# Instance-based Domain Adaptation via Multi-clustering Logistic Approximation

FENG XU, Nanjing University of Science and Technology
JIANFEI YU, Singapore Management University
RUI XIA[1], Nanjing University of Science and Technology
TAO LI, Florida International University

With the explosive growth of the Internet online texts, we could nowadays easily collect a large amount of labeled training data from different source domains. However, a basic assumption in building statistical machine learning models for sentiment analysis is that the training and test data must be drawn from the same distribution. Directly training a statistical model usually results in poor performance, when the training and test data have different distributions. Faced with the massive labeled data from different domains, it is therefore important to identify the source-domain training instances that are closely relevant to the target domain, and make better use of them. In this work, we propose a new approach, called multi-clustering logistic approximation (MLA), to address this problem. In MLA, we adapt the source-domain training data to the target domain via a framework of multi-clustering logistic approximation. Experimental results demonstrate that MLA has significant advantages over the state-of-the-art instance adaptation methods, especially in the scenario of multi-distributional training data.

## 1. INTRODUCTION

With the growing volume of online texts available through the Internet, nowadays we can easily obtain huge amounts of labeled training texts from different domains. But only some of them might be beneficial for training a target-domain-desired classifier for sentiment analysis. A lot of research works have been proposed to conduct domain adaptation for sentiment analysis [Jiang 2008; Xia et al., 2013b] and open-domain opinion mining and sentiment analysis [Cambria et al., 2014; Cambria, 2016].

As a special type of domain adaptation methods in sentiment analysis, instance-based domain adaptation (instance adaptation for short) aims to identify the training samples that are the most relevant to the target domain and make better use of them. Consider the following example: we want to learn a laptop sentiment classifier in the absence of labeled laptop reviews. Instead we can obtain a large set of labeled E-product reviews, which covers reviews from several kinds of E-products including phones, digital cameras, etc. In this case, the performance of a sentiment classifier simply trained with all labeled reviews might be unsatisfactory, since only a few samples in the training data are closely related to laptop. Therefore, methods are needed to identify and adapt the labeled reviews from different source domains (or sub-domains) to obtain a classifier that performs well on the laptop domain.

To address this problem, methods have been developed to conduct instance adaptation by reweighting the training instances and applying importance sampling. In the field of machine learning, instance adaptation was studied under the concepts of “covariate shift”, “sample selection bias”, and “importance sampling”. Most of the previous instance adaptation methods traditionally assume that the source-domain training data have only one underline distribution.

However, in many NLP applications including sentiment analysis, we are usually faced with a large amount of labeled data that might come from multiple source domains. Even if the data come from one source domain, they also might contain many sub-domains with different patterns in distribution. For example, the reviews collected from multiple domains (such as Moive, Book, E-product, etc) are multi-domain training data. Even if the training data might come from one domain, there are still many sub-domains or categories. E.g., the reviews come from the E-product

[1] Rui Xia is the corresponding author of this paper.

domain, they may still cover several smaller sub-domains (such as phones and laptops), and therefore have distinct differences in distribution.

In [Xia et al. 2014], an in-target-domain logistic approximation (ILA) approach has been proposed for single-domain instance adaptation. On this basis, we propose a multi-clustering logistic approximation (MLA) in this work, as an extension to ILA, to deal with domain adaption in case of multi-distributional training data. Moreover, we infer the instance weighting learning criterion based on the multinomial event model, which leads to more profound insights of instance-based domain adaptation. To fully evaluate MLA, we conduct experiments on two tasks including cross-domain sentiment analysis and cross-domain text categorization. The experimental results show that our MLA approach can significantly outperform the state-of-the-art instance adaptation methods, especially in case of multi-distributional training data.

## 2. RELATED WORK

In general, domain adaptation methods include feature-based domain adaptation, parameter-based domain adaptation, and instance-based domain adaptation [Jiang 2008; Pan and Yang 2010]. Note that different methods have different settings. In this work, we focus on instance-based domain adaptation.

