Transfer Learning with Active Queries from Source Domain*

Sheng-Jun Huang and **Songcan Chen**
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, 211106
{huangsj, s.chen}@nuaa.edu.cn

Abstract

To learn with limited labeled data, active learning tries to query more labels from an oracle, while transfer learning tries to utilize the labeled data from a related source domain. However, in many real cases, there is very few labeled data in both source and target domains, and the oracle is unavailable in the target domain. To solve this practical yet rarely studied problem, in this paper, we jointly perform transfer learning and active learning by querying the most valuable information from the source domain. The computation of importance weights for domain adaptation and the instance selection for active queries are integrated into one unified framework based on distribution matching, which is further solved with alternating optimization. The effectiveness of the proposed method is validated by experiments on 15 datasets for sentiment analysis and text categorization.

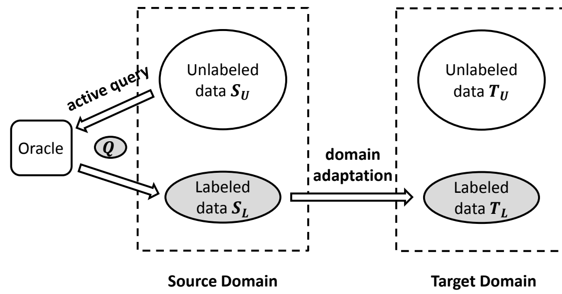


Figure 1: Problem setting: labeled data is insufficient in both domains, and oracle is available only in the source domain, from which a small batch of unlabeled data is iteratively selected to query their labels. The objective is to train an effective model for the target domain with least queries.

1 Introduction

In many applications, we have plenty of unlabeled data but very limited labeled data, making the learning task rather difficult. Transfer learning and active learning are two important approaches to overcoming this challenge. The former tries to utilize data from a related source domain [Pan and Yang, 2010]; while the latter tries to query labels for the most valuable unlabeled data from an oracle [Settles, 2009]. In transfer learning, information is transferred from source domain to target domain at feature level [Gong *et al.*, 2013; Tan *et al.*, 2015] or instance level [Sugiyama *et al.*, 2008; Xiao and Guo, 2015]. In active learning, unlabeled instances are actively queried based on informativeness or representativeness [Huang *et al.*, 2014].

In recent years, there are some studies try to combine the transfer learning and active learning to learn with limited labeled data, either in separating stages [Li *et al.*, 2013; Saha *et al.*, 2011] or in one unified framework [Wang *et al.*, 2014; Kale *et al.*, 2015]. A common assumption of existing methods is that there are plenty of labeled data in the source domain and labels can be further queried in the target domain. However, such an assumption does not always hold. In many real tasks, label acquisition is expensive in both source domain and target domain, and thus the labeled data is usually insufficient in both domains. Furthermore, we even cannot get any additional labeled data in the target domain because the oracle is available only in the source domain. For example, in the influenza prediction task, we want to make prediction for a new strain of flu (target domain) by transferring knowledge from a known flu strain (source domain). At the beginning stage, we may not be able to precisely diagnose patients infected by the new flu, i.e., cannot acquire labels from the target domain. Although there are experts for the known flu strain, the diagnosing process could be time consuming and expensive, we thus need to actively select and diagnose a small number of patients which are most helpful for predicting the new flu. Another case is that the data from target domain contains sensitive private information and thus cannot be posted to annotators for labeling. Instead, we can actively query informative labels from a related yet non-private domain. In summary, we consider the setting where labeled data is insufficient in both source and target domains, and no oracle is available in the target domain. The problem setting is summarized in Figure 1. To the best of our knowledge, this problem has not been studied before.

*This research was supported by NSFC (61503182, 61472186, 61572252), JiangsuSF (BK20150754, BK20150755), CCFTencent Open Research Fund and Jiangsu QingLan Project.

In this paper, we try to address this problem by jointly performing transfer learning and active learning with queries from source domain. On one hand, to compute the importance weight for domain adaptation, we minimize the distance between the distributions of the target domain and adapted source domain. On the other hand, to select the most valuable instances from source domain for label querying, we minimize the distance between distributions of labeled and unlabeled data. These two objectives are integrated into one unified framework, where the distribution distance is estimated with Maximum Mean Discrepancy (MMD). To utilize the supervised information for better active selection, we further incorporate an uncertainty term based on the model prediction. At last, the framework is implemented and optimized with alternating quadratic programming. We test our approach for sentiment analysis on Amazon product reviews and text categorization on Reuters-21578. Results on 15 datasets validated the effectiveness of the proposed approach.

2 Related Work

In recent years, there have been increasing interests in combining transfer learning with active learning to deal with tasks with insufficient labeled data. However, they usually assume that there are plenty of labeled data in the source domain, and perform active queries only in the target domain.

