Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Instance Weighting for Domain Adaptation via Trading off Sample Selection Bias and Variance

Rui Xia, Zhenchun Pan, Feng Xu
School of Computer Science and Engineering, Nanjing University of Science and Technology, China
rxia@njust.edu.cn, zhenchunpan@gmail.com, breezewing@126.com

Abstract

Domain adaptation is an important problem in natural language processing (NLP) due to the distributional difference between the labeled source domain and the target domain. In this paper, we study the domain adaptation problem from the instance weighting perspective. By using density ratio as the instance weight, the traditional instance weighting approaches can potentially correct the sample selection bias in domain adaptation. However, researchers often failed to achieve good performance when applying instance weighting to domain adaptation in NLP and many negative results were reported in the literature. In this work, we conduct an in-depth study on the causes of the failure, and find that previous work only focused on reducing the sample selection bias, but ignored another important factor, sample selection variance, in domain adaptation. On this basis, we propose a new instance weighting framework by trading off two factors in instance weight learning. We evaluate our approach on two cross-domain text classification tasks and compare it with eight instance weighting methods. The results prove our approach's advantages in domain adaptation performance, optimization efficiency and parameter stability.

1 Introduction

Instance weighting (also called importance sampling), a generalization of statistical bias correction techniques, can potentially correct the sample selection bias in the labeled training data. It is therefore considered as an important type of domain adaptation method, that improves the adaptive performance by re-weighting the source-domain training samples to approximate the target-domain distribution.

In the field of machine learning, instance weighting has been studied extensively under the concept of sample selection bias or covariate shift. It has been proven that the sample selection bias can be canceled by using the density ratio to re-weight the training samples [Shimodaira, 2000; Zadrozny, 2004]. Therefore, density ratio estimation (DRE) becomes the key problem in correcting sample selection bias. A series of methods were proposed in the literature to solve the DRE problem [Shimodaira, 2000; Huang *et al.*, 2006; Sugiyama *et al.*, 2008; Tsuboi *et al.*, 2009; Kanamori *et al.*, 2009]. The core idea is to learn the instance weights by minimizing the gap between the re-weighted source-domain distribution and the target-domain true distribution.

Instance weighting for domain adaptation has been widely discussed in the field of NLP. Most of the methods first learn the weight for each training instance based on sample bias correction, and then train an instance-weighted classifier for cross-domain classification. Although these methods were effective to estimate the density ratio, most of them do not perform well when applied to NLP domain adaptation tasks. There were even some negative reports in the literature toward using instance weighting for domain adaptation in NLP. For example, [Plank *et al.*, 2014] presented a negative result on instance weighting for cross-domain POS tagging. They stated that "the publication bias toward reporting positive results makes it hard to say whether researchers have tried." Due to the difficulty in using instance weighting, instance selection by setting the weight of all selected instances to 1 and 0 otherwise, is used instead in real practice for instance-based domain adaptation [Jiang and Zhai, 2007; Axelrod *et al.*, 2011; Xia *et al.*, 2013].

Till now, the causes to the failure and inefficiency of instance weighting for domain adaptation still remained unclear in the literature, and we have not seen studies that explicitly discussed this problem before.

In this work, we find that sample selection bias is actually not the only way the training data affect the domain adaptation performance. In addition to that, there is another important factor which we call "sample selection variance". However, most of the previous instance weighting approaches only aimed at correcting the sample selection bias, while ignored the influence of sample selection variance in domain adaptation. We find that as a result of correcting selection bias, the traditional instance weighting algorithms tend to make a small number of samples have much higher weights than the rest, which increases the risk of having a high sample selection variance, and it will accordingly cause poor domain adaptation performance.

In machine learning, the bias-variance tradeoff is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithms from generalizing beyond their training set. This tradeoff applies to all forms of

supervised learning, including instance weighting for domain adaptation.

Based on the aforementioned findings, we propose a new instance weighting framework, by taking both sample selection bias and sample selection variance into consideration for domain adaptation. We evaluate our approach on two NLP domain adaptation tasks including cross-domain sentiment classification and personalized spam filtering. We compare our approach with the state-of-the-art instance weighting algorithms. The experimental results clearly confirm our conjecture on the inefficiency of previous instance weighting methods and strongly prove our approach's effectiveness and stability for domain adaptation.

