., n_1$, as well as a large amount of unlabeled data with two views, $\{\mathbf{x}_j^{(1)}\}_{j=n_1+1}^{n}$ and $\{\mathbf{x}_j^{(2)}\}_{j=n_1+1}^{n}$, where $n_1 \mathcal{L} n$. Our goal is to learn the density ratio functions $r^{(1)}(\mathbf{x}^{(1)})$ and $r^{(2)}(\mathbf{x}^{(2)})$ on the two views simultaneously, and use the following form to make predictions,

$$r(\mathbf{x}) = \frac{1}{2}\left(r^{(1)}(\mathbf{x}^{(1)}) + r^{(2)}(\mathbf{x}^{(2)})\right), \tag{7}$$

Intuitively, for the same instance $\mathbf{x}$, density ratio functions learned from different views should have the mutual agreement on predicted values. By embedding this idea into the density-ratio-based PU learning method, our objective to multi-view PU learning can be written as follows,

$$\min_{r^{(1)},r^{(2)}} \sum_{v=1}^{2}\left(\frac{1}{2}\sum_{i=1}^{n} r^{(v)}(\mathbf{x}_i^{(v)})^2 - \frac{1}{n_1}\sum_{j=1}^{n_1} r^{(v)}(\mathbf{x}_j^{(v)}) + \lambda_1^{(v)} R(r^{(v)})\right) + \lambda_2 \sum_{i=1}^{n}\left(r^{(1)}(\mathbf{x}_i^{(1)}) - r^{(2)}(\mathbf{x}_i^{(2)})\right)^2,$$

where the fourth term in the objective is a co-regularization term to enforce the density ratio functions $r^{(1)}(\mathbf{x}^{(1)})$ and $r^{(2)}(\mathbf{x}^{(2)})$ to make agreement on the same instance $\mathbf{x}$, and $\lambda_2 > 0$ is a

tradeoff parameter. Similar to DRPU, we assume that the density ratio functions $r^{(1)}(\mathbf{x}^{(1)})$ and $r^{(2)}(\mathbf{x}^{(2)})$ can be represented by

$$r^{(v)}(\mathbf{x}^{(v)}) = \mathbf{k}(\mathbf{x}^{(v)})^T \boldsymbol{\theta}^{(v)},\ v \in \{1,2\}. \tag{8}$$

By substituting (8) into the above multi-view objective, we obtain

$$\min_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}} (\frac{1}{2}+\lambda_2) \sum_{v=1}^{2} \boldsymbol{\theta}^{(v)^T} \mathbf{H}^{(v)} \boldsymbol{\theta}^{(v)} - \sum_{v=1}^{2} \mathbf{h}^{(v)^T} \boldsymbol{\theta}^{(v)} + \sum_{v=1}^{2} \lambda_1^{(v)} \boldsymbol{\theta}^{(v)^T} \boldsymbol{\theta}^{(v)} - 2\lambda_2 \boldsymbol{\theta}^{(1)^T} \mathbf{H}^{(1,2)} \boldsymbol{\theta}^{(2)}, \tag{9}$$

where $\mathbf{H}^{(1,2)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{k}(\mathbf{x}_i^{(1)})\mathbf{k}(\mathbf{x}_i^{(2)})^T$. For $v \in \{1,2\}$, $\mathbf{H}^{(v)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{k}(\mathbf{x}_i^{(v)})\mathbf{k}(\mathbf{x}_i^{(v)})^T$, and $\mathbf{h}^{(v)} = \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbf{k}(\mathbf{x}_j^{(v)})$. The objective (9) can be further reformulated as the following matrix form,

$$\begin{bmatrix} \boldsymbol{\theta}^{(1)} \\ \boldsymbol{\theta}^{(2)} \end{bmatrix}^T \begin{bmatrix} \mathbf{S}_1 & -\lambda_2 \mathbf{H}^{(1,2)} \\ -\lambda_2 \mathbf{H}^{(2,1)} & \mathbf{S}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}^{(1)} \\ \boldsymbol{\theta}^{(2)} \end{bmatrix} - \begin{bmatrix} \mathbf{h}^{(1)} \\ \mathbf{h}^{(2)} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\theta}^{(1)} \\ \boldsymbol{\theta}^{(2)} \end{bmatrix}, \tag{10}$$

where $\mathbf{S}_1 = (\frac{1}{2}+\lambda_2)\mathbf{H}^{(1)} + \lambda_1^{(1)}\mathbf{I}$, $\mathbf{S}_2 = (\frac{1}{2}+\lambda_2)\mathbf{H}^{(2)} + \lambda_1^{(2)}\mathbf{I}$, $\mathbf{H}^{(2,1)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{k}(\mathbf{x}_i^{(2)})\mathbf{k}(\mathbf{x}_i^{(1)})^T$ and $\mathbf{I}$ is the identity matrix. It can be shown that the above optimization problem is a Quadratic Programming (QP) problem, where the global optimum exists. By taking the derivatives on (10) and setting it to zero, we can obtain the following equations,

$$2\begin{bmatrix} \mathbf{S}_1 & -\lambda_2 \mathbf{H}^{(1,2)} \\ -\lambda_2 \mathbf{H}^{(2,1)} & \mathbf{S}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}^{(1)} \\ \boldsymbol{\theta}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{h}^{(1)} \\ \mathbf{h}^{(2)} \end{bmatrix}.$$

