Journal of Artificial Intelligence Research 49 (2014) 501-525 Submitted 09/13; published 03/14

Information-Theoretic Multi-view Domain Adaptation: A Theoretical and Empirical Study

Pei Yang YANGPEI@SCUT.EDU.CN
South China University of Technology
Guangzhou, China

Wei Gao WGAO@QF.ORG.QA
Qatar Computing Research Institute
Qatar Foundation, Doha, Qatar

Abstract

Multi-view learning aims to improve classification performance by leveraging the consistency among different views of data. The incorporation of multiple views was paid little attention in the studies of domain adaptation, where the view consistency based on source data is largely violated in the target domain due to the distribution gap between different domain data. In this paper, we leverage multiple views for cross-domain document classification. The central idea is to strengthen the views' consistency on target data by identifying the associations of domain-specific features from different domains. We present an Information-theoretic Multi-view Adaptation Model (IMAM) using a multi-way clustering scheme, where word and link clusters can draw together seemingly unrelated features across domains, which boosts the consistency between document clusterings that are based on the respective word and link views. Moreover, we demonstrate that IMAM can always find the document clustering with the minimal disagreement rate to the overlap of view-based clusterings. We provide both theoretical and empirical justifications of the proposed method. Our experiments show that IMAM significantly outperforms traditional multi-view algorithm co-training, the co-training-based adaptation algorithm CODA, the single-view transfer model CoCC and the large-margin-based multi-view transfer model MVTL-LM.

1. Introduction

In many mission-critical applications of data mining, natural language processing and information retrieval, it is typically expensive and time-consuming to obtain appropriate training data to learn the needed models. For example, sentiment classifiers for online reviews need to work properly on data of different types of products; search engines must provide consistent quality of service on the Web data in the markets of different languages or verticals. However, the training data commonly exist only in a limited number of domains. Collecting and annotating data for all different domains would become practically prohibitive.

Domain adaptation is a task that utilizes the training data out of the domain (i.e., out-of-domain or source domain) to effectively transform the relevant knowledge to the domain where the task is performed (i.e., in-domain or target domain). Abundant labeled data may exist in a source domain such as webpage data for training a general Web search ranker, but they are not readily available in target domains such as the ranking systems for image search or music search. The out-of-domain data are commonly drawn from some form of feature distribution that is different from that of the in-domain counterpart. Bridging the domain gap is a challenging issue for the model learned from source domain to be generalized well in target domain. For practical reasons, domain adaptation is of great importance to many real-world applications, such as entity mention detection (Daumé III & Marcu, 2006), document classification (Sarinnapakorn & Kubat, 2007), sentiment classification (Blitzer, Dredze, & Pereira, 2007), part-of-speech tagging (Jiang & Zhai, 2007), and more recently Web search ranking (Gao, Cai, Wong, & Zhou, 2010; Cai, Gao, Zhou, & Wong, 2011a, 2011b; Gao & Yang, 2014).

Many types of data can be represented by multiple independent sets of features, reflecting the different views of the data. For example, in document classification, Web document features consist of not only the word-based features but also the features based on link structures among the documents (Blum & Mitchell, 1998); in Web search, document rankers accept both query-dependent features (e.g., tfidf, BM25, language-modeling IR scores, etc.) as well as query-independent features (e.g., page rank, inlink/outlink numbers, url click count, etc.) (Gao, Blitzer, Zhou, & Wong, 2009). Traditionally, the learning scheme called multi-view learning aims to improve classifiers by leveraging the redundancy and consistency among these distinct views (Blum & Mitchell, 1998; Rüping & Scheffer, 2005; Abney, 2002). Existing methods of multi-view learning were designed for the data from a single domain, which assumes that either view alone can predict the in-domain class consistently and accurately. However, this view-consistency assumption is largely violated in the setting of domain adaptation where training and test data are drawn from different distributions (which is empirically justified in the experiment section). In such a case, domain adaptation with multiple views of data needs to be investigated carefully.

Little research has been done on multi-view domain adaptation in the literature. Zhang, He, Liu, Si, and Lawrence (2011) proposed an instance-based multi-view transfer learning approach that integrates the loss of cross-domain classification and multi-view consistency in a large margin framework. However, the instance-level approach assumes that some useful source training examples can be identified and reused to train the target model. It cannot mine the relationships at feature level such as the correlation between source-specific and target-specific features, and may perform poorly since target-specific features are the key for good adaptation performance (Blitzer, Kakade, & Foster, 2011).

In this work, we present an Information-theoretical Multi-view Adaptation Model (IMAM) that combines the paradigms of multi-view learning and domain adaptation based on an co-clustering framework (Dhillon, Mallela, & Modha, 2003) and aims to transfer knowledge across domains in multiple subspaces of *features* complementarily. IMAM exploits a multi-way-clustering-based classification scheme to simultaneously cluster documents, words and links into their respective clusters. The word and link clusterings can automatically associate the specific features from different domains that seemingly may not be directly correlated. Such correlations can bridge the domain gap and then enhance the consistency of distinct views when clustering (i.e., classifying) the target data. The more consistent the views, the better the document clustering, and then the better the word and link clustering, which creates a cycle of positive feedback and gradually improves the adaptation performance. In essence, the enhanced consistency of views helps to bridge the domain gap (i.e., by finding more cross-domain feature correlations), and vice versa. We also provide theoretical justifications for the proposed approach regarding the objective, convergence property and the optimal solution. Our experimental results demonstrate that IMAM significantly outperforms the state-of-the-art baselines including the traditional single-domain multi-view algorithm co-training (Blum & Mitchell, 1998), the co-training-based domain adaptation algorithm CODA (Chen, Weinberger, & Blitzer, 2011), the single-view transfer learning algorithm CoCC (Dai, Xue, Yang, & Yu, 2007a) and the instance-level multi-view transfer learning algorithm MVTL-LM (Zhang et al., 2011).

The rest of the paper is organized as follows: Section 2 reviews the related work; Section 3 describes the background concepts on which we build our model; Section 4 presents the proposed the IMAM model and the corresponding algorithm; Section 5 analyses the realization of consistency between distinct views in our model; Section 6 discusses the experiments and results; Finally, we conclude in Section 7 with prospects on future work.

2. Literature Review

Domain adaptation assumes that multiple tasks can benefit from certain structures of data shared between different distributions. Existing methods can be divided into instance-based approach (Jiang & Zhai, 2007; Dai, Yang, Xue, & Yu, 2007b), feature-based approach (Blitzer et al., 2007; Dai et al., 2007a) and parameter-based approach (Dayanik, Lewis, Madigan, Menkov, & Genkin, 2006). Pan and Yang (2010) presented a comprehensive survey of transfer learning which described domain adaptation as a sub-category of transfer learning. We would not give comprehensive review on domain adaptation for this reason. Interested readers may refer to the survey paper (Pan & Yang, 2010) for details.

The work closely related to ours was done by Dai et al. (2007a), where they proposed a co-clustering-based classification (CoCC) algorithm to learn from the out-of-domain data and apply the learned classifier to the in-domain task. CoCC extended the information-theoretic co-clustering method proposed by Dhillon et al. (2003), where in-domain constraints were added to word clusters to provide a class structure and partial categorization knowledge. However, CoCC is a single-view algorithm which cannot leverage the complementary nature of multiple views. Our framework is an extension from single-view CoCC, and our algorithm is focused on strengthening the consistency of predictions between distinct views across two domains, which is considered the key to the success of multi-view domain adaptation.

Multi-view learning has been studied extensively under single-domain setting. Co-training is the first multi-view algorithm, which trained a learner on each view of labeled examples and then let each learner label the unlabeled examples that receive the highest confidence (Blum & Mitchell, 1998). It was proved that the two independent yet consistent views can be used to learn a concept in the PAC framework based on few labeled and many unlabeled examples. Many extensions were proposed following the idea of co-training. Collins and Singer (1999) introduced an explicit objective function that measures the compatibility of learned hypotheses and used boosting to optimize the function. Dasgupta, Littman, and McAllester (2001) provided PAC-like guarantees for co-training providing an upper bound for the error of classifiers learned from two views. Abney (2002) relaxed the view independence assumption and suggested that there may be an underlying principle which gives rise to a family of new methods: the disagreement rate of two independent hypotheses upper bounds the error rate of either hypothesis. Sridharan and Kakade (2008) proposed an information-theoretic framework for multi-view learning. They showed how to derive incompatibility functions for certain loss functions of interest so that minimizing this incompatibility over unlabeled data helps reduce expected loss on the test data. Nevertheless, multi-view learning generally is not effective for domain adaptation since they treat the domain divergence indiscriminately, which is empirically justified in our experiments (see Experiments and Results section).