The main contributions of this paper are as follows.

- To our knowledge, it is the first work that explicitly discusses the shortcomings of only correcting sample selection bias, and introduces the concept of sample selection variance as well as the bias-variance dilemma in the settings of instance weighting for domain adaptation;
- This work provides three ways of limiting the sample selection variance, and proposes a new framework that trades off sample selection bias and variance in instance weighting. It makes instance weighting more practical for domain adaptation tasks in NLP;
- In comparison with the previous instance weighting methods, the proposed approach has significant advantages in domain adaptive performance, optimization efficiency and parameter stability.

2 Related Work

Domain adaptation methods include feature-based adaptation and instance-based adaptation, etc [Jiang, 2008; Pan and Yang, 2010]. In this work, we focus on instance-based adaptation, which performs domain adaptation by instance weighting on the source-domain labeled training data.

In the field of machine learning, instance weighting is normally known as sample selection bias or covariate shift. This concept was first introduced in the field of econometrics by [Heckman, 1979], and brought into the field of machine learning by [Zadrozny, 2004]. It has been proven that the sample selection bias can be canceled by using the density ratio to re-weight the likelihood terms [Shimodaira, 2000; Zadrozny, 2004]. Hence, the key problem becomes density ratio estimation (DRE). A range of density estimation methods have been proposed to solve the DER problem. For example, kernel density estimation, maximum entropy density estimation and kernel mean matching, were used in [Shimodaira, 2000], [Dudík *et al.*, 2005] and [Huang *et al.*, 2006], respectively.

[Sugiyama *et al.*, 2008] proposed a K-L importance estimation procedure (KLIEP) to directly estimate the density ratio under a linear assumption. The instance weights were learned by minimizing the K-L divergence between the reweighted source distribution and the target distribution, based on a projection descent algorithm. The least square importance fitting (LSIF) and unconstrained LSIF (uLSIF) were further proposed by [Kanamori *et al.*, 2009]. [Tsuboi *et al.*, 2009] extended KLIEP by employing a log-linear model instead of the linear model, and made KLIEP feasible in large-scale test data set. [Xia *et al.*, 2014] proposed a logistic approximation instance adaptation model that performed instance adaptation more effectively than KLIEP. [Wen *et al.*, 2015] used the Frank-Wolfe algorithm to solve the projected gradient optimization in KMM and KLIEP. Since Frank-Wolfe is projection-free, its time complexity is in general smaller than KMM and KILEP.

In addition to the general approaches in machine learning, existing work has also proposed methods specifically designed for NLP tasks. For example, [Jiang and Zhai, 2007] first used instance weighting to select a subset of the source data for domain adaptation. [Axelrod *et al.*, 2011] proposed a pseudo in-domain data selection method to select source-domain training samples based on language model for machine translation. [Xia *et al.*, 2013] proposed an instance selection and instance weighting approach via PU learning for domain adaptation in sentiment classification.

In comparison with the generally positive results in machine learning, the reports of instance weighting on NLP domain adaptation tasks tend to be negative. For example, it seems that the instance selection technique performed better and more stably than instance weighting [Jiang and Zhai, 2007; Xia *et al.*, 2013]. [Plank *et al.*, 2014] presented a negative results of using instance weighting for domain adaptation in POS tagging.

The aforementioned methods in both machine learning and NLP only aimed at correcting sample selection bias. However, sample selection bias is not the only way the training data are biased. They ignored the influence of sample selection variance when applying instance weighting to domain adaptation. By contrast, our approach takes both factors into consideration in instance weighting.

3 Preliminaries

3.1 Instance Weight Estimation by Correcting Sample Selection Bias

Given the source-domain distribution $p_s(x)$ and the target-domain distribution $p_t(x)$, it has been proven that the sample selection bias can be corrected by using density ratio (DRE) $w(x) = \frac{p_t(x)}{p_s(x)}$ as the weights for instance weighting [Shimodaira, 2000; Zadrozny, 2004]. A range of density estimation methods have been proposed to solve the DRE problem.