The optimal solutions $\boldsymbol{\theta}^{(1)^*}$ and $\boldsymbol{\theta}^{(2)^*}$ has a closed-form, which can be obtained by solving following linear systems respectively,

$$\left(\mathbf{H}^{(1,2)^{-1}}\mathbf{S}_1 - \lambda_2^2 \mathbf{S}_2^{-1}\mathbf{H}^{(2,1)}\right)\boldsymbol{\theta}^{(1)} = \lambda_2 \mathbf{S}_2^{-1}\mathbf{h}^{(2)} + \mathbf{H}^{(1,2)^{-1}}\mathbf{h}^{(1)}$$
$$\left(\mathbf{H}^{(2,1)^{-1}}\mathbf{S}_2 - \lambda_2^2 \mathbf{S}_1^{-1}\mathbf{H}^{(1,2)}\right)\boldsymbol{\theta}^{(2)} = \lambda_2 \mathbf{S}_1^{-1}\mathbf{h}^{(1)} + \mathbf{H}^{(2,1)^{-1}}\mathbf{h}^{(2)}$$

After obtaining the optimal solution, similarly, we follow (Kanamori et al., 2009) to modify the solution $\boldsymbol{\theta}^{(1)^*} = \max\{\boldsymbol{\theta}^{(1)^*}, \mathbf{0}\}$, and $\boldsymbol{\theta}^{(2)^*} = \max\{\boldsymbol{\theta}^{(2)^*}, \mathbf{0}\}$. We further reconstruct the density ratio functions $r^{(1)}$ and $r^{(2)}$, and use the decision rule (7) to make predictions. In the sequel, we denote this method by *Co-regularized DRPU* or *Co-DRPU* in short.

4.2. Manifold Co-regularization for Multi-View PU Learning

It has been shown in the literature that when the manifold assumption holds on underlying data observations, a regularizer defined on data graphs can effectively propagate label information from a few labeled data to a large amount of unlabeled data, thus boost the classification performance (Belkin et al., 2006). Sindhwani and Niyogi (2005); Sindhwani and Rosenberg (2008) proposed to extend this manifold regularizer in a multi-view manner through a co-regularization framework such that predictive functions of different views are indirectly coupled through the multi-view regularizer. In this section, we present how to encode such a manifold Co-regularizer into our proposed multi-view PU learning framework.

Denote $\mathbf{L}^{(1)}$ and $\mathbf{L}^{(2)}$ the Laplacian matrices (Belkin et al., 2006) of the two views $\{\mathbf{x}_i^{(1)}\}_{i=1}^n$ and $\{\mathbf{x}_i^{(2)}\}_{i=1}^n$ respectively. Let $\mathbf{L} = (1-\alpha)\mathbf{L}^{(1)} + (\alpha)\mathbf{L}^{(2)}$, where $0 \le \alpha \le 1$ is

a tradeoff parameter which controls the influence of the two views. Motivated by previous work (Sindhwani and Niyogi, 2005), we use the following co-regularizer for multi-view PU learning,

$$\sum_{v=1}^{2} \mathbf{r}^{(v)^T} \mathbf{L} \mathbf{r}^{(v)}, \tag{11}$$

where $\mathbf{r}^{(v)} = (r^{(v)}(\mathbf{x}_1), \cdots, r^{(v)}(\mathbf{x}_n))^T$, for $v = 1, 2$. Minimizing (11) leads to making the predicted values $r^{(v)}(\mathbf{x})$ smooth with respect to the similarity structures encoded in $\mathbf{L}^{(v)}$ for each view. Furthermore, the two density ratio functions are coupled through the combination Laplacian matrix $\mathbf{L}$. By plugging (11) into (9), we obtain a new formula for multi-view PU learning as follows,

$$\begin{aligned} \min_{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}} \quad & \left(\frac{1}{2} + \lambda_2\right) \sum_{v=1}^{2} \boldsymbol{\theta}^{(v)^T} \mathbf{H}^{(v)} \boldsymbol{\theta}^{(v)} - \sum_{v=1}^{2} \mathbf{h}^{(v)^T} \boldsymbol{\theta}^{(v)} + \sum_{v=1}^{2} \lambda_1^{(v)} \boldsymbol{\theta}^{(v)^T} \boldsymbol{\theta}^{(v)} \\ & -2\lambda_2 \boldsymbol{\theta}^{(1)^T} \mathbf{H}^{(1,2)} \boldsymbol{\theta}^{(2)} + \lambda_3 \sum_{v=1}^{2} \boldsymbol{\theta}^{(v)^T} \mathbf{K}^v \mathbf{L} \mathbf{K}^{(v)^T} \boldsymbol{\theta}^{(v)} \end{aligned} \tag{12}$$