Many approaches perform transfer learning and active learning separately. The approach proposed in [Shi *et al.*, 2008] builds a classifier in the source domain to predict labels for the target domain, and queries the oracle only if the prediction is of low confidence. In [Li *et al.*, 2013], two individual classifiers are trained with labeled data from the source and target domains respectively, and then informative samples are selected from the target domain based on the Query By Committee (QBC) strategy. The method in [Saha *et al.*, 2011] builds a domain separator to distinguish between source and target domain data, and uses this separator to avoid querying labels for those target domain examples that are similar to examples from the source domain. Similar idea is implemented in another work [Rai *et al.*, 2010].

There are also some studies combining the two tasks in one framework. The method in [Wang *et al.*, 2014] relaxes the assumption to allow changes in both marginal and conditional distributions but assumes the changes are smooth between source and target domains. The authors incorporate active learning and transfer learning into a Gaussian Process based approach, and sequentially select query points from the target domain based on the predictive covariance. Kale and Liu [2013] present a principled framework to combine the agnostic active learning algorithm with transfer learning, and utilize labeled data from source domain to improve the performance of an active learner in the target domain. Kale et al. [2015] propose a hierarchical framework to exploit cluster structure shared between different domains, which is further utilized for both imputing labels for unlabeled data and selecting active queries in the target domain.

Xiao and Guo [Xiao and Guo, 2013] study the active transfer learning problem under the online and multi-view setting, where instances are assumed to have multiple feature views, and arrive online in pairs, one from source domain and one from target domain. The method selects one of them with a fixed probability and decides whether to query its label based on multi-view disagreement or uncertainty.

The JO-TAL method proposed in [Chattopadhyay *et al.*, 2013] is most related to our work. It also jointly performs transfer learning and active learning, and employs MMD to measure the distribution distance. However, it is significantly different from the proposed approach with the following reasons. Firstly, JO-TAL assumes that all source domain data are labeled, and tries to query more labels from the target domain, while in our approach, labeled data is insufficient in both domains and we need to query most valuable labels from source domain. Secondly, JO-TAL only minimizes the distribution distance between labeled and unlabeled data, while in our approach, we match the distributions of source and target data as well as labeled and unlabeled data, emphasizing the objectives of both transfer learning and active learning. Moreover, our approach explicitly incorporates the model prediction to further enhance the active selection with uncertainty. At last, different optimization techniques are used.

Transfer learning and active learning have been incorporated for various applications, such as cross-system recommendation [Zhao *et al.*, 2013], natural language parsing [Attardi *et al.*, 2013] and sentiment analysis [Luo *et al.*, 2012]. Theoretical analysis is also presented in [Yang *et al.*, 2013] with an upper bound on the sample complexity in sequential transfer learning settings.

3 The Method

We denote by $S = S_L \cup S_U$ the dataset in the source domain, where $S_L = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_{n_{S_L}}, y_{n_{S_L}})\}$ is the labeled set consisting of n_{S_L} instances, $S_U = \{\mathbf{x}_1, \cdots, \mathbf{x}_{n_{S_U}}\}$ is the unlabeled set consisting of n_{S_U} instances, and $n_S = n_{S_L} + n_{S_U}$. Similarly, the dataset in the target domain is denoted by $T = T_L \cup T_U$, with n_{T_L} labeled instances in T_L and n_{T_U} unlabeled instances in T_U, and $n_T = n_{T_L} + n_{T_U}$. It is assumed that $n_{S_L} \ll n_{S_U}$ and $n_{T_L} \ll n_{T_U}$, i.e., labeled data is insufficient in both source and target domains. We also assume that the oracle is available only in the source domain. This implies that the labeled data in the target domain is fixed in the whole learning procedure, and we need to actively query some informative labels from the source domain. This setting seems restricted, yet is common and practical as discussed in Section 1.

In this paper, we consider the covariate shift setting, where the margin distribution $P(\mathbf{x})$ is different in the source and the target domains, while the conditional distribution $P(y|\mathbf{x})$ is the same. It is well known that the key issue for covariate shift adaptation is to accurately estimate the importance weight for each instance $\mathbf{x}$, which is defined as $\beta(\mathbf{x}) = \frac{p_T(\mathbf{x})}{p_S(\mathbf{x})}$. Here p_T and p_S denote the density functions of target and source domains, respectively. To avoid the density estimation, we can directly optimize the importance weights by minimizing the distance between the distributions of the target domain and adapted source domain. Here we employ Maximum Mean Discrepancy (MMD) [Gretton *et al.*, 2006; Borgwardt *et al.*, 2006] as the criterion to estimate the distance between different distributions. Specifically, the empirical estimate of MMD between the target domain and the adapted source domain can be written as:

$$MMD(\hat{S},T) = ||\frac{1}{n_S}\sum_{\mathbf{x}\in S}\beta(\mathbf{x})\phi(\mathbf{x}) - \frac{1}{n_T}\sum_{\mathbf{x}\in T}\phi(\mathbf{x})||_{\mathcal{H}}, \tag{1}$$

where $\hat{S} = \{\beta(\mathbf{x})\mathbf{x} \mid \mathbf{x} \in S\}$ is the set of adapted source domain data, and $\phi : \mathcal{X} \to \mathcal{H}$ is a mapping from the feature space to a Reproducing Kernel Hilbert Space (RKHS). It is easy to observe that the MMD is actually measured with the distance between the means of the two samples mapped into a RKHS. The target of domain adaptation is to optimize the importance weights β by minimizing Eq. (1).

As discussed previously, there is few labeled data in both target and source domains, and more labels should be queried from an oracle in the source domain. Because the label acquisition could be very expensive, we need to actively select as few unlabeled examples as possible from the source domain for label querying. The active selection criterion should favor instances which are most helpful on improving the classification model in the target domain. This is essentially different from traditional active learning, which selects instances to improve the model in the same domain. It has been validated by previous research that margin distribution matching is an effective approach for active selection [Chattopadhyay *et al.*, 2012]. The basic idea is that after the label querying, the distributions of labeled data and unlabeled data should be close, such that the model trained will have good generalization ability. It is worth noticing that in our setting, the model is trained for predicting unseen instances in the target domain, while the queried instances along with existing labeled data are distributed in both domains. This implies that the importance weights for domain adaptation should be considered when performing distribution matching, consequently, making the active selection more challenging.

Formally, at each iteration of active learning, we select a small subset Q of size n_Q from S_U to query their labels. A vector $\alpha = \{0,\ 1\}^{n_{S_U}}$ is introduced to identify which instances are selected, where $\alpha(\mathbf{x}) = 1$ indicates the instance $\mathbf{x}$ in S_U is selected for query. In other words, we have $Q = \{\mathbf{x} \mid \mathbf{x} \in S_U, \alpha(\mathbf{x}) = 1\}$. Again, MMD is used to measure the distance between two distributions, and the following measurement should be minimized:

$$MMD(\hat{S}_L \cup \hat{Q} \cup T_L,\ T_U \cup \hat{S}_U). \tag{2}$$

Note that the labeled set consists of three parts: labeled data in the target domain T_L, labeled data in the source domain $\hat{S}_L$ and the queried data from source domain $\hat{Q}$. Here the symbol $\hat{}$ represents a data set adapted with importance weights. For example, $\hat{Q} = \{\beta(\mathbf{x})\mathbf{x} \mid \mathbf{x} \in Q\}$.

Noticing that MMD measures the distance between margin distributions, which means the label information is neglected during the active selection. We can thus incorporate an uncertainty term to further improve the active selection. Specifically, we first get the predictions of the current classification model g on all instances in S_U, denoted by g_{S_U}. Then the certainty of an instance $\mathbf{x}$ is simply estimated with $|g(\mathbf{x})|$, indicating that an instance with a prediction value closer to zero is more uncertain. Our target is to select a small batch of instances with larger uncertainty in the target domain. In other words, α should be optimized to achieve a minimal value on $\alpha\beta|g_{S_U}|$.

By combining the objectives for the domain adaption, the margin distribution matching based active selection and the uncertainty based active selection all together, we have the following framework for Transfer Learning with Active queries from Source domain (TLAS):

$$\begin{aligned}\min & MMD(\hat{S},T) + MMD(\hat{S}_L \cup \hat{Q} \cup T_L,\ T_U \cup \hat{S}_U)\\ &+ \lambda\alpha\beta|g_{S_U}|,\end{aligned} \tag{3}$$

where λ is a tradeoff parameter for balancing the contributions of distribution matching and uncertainty. This framework can be rewritten in more detail as the following optimization problem:

$$\begin{aligned}&\min_{\alpha,\beta} \left\|\frac{1}{n_S}\sum_{\mathbf{x}\in S}\beta(\mathbf{x})\phi(\mathbf{x}) - \frac{1}{n_T}\sum_{\mathbf{x}\in T}\phi(\mathbf{x})\right\|^2 + \left\|\frac{1}{n_L}\right.\\ &\left(\sum_{\mathbf{x}\in S_L}\beta(\mathbf{x})\phi(\mathbf{x}) + \sum_{\mathbf{x}\in S_U}\alpha(\mathbf{x})\beta(\mathbf{x})\phi(\mathbf{x}) + \sum_{\mathbf{x}\in T_L}\phi(\mathbf{x})\right)\\ &\left.-\frac{1}{n_U}\left(\sum_{\mathbf{x}\in S_U}(1-\alpha(\mathbf{x}))\beta(\mathbf{x})\phi(\mathbf{x}) + \sum_{\mathbf{x}\in T_U}\phi(\mathbf{x})\right)\right\|^2\\ &+\lambda\sum_{\mathbf{x}\in S_U}\alpha(\mathbf{x})\beta(\mathbf{x})|g(\mathbf{x})|\\ &s.t.\quad \alpha(\mathbf{x}) \in \{0,1\},\ \forall\mathbf{x}\in S_U;\quad \textstyle\sum_{\mathbf{x}\in S_U}\alpha(\mathbf{x}) = n_Q;\\ &\qquad \beta(\mathbf{x}) \in [0,1],\ \forall\mathbf{x}\in S\end{aligned} \tag{4}$$

where $n_L = n_{S_L} + n_Q + n_{T_L}$ and $n_U = n_{S_U} - n_Q + n_{T_U}$.

Note that the binary constraints on α make the above problem NP-Hard. By relaxing the constraints to let $\alpha(\mathbf{x}) \in [0,1]$, the problem in Eq. (4) is biconvex, and can be solved alternatingly with a guarantee on the convergence [Gorski *et al.*, 2007]. To optimize α with β fixed, we have the following quadratic programming problem:

$$\begin{aligned}&\min_{\alpha} \frac{1}{2}\alpha^\top A\alpha + \mathbf{a}^\top\alpha + \text{constant}\\ &s.t.\quad \alpha\in[0,1]^{n_{S_U}},\quad \alpha^\top\mathbf{1} = n_Q,\end{aligned} \tag{5}$$

where

$$\begin{aligned}A =& (\frac{1}{n_L} + \frac{1}{n_U})^2(\beta_{S_U}\beta_{S_U}^\top)\circ K_{S_U,S_U},\\ \mathbf{a} =& -(\frac{1}{n_U^2} + \frac{2}{n_L n_U})(\beta_{S_U}\beta_{S_U}^\top)\circ K_{S_U,S_U}\mathbf{1}\\ &+(\frac{1}{n_L^2} + \frac{2}{n_L n_U})(\beta_{S_U}\beta_{S_L}^\top)\circ K_{S_U,S_L}\mathbf{1}\\ &+(\frac{1}{n_L^2} + \frac{2}{n_L n_U})(\beta_{S_U}\mathbf{1}^\top)\circ K_{S_U,T_L}\mathbf{1}\\ &-(\frac{1}{n_U^2} + \frac{2}{n_L n_U})(\beta_{S_U}\mathbf{1}^\top)\circ K_{S_U,T_U}\mathbf{1}\\ &+\frac{\lambda}{2}\beta_{S_U}\circ|g_{S_U}|.\end{aligned}$$

Here K is the kernel matrix corresponding to the feature mapping $\phi(\cdot)$, and $\mathbf{1}$ is a vector with all entries being 1. We use $\circ$ to denote the element-wise product of vectors or matrixes, and refer to by subscript S_U the rows/columns in a vector or matrix for the unlabeled instances in S_U. Other subscripts S_L, T_U, T_L follow similar denotations.

To optimize β with α fixed, Eq. (4) can be rewritten as follows, which is also a quadratic programming problem.

$$\min_{\beta} \frac{1}{2}\beta^\top B\beta + \mathbf{b}^\top\beta + \text{constant} \tag{6}$$
$$s.t. \quad \beta \in [0,1]^{n_{S_L}+n_{S_U}},$$

where

$$B = \lambda\frac{1}{n_S}^2 K_{S,S} + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

$$\mathbf{b} = -\frac{\lambda}{n_S n_T} K_{S,T}\mathbf{1} + \left(\begin{array}{c} \frac{1}{n_L^2}K_{S_L,T_L}\mathbf{1} - \frac{1}{n_L n_U}K_{S_L,T_U}\mathbf{1} \\ \hline K_{S_U,T_L}\mathbf{1} \circ \left((\frac{1}{n_L^2} + \frac{1}{n_L n_U})\alpha - \frac{1}{n_L n_U}\mathbf{1} \right) \\ -K_{S_U,T_U}\mathbf{1} \circ \left((\frac{1}{n_U^2} + \frac{1}{n_L n_U})\alpha - \frac{1}{n_U^2}\mathbf{1} \right) \\ +\frac{\lambda}{2}\alpha \circ |g_{S_U}| \end{array} \right)$$