For example, the KMM method [Huang *et al.*, 2006] tries to match the mean elements in a kernel feature space Φ by solving

$$\min_{w_i} \left\| \frac{1}{N_s} \sum_{i=1}^{N_s} w_i \Phi(x_i) - \frac{1}{N_t} \sum_{j=1}^{N_t} \Phi(x_j) \right\|_{\mathcal{H}} \quad (1)$$

where $\mathcal{H}$ denotes the reproducing kernel Hilbert space (RKHS). N_s and N_t are the number of samples in the source-domain training set and the target-domain test set, respectively.

The KLIEP algorithm [Sugiyama *et al.*, 2008] assumes the

density ratio is a linear model

$$w(x) = \frac{p_t(x)}{p_s(x)} \triangleq \beta^{\mathrm{T}} x \tag{2}$$

and determine the parameters by minimizing the Kullback-Leibler divergence between the target-domain true distribution $p_t(x)$ and the re-weighted source-domain distribution $q_s(x) = w(x)p_s(x)$:

$$\begin{aligned} &\min_{\alpha,\beta} KL\left[p_t(x)||q_s(x)\right] \\ &\text{s.t.} \int_{x\in\mathcal{X}} q_s(x)dx = 1. \end{aligned} \tag{3}$$

which finally results in an equality-constrained optimization.

We adopt the In-target-domain Logistic Approximation (ILA) algorithm [Xia *et al.*, 2014] as the basis of this work. In ILA, the density ratio was defined as a normalized logistic function:

$$w(x) = \frac{p_t(x)}{p_s(x)} \triangleq \frac{\alpha}{1 + e^{-\beta^{\mathrm{T}} x}}, \tag{4}$$

where α is a normalization factor. Due to the constraint $q_s(x)$ sum up to one, α can be represented by a function of β:

$$\alpha = \frac{N_s}{\sum_{j=1}^{N_s} \frac{1}{1+e^{-\beta^{\mathrm{T}} x_j}}} \tag{5}$$

and the equality-constrained K-L minimization in Equation (3) can be converted to a unconstrained optimization:

$$\min_{\beta} J_{MSD} = \log \sum_{i=1}^{N_s} \frac{1}{1 + e^{-\beta^{\mathrm{T}} x_i}} - \frac{1}{N_t} \sum_{j=1}^{N_t} \log \frac{1}{1 + e^{-\beta^{\mathrm{T}} x_j}}. \tag{6}$$

3.2 Instance-weighted Classification for Domain Adaptation

Once the instance weight $w(x)$ have been estimated, the next step is then to build an instance weighted classifier for domain adaptation.

Let θ denote the parameters of the classification model. The goal is then to find the best parameter θ^* that maximizes the likelihood of the data drawn from the target domain

$$\theta^* = \arg\max_{\theta} \int_{x\in\mathcal{X}_t} p_t(x) \sum_{y\in\mathcal{Y}} p_t(y|x) \log p(x, y|\theta) dx. \tag{7}$$

Since there is not labeled data in the target domain, we maximize the likelihood of the data drawn from the source domain

$$\begin{aligned} \theta^* &\approx \arg\max_{\theta} \int_{x\in\mathcal{X}_s} w(x)p_s(x) \sum_{y\in\mathcal{Y}} p_s(y|x) \log p(x, y|\theta) dx \\ &\approx \arg\max_{\theta} \frac{1}{N_s} \sum_{i=1}^{N_s} w(x_i) \log p(x_i, y_i|\theta), \end{aligned} \tag{8}$$

by assuming that $p_t(x) = w(x)p_s(x)$ and $p_t(y|x) = p_s(y|x)$. This results an instance-weighted classification.

In short, the process of instance weighting for domain adaptation is a pipeline of two steps:

- **Step 1 (density ratio estimation, DRE)**: Using density ratio estimation methods to learn the instance weights;
- **Step 2 (instance-weighted classification, IWC)**: Based on the instance weights to learn an instance-weighted classifier.