where $\mathbf{K}^{(v)} = (\mathbf{k}(\mathbf{x}_1^{(v)}), \cdots, \mathbf{k}(\mathbf{x}_n^{(v)}))$, for $v = 1, 2$. The objective (12) can be further rewritten as a matrix form,

$$\begin{bmatrix} \boldsymbol{\theta}^{(1)} \\ \boldsymbol{\theta}^{(2)} \end{bmatrix}^T \begin{bmatrix} \mathbf{W}_1 & -\lambda_2 \mathbf{H}^{(1,2)} \\ -\lambda_2 \mathbf{H}^{(2,1)} & \mathbf{W}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}^{(1)} \\ \boldsymbol{\theta}^{(2)} \end{bmatrix} - \begin{bmatrix} \mathbf{h}^{(1)} \\ \mathbf{h}^{(2)} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\theta}^{(1)} \\ \boldsymbol{\theta}^{(2)} \end{bmatrix} \tag{13}$$

where $\mathbf{W}_1 = (\frac{1}{2} + \lambda_2)\mathbf{H}^{(1)} + \lambda_1^{(1)}\mathbf{I} + \lambda_3 \mathbf{K}^{(1)} \mathbf{L} \mathbf{K}^{(1)^T}$ and $\mathbf{W}_2 = (\frac{1}{2} + \lambda_2)\mathbf{H}^{(2)} + \lambda_1^{(2)}\mathbf{I} + \lambda_3 \mathbf{K}^{(2)} \mathbf{L} \mathbf{K}^{(2)^T}$. Similar to (10), the optimization problem (13) is also a QP problem. By taking the derivatives on (13) and setting it to zero, we can obtain the following equations,

$$2 \begin{bmatrix} \mathbf{W}_1 & -\lambda_2 \mathbf{H}^{(1,2)} \\ -\lambda_2 \mathbf{H}^{(2,1)} & \mathbf{W}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}^{(1)} \\ \boldsymbol{\theta}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{h}^{(1)} \\ \mathbf{h}^{(2)} \end{bmatrix}$$

Closed-form solutions can be obtained by solving the following linear systems respectively,

$$\begin{aligned} 2(\mathbf{H}^{(1,2)^{-1}} \mathbf{W}_1 - \lambda_2^2 \mathbf{W}_2^{-1} \mathbf{H}^{(2,1)}) \boldsymbol{\theta}^{(1)} &= \lambda_2 \mathbf{W}_2^{-1} \mathbf{h}^{(2)} + \mathbf{H}^{(1,2)^{-1}} \mathbf{h}^{(1)} \\ 2(\mathbf{H}^{(2,1)^{-1}} \mathbf{W}_2 - \lambda_2^2 \mathbf{W}_1^{-1} \mathbf{H}^{(1,2)}) \boldsymbol{\theta}^{(2)} &= \lambda_2 \mathbf{W}_1^{-1} \mathbf{h}^{(1)} + \mathbf{H}^{(2,1)^{-1}} \mathbf{h}^{(2)} \end{aligned}$$

We use the truncation rule $\boldsymbol{\theta}^{(i)} = \max\{\boldsymbol{\theta}^{(i)}, \mathbf{0}\}$, $i = 1, 2$, to modify solution to guarantee the non-negativity of density ratio.

In the sequel, we denote this method by *Co-regularized Laplician DRPU (Co-LapDRPU)*. We can obtain the density ratio estimations of two views respectively for the Co-DRPU and Co-LapDRPU methods. For either of above two approaches, the LOOCV scores can be computed for each view as single-views setting. Heuristically, we define the LOOCV under multi-view setting ($\text{LOOCV}_{\text{multiview}}$) as the sum of LOOCV scores for different views as,

$$\text{LOOCV}_{\text{multiview}} = \frac{1}{n_1} \sum_{v=1}^{2} \sum_{j=1}^{n_1} \left(\frac{1}{2} (\widehat{r}_{(j)}^{(v)}(\mathbf{x}_j))^2 - \widehat{r}_{(j)}^{(v)}(\mathbf{x}_j) \right), \tag{14}$$

where $\widehat{r}^{(v)}_{(j)}$ is the estimator of $\widehat{r}^{(v)}$ based on training samples except $\mathbf{x}_j$. The hyper-parameters achieving the minimum value of $\text{LOOCV}_{\text{multiview}}$ are chosen for Co-DRPU and Co-LapDRPU. The complexity of computing $\text{LOOCV}_{\text{multiview}}$ is the same order as computing a single solution.