Multi-view adaptation is not well studied in the literature. Daumé III, Kumar, and Saha (2010) proposed a co-regularization based approach (EA++) to semi-supervised domain adaptation. EA++ builds on the feature augmentation and harnesses unlabeled data in target domain to assist the transfer of information from source to target. Different from EA++ that aims to make the different hypotheses learned from different distributions agree on unlabeled data, we consider a true multi-view setting and try to make the hypotheses learned from different views consistent with each other. Furthermore, EA++ builds the classifier on the transformed feature space via feature augmentation, while our proposed method learns the hypotheses on the mapped feature space via multi-way clustering. Chen et al. (2011) proposed CODA for adaptation based on co-training (Blum & Mitchell, 1998), which is however a pseudo multi-view algorithm for the original data that have only one view. In order to apply CODA for the real multi-view data, the views have to be first concatenated and then split into multiple pseudo-views. Therefore, it is not suitable nor natural for the true multi-view case as ours. He and Lawrence (2011) proposed a graph-based learning framework to tackle the problems with both feature heterogeneity and task heterogeneity. Their algorithm is a transductive learning approach. Zhang and Huan (2012) proposed an inductive multi-view learning algorithm for multiple related tasks. They used co-regularization to obtain view-based classifiers that agree with each other on unlabeled data and ensure that the learned functions are similar in each view across different tasks. Both of these two algorithms were designed for multi-task learning rather than transfer learning. Zhang et al. (2011) proposed an instance-level multi-view transfer algorithm that integrates classification loss and view consistency terms based on large margin framework. The instance-level approach assumes that some similar source training examples can be identified and reused to train the target model. However, the performance of instance-based approach is generally poor when new target features lack support from source data (Blitzer et al., 2011). We focus on feature-level multi-view adaptation, where adaptation takes place in the multiple transformed feature spaces simultaneously and complementarily. To the best of our knowledge, there are no existing work focused on the feature-level multi-view domain adaptation except for our preliminary study recently published (Yang, Gao, Tan, & Wong, 2012). This paper extends the work of Yang et al. (2012) substantially by providing the detailed algorithm, theoretical justification and comprehensive empirical evaluation, which were not specifically presented in the preliminary version.

3. Background Concepts

Our multi-view approach is based on the co-clustering (Dhillon et al., 2003) and co-clustering-based classification (CoCC) model (Dai et al., 2007a) for building the underlying clusters of each view. Before going to the details of our model, we will briefly describe some background concepts and lemmas related to the co-clustering techniques in this section.

Mutual information is a fundamental measure to quantify the mutual dependence of two random variables. Let $\mathcal{I}(X,Y)$ be the mutual information of variables X and Y, which is defined as $\mathcal{I}(X,Y) = \sum_x \sum_y p(x,y) log \frac{p(x,y)}{p(x)p(y)}$ (Cover & Thomas, 1991). Mutual information can also be expressed in the form of Kullback-Leibler (KL) divergence, i.e., $\mathcal{I}(X,Y) = \mathcal{D}\left(p(x,y)||p(x)p(y)\right)$. Given two discrete random variables X and Y with joint probability distribution $p(x,y)$, co-clustering approach (Dhillon et al., 2003) aims to simultaneously cluster X into disjoint clusters $\hat{X}$, and Y into disjoint clusters $\hat{Y}$. The quality of co-clustering is measured by the resulting loss based on mutual information:

$$\mathcal{I}(X,Y) - \mathcal{I}(\hat{X},\hat{Y})$$

For the given X and Y, since $\mathcal{I}(X,Y)$ is fixed, minimizing the above equation is equivalent to maximizing $\mathcal{I}(\hat{X},\hat{Y})$.

For the simplicity of expression, a joint distribution $q(x,y) = p(\hat{x},\hat{y})\frac{p(x)}{p(\hat{x})}\frac{p(y)}{p(\hat{y})}$ is defined to approximate the probability $p(x,y)$ under co-clustering $(\hat{X},\hat{Y})$. Note that the distribution $q(x,y)$ preserves the marginals of $p(x,y)$. That is, for any $x \in \hat{x}$, $y \in \hat{y}$, we have $q(x) = p(x)$ because

$$q(x) = \sum_{y} q(x,y) = \sum_{\hat{y}}\sum_{y\in\hat{y}} p(\hat{x},\hat{y})\frac{p(x)}{p(\hat{x})}\frac{p(y)}{p(\hat{y})} = \sum_{\hat{y}} p(\hat{x},\hat{y})\frac{p(x)}{p(\hat{x})} = p(x).$$

Likewise we have $q(y) = p(y)$.

Dhillon et al. (2003) proved that the loss in mutual information between pre- and post-clustering can be reformulated as the KL-divergence between $p(x,y)$ and an approximation $q(x,y)$, which is given as the following lemma:

Lemma 3.1. *For a fixed co-clustering $(\hat{X},\hat{Y})$, the loss in mutual information can be expressed as*

$$\mathcal{I}(X,Y) - \mathcal{I}(\hat{X},\hat{Y}) = \mathcal{D}\left(p(x,y)||q(x,y)\right),$$

where $\mathcal{D}(\cdot||\cdot)$ is the KL-divergence, and $q(x,y)$ is the distribution of the form

$$q(x,y) = p(\hat{x},\hat{y})\frac{p(x)}{p(\hat{x})}\frac{p(y)}{p(\hat{y})},$$

where $x \in \hat{x}$ and $y \in \hat{y}$.

For completeness and clarity, we reproduce the illustrative example given by Dhillon et al. (2003) for interpreting Lemma 3.1. Consider the joint distribution of (X,Y) represented by a 6*6 matrix below:

$$\begin{pmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{pmatrix}$$

It follows naturally that the rows are divided into three clusters: $\hat{x}_1 = \{x_1, x_2\}$, $\hat{x}_2 = \{x_3, x_4\}$ and $\hat{x}_3 = \{x_5, x_6\}$, and the columns clustering is: $\hat{y}_1 = \{y_1, y_2, y_3\}$, $\hat{y}_2 = \{y_4, y_5, y_6\}$. The resulting joint distribution of $(\hat{X},\hat{Y})$ is given by:

$$\begin{pmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{pmatrix}$$

It can be verified that the mutual information loss in this co-clustering is .0957, which is the minimum among all the possible co-clusterings.

4. Information-Theoretic Multi-view Adaptation Model (IMAM)

We will first introduce the motivation, and then will describe our model and its algorithm.

4.1 Motivation

Traditional multi-view learning such as co-training framework (Blum & Mitchell, 1998) employs two basic assumptions: (1) the target functions in each view agree on the labels of most examples (consistency assumption); and (2) the views are independent given the class label (independence assumption). The first assumption reduces the complex learning problem to the search of compatible functions; and the second assumption allows the model to achieve high-confidence predictions since it becomes unlikely for consistent classifiers trained on independent views to agree on an incorrect label.

Considering the training and test data drawn from different distributions, nonetheless, the consistency assumption is mostly violated because the distinct views agreeing on the labels of source data are unnecessarily compatible on the labels of target examples due to the domain gap. Therefore, it can be expected that traditional multi-view learning framework will not work effectively across different domains, which can be empirically justified in the comparison experiments. Hence, how to enhance the consistency among multiple views and bridge the gap among different domains simultaneously is the key issue for the multi-view domain adaptation approach to succeed.

Without loss of generality, we will focus on cross-domain document classification in this paper where the document representation consists of two views such as word and link. Given text documents from two domains, there would be a set of common word features available on both domains, considered as domain-independent features, and the remaining words would be regarded as either source-specific or target-specific features. The same taxonomy regarding domain-independent and domain-specific features also apply to the inter-document links, e.g., the hyperlinks or citations features.

From a single view's perspective, source-specific and target-specific features can be drawn together by mining their co-occurrence with domain-independent features. IMAM exploits multi-way clustering to correlate those seemingly unrelated domain-specific features via the domain-independent features which act as a bridge. Such correlations help bridge the domain gap and facilitate the adaptation (Dai et al., 2007a). From multiple view's perspective, if the word and link clusters constructed over the two domains are of high quality, the corresponding target document clustering resulted from either view can be subsequently improved due to the effect of co-clustering (Dhillon et al., 2003). It can be expected that the predictive power of distinct views on the target data tends to become more concordant and approaches to the optimal solution. Our model leverages complementary cooperation between different views to yield better adaptation performance.

Next, we will present some representational preliminaries and the objective function of our multi-view adaptation model, and then an iterative two-phase algorithm is presented to optimize the objective.