4 Our Approach

4.1 Why Only Correcting Sample Selection Bias is not Enough?

In machine learning, there are three types of prediction error: bias, variance, and irreducible noise. According to bias-variance decomposition [Geman *et al.*, 1992], the expected error on an unseen example x can be decomposed as follows

$$\begin{aligned} Error &= E\left[\hat{h}(x) - h(x)\right]^2 + E\left[\hat{h}(x)^2\right] - E\left[\hat{h}(x)\right]^2 + \delta^2 \\ &= Bias\left[\hat{h}(x)\right] + Var\left[\hat{h}(x)\right] + \delta^2 \\ &= Bias + Variance + Noise. \end{aligned} \tag{9}$$

The bias is an error from erroneous assumptions in the learning algorithm; the variance is an error from sensitivity to small fluctuations in the training set. Ideally, one wants to choose a model that both accurately capture the regularities in its training data, but also generalizes well to unseen data. Unfortunately, it is typically impossible to do both simultaneously. High-variance learning methods may be able to represent their training set well but are at risk of overfitting to noisy or unrepresentative training data. In contrast, algorithms with high bias typically produce simpler models that don't tend to overfit but may underfit their training data, failing to capture important regularities.

The bias-variance tradeoff applies to all supervised learning, including both steps summarized in Section 3.2. To trade off the bias and variances in Step 2, people normally added a L_2 penalty to the objective function in Equation (8). And in Step 1, the sample selection bias can be corrected by using density ratio as the instance weight. However, the variance in Step 1 was normally ignored in previous instance weighting approaches. We name the bias and variance in Step 1 as **Sample Selection Bias** and **Sample Selection Variance**, respectively.

Given a training instance x, the instance weight $w(x) = \frac{p_t(x)}{p_s(x)}$ has roughly three situations:

- If $p_t(x)$ is much higher than $p_s(x)$, $w(x)$ will be very large;
- If $p_t(x)$ is much lower than $p_s(x)$, $w(x)$ will be very small;
- If the difference of $p_t(x)$ and $p_s(x)$ is not big, $w(x)$ will be moderate.

In NLP, the multinomial distribution is widely used to model text. The probability of a document $p(x)$ is normally calculated in terms of the product of each term's probability: $p(x) = \prod_{i=1}^{|x|} p(t_i)$, and is apt to be arbitrarily close to 1 or 0. As a result, the density ratio might be very large (e.g., $p_t(x) = 0.999$, $p_s(x) = 0.002$, $w(x) = \frac{0.999}{0.002} = 499.5$),

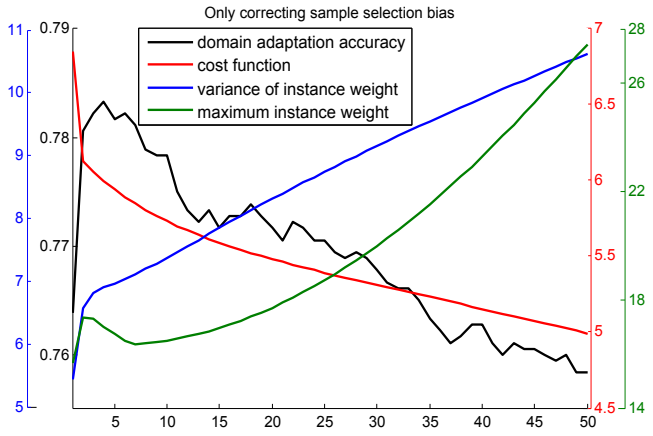

Figure 1: The instance weight learning process when correcting only sample selection bias. The x-axis denotes the iteration step in gradient descent optimization. The y-axis are the cost of sample selection bias, the domain adaptation accuracy, the instance weight variance and the maximum instance weight, respectively.

or very small (e.g., $p_t(x) = 0.002$, $p_s(x) = 0.999$, $w(x) = \frac{0.002}{0.999} = 0.002$). That is to say, although the sample selection bias can be corrected by using density ratio as the instance weight, the value of the instance weights may vary widely.

Hence, the **Bias-Variance Dilemma in Instance Weighting for Domain Adaptation** can be summarized as follows:

- To reduce the sample selection bias in Step 1, the density ratio needs to be used as the instance weight which may vary widely and accordingly cause poor classification performance in Step 2 due to high variance;
- To reduce the sample selection variance in Step 1, the more uniform instance weights are, the better. This will increase the sample selection bias and finally result in a domain non-adapted classifier in Step 2.

Most of the previous instance weighting approaches only sought correcting the sample selection bias while ignore reducing the sample selection variance. We infer that it is the most important reason of their failure in instance-based domain adaptation.