5. Experiments

In this section, we first present the model selection criterion for compared methods, and then conduct extensive experiments in both single-view setting and multi-view setting on the toy and real-world datasets.

5.1. PUAUC Criterion

As aforementioned, information retrieval or ranking is one of the important applications of PU learning. In this case, AUC score is adequate to be considered as the proper criterion for evaluating performance of the compared learning methods. On the other hand, there is no negative training data in the PU learning setting, so it is challenging to tune parameters of learning methods by using AUC directly. Mineiro (2012) uncovers the relationship between AUC scores over positive and unlabeled data (PUAUC) and AUC scores over the corresponding data with positive and negative instances. Let $\mathbf{x}$ be an instance and y be its corresponding label. The pair $(\mathbf{x}, y)$ follows the joint distribution $D = D_{\mathbf{x}} \times D_{y|\mathbf{x}} = D_y \times D_{\mathbf{x}|y}$ where $D_{\mathbf{x}|1}$ is positive label distribution, $D_{\mathbf{x}}$ is unlabeled instance distribution and $D_{\mathbf{x}|-1}$ is negative instance distribution. The relationship between PUAUC and true AUC is:

$$PUAUC(r) - \frac{1}{2} \propto AUC(r) - \frac{1}{2},$$

where $PUAUC(r) = E_{(\mathbf{x}_+,\mathbf{x}_-)\sim D_{x|1}\times D_{\mathbf{x}}}[1_{r(\mathbf{x}_+)>r(\mathbf{x}_-)} + \frac{1}{2}1_{r(\mathbf{x}_+)=r(\mathbf{x}_-)}]$ and $AUC(r) = E_{(\mathbf{x}_+,\mathbf{x}_-)\sim D_{x|1}\times D_{\mathbf{x}|-1}}[1_{r(\mathbf{x}_+)>r(\mathbf{x}_-)} + \frac{1}{2}1_{r(\mathbf{x}_+)=r(\mathbf{x}_-)}]$. Since PUAUC is proportional to true AUC, we can simply set all the unlabeled instance to be negative and calculate AUC instead. Note, the complexity of computing PUAUC is $O(n \log n)$.

5.2. Single-view PU Leaning

In the single-view PU learning setting, we compare our proposed method with state-of-the-art PU learning methods such as B-SVM (Liu et al., 2003), S-EM (Liu et al., 2002), and B-Pr (Zhang, 2005) on two real-world datasets: Letters [2] and USPS [3], where instances with the first class are set to be positive and the rest to be negative. The standard training and testing splits for these two datasets provided by their websites are used in experiments. Since the LOOCV score is inherently provided by DRPU, LOOCV is adopted for its model selection. However, other PU learning methods cannot apply this scheme directly, so we randomly select 30% of the training set as the validation set in each experiment for the model selection of B-Pr and B-SVM, while S-EM does not need any validation set for model selection. Since we use PUAUC for performance evaluation instead of the criteria used in (Zhang, 2005; Liu et al., 2003), the model selection for all PU learning methods except DRPU are all based on the maximizing PUAUC score on the validation set. Following the common assumption in PU learning that the true positive instances are randomly labeled with a positive label ratio γ from the training data (Liu et al., 2003; Zhang, 2005; Elkan and Noto, 2008) and the rest are used as the unlabeled data. Experimental results are reported

2. http://archive.ics.uci.edu/ml/datasets/Letter+Recognition/
3. http://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/multiclass.html

Table 1: Comparison results on Letters and USPS in terms of AUC scores under varying positive label ratio

Datasets	Methods	$\gamma = 0.5$	$\gamma = 0.3$	$\gamma = 0.1$	$\gamma = 0.07$	$\gamma = 0.05$	$\gamma = 0.03$
Letters	B-SVM	**0.9951**	**0.9900**	**0.9762**	**0.9585**	0.9325	0.9276
	B-Pr	0.9553	0.9532	0.9420	0.9128	0.8960	0.8772
	S-EM	0.9467	0.9472	0.9336	0.9419	0.9397	0.9150
	DRPU	0.9506	0.9516	0.9467	0.9432	**0.9567**	**0.9462**
USPS	B-SVM	**0.9943**	**0.9861**	0.9751	0.9725	0.9635	0.9311
	B-Pr	0.9553	0.9407	0.9056	0.9127	0.9043	0.8953
	S-EM	0.9242	0.9234	0.9132	0.9324	0.9220	0.9156
	DRPU	0.9819	0.9824	**0.9752**	**0.9767**	**0.9767**	**0.9653**

by the average over the 10 random runs with $\gamma \in [0.03, 0.5]$. In addition, since S-EM and B-Pr can only deal with discrete features as the input, we employ the discretization tool provided by WEKA [4] to obtain the discrete features for the data with continuous features. B-Pr method is reduplicated by ourselves and S-EM code is available online [5]. Gaussian kernel is used for B-SVM and DRPU on all the datasets.