Instance adaptation learns the importance of labeled data in the source domain by instance re-weighting and importance sampling. The re-weighed instances are then used for training a target-domain model. In the machine learning community, instance adaptation is also known as the "covariate shift" or "instance selection bias". This concept was first introduced in the field of econometrics by [Heckman 1979], and brought into the field of machine learning by [Zadrozny 2004]. The key problem in instance selection bias is density ratio estimation (DRE).

There was a line of kernel-based methods to solve the DER problem, such as kernel density estimation [Shimodaira, 2000], maximum entropy density estimation [Dudik et al. 2005], kernel mean matching [Huang et al. 2007], etc. [Sugiyama et al. 2007] proposed a KLIEP algorithm to directly estimate the density ratio by using a linear model in a Gaussian kernel space. Parameters were learnt by minimizing the K-L divergence between the true and approximated distributions. The following work included [Tsuboi et al. 2008], [Kanamori et al. 2009], etc.

[Bickel et al. 2007; Bickel et al. 2009] utilized a logistic regression model to learn the density ratio together with the classification parameters, under the multi-task learning framework. In the field of NLP, [Xia et al. 2013] proposed an instance selection and instance weighting approach via PU learning (PUIS and PUIW) for the task of cross-domain sentiment classification, [Xia et al. 2014] proposed an instance-based domain adaptation in NLP via in-target-domain logistic approximation.

As mentioned above, in many applications in NLP, the source-domain training data may come from many sub-domains and have different distributions. While most of the previous work conducted domain adaptation in case of single-distributional training data. By contrast, in this work, we proposed a multi-clustering logistic regression model, to address this issue.

## 3. MULTI-CLUSTERING LOGISTIC APPROXIMATION

### 3.1 Logistic approximation for single source-domain instance adaptation

[Xia et al. 2014] have introduced in-target-domain logistic approximation (ILA) for instance adaptation from a single source domain. They assumed that the target-domain data are generated as follows:

- An instance $\mathbf{x}$ is first drawn from the source domain distribution $p_s(\mathbf{x})$;

- An in-target-domain selector $p(d = 1|x) = \frac{1}{1+e^{-\beta^T x}}$ then adapts $x$ from the source to the target domain, where $d$ denotes a domain indicator ($d = 1$ represents target domain, and $d = 0$ represents source domain).

The approximated target-domain distribution $q_t(x)$ was then be formulated as $q_t(x) = p(d = 1|x)p_s(x) = \frac{\alpha}{1 + e^{-\beta^T x}} p_s(x)$.

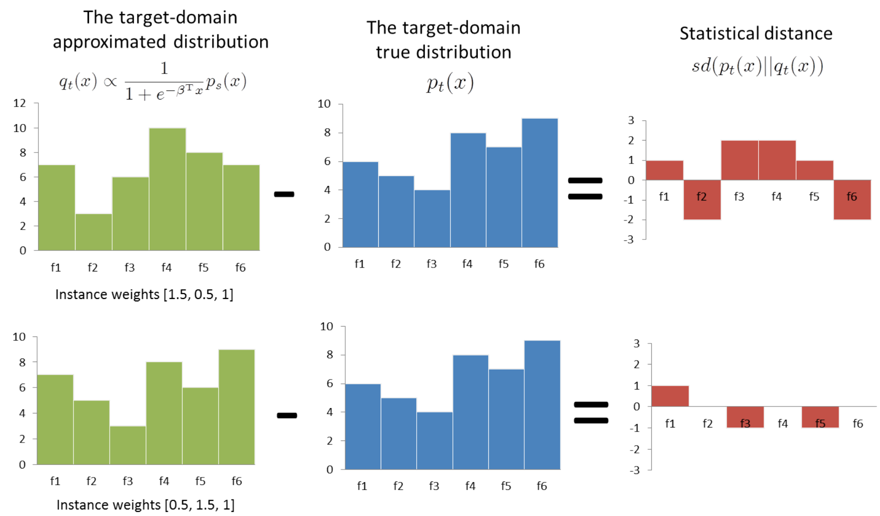


Fig. 1: An illustration of the minimal statistical distance criterion for instance weight learning in ILA.