Figure 1 reports a failed result of only correcting sample selection bias on the "movie→book" cross-domain sentiment classification task (see Section 5.1 in detail). As can be seen, directly reducing sample selection bias causes a rapid growth of the instance weight variance. It consequently weakens the generalization ability in instance weighted classification, and finally leads to poor domain adaptation performance. Therefore, a reasonable instance weighting algorithm should trade off sample selection bias and variance.

4.2 Trading off Sample Selection Bias and Variance

Although Section 3.1 can correct the sample selection bias, it tends to increase the sample selection variance, and finally reduce the generalization performance. A simple way to limit the sample selection variance is to protect the instance weight from varying too much.

To this end, we develop the following three kinds of penalty functions, as an extension to the MSD cost function, to reduce sample selection bias.

(1) Limiting the Maximum Instance Weight (MIW)

$$J_{MIW} = (w_{max}(x) - 1)^2 = \left(\frac{N_s \delta(\beta^{\mathrm{T}} x_{max})}{\sum_{j=1}^{N_s} \delta(\beta^{\mathrm{T}} x_j)} - 1 \right)^2, \tag{10}$$

where $w_{max}(x)$ and x_{max} respectively denote the maximum weight and the corresponding instance. $\delta(\cdot)$ denotes the sigmoid function.

(2) Limiting the Variance of Instance Weights (VIW)

$$J_{VIW} = \frac{1}{N_s} \sum_{i=1}^{N_s} (w(x_i) - \bar{w})^2, \tag{11}$$

where $w(x_i)$ denotes the weight of x_i, and $\bar{w} = \frac{1}{N_s} \sum_j w(x_j)$ represents the average weight of all training instances.

(3) Limiting the Top-ranked Instance Weights (TIW)

$$J_{TIW} = \frac{1}{[pN_s]} \sum_{i=1}^{[pNs]} \left(\frac{N_s \delta(\beta^{\mathrm{T}} x_i)}{\sum_{j=1}^{N_s} \delta(\beta^{\mathrm{T}} x_j)} - 1 \right)^2, \tag{12}$$

where p denotes the percentage of top-ranked instance weights we want to limit, and $[\cdot]$ denotes the integer part.

We further have the following proposition that MIW and VIW are special cases of TIW.

Proposition 1. *TIW equals to MIW when* $p = \frac{1}{N_s}$, *and equals to VIW when* $p = 1$.

Proof:

Firstly, by setting $p = \frac{1}{Ns}$ in Equation (12), we get Equation (10) directly.

Secondly, according to Equations (4) and (5), we have

$$\bar{w} = \frac{1}{N_s} \sum_{i=1}^{Ns} w(x_i) = 1, \tag{13}$$

and we can therefore re-write the VIW penalty as follows:

$$\begin{aligned} J_{VIW} &= \frac{1}{N_s} \sum_{i=1}^{N_s} (w(x_i) - \bar{w})^2 \\ &= \frac{1}{N_s} \sum_{i=1}^{N_s} \left(\frac{N_s \delta(\beta^{\mathrm{T}} x_i)}{\sum_j \delta(\beta^{\mathrm{T}} x_j)} - 1 \right)^2, \end{aligned} \tag{14}$$

which is equal to J_{TIW} when $p = 1$. ■

Finally, we add the penalty term to the MSD cost function in Equation (6) for a joint optimization:

$$\min_{\beta} J = J_{MSD} + \lambda J_{TIW}, \tag{15}$$

where λ is a tradeoff parameter. Since the cost function is unconstrained, we can use either gradient descent or L-BFGS for optimization. In each iteration, the indices of the optimal top-ranked instance weights were computed according to the results at iteration $(t-1)$. After that, we update the parameters at iteration (t). It is reasonable since we just need to penalize the (previously) top-weighted instances.

5 Experiments

5.1 Experimental Settings and Datasets

We evaluate our approach on the following two tasks.

- **Cross-domain sentiment classification**. The Movie review dataset[1] is used as the source domain, and each domain of Multi-domain sentiment dataset[2] (Book, DVD, Electronics, and Kitchen) serves as the target domain. "A→B" denotes the task where A is the source domain and B is the target domain.
- **Cross-domain personalized spam filtering**. The dataset comes from the ECML/PKDD 2006 discovery challenge[3]. The goal is to adapt a spam filter trained on a common pool of 4000 labeled emails to three individual users' personal inboxes (u00, u01, u02), each containing 2500 emails.