and

$$B11 = \frac{1}{n_L^2} K_{S_L,S_L},$$

$$B12 = \left((\frac{1}{n_L^2} + \frac{1}{n_L n_U})\mathbf{1}\alpha^\top - \frac{1}{n_L n_U}\mathbf{1} \right) \circ K_{S_L,S_U},$$

$$B21 = \left((\frac{1}{n_L^2} + \frac{1}{n_L n_U})\alpha\mathbf{1}^\top - \frac{1}{n_L n_U}\mathbf{1} \right) \circ K_{S_U,S_L},$$

$$B22 = \Bigg((\frac{1}{n_L} + \frac{1}{n_U})^2\alpha\alpha^\top - (\frac{1}{n_U^2} + \frac{1}{n_L n_U})(\mathbf{1}\alpha^\top + \alpha\mathbf{1}^\top) + \frac{1}{n_U^2}\mathbf{1} \Bigg) \circ K_{S_U,S_U}.$$

Note here we denote by $\mathbf{1}$ a vector or a matrix with all entries being 1, and assume that the unlabeled instances are ordered before the labeled data in S for simplicity of presentation.

The alternating optimization process is repeated until convergence. In our experiments, it converges in two iterations for most cases, and the algorithm is in general efficient. The pseudo-code of the proposed TLAS approach is summarized in Algorithm 1. Note that β can be simply initialized with all ones or by kernel mean matching. Because $\alpha(\mathbf{x})$ is relaxed from binary to a real value in $[0, 1]$, it cannot be directly identified which instances should be selected, we thus instead sort the instances of S_U in descending order of α, and select the top n_Q instances to query their labels.

Algorithm 1 The TLAS Algorithm

Input:
- $S = S_L \cup S_U$: source domain data;
- $T = T_L \cup T_U$: target domain data;
- n_Q: batch size of active query; λ: tradeoff parameter.

Calculate the kernel matrix K and initialize β;
For each active querying iteration:
Repeat until convergence
- Update α by solving Eq. (5);
- Update β by solving Eq. (6);

$Q \leftarrow$ top n_Q instances of S_U with largest α values;
$S_U = S_U \setminus Q$; $S_L = S_L \cup Q$;
Train the model based on T_L and adapted S_L with β.