4.2 The Graphical Representation

Let D_S be the training documents of source domain and D_T be the unlabeled documents of the target domain. The source and target data are assumed to draw from different feature spaces where the *i.i.d.* assumption no longer holds. Some features are defined in source or target domain only while some others are defined in both domains. We simply expand the feature space to include all features of both domains where the missing features in either domain are replenished as 0. Let W be the vocabulary of the entire document collection $D = D_S \cup D_T$. Each $d \in D$ can be represented by a bag-of-words set $\{w | w \in d \wedge w \in W\}$. Let L be the set of all links (hyperlinks or citations) in the collection. Each d can be also represented by a bag-of-links set $\{l|l \in d \wedge l \in L\}$. D and L naturally form independent sets of features respectively corresponding to word view and link view. Let C denote the set of class labels shared between the two domains. Each source document $d_s \in D_S$ is labeled with a unique class label $c \in C$. Our objective is to assign the appropriate class label to target document $d_t \in D_T$ as accurately as possible. Note that we assume there is no labeled data available in target domain, which follows the transductive learning scheme. Transductive approach is a typical domain adaptation setting, which is more general and widely applicable to different scenarios including the inductive setting where only a small number of labeled target data exist.

Figure 1 shows the graphical multi-view adaption model representation, where $\hat{D}$, $\hat{W}$ and $\hat{L}$ are the respective clusterings of documents, words and links. Additionally, the multi-way clusterings mutually constrain each other and are subject to various explicit and implicit association relationships. Explicit association includes two types of constraints: (1) Document clustering is constrained by word clustering and link clustering; (2) Word or link clustering is constrained by document clustering and class labels. Implicit association means that the class label knowledge is transferred from source documents to target documents through word and link clusters.

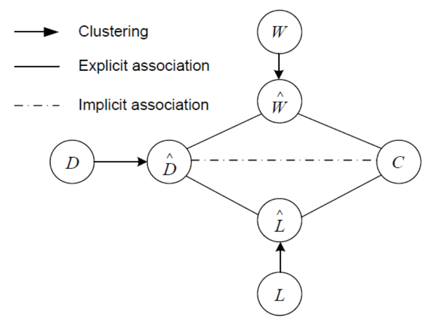

Figure 1: The graphical representation of the proposed multi-view adaptation model.

Our model incorporates such a multi-way clustering scheme that simultaneously clusters documents, words and links. The clustering functions are defined as $\mathcal{C}_D(d) = \hat{d}$ for documents, $\mathcal{C}_W(w) = \hat{w}$ for words and $\mathcal{C}_L(l) = \hat{l}$ for links, where $\hat{d}$, $\hat{w}$ and $\hat{l}$ represent the corresponding clusters.

4.3 Preliminaries – Co-clustering-Based Classification

Dai et al. (2007a) proposed a co-clustering-based classification framework, namely CoCC, to learn a classifier from source-domain documents and then use it to classify target-domain documents. In their approach, co-clustering was leveraged as a bridge to transfer the knowledge from source to target.

Co-clustering aims to simultaneously cluster target documents D_T into clusters $\hat{D}_T$ and words W into clusters $\hat{W}$. Since the problem is to classify target-domain documents, the key is to make use of the knowledge about classes in the data of source domain for the co-clustering process. Such kind of correlation between the source document class knowledge and the target document clustering can be established by considering their respective relationship with the word clusters as an intermediary. A good word clustering should minimize the loss in mutual information between class labels and words before and after clustering for the source data, and meanwhile it should minimize the same loss between documents and words for the target data. Therefore, the loss function of CoCC (Dai et al., 2007a) is formulated as follows:

$$\mathcal{I}(D_T, W) - \mathcal{I}(\hat{D}_T, \hat{W}) + \lambda \left[\mathcal{I}(C, W) - \mathcal{I}(C, \hat{W}) \right]$$

where λ is a trade-off parameter that balances the effect to word clusters from co-clustering and word clustering.

4.4 Objective Function

We extend the information-theoretic framework for co-clustering (Dhillon et al., 2003) and co-clustering-based classification (Dai et al., 2007a) by incorporating the loss terms from multiple views. Co-clustering aims to minimize the loss of mutual information between pre- and post-clustering with respect to a pair of clustering variables, such as documents and words. The objective of our Information-theoretic Multi-view Adaptation Model (IMAM) is to minimize the loss by trading off different views:

$$\Theta = \alpha \Theta_W + (1 - \alpha) \Theta_L \tag{1}$$

where

$$\Theta_W = \mathcal{I}(D_T, W) - \mathcal{I}(\hat{D}_T, \hat{W}) + \lambda \left[\mathcal{I}(C, W) - \mathcal{I}(C, \hat{W}) \right] \tag{2}$$

$$\Theta_L = \mathcal{I}(D_T, L) - \mathcal{I}(\hat{D}_T, \hat{L}) + \lambda \left[\mathcal{I}(C, L) - \mathcal{I}(C, \hat{L}) \right]. \tag{3}$$

Θ_W and Θ_L are the loss terms based on word view and link view, respectively, and α is the trade-off coefficient. In Eq. 2, $\mathcal{I}(D_T, W) - \mathcal{I}(\hat{D}_T, \hat{W})$ measures the loss of word-document co-clustering, $\mathcal{I}(C, W) - \mathcal{I}(C, \hat{W})$ measures the loss between vocabulary and class labels, and λ is the weight of the loss for word clustering. Class labels act as indirect constraints added on vocabulary via source documents and are propagated to target documents through co-clustering. In Eq. 3, we have the similar loss term for the link view. When $\alpha = 1$, the function relies on text information only, which reduces to CoCC (Dai et al., 2007a). But unlike CoCC (Dai et al., 2007a), we aim to learn the cross-domain classifiers for multi-view data.

It is worth noting that by substituting Eq. 2 and 3 in Eq. 1 and ignoring the constant terms, we can reformulate the problem as the following maximization, which is kind of easier to interpret:

$$\alpha \mathcal{I}(\hat{D}_T, \hat{W}) + (1 - \alpha) \mathcal{I}(\hat{D}_T, \hat{L}) + \lambda \left[\alpha \mathcal{I}(C, \hat{W}) + (1 - \alpha) \mathcal{I}(C, \hat{L}) \right]$$

where the first two terms enforce that the view consistency on $\hat{D}_T$, which means that the document clusters $\hat{D}_T$ should preserve their mutual information with both words and links as much as possible, and the last two terms enforce transfer of information from source to target via agreement with labels C, which indicates that the source label knowledge should be maximally preserved by both word and link clusters.

Given the multi-view data data from different domains, the central problem would be how different views could cooperate each other to form consistent target class output in the scenario where different domain data follow different distributions. This is challenging because the view consistency based on source data is largely violated in the target domain due to the domain gap. To tackle this problem, we aim to simultaneously enhance the consistency among multiple views and bridge the gap among different domains in a unified objective. IMAM exploits multi-way clustering to enrich common words (and links) by drawing together those seemingly unrelated source-specific and target-specific words (and links). Such correlations bridge the domain gap and facilitate the adaptation process. On the other hand, IMAM takes the weighted combination of view-based loss of mutual information. As pointed out in Section 5 (Consistency of Multiple Views), the optimal document clustering is to optimize the weighted sum of word-view and link-view document clustering functions, and try to minimize the disagreement between different views. Moreover, the multi-way clustering scheme imposes the constraints on all of document and word/link clustering, which can make them mutually benefit from each other. In summary, IMAM uses such a boosting procedure to enhance the view consistency and bridge domain gap simultaneously, and can be expected to improve the adaptation performance on the multi-view data.

4.5 IMAM Algorithm

Based on $q(x, y)$ defined in Section 3, we can also define the corresponding conditional distribution $q(x|\hat{y}) = \frac{q(x,y)}{p(y)}$ under co-clustering. For any $x \in \hat{x}$, we can easily prove that $q(x|\hat{y}) = p(x|\hat{x})p(\hat{x}|\hat{y})$. Therefore, for any $w \in \hat{w}$, $l \in \hat{l}$, $d \in \hat{d}$ and $c \in C$, we can calculate a set of conditional distributions including $q(w|\hat{d})$, $q(d|\hat{w})$, $q(l|\hat{d})$, $q(d|\hat{l})$, $q(c|\hat{w})$, $q(c|\hat{l})$.

The objective of Eq. 1 is hard to optimize directly because it contains mutual information of two clusterings, which is a combinatorial optimization problem. Therefore, we transform it to the form of KL-divergence between two conditional distributions in Lemma 4.1 in order to facilitate our search for the optimal value. Let $\mathcal{D}(p(x|y)||q(x|y))$ denote KL-divergence between $p(x|y)$ and $q(x|y)$, which is defined as

$$\mathcal{D}(p(x|y)||q(x|y)) = \sum_x p(x|y) log \frac{p(x|y)}{q(x|y)}.$$

We have the following lemma, and using the similar technique as in Dhillon et al. (2003), we provide its proof in the Appendix A.