The normalized in-target-domain probability $w(x) = \frac{\alpha}{1 + e^{-\beta^T x}}$ was used as the instance weight for training a weighted classification model after instance adaptation. Different instance weights can yield different target-domain approximated distributions. The instance weight can be learnt by minimizing the statistical distance (such as K-L distance) between the target-domain true distribution $p_t(x)$ and the approximated distribution $q_t(x)$.

In Fig. 1, we give an illustration of the minimal statistical distance criterion for instance weight learning. In the left column there are two approximated target-domain distributions that are generated by sampling the source-domain training data with different instance weights. We subtract the target-domain true distribution (in the middle column) from the approximated one, and obtain the distributional distance (in the right column). Obviously, the instances weights that yield the minimal distributional distance (i.e., the second row) are preferred for instance adaptation.

However, when the training data contain many sub-domains and have distinct distributions, ILA might lead to poor adaptation performance, especially when the gap between sub-domains is large.

Let us use an artificial example to illustrate the motivation. In Fig. 2, the red dots denote the target-domain test data. The blue and black crosses denote the training data, which are drawn from two different distributions. In instance adaptation, we learn an in-target-domain selector and assign an in-target-domain probability as the weight to each training instance. The size of the cross denotes the weight. The crosses that are closer to the separating line will have larger weights. Fig. 2 illustrates how ILA conducts instance adaptation. It treats the multi-distributional training data as a single domain, and learns a single in-target-domain separating line. The crosses that are closer to the separating line will have larger weights.

### 3.2 Multi-clustering Logistic Approximation

Unlike the standard ILA, in this work we propose a multi-clustering logistic approximation (MLA) to address this issue. Fig. 3 shows how MLA works in this case. It first clusters the training data into several categories and treats each category as a single domain. Then, it learns the in-target-domain separating lines for each cluster, and conduct instance adaptation respectively. A weighted combination of the instance weights learnt in each cluster is finally utilized as the overall instance weight. In comparison with ILA, MLA can capture various patterns in the massive training data. Therefore, it would be more reasonable in case of multi-domain or multi-distributional instance adaptation.

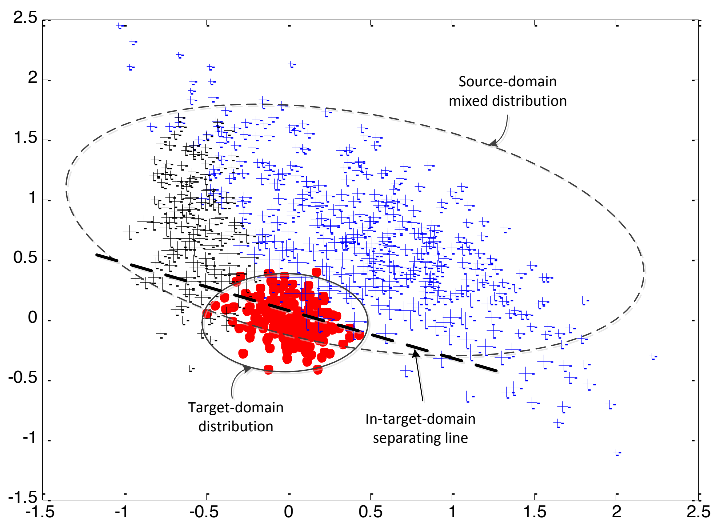


Fig. 2: An illustration of ILA instance adaptation for multi-distributional training data.