Table 1 shows the AUC scores of the compared methods by varying the positive label ratio γ on Letters and USPS datasets. We have the following observations: 1) B-SVM performs the best in the case of large positive label ratio γ, but degrades greatly when γ decreases. 2) DRPU performs well in most range of γ and demonstrates the best results for small γ. 3) B-Pr and S-EM are consistently worse than others in most range of γ. The first two observations imply that the proposed DRPU is more stable than B-SVM with varying γ. This is mainly because the decision boundary learned by B-SVM is not robust when the proportion of positive data is small. Besides, when the label ratio is small, the set of negative instances selected by S-EM is not reliable, and the simple counting or empirical average may not be statistically meaningful for B-Pr as well. All these poor intermediate estimations lead to the worse results of S-EM and B-Pr compared to DRPU, which bypasses the above two drawbacks by estimating density ratio directly. Recall that positive labeled instances are hard and expensive to collect, but unlabeled data are usually abundant and freely available. This means that the setting of small γ fits for practice, and thus DRPU is more reliable than other PU learning methods considered in the present paper for real-world information retrieval applications.

5.3. Multi-view PU Learning

In this section, we compare our proposed multi-view PU learning methods, Co-DRPU and Co-LapDRPU, with other two baseline methods, B-SVM and PNCT. Two-views, denoted by View1 and View2, are studied in this section. Furthermore, to thoroughly study the performance of multi-view settings, the single view setting of various methods is also included for comparison. In other words, eight baseline settings will be studied, including 1) B-SVM on View1 (B-SVM$_1$), 2) B-SVM on View2 (B-SVM$_2$), 3) B-SVM on the concatenation of View1 and View2 (B-SVM$_{con}$), 4) B-SVM after CCA (Hardoon et al., 2004) (B-SVM$_{cca}$), 5) DRPU on View1 (DRPU$_1$), 6) DRPU on View2 (DRPU$_2$), 7) DRPU on the concatenation of View1 and View2 (DRPU$_{con}$), 8) PNCT. Experiments are conducted on both toy dataset

4. http://www.cs.waikato.ac.nz/ml/weka/
5. http://www.cs.uic.edu/∼liub/S-EM/S-EM-download.html

and real world datasets, where View1 and View2 are distinct. The linear kernel is used for our methods and B-SVM.

5.3.1. Experiments on Multi-view Toy Dataset

The Ionosphere [6] dataset with 351 features are used to generate the multi-view dataset. In this toy experiment, we want to show the performance of various methods over different positive label ratios as well as the increasing noise and redundant features. Specifically, let $\gamma \in (0,1)$ be the positive label ratio, $\delta \in (0,1)$ be the ratio of the number of noise features over the total number of features, and $\epsilon \in (0,1)$ be the ratio of the number of redundant features over the total number of features of View1. And the generation of multi-view dataset can be conducted as follows. For generating positive instances, we randomly select γ portion of positive labeled instances to construct the set of positive instances. For generating two views, we form the View1 dataset with noise features by combining the randomly selected 117 features from the original 351 features with $\lfloor 351\delta \rfloor$ noise features that follow the standard normal distribution. Similarly, we can form the View2 dataset by combining randomly selected $\lfloor \epsilon(117+351\delta) \rfloor$ features from View1 as redundant features with $\lfloor (117+351\delta)(1-\epsilon) \rfloor$ features selected from the original dataset by excluding those features in View1. In both views, noise features are added into View1 and View2 such that each view cannot learn a good classifier alone. Moreover, the added redundant features between two views can lead to degraded performance of concatenated views. In such settings, the multi-view learning is expected to obtain better performance over the performance on single-view and concatenated view datasets.

In this experiment, we test how positive label ratio, noise ratio and redundancy ratio, affect the overall performance of different methods respectively. Specifically, we study the influence of one factor by fixing the rest two factors. We randomly divide the original dataset into two parts: 70% as training set and 30% as test dataset. For B-SVM, we further randomly select 30% instances from the training set as the validation set for model selection using PUAUC. For our proposed methods, the model selections are done by the leave-one-out scheme using $\text{LOOCV}_{\text{multiview}}$ on the whole training set. We implement the PNCT method (Denis et al., 2003) by ourselves, where parameter $P(1)$ is set as 0.5 such that all the classes share the same proportion, seed size is equal to the number of positive labeled instances, and the algorithm terminates when either the number of maximum iteration reaching 100 or UD^{train} larger than or equal to UD is satisfied.