Lemma 4.1 (Objective functions). *Equation 1 can be turned into the form of alternate minimization between two objectives:*
(i) For document clustering while keeping word and link clustering fixed, we minimize

$$\Theta = \sum_d p(d)\phi_D(d, \hat{d}) + \phi_C(\hat{W}, \hat{L})$$

where $\phi_C(\hat{W}, \hat{L})$ *is a constant*[1] *and*

$$\phi_D(d, \hat{d}) = \alpha\mathcal{D}(p(w|d)||q(w|\hat{d})) + (1-\alpha)\mathcal{D}(p(l|d)||q(l|\hat{d})).$$

(ii) For word and link clustering while keeping document clustering fixed, we minimize

$$\Theta = \alpha \sum_w p(w)\phi_W(w, \hat{w}) + (1-\alpha)\sum_l p(l)\phi_L(l, \hat{l})$$

1. We can prove that $\phi_C(\hat{W}, \hat{L}) = \lambda \left[\alpha(\mathcal{I}(C, W) - \mathcal{I}(C, \hat{W})) + (1-\alpha)(\mathcal{I}(C, L) - \mathcal{I}(C, \hat{L}))\right]$, where MI between class label and other variables is constant.

Algorithm 1 Algorithm for IMAM

Input:

Document-term matrices $D_S \times W$ and $D_T \times W$;
Document-link matrices $D_S \times L$ and $D_T \times L$;
Class label $c \in C$ assigned to each doc $d \in D_S$;
# of document clusters (i.e., # of classes);

Output:

Class label assigned to each document $d \in D_T$;

1: Set $t = 0$. Initialize document clustering $\mathcal{C}_D^{(0)}$ using NBC. Initialize word clustering $\mathcal{C}_W^{(0)}$ and link clustering $\mathcal{C}_L^{(0)}$ randomly;
2: Initialize distributions $q^{(0)}(w|\hat{d})$, $q^{(0)}(l|\hat{d})$, $q^{(0)}(d|\hat{w})$, $q^{(0)}(d|\hat{l})$, $q^{(0)}(c|\hat{w})$, $q^{(0)}(c|\hat{l})$;
3: **repeat**
4: Document clustering: For each d, find its new cluster index using Eq. 4;
5: Keep $q^{(t+1)}(c|\hat{w}) = q^{(t)}(c|\hat{w})$ and $q^{(t+1)}(c|\hat{l}) = q^{(t)}(c|\hat{l})$;
Update $q^{(t+1)}(w|\hat{d})$, $q^{(t+1)}(l|\hat{d})$, $q^{(t+1)}(d|\hat{w})$, $q^{(t+1)}(d|\hat{l})$;
6: Word clustering: For each word w, find its new cluster index using Eq. 5;
Link clustering: For each link l, find its new cluster index using Eq. 6;
7: Update $q^{(t+2)}(w|\hat{d})$, $q^{(t+2)}(l|\hat{d})$, $q^{(t+2)}(d|\hat{w})$, $q^{(t+2)}(d|\hat{l})$, $q^{(t+2)}(c|\hat{w})$ and $q^{(t+2)}(c|\hat{l})$;
8: $t = t + 2$;
9: **until** no document's cluster index needs to adjust
10: **for** each unlabeled $d \in D_T$ **do**
11: Assign d the class label based on Eq. 7;
12: **end for**

where for any feature v (e.g., w and l) in feature set V (e.g., W and L)

$$\phi_V(v, \hat{v}) = \mathcal{D}(p(d|v)||q(d|\hat{v})) + \lambda \mathcal{D}(p(c|v)||q(c|\hat{v})).$$

The intuition of the optimization is that given the document-word and document-link matrices, let us simultaneously re-order documents in the two matrices such that all documents mapping to the first document cluster are arranged first, followed by all documents mapping to the second cluster, and so on. A good document clustering tries to ensure the consistency between different views. Next, let us simultaneously re-order words and links in document-word and document-link matrices in a similar way. A good word (or link) clustering draws indirectly related domain-specific words (or links) together since both of them may co-occur with domain-independent words (or links) in the documents. The document-word-link interaction helps finding an optimal multi-way clustering.

Lemma 4.1 allows us to alternately reorder either documents or both words and links, which is shown as Algorithm 1, in such a way that the mutual information loss decreases monotonically (see Lemma 4.2).

4.5.1 ALGORITHM

The algorithm starts with an initial multi-clustering $(\mathcal{C}_D^{(0)}, \mathcal{C}_W^{(0)}, \mathcal{C}_L^{(0)})$ and iteratively refines it until the algorithm converges. The algorithm uses a two-phase iterative procedure to minimize the loss, in which it first searches for the best document clustering while keeping word and link clustering unchanged, and then clusters words and links while document clustering remains fixed.

In step 1, Naive Bayes classifier (NBC) is trained on source data D_S and used to predict the class of target data D_T, which produces the initial document clustering of entire D. Note that the cluster index of source documents is fixed with class labels. Thus, the allocation of each target document to certain cluster also means that the document is assigned with the corresponding class label. It is worth noting since the objective Eq. 1 is non-convex, it will be somewhat sensitive to the initialization. Hence, instead of random initialization, we use NBC to generate the initial document clusterings so as to keep it start from some good points.

Step 4 updates the cluster index for each d:

$$\mathcal{C}_D^{(t+1)}(d) = \arg\min_{\hat{d}} \left[\alpha\mathcal{D}(p(w|d)||q^{(t)}(w|\hat{d})) + (1-\alpha)\mathcal{D}(p(l|d)||q^{(t)}(l|\hat{d}))\right] \tag{4}$$

Step 6 updates the cluster index of each w:

$$\mathcal{C}_W^{(t+2)}(w) = \arg\min_{\hat{w}} \left[\mathcal{D}(p(d|w)||q^{(t+1)}(d|\hat{w})) + \lambda\mathcal{D}(p(c|w)||q^{(t+1)}(c|\hat{w}))\right] \tag{5}$$

and then updates the cluster index of each l:

$$\mathcal{C}_L^{(t+2)}(l) = \arg\min_{\hat{l}} \left[\mathcal{D}(p(d|l)||q^{(t+1)}(d|\hat{l})) + \lambda\mathcal{D}(p(c|l)||q^{(t+1)}(c|\hat{l}))\right] \tag{6}$$

Note that Algorithm 1 does not separately update the membership of each word and link since there are implicit association relationships between the word clustering and link clustering via document clustering. The document clustering acts as the bridge to make word clustering and link clustering mutually affect each other.

After finishing the multi-way clustering procedure, we assign each target document $d \in D_T$ with the class label predicted by

$$c^* = \arg\min_{c \in C} \left[\alpha\mathcal{D}(p(w|c)||q(w|\hat{d})) + (1-\alpha)\mathcal{D}(p(l|c)||q(l|\hat{d}))\right] \tag{7}$$

Lemma 4.2 below guarantees the convergence of the algorithm, and its proof is given in the Appendix B by borrowing the similar technique from Dhillon et al. (2003). Note that finding a global minimum for multi-way clustering is NP-hard, and IMAM uses a greedy approach to find a local minimum, which does not guarantee the global optimum. But usually we can run experiments multiple times and then average over the performance of different runs.

Lemma 4.2 (Convergence)**.** *IMAM monotonically reduces the objective given in Equation 1. That is,*

$$\Theta^{(t)} \geq \Theta^{(t+1)}$$
$$\Theta^{(t+1)} \geq \Theta^{(t+2)}$$

where $t = 0, 2, 4, \ldots$

5. Consistency of Multiple Views

In this section, we present how the consistency of document clustering on target data could be enhanced among multiple views, which is the key issue of our multi-view adaptation method. We particularly discuss the relationship between the disagreement rate of views and the optimal document clustering function.

In each iteration of Algorithm 1, the optimal document clustering function $\mathcal{C}_D^{(t+1)}$ (see Eq. 4) is to minimize the weighted sum of KL-divergences used in optimal word-view and link-view document clustering functions as shown above. The optimal word-view clustering functions can be denoted as follows:

$$\mathcal{C}_{D_W}^{(t+1)}(d) = \arg\min_{\hat{d}} \mathcal{D}(p(w|d)||q^{(t)}(w|\hat{d})) \tag{8}$$

and similarly the link-view function as

$$\mathcal{C}_{D_L}^{(t+1)}(d) = \arg\min_{\hat{d}} \mathcal{D}(p(l|d)||q^{(t)}(l|\hat{d})) \tag{9}$$

Our central idea is that the document clusterings $\mathcal{C}_{D_W}^{(t+1)}$ and $\mathcal{C}_{D_L}^{(t+1)}$ based on the two views are drawn closer in each iteration due to the word and link clusterings (Eq. 5 and 6) that bring together seemingly unrelated source-specific and target-specific features. Meanwhile, $\mathcal{C}_D^{(t+1)}$ combines the two views and reallocates the documents so that it maintains the consistency with the view-based clusterings as much as possible.