4 Experiments

4.1 Settings

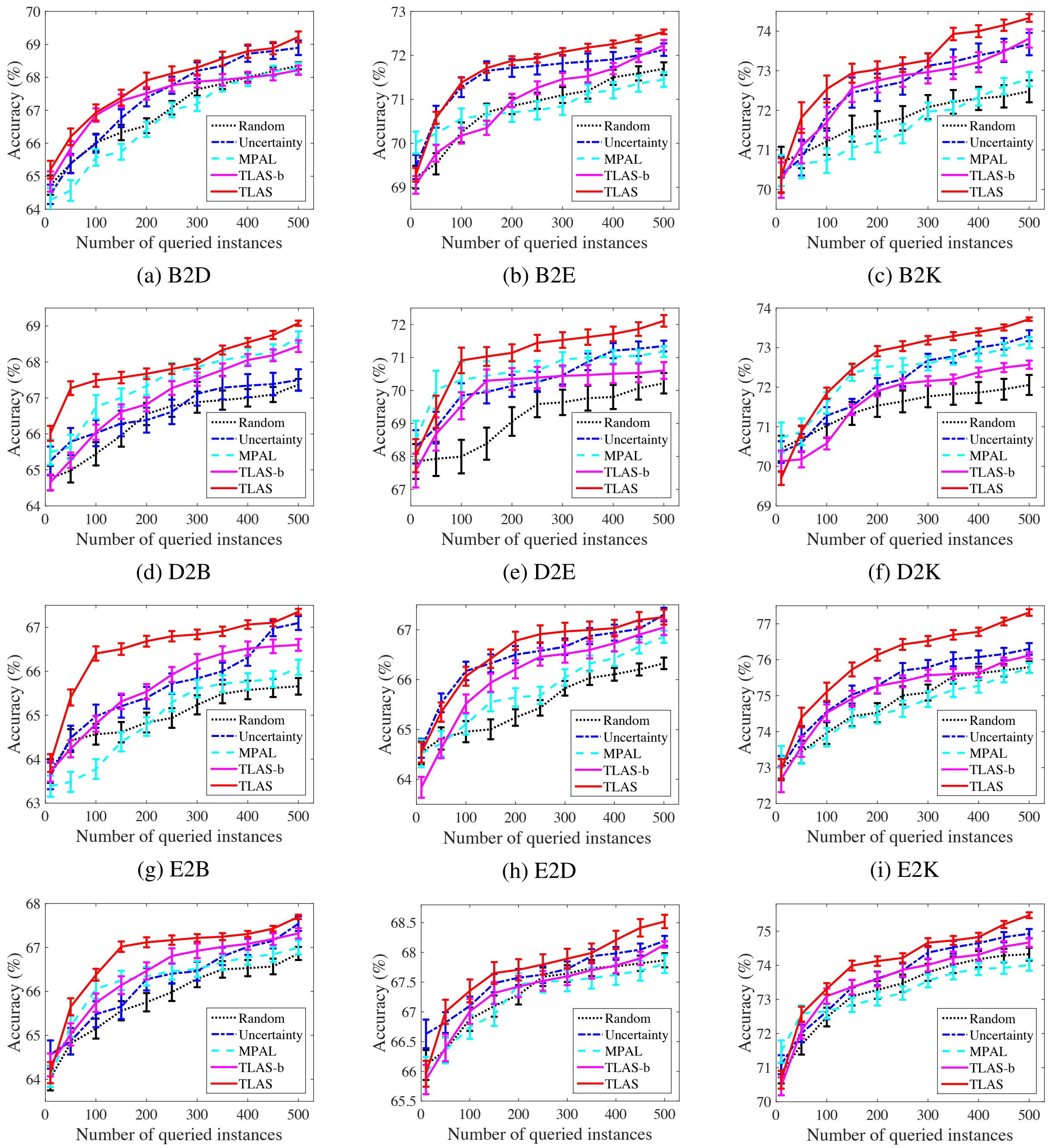
The proposed TLAS approach is evaluated on two tasks: sentiment analysis and text categorization. The Sentiment Analysis dataset[1] contains product reviews on Amazon from four domains: Book, DVD, Electronics and Kitchen. For each domain, 1000 positive reviews and 1000 negative reviews are collected. Each review text is represented by a 200 dimensional feature vector according to [Chattopadhyay *et al.*, 2013]. By taking each domain as source or target domain, we have in all 12 domain pairs: B2D, B2E, B2K, D2B, D2E, D2K, E2B, E2D, E2K, K2B, K2D and K2E. For the text categorization task, we use a preprocessed subset of Reuters-21578[2] as in [Dai *et al.*, 2007]. Reuters-21578 is a collection of Reuters news articles, which are organized in a hierarchical structure. Following the method in [Dai *et al.*, 2007], three top categories are selected: Orgs, People and Places. Each category has different sub-categories, and thus we can dived the data with different sub-categories into source and target domains. Then three binary classification tasks between top categories are constructed: Orgs vs People, Orgs vs Places and People vs Places.

For each dataset, we randomly divide the source domain data into two parts: 10% as the labeled set S_L, and the rest 90% as the unlabeled set S_U. Similarly, the target domain data is randomly divided into three parts: 50% for testing, 10% as the labeled set T_L, and the rest 40% as the unlabeled set T_U. We perform active queries iteratively. At each iteration, n_Q instances are selected from S_U, and are added into S_L with label assignments. After each query, the classification model is trained based on T_L along with adapted S_L. The classification accuracy on the test data is recorded at each iteration. The data partition is repeated randomly for 30 times, and the average results are reported. We employ LibSVM [Chang and Lin, 2011] with default parameters to implement the classification model. In our experiments, we set $n_Q = 10$ and $\lambda = 10$ as default for all datasets, and compute the kernel matrix K using RBF kernel with default parameters.

To the best of our knowledge, there is no existing study can be directly applied to our setting. The following methods are compared in our experiments:

- **Random**: Randomly selects instances from unlabeled source domain data S_U, and performs domain adaptation with kernel mean matching (KMM) [Huang *et al.*, 2006];

[1] http://www.cs.jhu.edu/ mdredze/datasets/sentiment
[2] http://www.cse.ust.hk/TL/dataset/Reuters.zip



(a) B2D (b) B2E (c) B2K

(d) D2B (e) D2E (f) D2K

(g) E2B (h) E2D (i) E2K

(j) K2B (k) K2D (l) K2E

Figure 2: Performance comparison on Sentiment Analysis

- **Uncertainty**: Selects the most uncertain instances from the source domain, and performs domain adaptation with KMM;