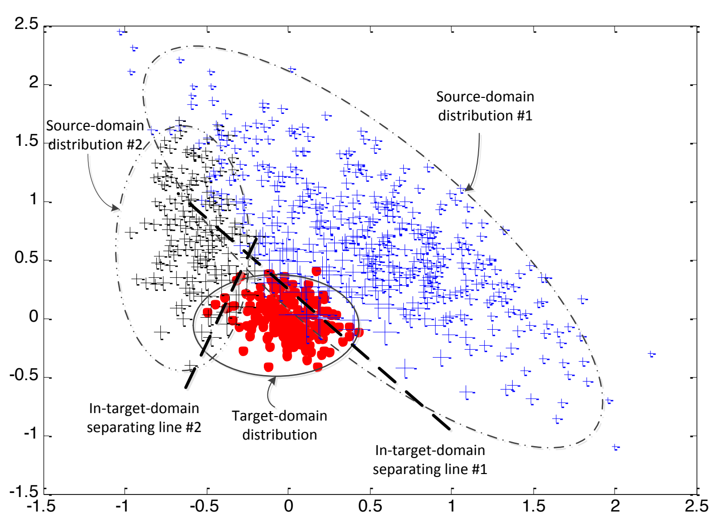


Fig. 3: An illustration of MLA instance adaptation for multi-distributional training data.

We first apply a clustering algorithm to cluster the training data into several sub-domains. Thereafter, we develop a multi-clustering in-target-domain logistic approximation model, to adapt data from different clusters, in the manner of a weighted combination.

In MLA, the $k$-means clustering algorithm is employed to split the source domain data into $k$ sub-domains. The cosine distance is used as the similarity measure. After clustering, the cosine distance between different instances within a domain is small, while the distance between different domains is large.

Based on the $k$ clusters of training data, we propose to conduct multi-clustering logistic approximation. Let $p_{s_d}(x)$ be the distribution of the $d$-th cluster of the training data, and $q_{t_d}(x)$ be the (component) target-domain approximated distribution adapted from $p_{s_d}(x)$. We suppose the target-domain approximated distribution $q_t(x)$ in MLA is a weighted sum of $q_{t_d}(x)$:

$$q_t(x) = \sum_{d=1}^{k} \lambda_d q_{t_d}(x) = \sum_{d=1}^{k} \lambda_d \frac{\alpha_d}{1+e^{-\beta_d^T x}} p_{s_d}(x) \tag{1}$$

where $\alpha_d$ and $\beta_d$ are respectively the normalization factor and feature weight of the $d$-th cluster, and $\lambda_d$ is a tradeoff parameter controlling the importance of each cluster. The weighted ensemble of the normalized in-target-domain probabilities

$$w(x) = \sum_{d=1}^{K} \lambda_d \frac{\alpha_d}{1+e^{-\beta_d^T x}} \tag{2}$$

will be used as sampling weights for training an weighted classification model.

Under this assumption, the K-L distance between $p_t(x)$ and $q_t(x)$ can be written approximately as

$$\begin{aligned} KL(p_t||q_t) &\approx \frac{1}{N_t} \sum_{d=1}^{K} \lambda_d \sum_{i=1}^{N_t} \log \frac{p_t(x)}{q_{t_d}(x)} \\ &= \frac{1}{N_t} \sum_{d=1}^{K} \lambda_d \sum_{i=1}^{N_t} \log \frac{\prod_{i=1}^{V} \theta_i^{x_{ik}}}{\frac{\alpha_d}{1+e^{-\beta_d^T x_i}} \prod_{i=1}^{V} \phi_{di}^{x_{ik}}} \\ &= \frac{1}{N_t} \sum_{d=1}^{K} \lambda_d \sum_{i=1}^{N_t} \log \frac{\prod_{i=1}^{V} \theta_i^{x_{ik}}}{\prod_{i=1}^{V} \phi_{di}^{x_{ik}}} - \frac{1}{N_t} \sum_{d}^{K} \lambda_d \sum_{i=1}^{N_t} \log \frac{\alpha_d}{1+e^{-\beta_d^T x_i}} \end{aligned} \tag{3}$$

The parameters are learnt by minimizing $KL(p_t||q_t)$ subject to the normalization constraint

$$\begin{aligned} &\min_{\alpha_d;\beta_d} KL(p_t||q_t) = \min_{\alpha_d;\beta_d} -\frac{1}{N_t} \sum_{d}^{K} \lambda_d \sum_{i=1}^{N_t} \log \frac{\alpha_d}{1+e^{-\beta_d^T x_i}} \\ &s.t. \int_{x \in X} q_{t_d}(x)dx = \frac{1}{N_d} \sum_{j=1}^{N_d} \frac{\alpha_d}{1+e^{-\beta_d^T x_j}} = 1; d = 1; \ldots; k \end{aligned} \tag{4}$$

where $N_d$ is the number of training instances in the $d$-th cluster.