In this work, we used the instance-weighted naïve Bayes model for cross-domain text classification. Unigrams are used as features in instance weight learning. Both unigrams and bigrams with term frequency no less than 4 are used as features for instance weighted classification. The results are reported in terms of the average of 10 random repeats.

5.2 The Performance of Instance Weighting for Domain Adaptation

In this paper we focus on instance weighting for domain adaptation and have not compared our approach with the feature-based domain adaptation algorithms such as SCL [Blitzer *et al.*, 2007] and TCA [Pan *et al.*, 2011]. We implement the following eight instance weighting systems for comparison.

- **No-Adapt** (the standard machine learning method without adaptation);
- **KMM** (kernel mean matching) [Huang *et al.*, 2006];
- **KLIEP** (Kullback-Leibler importance estimation procedure) [Sugiyama *et al.*, 2008];
- **uLSIF** (unconstrained least-squares importance fitting) [Kanamori *et al.*, 2009];
- **DALR** (domain adaptive logistic regression) [Jiang and Zhai, 2007];
- **PUIS** (instance selection via PU learning [Xia *et al.*, 2013];
- **PUIW** (instance weighting via PU learning) [Xia *et al.*, 2013];
- **FW** (Frank-Wolf algorithm for correcting covariate shift) [Wen *et al.*, 2015].

In our approach, the tradeoff parameter λ is set to be 0.01. The percentage parameter p is set as 5% in TIW. We use the L-BFGS algorithm to optimize the parameters, and set the maximum iterations steps as 20. The paired t-test is used for the significance testing.

[1]http://www.cs.cornell.edu/people/pabo/movie-review-data/

[2]https://www.cs.jhu.edu/ mdredze/datasets/sentiment/

[3]http://www.ecmlpkdd2006.org/challenge.html

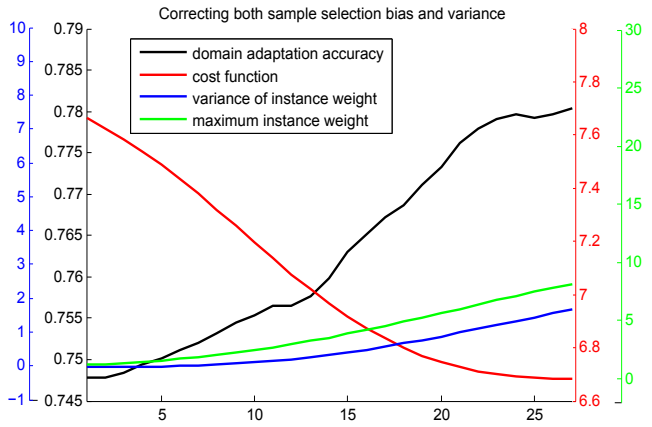

Figure 2: The instance weight learning process when reducing sample selection bias and variance together. The x-axis and y-axis are the same as Figure 1.

In Table 1, we report the domain adaptation performance of nine systems on two tasks. It can be seen that most of the instance weighting approaches outperform No-Adapt. But it is somehow surprising that KMM almost fails in domain adaptation. uLSIF performs significantly better than KLIEP, but the performance is not stable across different datasets. PUIW has a similar performance to uLSIF. Their performance is better than PUIS and DALR in sentiment classification but worse in personalized spam filtering. FW, as a modification to KLIEP, performs slightly better than the latter. Our approach obtains generally the best performance across different settings. The improvements are stable and significant according to the paired t-test.

5.3 The Effect of Trading off Sample Selection Bias and Variance

In Figure 1, it is already shown that when we only correct sample selection bias, the decrease of the cost function does not necessarily increase the cross-domain classification performance. For comparison, in Figure 2 we display the results of trading off sample selection bias and variance. It can be seen that, due to the penalty term that limits the sample selection variance, the instance weight variance and maximum instance weight have been controlled. Consequently, the domain adaptation accuracy increases gradually as the cost function decreases.