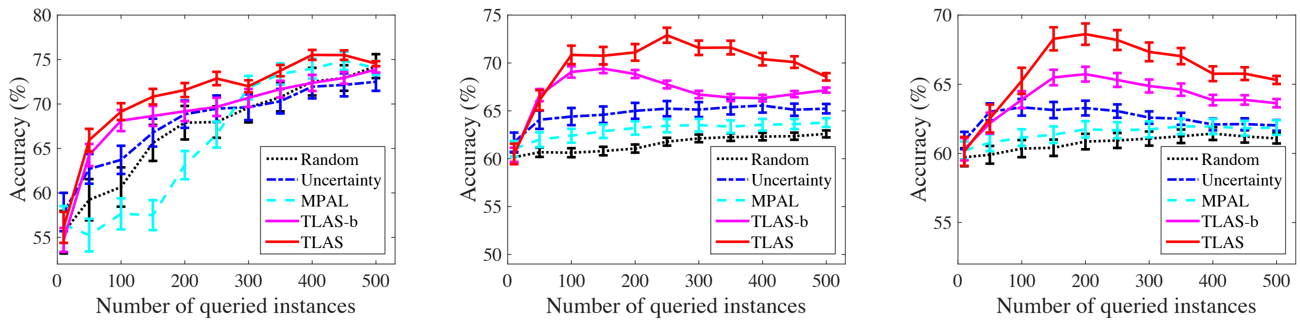
- **MPAL**: Selects instances in the source domain by distribution matching according to the active learning method proposed in [Chattopadhyay *et al.*, 2012], and performs domain adaptation with KMM;

- **TLAS-b**: A baseline of our method, which fixes β with KMM, and optimizes α for active selection with Eq. (5);

- **TLAS**: The method proposed in this paper.

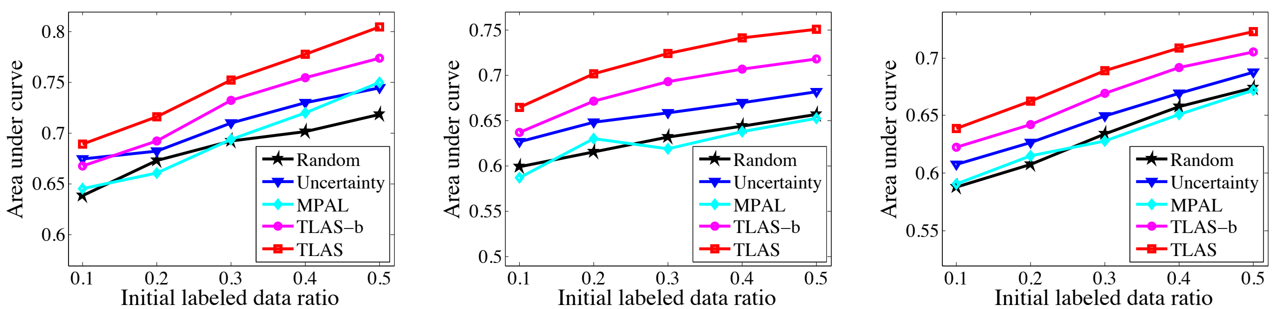
4.2 Performance comparison

We perform active queries iteratively, and record the classification performance in the target domain after updating the model with the queried labels. The performance curves with



(a) Orgs vs People (b) Orgs vs Places (c) People vs Places

Figure 3: Performance comparison on Reuters



(a) Orgs vs People (b) Orgs vs Places (c) People vs Places

Figure 4: Performance comparison with different ratios of labeled data on Reuters

increasing queries are plotted in Figure 2 and Figure 3 respectively for the Sentiment and Reuters datasets. In Figure 2, we can observe that the proposed method TLAS achieves the best performance in most cases. As expected, Random sampling leads to the worst performance on most datasets. Uncertainty sampling usually achieves decent performance, but is less effective than TLAS. The performance of MPAL is not very stable. It works well on some datasets but fails on the others, suggesting that a valuable query for the source domain may be less helpful for the target domain. When comparing TLAS with TLAS-b, the proposed method is always superior to its baseline, validating that iteratively optimizing the importance weights β is useful on improving the performance. In Figure 3, we get similar results on the Reuters dataset. The superiority of TLAS is more obvious on *Orgs vs Places* and *People vs Places*, where even the baseline TLAS-b outperforms all the other compared methods. We also notice that the performance could be degenerated with more queries on Orgs vs Places and People vs Places. One possible reason is the negative transfer, which is an interesting challenge deserve to be overcome in the future.

4.3 Study with different labeled ratios

In this subsection, we examine the performance of the compared approaches with varying numbers of initially labeled data in the target domain. The experiments are performed with the ratio of initial labeled data (n_{T_L}/n_T) increasing from 10% to 50%. Due to space limitation, for each ratio, we report the area under the performance curve on Reuters, instead of plotting the whole performance curve. The comparison results are plotted in Figure 4. Note that the area under curve is normalized by the area of the full rectangle, such that the value is in the interval of 0 to 1. It can be observed that all the compared methods achieve better performance with more initial labeled data in the target domain. The superiority of TLAS over other methods is consistent with different ratios of initial labeled data. Surprisingly, even the baseline version TLAS-b can outperform the other methods in most cases, which suggests that our strategy of active selection is effective even with fixed importance weights for domain adaptation.