Similarly, we get the final cost function by solving $\beta_d$ analytically and plugging it back:

$$
\begin{aligned}
J_{mila} &= -\frac{1}{N_t}\sum_{d=1}^{K}\alpha_d\sum_{i=1}^{N_t}\log\frac{\frac{N_d}{\sum_{j=1}^{N_d}\frac{1}{1+e^{-\theta_d^T x'_j}}}}{1+e^{-\theta_d^T x_i}} \\
&= \frac{1}{N_t}\sum_{d=1}^{K}\alpha_d\sum_{i=1}^{N_t}\left[\log(1+e^{-\theta_d^T x_i})+\log\sum_{j=1}^{N_d}\frac{1}{1+e^{-\theta_d^T x'_j}}\right].
\end{aligned}
\tag{5}
$$

For each clustering, once the parameter $\beta_d$ and $\theta_d$ is learnt, we could calculate the instance weights according to Equation (2). Based on this, we train an instance-weighted classifier for the cross-domain classification task.

## 4. EXPERIMENTAL STUDY

### 4.1 Tasks, Datasets and Experimental Settings

To fully evaluate the performance of MLA, we conduct experiments on two NLP tasks: 1) cross-domain text categorization; 2) cross-domain sentiment classification.

For cross-domain text categorization, we use the 20 Newsgroups dataset for experiments. It contains seven top categories, under which there are 20 subcategories. We use four top categories as the class labels, and generate source and target domains based on subcategories. Taking “sci vs talk” as an example, the top categories “sci” and “talk” are the class labels. The subcategories “med”, ”elec” (under category “sci”) and “ guns”, ”mideast” (under category “talk”) are used as the multi-distributional training data. Subcategories “crypt” (under category “sci”) and “misc” (under category “talk”) are used as the target-domain test data.

For cross-domain sentiment classification, we randomly choose two domains from four Multi-domain sentiment datasets as the source domain, and choose one from the remaining two domains as the target domain. For example, “book + kitchen → dvd” represents that “book” and “kitchen” are the two sub-domains in the source domain, and “dvd” is the target domain. It is worth pointing out that here, we only present the result of MLA when the number of sub-domains is two. Similar behaviors can be observed when the number of sub-domains increases.

In both tasks, unigrams and bigrams with term frequency no less than four are used as features for classification. We randomly repeat the experiments for 10 times, and report the average results. The paired *t*-test [Yang and Liu 1999] is employed for significance testing, with a default significant level of 0.05.

### 4.2 The comparison of system performance

The following systems are implemented for comparison:

- **No-Adaptation**: the standard approach using all training data without domain adaptation;
- **KLIEP-Gaussian**: the KLIEP model with a Gaussian kernel [Sugiyama et al. 2007];
- **PUIW**: the instance weighting model via PU learning [Xia et al. 2013b].
- **ILA**: the standard in-target-domain logistic approximation algorithm [Xia et al. 2014]

We compare the system performance on two tasks respectively.

***Text categorization***: In Table 1, we can observe that in comparison with the No-Adaptation system, the improvements of KLIEP-Gaussian and PUIW are very slight.

ILA is a bit more effective. But the effect is quite limited compared with its performance in single-domain instance adaptation [Xia et al., 2014]. In contrast, MLA outperforms No-Adaptation, KLIEP-Gaussian, PUIW and ILA significantly.

***Sentiment classification***: In comparison with No-adaptation, three instance adaptation methods (KLIEP-Gaussian, PUIW and ILA) exhibit effective performance. ILA does not show significant superiority when compared with KLIEP-Gaussian and PUIW. By contrast, MLA outperforms all the other methods significantly.

In general, ILA does not yield significant improvements in comparison with KLIEP-Gaussian and PUIW, in multi-distributional case. It suggests that the effect of ILA is limited for multi-distributional instance adaptation. By contrast, the MLA algorithm is rather effective under this setting.