In the first experiment, we study the performance of various methods under different positive label ratios γ by fixing $\epsilon = 0.1$ and $\delta = 0.5$. Table 2 shows the variation of AUC score under different $\gamma \in [0.03, 0.5]$. From Table 2, we can observe that: 1) For DRPU and B-SVM methods, concatenated views can help to improve the performance over single view in the range $\gamma \in [0.03, 0.5]$, especially when the label ratio is relatively small. 2) Co-DRPU and Co-LapDRPU show more stable performance over most range of label ratio, which demonstrates the validity of our proposed models. 3) Although B-SVM on concatenated views performs better than Co-DRPU and Co-LapDRPU when the label ratio is 0.5, our proposed Co-DRPU and Co-LapDRPU can achieve better performance than B-SVM on concatenated views when the label ratio decreases. This observation may be due to the evidence that *multi-view agreement regularizer and manifold regularizer* can make use of

6. http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html.

Table 2: Comparison on the toy dataset in terms of AUC under varying positive label ratio

Methods	$\gamma = 0.5$	$\gamma = 0.3$	$\gamma = 0.1$	$\gamma = 0.07$	$\gamma = 0.05$	$\gamma = 0.03$
B-SVM$_1$	0.6889	0.5775	0.5958	0.5772	0.5642	0.5457
B-SVM$_2$	0.7225	0.6969	0.7406	0.5189	0.5077	0.4859
B-SVM$_{con}$	**0.8327**	0.7282	0.6746	0.6805	0.6541	0.6370
B-SVM$_{cca}$	0.7207	0.6964	0.6946	0.7361	0.6530	0.4911
DRPU$_1$	0.7523	0.8047	0.6901	0.7639	0.5503	0.6447
DRPU$_2$	0.7424	0.7469	0.7902	0.7714	0.6850	0.6877
DRPU$_{con}$	0.7954	**0.8232**	0.7528	0.7823	0.7540	0.7077
PNCT	0.5361	0.4761	0.4701	0.4731	0.4528	0.4161
Co-DRPU	0.7665	0.8186	**0.7925**	**0.7942**	0.7512	0.7115
Co-LapDRPU	0.7699	0.8159	0.7893	0.7929	**0.7624**	**0.7160**

information from unlabeled data. 4) PNCT shows poor results on this toy dataset. It may result from poor estimation of positive label ratio which is set to 0.5 as the prior for this method or the loss of information due to discretization of continuous features.

Table 3: Comparison on the toy dataset in terms of AUC under varying noise ratio

Methods	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.5$
B-SVM$_1$	0.8121	0.4385	0.7417	0.6555	0.5772
B-SVM$_2$	0.8025	0.5853	0.3857	0.5815	0.5189
B-SVM$_{con}$	0.8132	0.5981	0.6708	0.6859	0.6805
B-SVM$_{cca}$	0.7983	0.6833	0.7072	0.7256	0.7361
DRPU$_1$	0.7672	0.7809	0.7988	0.7684	0.7639
DRPU$_2$	0.7705	0.7528	0.7664	0.7581	0.7714
DRPU$_{con}$	0.7991	0.7841	0.8007	0.7816	0.7823
PNCT	0.5350	0.4885	0.4669	0.4307	0.4731
Co-DRPU	0.8104	0.8028	**0.8114**	**0.7933**	**0.7942**
Co-LapDRPU	**0.8175**	**0.8074**	0.7986	0.7871	0.7929

Table 4: Comparison on the toy dataset in terms of AUC under varying redundancy ratio

Methods	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$
B-SVM$_1$	0.5772	0.5772	0.5772	0.5772
B-SVM$_2$	0.5189	0.5623	0.5872	0.5770
B-SVM$_{con}$	0.6805	0.6717	0.6082	0.5518
B-SVM$_{cca}$	0.7361	0.7178	0.5876	0.6192
DRPU$_1$	0.7639	0.7639	0.7639	0.7639
DRPU$_2$	0.7714	0.7688	0.7718	0.7599
DRPU$_{con}$	0.7823	0.7812	0.7810	0.7745
PNCT	0.4731	0.4838	0.4618	0.5077
Co-DRPU	**0.7942**	**0.7929**	0.7845	0.7586
Co-LapDRPU	0.7929	0.7868	**0.7865**	**0.7773**

In the second experiment, we test the performance of all methods over different number of noise features by setting $\epsilon = 0.1$ and $\gamma = 0.07$. In Table 3, we record the AUC score by varying noise feature ratio $\delta \in [0.1, 0.5]$. Several observations can be obtained from this table: 1) B-SVM's performance varies vastly and is sensitive to the noise. When the noise ratio increases, B-SVM on single view and concatenated views cannot work well. 2) B-SVM with CCA procedure can enhance learning performance because CCA can eliminate the noise features and extract the discriminative feature components (Hardoon et al., 2004).