In Table 2, we further compare the performances of using different ways to limit the sample selection variance. The results are obtained with L-BFGS, except for Early Stopping which is based on gradient descent. First, it can be seen that only correcting the sample selection bias without considering the variance fails in domain adaptation. It confirms again that limiting sample selection bias is of great necessity. Second, early stopping may be a choice to prevent the sample selection variance from being too large. But it is infeasible in L-BFGS due to the fast convergence speed. The difficulty in determining "when to stop" also makes Early Stopping less efficient in practice. Finally, the three ways to limit sample selection variance (MIW, VIW, TIW) perform significantly

	Cross-domain Sentiment Classification					Cross-domain Personalized Spam Filtering			
	M→B	M→D	M→E	M→K	Avg.	train→u00	train→u01	train→u02	Avg.
No-Adapt	0.756	0.762	0.697	0.709	0.731	0.822	0.876	0.869	0.856
KMM	0.728	0.735	0.669	0.616	0.687	0.823	0.873	0.865	0.854
KLIEP	0.737	0.738	0.673	0.626	0.694	0.839	0.872	0.870	0.860
uLSIF	0.783	0.796	**0.770**	0.747	0.775	0.830	0.881	0.893	0.868
DALR	0.761	0.766	0.731	0.745	0.751	0.847	0.886	0.882	0.872
PUIS	0.757	0.762	0.726	0.743	0.747	0.851	0.893	0.888	0.877
PUIW	0.774	0.782	0.750	0.777	0.771	0.825	0.873	0.861	0.853
FW	0.744	0.758	0.661	0.664	0.707	0.810	0.872	0.852	0.845
Our Approach	**0.785**	**0.798**	0.767	**0.787**	**0.784**	**0.862**	**0.897**	**0.909**	**0.889**

Table 1: Domain adaptation performance of nine instance weighting methods.

Correcting Sample Selection Bias	Correcting Sample Selection Variance	Cross-domain Sentiment Classification				Cross-domain Personalized Spam Filtering		
		M→B	M→D	M→E	M→K	train→u00	train→u01	train→u02
MSD	None	0.501	0.577	0.501	0.501	0.500	0.500	0.866
MSD	Early Stopping	0.780	0.791	0.765	0.780	0.855	0.889	0.900
MSD	MIW	0.784	0.792	**0.768**	0.783	0.835	0.891	0.901
MSD	VIW	0.784	0.794	0.762	0.778	0.842	0.888	0.893
MSD	TIW	**0.785**	**0.798**	0.767	**0.787**	**0.862**	**0.897**	**0.909**

Table 2: Domain adaptation performance of different ways to correct sample selection bias and variance

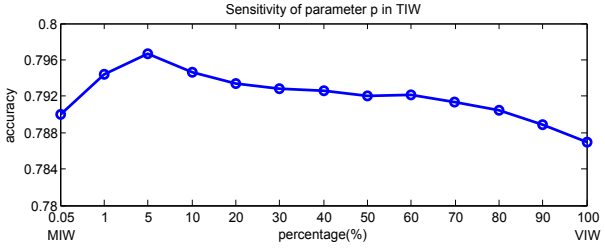

Figure 3: The domain adaptation accuracy curve by tuning p in TIW (M→K Task).

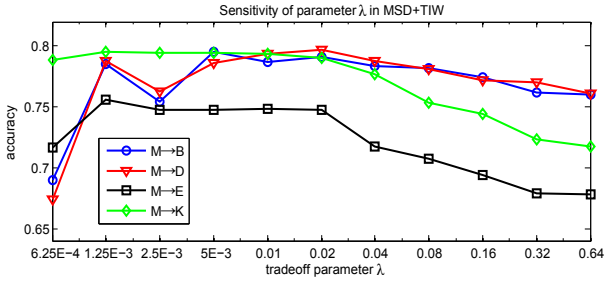

Figure 4: The domain adaptation accuracy curve by tuning the tradeoff parameter λ in MSD+TIW.

better than either ignoring the variance or using Early Stopping to correct it.

5.4 Tuning the Percentage Parameter p in TIW

It has been proven in Section 4.2 that MIW and VIW are special cases of TIW. In Figure 3, we display the domain adaptation accuracy curve by tuning p in TIW on the M→K task. TIW equals to VIW when $p = 1$, and MIW when $p = \frac{1}{N_s}$. We can find the performance is quite stable. TIW can obtain the best performance when p is at [1%, 10%]. In this work, p is set as 5%.