5 Conclusion

In this paper, we propose a novel and practical setting for active transfer learning, where labeled data is insufficient in both source and target domains, and further labels can be actively queried only from the source domain. We jointly perform domain adaptation and active selection in one framework, aiming to train an effective model for the target domain with least queries from the source domain. Experiments on 15 datasets validated the effectiveness of the proposed approach. In the future, we plan to extend the framework for transfer learning with multiple source domains. Also, other active selection strategies will be studied under the proposed setting.

References

[Attardi *et al.*, 2013] Giuseppe Attardi, Maria Simi, and Andrea Zanelli. Domain adaptation by active learning. In *Evaluation of Natural Language and Speech Tools for Italian*, pages 77–85. 2013.

[Borgwardt *et al.*, 2006] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):49–57, 2006.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[Chattopadhyay *et al.*, 2012] R. Chattopadhyay, Z. Wang, W. Fan, I. Davidson, S. Panchanathan, and J. Ye. Batch mode active sampling based on marginal probability distribution matching. In *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining*, pages 741–749, 2012.

[Chattopadhyay *et al.*, 2013] Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Joint transfer and batch-mode active learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 253–261, 2013.

[Dai *et al.*, 2007] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 193–200, 2007.

[Gong *et al.*, 2013] Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *Proceedings of the 30th International Conference on Machine Learning*, pages 222–230, 2013.

[Gorski *et al.*, 2007] Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, 2007.

[Gretton *et al.*, 2006] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2006.

[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*, pages 601–608, 2006.

[Huang *et al.*, 2014] S.-J. Huang, R. Jin, and Z.-H. Zhou. Active learning by querying informative and representative examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (10):1936–1949, 2014.

[Kale and Liu, 2013] David Kale and Yan Liu. Accelerating active learning with transfer learning. In *Proceedings of the IEEE 13th International Conference on Data Mining*, pages 1085–1090, 2013.

[Kale *et al.*, 2015] David C. Kale, Marjan Ghazvininejad, Anil Ramakrishna, Jingrui He, and Yan Liu. Hierarchical active transfer learning. In *Proceedings of the SIAM International Conference on Data Mining*, pages 514–522, 2015.

[Li *et al.*, 2013] Shoushan Li, Yunxia Xue, Zhongqing Wang, and Guodong Zhou. Active learning for cross-domain sentiment classification. In *Proceedings of the 23th International Joint Conference on Artificial Intelligence*, pages 2127–2133, 2013.

[Luo *et al.*, 2012] Chunyong Luo, Yangsheng Ji, Xinyu Dai, and Jiajun Chen. Active learning with transfer learning. In *Proceedings of ACL Student Research Workshop*, pages 13–18, 2012.

[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.

[Rai *et al.*, 2010] Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing*, pages 27–32, 2010.

[Saha *et al.*, 2011] Avishek Saha, Piyush Rai, Hal Daumé III, Suresh Venkatasubramanian, and Scott L DuVall. Active supervised domain adaptation. In *Machine Learning and Knowledge Discovery in Databases*, pages 97–112. 2011.

[Settles, 2009] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison, 2009.

[Shi *et al.*, 2008] Xiaoxiao Shi, Wei Fan, and Jiangtao Ren. Actively transfer domain knowledge. In *Machine Learning and Knowledge Discovery in Databases*, pages 342–357. 2008.

[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*, pages 1433–1440, 2008.

[Tan *et al.*, 2015] Ben Tan, Yangqiu Song, Erheng Zhong, and Qiang Yang. Transitive transfer learning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1155–1164, 2015.

[Wang *et al.*, 2014] Xuezhi Wang, Tzu-Kuo Huang, and Jeff Schneider. Active transfer learning under model shift. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1305–1313, 2014.

[Xiao and Guo, 2013] Min Xiao and Yuhong Guo. Online active learning for cost sensitive domain adaptation. In *Proceedings of the 17th Conference on Computational Natural Language Learning*, pages 1–9, 2013.

[Xiao and Guo, 2015] Min Xiao and Yuhong Guo. Feature space independent semi-supervised domain adaptation via kernel matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):54–66, 2015.

[Yang *et al.*, 2013] Liu Yang, Steve Hanneke, and Jaime Carbonell. A theory of transfer learning with applications to active learning. *Machine learning*, 90(2):161–189, 2013.

[Zhao *et al.*, 2013] Lili Zhao, Sinno Jialin Pan, Evan Wei Xiang, Erheng Zhong, Zhongqi Lu, and Qiang Yang. Active transfer learning for cross-system recommendation. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 1205–1211, 2013.