5.1 Disagreement Rate of Views

Suppose $\Omega = \{\mathcal{F}_i|\mathcal{F}_i(d) = \hat{d}, \hat{d} \in \hat{D}\}$ is the set of all document clustering functions where the number of clusters is fixed. For any document, a consistency indicator function with respect to any two clustering functions can be defined as follows (Round indicator t is omitted for simplicity):

Definition 1 (Indicator function) For any $d \in D$, and any $\mathcal{F}_i \in \Omega$, $\mathcal{F}_j \in \Omega$

$$\delta_{\mathcal{F}_i,\mathcal{F}_j}(d) = \begin{cases} 1, & \text{if} \quad \mathcal{F}_i(d) = \mathcal{F}_j(d); \\ 0, & \text{otherwise} \end{cases}$$

Then we define the disagreement rate between two view-based clustering functions:

Definition 2 (View disagreement rate) For any $\mathcal{F}_i \in \Omega$ and $\mathcal{F}_j \in \Omega$

$$\eta(\mathcal{F}_i, \mathcal{F}_j) = 1 - \frac{\sum_{d\in D} \delta_{\mathcal{F}_i,\mathcal{F}_j}(d)}{|D|} \tag{10}$$

Obviously, $\eta(\mathcal{C}_{D_W}, \mathcal{C}_{D_L})$ denotes the disagreement rate between the word-view and link-view clustering functions. Abney (2002) suggests that the disagreement rate of two independent hypotheses upper-bounds the error rate of either hypothesis. By minimizing the disagreement rate on unlabeled data, the error rate of each view can be minimized (so does the overall error). However, the disagreement rate function is not continuous nor convex, which is difficult to optimize directly[2]. Alternatively, we minimize the mutual information loss in Eq. 1 as a surrogate for the disagreement rate function. We believe that the mutual information loss is a good surrogate because, as discussed in Section 4.4, Eq. 1 aims to enhance the view consistency, which is equivalent to minimizing the disagreement rate of views. Moreover, we show empirically that by optimizing Eq. 1 the disagreement rate $\eta(\mathcal{C}_{D_W}, \mathcal{C}_{D_L})$ is indeed monotonically decreased with the iterations in our experiments (see Section 6).

2. Abney (2002) used a greedy approach.

5.2 View Combination

Note that in practice the view-based document clusterings in Eq. 8 and Eq. 9 are not computed explicitly. Instead, Eq. 4 directly optimizes the view combination and produces the document clustering. Therefore, it is necessary to disclose how consistent the combined view-based clustering could be with the individual view-based clusterings.

For any $\mathcal{F}_i \in \Omega$, we obtain the disagreement rate $\eta(\mathcal{F}_i, \mathcal{C}_{D_W} \cap \mathcal{C}_{D_L})$, where $\mathcal{C}_{D_W} \cap \mathcal{C}_{D_L}$ denotes the clustering resulting from the overlap of the individual view-based clusterings. Note that the co-training style algorithms usually assume that the multiple views are redundant. Thus, the intersection of them would not be empty. We obtain Lemma 5.1 as below, and its proof is given in the Appendix C.

Lemma 5.1. *The optimal document clustering function $\mathcal{C}_D$ in IMAM model always minimizes the disagreement rate for any $\mathcal{F}_i \in \Omega$ such that*

$$\eta(\mathcal{C}_D, \mathcal{C}_{D_W} \cap \mathcal{C}_{D_L}) = \min_{\mathcal{F}_i \in \Omega} \eta(\mathcal{F}_i, \mathcal{C}_{D_W} \cap \mathcal{C}_{D_L})$$

And meanwhile, $\eta(\mathcal{C}_D, \mathcal{C}_{D_W} \cap \mathcal{C}_{D_L}) = \eta(\mathcal{C}_{D_W}, \mathcal{C}_{D_L})$.

Lemma 5.1 suggests that IMAM always finds the document clustering with the minimal disagreement rate to the overlap of the individual view-based clusterings, and the minimal value of disagreement rate equals to the disagreement rate of the individual view-based clusterings.

6. Experiments and Results

In this section, we empirically evaluate the IMAM algorithm for the cross-domain document classification tasks in comparison with the state-of-the-art baselines.

6.1 Data and Setup

Cora (McCallum, Nigam, Rennie, & Seymore, 2000) is an online archive which contains approximately 37,000 computer science research papers and over 1 million links among documents. The documents are categorized into a hierarchical structure. We selected a subset of Cora, which contains 5 top categories and 10 sub-categories (the numbers are in the parenthesis):

- DA_1: /data_structures_algorithms_and_theory/computational_complexity/ (711)
- DA_2: /data_structures_algorithms_and_theory/computational_geometry/ (459)
- EC_1: /encryption_and_compression/encryption/ (534)
- EC_2: /encryption_and_compression/compression/ (530)
- NT_1: /networking/protocols/ (743)
- NT_2: /networking/routing/ (477)
- OS_1: /operating_systems/realtime/ (595)
- OS_2: /operating_systems/memory_management/ (1,102)
- ML_1: /machine_learning/probabilistic_methods/ (687)
- ML_2: /machine_learning/genetic_algorithms/ (670)

Based on this dataset, we used a similar way as Dai et al. (2007a) to construct our training and test sets. For each set, we chose two top categories, one as positive class and the other as the negative. Different sub-categories were deemed as different domains. The task is defined as top category classification. For example, the subset denoted as DA-EC consists of source domain:

DA_1(+), EC_1(-); and target domain: DA_2(+), EC_2(-). The method ensures the domains of labeled and unlabeled data related due to same top categories, but the domains are different because they are drawn from different sub-categories. Such preprocessing is a common practice for data preparation for adaptation purpose. Some previous work (Ling, Dai, Xue, Yang, & Yu, 2008; Dai et al., 2007a) found that baseline SVM as well as transductive SVM classifiers trained on source-domain data performed much worse on the target domain, implying large domain gap between them. We have the same finding on this dataset by using transductive SVM.

We preprocessed the data for both text and link information. For the texts, we removed stop words and low-frequency words with count less than 5. For the links, we removed the links with less than 5 citation counts. Then the standard TF-IDF (Salton & Buckley, 1988) technique was applied to both the text and link datasets. Moreover, we generated the merged dataset by concatenating both the word and link features together.

Reuters-21578 (Lewis, 2004) is widely used for the evaluation of automatic text categorization algorithms. Reuters-21578 corpus also has a hierarchical structure, which contains 5 top categories. We used the pre-processed version of the corpus that is public accessible[3]. The statistics of this dataset can be seen in Table 1. Based on these data, we generated separate information representing two views: the first view corresponds to the features using the TF-IDF scores of terms; the second view corresponds to the topic-based features (i.e. document-topic distributions) obtained by applying probabilistic Latent Semantic Analysis (pLSA)[4] on the term counts information, where the topic number was set to 200.

Subset	Source	Target
Orgs-People	OrgsPeople.src (1,237)	OrgsPeople.tar (1,208)
Orgs-Places	OrgsPlaces.src (1,016)	OrgsPlaces.tar (1,043)
People-Places	PeoplePlaces.src (1,077)	PeoplePlaces.tar (1,077)

Table 1: The statistics of Reuters-21578 dataset.

Using Cora dataset, we conducted experiments with IMAM for studying the influence of different parameters and the manifestation of view disagreement rate. Also, we compared IMAM with various state-of-the-art domain adaptation algorithms on both Cora and Reuters datasets. In order to avoid the infinity values, we applied Laplacian smoothing when computing the KL-divergence.

6.2 Parameter Sensitivity

We first studied the influence of some important parameters, i.e., the number of word/link clusters, α, and λ.

6.2.1 INFLUENCE OF CLUSTER NUMBER

Figure 2 shows the error rate curves varying with different number of word (and link) clusters on the 4 subsets: DA-EC, DA-NT, DA-OS and EC-NT. The X-axis represents the number of word (and link) clusters which is tuned from 32 to 512. According to the performance shown in the figure, we empirically set the number of word (and link) clusters to 128.

3. `http://www.cse.ust.hk/TL/dataset/Reuters.zip`.
4. `http://lear.inrialpes.fr/people/verbeek/code/plsa.tar.gz`.

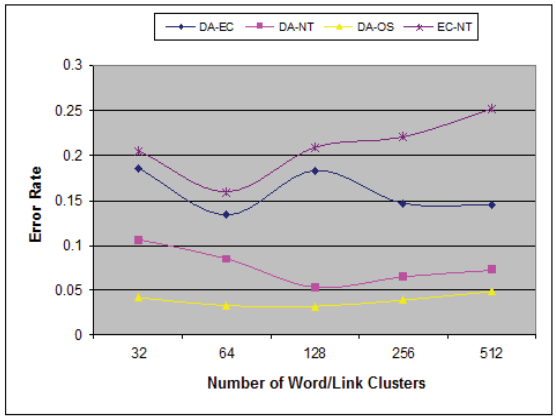

Figure 2: Error rate curves varying with different number of word/link clusters.