Table 5: Comparison results on Corel and Web-KB in terms of AUC scores under varying positive label ratio

Datasets	Methods	$\gamma = 0.5$	$\gamma = 0.3$	$\gamma = 0.1$	$\gamma = 0.07$	$\gamma = 0.05$	$\gamma = 0.03$
Corel	B-SVM$_1$	0.9624	0.8925	0.8334	0.7207	0.8178	0.6948
	B-SVM$_2$	0.9707	0.9489	0.9201	0.8468	0.8483	0.7830
	B-SVM$_{con}$	**0.9707**	0.9291	0.9201	0.8469	0.8483	0.7830
	B-SVM$_{cca}$	0.7608	0.6275	0.6267	0.6748	0.5624	0.6059
	DRPU$_1$	0.8733	0.8003	0.8486	0.8062	0.8359	0.6488
	DRPU$_2$	0.9087	0.9141	0.9044	0.8983	0.8921	0.8686
	DRPU$_{con}$	0.9087	0.9141	0.8645	0.8983	0.8921	0.8686
	PNCT	0.5309	0.4946	0.4835	0.4827	0.4649	0.5026
	Co-DRPU	0.9622	0.9214	0.9265	0.8809	0.9046	0.8702
	Co-LapDRPU	0.9623	**0.9410**	**0.9621**	**0.9309**	**0.9150**	**0.9017**
Web-KB	B-SVM$_1$	0.8884	0.9001	0.8615	0.7661	0.7868	0.6994
	B-SVM$_2$	0.8989	0.6818	0.7394	0.6198	0.5438	0.3985
	B-SVM$_{con}$	0.9847	0.9084	0.7868	0.6294	0.6967	0.5714
	B-SVM$_{cca}$	0.9018	0.7657	0.7089	0.5486	0.6142	0.6130
	DRPU$_1$	0.9445	0.9302	0.9373	0.9003	0.9225	0.8752
	DRPU$_2$	0.9848	0.9621	0.9588	0.9239	0.9461	0.7560
	DRPU$_{con}$	**0.9939**	0.9830	0.9529	0.9146	0.9469	0.7496
	PNCT	0.9445	0.9559	0.9109	0.8969	0.6222	0.5032
	Co-DRPU	0.9768	**0.9926**	0.9884	0.9607	0.9646	0.9224
	Co-LapDRPU	0.9911	0.9857	**0.9884**	**0.9722**	**0.9751**	**0.9520**

However, the results are still worse than our proposed DRPU. In conclusion, our proposed PU learning methods show very stable performance and are insensitive to noises.

In the third experiment, the performances of compared methods over different redundancy ratio are studied, where we keep $\gamma = 0.07$ and $\delta = 0.5$ and vary the redundancy ratio $\epsilon \in [0.1, 0.4]$. The AUC scores under different redundancy ratio are recorded in Table 4. From Table 4, we can observe that: 1) When the redundancy ratio increases, the performance of two views becomes more similar, which can explain that the larger redundancy ratio is, the closer AUC scores of two views are. 2) Simple concatenation of two views cannot enhance the performance. Particularly, it may be even worse than the single view for B-SVM when the redundancy ratio becomes relatively large. 3) On the contrary, the proposed multi-view learning methods, Co-DRPU and Co-LapDRPU, show better performance than PNCT.

In summary, based on the above empirical studies, we can see our proposed Co-DRPU and Co-LapDRPU are more tolerant to the variations of positive label ratio, noise ratio and redundancy ratio than other baseline settings and remain better performance.

5.3.2. Experiments on Real-world Multi-view Datasets

We further conduct experiments on two real-world multi-view datasets, namely Corel dataset and Web-KB dataset, to demonstrate the effectiveness of our proposed multi-view PU learning methods by comparing with baselines. Specifically, Corel dataset (Lu and Ip, 2009) is an image dataset with 1000 instances in total, and whose features are extracted from Corel image collection. The View1 is generated by Bag-of-Words (BoW) histogram with 500 dimensions; View2 is generated by Color Moment (CM). Each image is divided into 8×8 subregions and then color moment with 576 dimensions are extracted in LUV space. Web-

KB dataset (Sindhwani and Niyogi, 2005) is the most common dataset used in multi-view learning problems. One view is page which has 3000 features and the other view is link which has 1840 features. Total number of instances for this dataset is 1051. All the splitting schemes remain the same as the multi-view toy dataset.

Table 5 shows the AUC scores of compared methods on the Corel and Web-KB datasets. From Table 5, we can observe that our proposed single-view PU learning DRPU on either of views and the concatenated view can consistently outperform B-SVM when the label ratio is small (less than 0.1) on the both datasets. By comparing the performances of the two views and the concatenated view, we can also obtain that B-SVM and DRPU methods performed on the concatenated views are dominated by one of two views or the performance have a slight improvement, sometimes even worse than single-views. On the Web-KB dataset, we can observe PNCT performs well when the positive label ratio is not too small. However, when the label ratio is less than 0.1, its AUC score drops very quickly. On the Corel dataset, PNCT cannot perform well no matter how large the label ratio is. Similar to the toy experiment, this observation is probably due to the loss of information after discretization on the dense continuous dataset. In conclusion, all the observations imply that our proposed methods are effective for multi-view PU learning when the positive label ratio is small.