Table 1. Multi-domain instance adaptation performance of different systems on text categorization

| **Dataset** | **No adaptation** | **KLIEP Gaussian** | **PUIW** | **ILA** | **MLA** |
|---|---|---|---|---|---|
| talk vs rec | 0.809 | 0.821 | 0.826 | 0.851 | **0.857** |
| talk vs com | 0.952 | **0.961** | 0.954 | 0.958 | **0.961** |
| sci vs talk | 0.720 | 0.728 | 0.725 | 0.727 | **0.735** |
| sci vs com | 0.727 | 0.718 | 0.728 | 0.725 | **0.738** |
| rec vs com | 0.880 | 0.883 | 0.885 | **0.890** | 0.887 |
| rec vs sci | 0.770 | 0.770 | 0.784 | 0.801 | **0.807** |
| Average | 0.809 | 0.814 | 0.817 | 0.825 | **0.831** |

Table 2. Multi-domain instance adaptation performance of different systems on sentiment classification

| **Dataset** | **No adaptation** | **KLIEP Gaussian** | **PUIW** | **ILA** | **MLA** |
|---|---|---|---|---|---|
| book + dvd → kitchen | 0.797 | 0.809 | 0.802 | 0.810 | **0.818** |
| book + dvd → elec | 0.733 | 0.782 | 0.772 | 0.778 | **0.791** |
| book + elec → kitchen | 0.84 | 0.846 | 0.847 | 0.846 | **0.854** |
| book + elec → dvd | 0.801 | 0.807 | 0.811 | 0.811 | **0.818** |
| book + kitchen → dvd | 0.81 | 0.811 | 0.815 | 0.816 | **0.817** |
| book + kitchen → elec | 0.822 | 0.828 | 0.833 | 0.827 | **0.843** |
| dvd + elec → kitchen | 0.85 | 0.851 | 0.852 | 0.853 | **0.854** |
| dvd + elec → book | 0.789 | 0.79 | 0.798 | 0.798 | **0.800** |
| dvd + kitchen → book | 0.798 | 0.802 | 0.799 | 0.780 | **0.820** |
| dvd + kitchen → elec | 0.814 | 0.820 | 0.819 | 0.814 | **0.842** |
| elec + kitchen → dvd | 0.781 | 0.783 | 0.782 | 0.783 | **0.787** |
| elec + kitchen → book | 0.753 | 0.756 | 0.758 | 0.752 | **0.763** |
| Average | 0.799 | 0.807 | 0.807 | 0.806 | **0.818** |

### 4.3 Parameter stability of the clustering weight

In this part, we discuss the sensitivity of the domain-based tradeoff parameter ¥$_d$ in MLA. In Fig. 4 and Fig. 5, we present the results of eight datasets in two tasks. It can be seen that all the accuracy curves are bimodal and stable. When ¥$_d$ is relatively small, i.e., ¥$_d$ < 0:5, the best accuracy is obtained around 0.1 to 0.4. When ¥$_d$ becomes larger, the best accuracy is obtained when ¥$_d$ is located at 0.7 to 0.9. It suggests that our approach MLA inclines to choose more target-domain-likely samples from one domain, but less from the others.

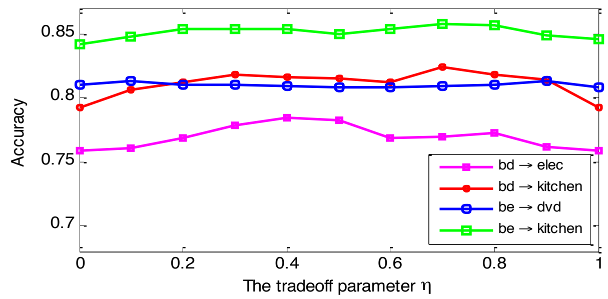


Fig. 4: Sensitivity of the tradeoff parameter in MILA in sentiment classification.