6.2.2 INFLUENCE OF α

Figure 3 shows that the performance curves vary with different values of α. The error rate generally decreases first and then increases when α is augmented. As always, the algorithm performs worst when the model heavily relies on either the text information ($0.9 \leq \alpha \leq 1.0$) or the link structure ($0 \leq \alpha \leq 0.1$). And setting α between 0.5 and 0.8 achieved the best results on most of the subsets. This implies that the two views of document are complementary. Therefore, in the remaining experiments, we set the value of α to 0.7.

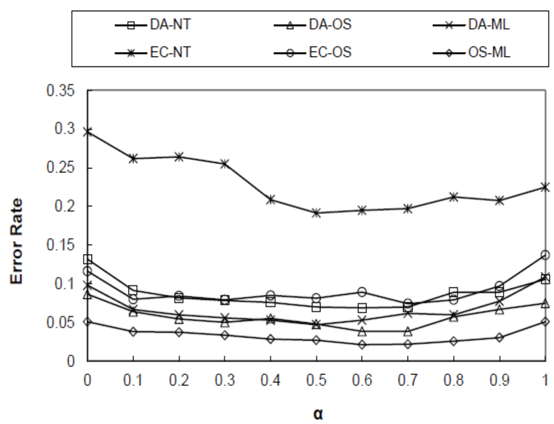

Figure 3: Error rate curves varying with different settings of α.

6.2.3 INFLUENCE OF λ

λ is used for propagating class labels from source document class to target document clustering through word and link clusters. Surprisingly, we did not observe its significant influence on most of the subsets. This is because we used NBC to initialize document clusterings for a good starting point, and the class information, though not accurately, could be largely propagated to the words and link clusters at the next iteration. This observation is similar to that of Dai et al. (2007a) when the number of their word clusters was appropriately provided. We empirically set λ to 0.5 after trying 0, 0.25, 0.5, 1, 2 and 4.

	DA-EC	DA-NT	DA-OS	DA-ML	EC-NT	Average
η on source	0.179	0.157	0.188	0.184	0.210	0.184
η on target	0.251	0.224	0.275	0.211	0.234	0.239

Table 2: The view disagreement rates under different domains using co-training.

Iteration		1	2	3	4	5	γ
DA-EC	ϵ	0.194	0.153	0.149	0.144	0.144	0.998
	η	0.340	0.132	0.111	0.101	0.095	
DA-NT	ϵ	0.147	0.083	0.071	0.065	0.064	0.996
	η	0.295	0.100	0.076	0.069	0.064	
DA-OS	ϵ	0.129	0.064	0.052	0.047	0.041	0.998
	η	0.252	0.092	0.068	0.060	0.052	
DA-ML	ϵ	0.166	0.102	0.071	0.065	0.064	0.984
	η	0.306	0.107	0.076	0.062	0.054	
EC-NT	ϵ	0.311	0.250	0.228	0.219	0.217	0.988
	η	0.321	0.137	0.112	0.096	0.089	

Table 3: View disagreement rate (η) and error rate (ϵ) both decrease with iterations. Their correlation is denoted as γ.

6.3 View Disagreement Rate η

In this section, we studied the view disagreement rate for two different purposes: (1) we experimentally verified that the view consistency assumption was violated due to distinct domains for the traditional multi-view learning using co-training, which justified our motivation to reduce the view disagreement rate; (2) we examined the property of view disagreement rate based on our method and revealed its relationship with the cross-domain classification performance.

6.3.1 η WITH CO-TRAINING

In this experiment, for each subset, the source data were splitted into two portions, one portion for training and the other for testing. The traditional multi-view algorithm co-training (Blum & Mitchell, 1998) was trained on the source training set, and then the model was evaluated on the source test set and the target test set separately. The first result corresponds to the single-domain performance and the second corresponds to cross-domain performance.

As shown in Table 2, it is clear that the view disagreement rate on the target domain is considerably higher than that on the source domain. It implies that the domain gap is likely to deteriorate view consistency. As Abney (2002) pointed out, view consistency is directly related to classification error rate, which is upper bounded by the view disagreement rate. Our finding from this experiment seems consistent with this claim, and furthermore, it implies that it would be helpful to overcome domain gap by enhancing the view consistency on target data.

6.3.2 η WITH IMAM

Here we examined the variance of disagreement rate $\eta(\mathcal{C}_{D_W}, \mathcal{C}_{D_L})$ between view-based clusterings and its correlation with the error rate ϵ.

We used the Pearson's correlation to measure the dependence of the disagreement rate and error rate, which takes the value between -1 (perfect negative correlation) and 1 (perfect positive correlation). Table 3 shows the monotonic decrease of disagreement rate η and error rate ϵ with the iterations, and their correlation γ is nearly perfectly positive. This indicates that IMAM may gradually improves adaptation performance by strengthening the consistency between different views, and alternatively, IMAM increases classification performance, which then causes the different views to be more consistent. Both procedures are therefore reciprocal causation. This is achieved by the mutual reinforcement of word and link clustering that draws together those target-specific and source-specific features, which are originally unrelated but could co-occur with the common features across the two domains.

6.4 Convergence

The convergence property of IMAM is shown as Figure 4. IMAM uses a two-phase iterative procedure to find a local optimal point. The convergence is guaranteed by Lemma 4.2. We can see that the number of documents needed to be reassigned into different clusters decreases very fast during the first 5 iterations and reaches 0 after 10 iterations. Thus, we terminate the algorithm after a maximum of 15 iterations.

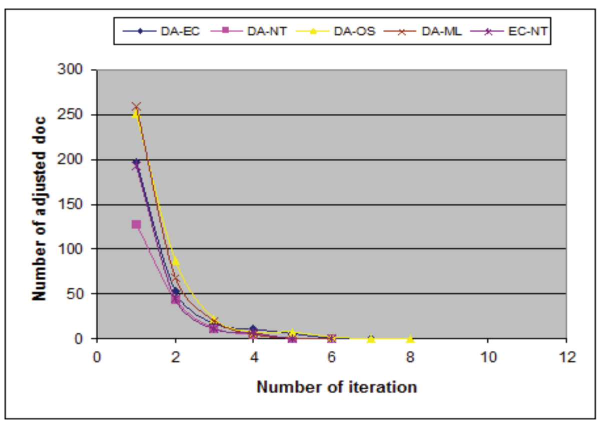

Figure 4: Number of documents needed to be reassigned into different clusters varies with iterations.

6.5 Algorithms for Comparison

We compared IMAM with a variety of the state-of-the-art algorithms including Transductive SVM[5] (TSVM) (Joachims, 1999) which is a semi-supervised classifier, co-training (Co-Train) (Blum & Mitchell, 1998), the co-clustering-based single-view transfer learning CoCC (Dai et al., 2007a), the large-margin-based multi-view transfer learning MVTL-LM (Zhang et al., 2011) and the co-training-based adaptation algorithm CODA[6] (Chen et al., 2011). We used both Cora and Reuters datasets for the comparative study.

On both datasets, for the ease of presentation, we used the postfix -C, -L and -CL to denote that the classifier was fed with data of different views. For Cora dataset, -C, -L and -CL represent the text

5. `http://svmlight.joachims.org/`.
6. `http://www1.cse.wustl.edu/~mchen/code/coda.tar`.

Subset	TSVM-C	TSVM-L	TSVM-CL	Co-Train	MVTL-LM	CODA	CoCC-C	CoCC-L	CoCC-CL	IMAM
DA-EC	0.293	0.157	0.214	0.230	0.192	0.234	0.149	0.227	0.187	**0.138**
DA-NT	0.175	0.137	0.114	0.163	0.108	0.076	0.106	0.132	0.115	**0.069**
DA-OS	0.276	0.261	0.262	0.175	0.068	0.109	0.075	0.086	0.067	**0.039**
DA-ML	0.217	0.114	0.107	0.171	0.183	0.150	0.109	0.098	0.095	**0.047**
EC-NT	0.305	0.220	**0.177**	0.296	0.261	0.178	0.225	0.296	0.239	0.191
EC-OS	0.355	0.201	0.245	0.175	0.176	0.187	0.137	0.116	0.125	**0.074**
EC-ML	0.333	0.205	**0.168**	0.206	0.264	0.322	0.203	0.269	0.237	0.173
NT-OS	0.364	0.501	0.396	0.220	0.288	0.240	0.107	0.142	0.115	**0.070**
NT-ML	0.205	0.106	0.101	0.132	0.071	**0.025**	0.054	0.094	0.047	0.031
OS-ML	0.202	0.170	0.179	0.128	0.126	0.087	0.051	0.051	0.062	**0.021**
Average	0.272	0.207	0.196	0.190	0.174	0.161	0.122	0.151	0.129	**0.085**

Table 4: Error rate of classification adaptation on Cora dataset.