6. Conclusion and Future Work

In this paper, we firstly propose a Density-Ratio-based PU (DRPU) learning to learn models from positive and unlabeled examples. Secondly, we extend it into multi-view setting and further two new PU learning methods Co-DRPU and Co-lapDRPU are developed under this setting. Last but not least, extensive experiments demonstrate the effectiveness and stability of the proposed methods under the small label ratio. In future work, we will further extend our proposed multi-view PU learning methods to handle more than two views and partial correspondence.

References

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, December 2006.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th annual conference on Computational learning theory*, pages 92–100, 1998.

Nello Cristianini and John Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press, New York, NY, USA, 2000.

François Denis. PAC learning from positive statistical queries. In *Proceedings of the 9th International Conference on Algorithmic Learning Theory*, pages 112–126, 1998.

François Denis, Rémi Gilleron, and Marc Tommasi. Text classification from positive and unlabeled examples. In *Proceedings of the Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 1927–1934, 2002.

François Denis, Anne Laurent, Rémi Gilleron, and Marc Tommasi. Text classification and co-training from positive and unlabeled examples. In *Proceedings of the ICML 2003 Workshop: The Continuum from Labeled to Unlabeled Data*, pages 80–87, 2003.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220, 2008.

Gene H. Golub and Charles F. Van Loan. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

David R. Hardoon, Sándor Szedmák, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12): 2639–2664, December 2004.

Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Inlier-based outlier detection via direct density ratio estimation. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 223–232, 2008.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, March 2012.

Xiaoli Li and Bing Liu. Learning to classify texts using positive and unlabeled data. In *Proceedings of the 18th international joint conference on Artificial intelligence*, pages 587–592, 2003.

Xiaoli Li, Philip S. Yu, Bing Liu, and See-Kiong Ng. Positive unlabeled learning for data stream classification. In *Proceedings of the SIAM International Conference on Data Mining*, pages 257–268, 2009.

Bing Liu, Wee Sun Lee, Philip S. Yu, and Xiaoli Li. Partially supervised classification of text documents. In *Proceedings of the 19th International Conference on Machine Learning*, pages 387–394, 2002.

Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S. Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 179–188, 2003.

Zhiwu Lu and Horace Ho-Shing Ip. Image categorization with spatial mismatch kernels. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 397–404, 2009.

Paul Mineiro. PU Learning and AUC. March 2012. URL http://www.machinedlearnings.com/2012/03/pu-learning-and-auc.html.

Tom Mitchell. *Machine Learning*. McGraw-Hill Education, 1st edition, October 1997.

Minh Nhut Nguyen, Xiaoli Li, and See-Kiong Ng. Ensemble based positive unlabeled learning for time series classification. In *Proceedings of the 17th International Conference on Database Systems for Advanced Applications*, pages 243–257, 2012.

Rong Pan, Yunhong Zhou, Bin Cao, Nathan Nan Liu, Rajan M. Lukose, Martin Scholz, and Qiang Yang. One-class collaborative filtering. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 502–511, 2008.

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, July 2001.

Vikas Sindhwani and Partha Niyogi. A co-regularized approach to semi-supervised learning with multiple views. In *Proceedings of the ICML Workshop on Learning with Multiple Views*, 2005.

Vikas Sindhwani and David S. Rosenberg. An RKHS for multi-view learning and manifold co-regularization. In *Proceedings of the 25th international conference on Machine learning*, pages 976–983, 2008.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Von Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems 20*, pages 1433–1440. 2008.

Masashi Sugiyama, Makoto Yamada, Manabu Kimura, and Hirotaka Hachiya. On information-maximization clustering: Tuning parameter selection and analytic solution. In *Proceedings of the 28th International Conference on Machine Learning*, pages 65–72, 2011.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density ratio matching under the bregman divergence: A unified framework of density ratio estimation. In *Annals of the Institute of Statistical Mathematics*, 2012.

Dell Zhang. A simple probabilistic approach to learning from positive and unlabeled examples. In *Proceedings of the 5th Annual UK Workshop on Computational Intelligence*, 2005.

Dell Zhang and Wee Sun Lee. Learning classifiers without negative examples: A reduction approach. In *In Proceedings of the 3rd IEEE International Conference on Digital Information Management*, pages 638–643, 2008.

Yuchen Zhao, Xiangnan Kong, and Philip S. Yu. Positive and unlabeled learning for graph classification. In *Proceedings of the 11th IEEE International Conference on Data Mining*, pages 962–971, 2011.