5.5 Tuning the Tradeoff Parameter λ

We further discuss the sensitivity of the tradeoff parameter λ in Equation (14). Figure 4 presents the corresponding results via tuning λ on cross-domain sentiment classification tasks. When $\lambda = 0$, it becomes the MSD cost only; when $\lambda = 1$, it only has the sample selection variance regularizer. Both cases are not effective. It again proves the necessity to correct sample selection bias and variance together. The best performance can be obtained when λ is located at [0.005, 0.02]. It suggests that the MSD and TIW have distinct strength, and an appropriate combination of them is reasonable for domain adaptation.

6 Conclusions

Instance weighting is one kind of domain adaptation method due to its capacity of correcting the bias in the labeled training data. However, in practice the instance weighting methods often fail to achieve good domain adaptation performance in many NLP tasks. In this work, we conduct in-depth discussion on such failure and point out there are two kinds of important factors (i.e., sample selection bias and sample selection variance) in domain adaptation, and only correcting sample selection bias is not enough. On this basis, we propose a new instance weighting framework that jointly limits the sample selection bias and sample selection variance. We evaluate our approach on two NLP domain adaptation tasks and compare it with eight strong instance weighting systems. The experimental results confirm our conjecture and prove our approach's effectiveness and stability in instance weighting methods for domain adaptation.

Although this work can make instance weighting more stable for domain adaptation, a major limitation is that it should be based on condition that supports of features spaces between domains are similar. Instance-based domain adaptation only takes charge of the marginal distribution's transfer $p_s(x) \rightarrow p_t(x)$, but is unable to model the discriminative function's transfer $p_s(y|x) \rightarrow p_t(y|x)$. Another limitation is that the framework of instance-based domain adaptation is a pipeline of two steps where instance weight learning and cross-domain classification are separated. Therefore, a instance/feature joint end-to-end domain adaptation learning framework under neural networks might be a promising direction for future work.

Acknowledgments

The work was supported by the Natural Science Foundation of China (No. 61672288), and the Natural Science Foundation of Jiangsu Province for Excellent Young Scholars (No. BK20160085).

References

[Axelrod *et al.*, 2011] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 355–362, 2011.

[Blitzer *et al.*, 2007] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Annual Conference of Association of Computational Linguistics (ACL)*, pages 440–447, 2007.

[Dudík *et al.*, 2005] Miroslav Dudík, Steven J Phillips, and Robert E Schapire. Correcting sample selection bias in maximum entropy density estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 323–330, 2005.

[Geman *et al.*, 1992] S Geman, E Bienenstock, and R Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[Heckman, 1979] James J Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

[Huang *et al.*, 2006] Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 601–608, 2006.

[Jiang and Zhai, 2007] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 264–271, 2007.

[Jiang, 2008] Jing Jiang. A literature survey on domain adaptation of statistical classifiers. *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey*, 2008.

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

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[Pan *et al.*, 2011] SJ Pan, IW Tsang, JT Kwok, and Q Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[Plank *et al.*, 2014] Barbara Plank, Anders Johannsen, and Anders Søgaard. Importance weighting and unsupervised domain adaptation of pos taggers: a negative result. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–973, 2014.

[Shimodaira, 2000] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[Sugiyama *et al.*, 2008] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1433–1440, 2008.

[Tsuboi *et al.*, 2009] Yuta Tsuboi, Hisashi Kashima, Shohei Hido, Steffen Bickel, and Masashi Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Information and Media Technologies*, 4(2):529–546, 2009.

[Wen *et al.*, 2015] Junfeng Wen, Russell Greiner, and Dale Schuurmans. Correcting covariate shift with the frank-wolfe algorithm. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1010–1016, 2015.

[Xia *et al.*, 2013] Rui Xia, Xuelei Hu, Jianfeng Lu, Jian Yang, and Chengqing Zong. Instance selection and instance weighting for cross-domain sentiment classification via pu learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2176–2182, 2013.

[Xia *et al.*, 2014] Rui Xia, Jianfei Yu, Feng Xu, and Shumei Wang. Instance-based domain adaptation in nlp via in-target-domain logistic approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1600–1606, 2014.

[Zadrozny, 2004] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 114–121, 2004.