Subset	TSVM-C	TSVM-L	TSVM-CL	Co-Train	MVTL-LM	CODA	CoCC-C	CoCC-L	CoCC-CL	IMAM
OrgsPeople	0.246	0.263	0.227	0.251	0.230	0.177	0.185	0.219	0.191	**0.153**
OrgsPlaces	0.278	0.304	0.263	0.270	0.249	0.226	0.214	0.235	0.221	**0.192**
PeoplePlaces	0.294	0.335	0.286	0.318	0.260	0.275	0.245	0.262	0.248	**0.218**
Average	0.273	0.301	0.259	0.280	0.246	0.226	0.215	0.239	0.220	**0.188**

Table 5: Error rate of classification adaptation on Reuters-21578 dataset.

view, link view and two views, respectively; for Reuters dataset, they correspond to term view, topic view and two views. If the examined classifier is inherently multi-view, both of the views's data were fed to it. Such algorithm include TSVM-CL, co-training, MVTL-LM, CoCC-CL, and IMAM. Since CODA is a pseudo multi-view adaptation algorithm, to fit our scenario, the CODA was fed with the merged view which could be automatically split into the sub-views. For each algorithm, the parameters were tuned by using five-fold cross-validation on training data. To cancel out local optimal results, we repeated the algorithms five times for each subset and reported the average error rate.

All the algorithms were trained on the source data and then tested on the target data. The classification error rate on target data is used as evaluation metric, which is defined as the ratio of the number of misclassified documents to that of total documents.

6.6 Performance Comparison

Table 4 shows the results of comparison on Cora dataset, and Table 5 shows the same on Reuters-21578. We have consistent findings on the two datasets.

On both datasets, TSVM performed poorly for adaptation when using either content or link features alone. Simply merging the two sets of features makes some improvements, implying that text and link in Cora data (or, term and topic in Reuters data) can be complementary, but it may degrade the confidence of the classifier on some instances whose features become conflicting because of merging. Co-training can avoid this problem by boosting the confidence of classifiers built on the distinct views in a complementary way, and its performance is comparable with TSVM though it uses a weaker base classifier. Since both TSVM and co-training do not consider the distribution gap, they performed clearly worse than CoCC even though CoCC is a single-view approach.

On both datasets, CODA outperformed co-training and MVTL-LM by splitting the feature space into multiple pseudo views and iteratively adding the shared source and target features based on their compatibility across domains. However, it could not be comparably effective than IMAM. It seems that the pseudo views automatically generated by CODA are not as complementary as the original view partition on these two datasets. It performed even worse than the COCC under single-view setting, indicating that sometimes pseudo views might be detrimental. The relatively lower performance of CODA may be explained as follows. It might happen that the original formation of the two views on our data was reasonably good, but after they were combined into one view, it was likely that CODA could be stuck in a poor locally optimal decomposition of features due to the non-smooth, non-convex nature of its objective function. Since its model parameters were initialized randomly, repeating the algorithm did not guarantee a better solution. In contrast, the objective function of IMAM, although non-convex, is smooth, and also, instead of using random initialization we used NBC to initialize the document clusters to ensure a good starting point.

IMAM significantly outperformed both CoCC-C and CoCC-L on all the subsets. In average, the error rate of IMAM is 30.3% lower than that of CoCC-C (or 43.7% lower than that of CoCC-L). This is because IMAM effectively leverages distinct and complementary views. Compared to CoCC, using source training data to improve the view consistency on target data is the key competency of IMAM. Moreover, IMAM performed much better than the CoCC-CL. Unlike CoCC-CL which simply concatenates the two-view data, our technique is to strengthen the view consistency by bootstrapping two CoCC models iteratively and complementarily. In our model the two CoCC models communicate complementarily in each iteration, which consequently boosts the consistency between the two views.

The result shows that multi-view adaptation using MVTL-LM performs worse than IMAM on most subsets. A general explanation suggests that instance-based approach relying on instance weighting are not effective when the data of different domains are drawn from different feature spaces. Although MVTL-LM regulates view consistency on both domains' instances, it cannot identify the useful correlation between the target-specific and source-specific features, which is the key to the success of adaptation especially when the domain gap is large and little commonality could be found. In contrast, CoCC and IMAM can use co-clustering or multi-way clustering to find such correlation.

Note that we use different ways to generate the multi-view data for the two datasets. Different from Cora dataset which has natural multiple views, i.e., text and link, we generate the term and topic views for Reuters-21578 dataset based on the text information only. Nevertheless, the results on both datasets show that IMAM works well on different types of multi-view data by using the multi-way clustering to enhance the view consistency.

7. Conclusion and Future Work

We presented a novel feature-level multi-view adaptation approach called IMAM for cross-domain document classification. The thrust of our technique is to incorporate distinct views of document features into the multi-way clustering framework and gradually strengthen the view consistency for classifying target documents. The improvements over the state-of-the-art baselines are substantial. We provided both theoretical and empirical justifications regarding the properties of the proposed algorithm. Experiments show that it considerably outperforms the state-of-the-art baselines including the multi-view single-domain algorithm co-training, the co-training-based adaptation CODA, the single-view adaptation CoCC as well as the instance-level multi-view adaptation MVLT-LM.

Multi-view domain adaptation is a promising direction since its underlying principle and practice are still open questions. As part of our ongoing work, we will further explore the foundations and limitations of multi-view domain adaptation. For example, multiple views might hurt adaptation performance when domains or views are very "dissimilar". Although it was not observed in our experiments, it needs to be analyzed more deeply. In addition, due to practical reasons, we did not directly optimize the consistency measure, i.e., view disagreement rate. Instead, we adopted the information-theoretical framework to optimize the mutual information loss, which worked well but may not be the ideal solution. In the future, we will study the techniques of directly optimize the consistency measure of views.

Appendix A. Proof of Lemma 4.1

Proof. The proof of Lemma 4.1 can be divided into two parts.
(i) For document clustering:

Note that the word and link clusterings keep fixed in this phase. Thus the mutual information between class label and word (or link) clusters remains unchanged during the document clustering phase, that is,

$$\phi_C(\hat{W}, \hat{L}) = \lambda\left[\alpha(\mathcal{I}(C, W) - \mathcal{I}(C, \hat{W})) + (1-\alpha)(\mathcal{I}(C, L) - \mathcal{I}(C, \hat{L}))\right]$$

is a constant. By using Eq. 1, we can obtain

$$\begin{aligned}
&\Theta - \phi_C(\hat{W}, \hat{L}) \\
&= \alpha\Theta_W + (1-\alpha)\Theta_L - \phi_C(\hat{W}, \hat{L}) \\
&= \alpha\left[I(D_T; W) - I(\hat{D}_T; \hat{W})\right] + (1-\alpha)\left[I(D_T; L) - I(\hat{D}_T; \hat{L})\right] \\
&= \alpha\left[\sum_{\hat{d}}\sum_{\hat{w}}\sum_{d\in\hat{d}}\sum_{w\in\hat{w}} p(d,w)\log\frac{p(d,w)}{p(d)p(w)} - \sum_{\hat{d}}\sum_{\hat{w}}\left(\sum_{d\in\hat{d}}\sum_{w\in\hat{w}} p(d,w)\right)\log\frac{p(\hat{d},\hat{w})}{p(\hat{d})p(\hat{w})}\right] \\
&\quad + (1-\alpha)\left[\sum_{\hat{d}}\sum_{\hat{l}}\sum_{d\in\hat{d}}\sum_{l\in\hat{l}} p(d,l)\log\frac{p(d,l)}{p(d)p(l)} - \sum_{\hat{d}}\sum_{\hat{l}}\left(\sum_{d\in\hat{d}}\sum_{l\in\hat{l}} p(d,l)\right)\log\frac{p(\hat{d},\hat{l})}{p(\hat{d})p(\hat{l})}\right] \\
&= \alpha\sum_{\hat{d}}\sum_{\hat{w}}\sum_{d\in\hat{d}}\sum_{w\in\hat{w}} p(d,w)\log\frac{p(d,w)p(\hat{d})p(\hat{w})}{p(\hat{d},\hat{w})p(d)p(w)} + (1-\alpha)\sum_{\hat{d}}\sum_{\hat{l}}\sum_{d\in\hat{d}}\sum_{l\in\hat{l}} p(d,l)\log\frac{p(d,l)p(\hat{d})p(\hat{l})}{p(\hat{d},\hat{l})p(d)p(l)} \\
&= \alpha\sum_{\hat{d}}\sum_{\hat{w}}\sum_{d\in\hat{d}}\sum_{w\in\hat{w}} p(d,w)\log\frac{p(d,w)}{q(d,w)} + (1-\alpha)\sum_{\hat{d}}\sum_{\hat{l}}\sum_{d\in\hat{d}}\sum_{l\in\hat{l}} p(d,l)\log\frac{p(d,l)}{q(d,l)} \\
&= \alpha\sum_{\hat{d}}\sum_{\hat{w}}\sum_{d\in\hat{d}}\sum_{w\in\hat{w}} p(d)p(w|d)\log\frac{p(w|d)}{q(w|\hat{d})} + (1-\alpha)\sum_{\hat{d}}\sum_{\hat{l}}\sum_{d\in\hat{d}}\sum_{l\in\hat{l}} p(d)p(l|d)\log\frac{p(l|d)}{q(l|\hat{d})} \\
&= \alpha\sum_{\hat{d}}\sum_{d\in\hat{d}} p(d)\sum_{\hat{w}}\sum_{w\in\hat{w}} p(w|d)\log\frac{p(w|d)}{q(w|\hat{d})} + (1-\alpha)\sum_{\hat{d}}\sum_{d\in\hat{d}} p(d)\sum_{\hat{l}}\sum_{l\in\hat{l}} p(l|d)\log\frac{p(l|d)}{q(l|\hat{d})} \\
&= \sum_{\hat{d}}\sum_{d\in\hat{d}} p(d)\left[\alpha\mathcal{D}(p(w|d)||q(w|\hat{d})) + (1-\alpha)\mathcal{D}(p(l|d)||q(l|\hat{d}))\right] \\
&= \sum_{d} p(d)\phi_D(d, \hat{d})
\end{aligned}$$