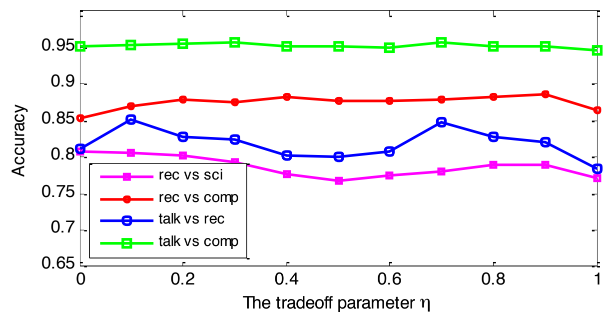


Fig. 5: Sensitivity of the tradeoff parameter in MILA in text categorization.

### 4.4 The relation between K-L distance and instance adaptation performance

We further investigate the relation between the K-L distances and domain adaptation performance. In Table 3, we report three kind of K-L distance between the data of source and target domain respectively:

1) KLD-1 represents the K-L distance between source domain #1 and the target domain;
2) KLD-2 represents the K-L distance between source domain #2 and the target domain;
3) KLD-3 represents the K-L distance between the source domains.

Table 3. Domain adaptation performance of different systems on sentiment classification

| Dataset | KLD-1 | KLD-2 | KLD-3 | No adaptation | ILA | MLA |
|---|---|---|---|---|---|---|
| book + dvd → kitchen | 31.48 | 31.98 | 19.15 | 0.797 | 0.810 | **0.818** |
| book + dvd → elec | 31.44 | 31.59 | 19.66 | 0.733 | 0.778 | **0.791** |
| book + elec → kitchen | 35.38 | 30.33 | 74.32 | 0.840 | 0.846 | **0.854** |
| book + elec → dvd | 18.07 | 60.88 | 73.32 | 0.801 | 0.811 | **0.818** |
| book + kitchen → dvd | 17.35 | 16.26 | 80.28 | 0.810 | 0.816 | **0.817** |
| book + kitchen → elec | 33.24 | 34.28 | 81.76 | 0.822 | 0.827 | **0.843** |
| dvd + elec → kitchen | 36.83 | 30.24 | 73.65 | 0.850 | 0.853 | **0.854** |
| dvd + elec → book | 56.54 | 60.52 | 71.15 | 0.789 | 0.798 | **0.800** |
| dvd + kitchen → book | 19.99 | 101.24 | 51.14 | 0.798 | 0.780 | **0.820** |
| dvd + kitchen → elec | 32.87 | 56.42 | 54.73 | 0.814 | 0.814 | **0.842** |
| elec + kitchen → dvd | 46.44 | 52.96 | 49.10 | 0.781 | 0.783 | **0.787** |
| elec + kitchen → book | 49.28 | 54.44 | 51.40 | 0.753 | 0.752 | **0.763** |

Taking “book + kitchen → dvd” for example, KLD-1 is the K-L distance beween “book” and “dvd”, KLD-2 is the K-L distance between “kitchen” and “dvd”, and KLD-3 is the K-L distance between “book” and “kitchen”.

First, it can be seen that when the K-L distance between the two source domains is small (KLD-3<10), the improvement of MLA is also limited. It shows that when the K-L distance between source domains is small, ILA and MLA exhibit competitive performance. The reason is that when KLD-3 is relatively small, i.e., the two source sub-domains have similar distributional difference.

Second, when the K-L distance between the two source domains is relatively large (KLD-3>40), we can observe MLA perform much better than ILA. Compared to the ILA method, the average improvements on the last ten subtasks are 1.3%. But it is also worth noting that when KLD-1 and KLD-2 are close to each other, the improvement is less than 1%, such as “book + kitchen □ dvd” and “dvd + electronics □ kitchen”; by contrast, if the difference between KLD-1 and KLD-2 is large, the improvement is more significant, such as “dvd + kitchen □ book” and “dvd + kitchen □ elec”. The reason is that if the distance between the distribution of the sub-domains and target domain is large, ILA model will assign large weights to few samples, while the weights of other samples in the source domain are close to zero. This might lead to the over-fitting of domain adaptation. However, MLA model could avoid the over-fitting problem by tuning the parameter assign large weights to more samples from the source domain whose distribution is much more similar to the target domain.

This result confirms our motivation of MLA very well. When KLD-3 is small, i.e., the distributions of the two source domains are similar, the effects of ILA (viewing the two distributions as a whole distribution, and choosing samples from it) and MLA (choosing samples from the two distributions respectively) are similar. When KLD-3 is large, it means the difference between the two distributions is large. For MLA, it could tune the parameter to pay more attention to the source domain whose distribution is much more similar to the target domain, and assign large weights to more samples from this source domain.

## 5. CONCLUSIONS

The traditional instance adaptation methods including in-target-domain logistic approximation (ILA) normally conducted domain adaptation for single source domain. In this work, a multi-clustering logistic approximation (MLA) model is proposed, to conduct instance adaptation for multi-distributional labelled training data, where the training data might come from many sub-domains. MLA extends the ILA algorithm and is more suitable and more efficient in the multi-distributional case. We conduct systematic experiments on the tasks of cross-domain sentiment classification and text categorization. The results indicate that MLA has significant advantages over traditional instance adaptation methods, especially when the gap between each sub-domain in the training data is large.

## ACKNOWLEDGMENTS

This 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). Rui Xia is the corresponding author of this paper.

## REFERENCES

A. Axelrod, X. He and J. Gao. 2011. Domain Adaptation via Pseudo In-target-domain data selection. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 355–362, 2011.

S. Bickel, M. Brückner, and T. Scheffer. 2007. Discriminative learning for differing training and test distributions. *Proceedings of the International Conference on Machine Learning (ICML)*, 2007.

S. Bikckle, M. Brückner, and T. Scheffer. 2009. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10:2137–2155, 2009.

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 440–447, 2007.

E. Cambria, Y. Song, H. Wang, and N. Howard. 2014. Semantic multidimensional scaling for open-domain sentiment analysis[J]. *IEEE Intelligent Systems*, 2014, 29(2): 44-51.

E. Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems* 31(2), pp. 102-107.

C. Chelba and A. Acero. 2006. Adaptation of Maximum Entropy Capitalizer: Little Data Can Help a Lot. *Computer Speech & Language*, 20(4): 382–399, 2006.

M. Dudik, R. Schapire, and S. Phillips. 2005. Correcting sample selection bias in maximum entropy density estimation. *Advances in Neural Information Processing Systems (NIPS)*, 2005.

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

J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. 2007. Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems (NIPS)*, 2007.

J. Jiang and C. Zhai. Instance weighting for domain adaptation in NLP. 2007. *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 264–271, 2007.

J. Jiang. A Literature Survey on Domain Adaptation of Statistical Classifiers. http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/da_survey.pdf, 2008.

T. Kanamori, S. Hido, and M. Sugiyama. 2009. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, vol.10, pp.1391-1445, 2009.

J. Pan and Q. Yang. A survey on transfer learning. 2010. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

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

M. Sugiyama, S. Nakajima, H. Kashima, P. von Bunau, and M. Kawanabe. 2007. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in Neural Information Processing Systems (NIPS)*, pages 1433–1440, 2007.

Y. Tsuboi, H. Kashima,S. Hido, S. Bickel, and M. Sugiyama. 2009. Direct Density Ratio Estimation for Large-scale Covariate Shift Adaptation. *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 2009.

R. Xia, X. Hu, J. Lu, J. Yang, and C. Zong. 2013a. Instance Selection and Instance Weighting for Cross-domain Sentiment Classification via PU Learning. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2013, pp. 2276–2182.

R. Xia, C. Zong, X. Hu, and E. Cambria. 2013b. Feature Ensemble plus Sample Selection: Domain Adaptation for Sentiment Classification. *IEEE Intelligent Systems*, 2013, volume 28, issue 3, pages 10–18.

R. Xia, J. Yu, F. Xu, and S. Wang. 2014. Instance-based Domain Adaptation in NLP via In-target-domain Logistic Approximation. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1600–1606, 2014.

Y. Yang and X. Liu. 1999. A re-examination of text categorization methods. *Proceedings of the Annual International SIGIR conference (SIGIR)*, 1999.

B. Zadrozny. 2004. Learning and evaluating classifiers under sample selection bias. *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.