(ii) For word and link clustering:

Note that the document clusterings remains unchanged in this phase. Using the similar technique as above, we can obtain

$$\Theta = \alpha\sum_{w} p(w)\phi_W(w, \hat{w}) + (1-\alpha)\sum_{l} p(l)\phi_L(l, \hat{l})$$

By combining steps (i) and (ii), Lemma 4.1 can be proved. □

Appendix B. Proof of Lemma 4.2

Proof. The proof of Lemma 4.2 can be divided into two parts.
(i) For document clustering: Note that the word and link clusterings keep fixed in this phase.

$$
\begin{aligned}
&\Theta^{(t)} - \phi_C^{(t)}(\hat{W}, \hat{L}) \overset{(a)}{=} \sum_{\hat{d}} \sum_{d:\mathcal{C}_D^{(t)}(d)=\hat{d}} p(d) \left[\alpha \mathcal{D}(p(w|d)||q^{(t)}(w|\hat{d})) + (1-\alpha)\mathcal{D}(p(l|d)||q^{(t)}(l|\hat{d}))\right] \\
&= \sum_{\hat{d}} \sum_{d:\mathcal{C}_D^{(t)}(d)=\hat{d}} p(d) \left[\alpha \sum_w p(w|d) \log \frac{p(w|d)}{q^{(t)}(w|\hat{d})} + (1-\alpha) \sum_l p(l|d) \log \frac{p(l|d)}{q^{(t)}(l|\hat{d})}\right] \\
&\overset{(b)}{\geq} \sum_{\hat{d}} \sum_{d:\mathcal{C}_D^{(t)}(d)=\hat{d}} p(d) \left[\alpha \sum_w p(w|d) \log \frac{p(w|d)}{q^{(t)}(w|\mathcal{C}_D^{(t+1)}(d))} + (1-\alpha) \sum_l p(l|d) \log \frac{p(l|d)}{q^{(t)}(l|\mathcal{C}_D^{(t+1)}(d))}\right] \\
&\overset{(c)}{=} \sum_{\hat{d}} \sum_{d:\mathcal{C}_D^{(t+1)}(d)=\hat{d}} p(d) \left[\alpha \sum_w p(w|d) \log \frac{p(w|d)}{q^{(t)}(w|\hat{d})} + (1-\alpha) \sum_l p(l|d) \log \frac{p(l|d)}{q^{(t)}(l|\hat{d})}\right] \\
&\overset{(d)}{=} \sum_{\hat{d}} \sum_{d:\mathcal{C}_D^{(t+1)}(d)=\hat{d}} p(d) \left[\alpha \sum_{\hat{w}} \sum_{w:\mathcal{C}_W^{(t+1)}(w)=\hat{w}} p(w|d) \log \frac{p(w|d)}{q^{(t)}(w|\hat{w})q^{(t)}(\hat{w}|\hat{d})} + (1-\alpha) \sum_{\hat{l}} \sum_{l:\mathcal{C}_L^{(t+1)}(l)=\hat{l}} p(l|d) \log \frac{p(l|d)}{q^{(t)}(l|\hat{l})q^{(t)}(\hat{l}|\hat{d})}\right] \\
&= \underbrace{\sum_{\hat{d}} \sum_{d:\mathcal{C}_D^{(t+1)}(d)=\hat{d}} p(d) \left[\alpha \sum_{\hat{w}} \sum_{w:\mathcal{C}_W^{(t+1)}(w)=\hat{w}} p(w|d) \log \frac{p(w|d)}{q^{(t)}(w|\hat{w})} + (1-\alpha) \sum_{\hat{l}} \sum_{l:\mathcal{C}_L^{(t+1)}(l)=\hat{l}} p(l|d) \log \frac{p(l|d)}{q^{(t)}(l|\hat{l})}\right]}_{I} \\
&\quad + \sum_{\hat{d}} \sum_{d:\mathcal{C}_D^{(t+1)}(d)=\hat{d}} p(d) \left[\alpha \sum_{\hat{w}} \sum_{w:\mathcal{C}_W^{(t+1)}(w)=\hat{w}} p(w|d) \log \frac{1}{q^{(t)}(\hat{w}|\hat{d})} + (1-\alpha) \sum_{\hat{l}} \sum_{l:\mathcal{C}_L^{(t+1)}(l)=\hat{l}} p(l|d) \log \frac{1}{q^{(t)}(\hat{l}|\hat{d})}\right] \\
&= I + \sum_{\hat{d}} \left[\alpha \sum_{\hat{w}} \left(\sum_{d:\mathcal{C}_D^{(t+1)}(d)=\hat{d}} \sum_{w:\mathcal{C}_W^{(t+1)}(w)=\hat{w}} p(d)p(w|d) \right) \log \frac{1}{q^{(t)}(\hat{w}|\hat{d})} \right. \\
&\quad \left. + (1-\alpha) \sum_{\hat{l}} \left(\sum_{d:\mathcal{C}_D^{(t+1)}(d)=\hat{d}} \sum_{l:\mathcal{C}_L^{(t+1)}(l)=\hat{l}} p(d)p(l|d) \right) \log \frac{1}{q^{(t)}(\hat{l}|\hat{d})} \right] \\
&= I + \sum_{\hat{d}} \left[\alpha \sum_{\hat{w}} q^{(t+1)}(\hat{d}, \hat{w}) \log \frac{1}{q^{(t)}(\hat{w}|\hat{d})} + (1-\alpha) \sum_{\hat{l}} q^{(t+1)}(\hat{d}, \hat{l}) \log \frac{1}{q^{(t)}(\hat{l}|\hat{d})}\right] \\
&\overset{(e)}{\geq} I + \sum_{\hat{d}} q^{(t+1)}(\hat{d}) \left[\alpha \sum_{\hat{w}} q^{(t+1)}(\hat{w}|\hat{d}) \log \frac{1}{q^{(t+1)}(\hat{w}|\hat{d})} + (1-\alpha) \sum_{\hat{l}} q^{(t+1)}(\hat{l}|\hat{d}) \log \frac{1}{q^{(t+1)}(\hat{l}|\hat{d})}\right] \\
&= \sum_{\hat{d}} \sum_{d:\mathcal{C}_D^{(t+1)}(d)=\hat{d}} p(d) \left[\alpha \sum_{\hat{w}} \sum_{w:\mathcal{C}_W^{(t+1)}(w)=\hat{w}} p(w|d) \log \frac{p(w|d)}{q^{(t)}(w|\hat{w})q^{(t+1)}(\hat{w}|\hat{d})} \right. \\
&\quad \left. + (1-\alpha) \sum_{\hat{l}} \sum_{l:\mathcal{C}_L^{(t+1)}(l)=\hat{l}} p(l|d) \log \frac{p(l|d)}{q^{(t)}(l|\hat{l})q^{(t+1)}(\hat{l}|\hat{d})}\right] \\
&\overset{(f)}{=} \sum_{\hat{d}} \sum_{d:\mathcal{C}_D^{(t+1)}(d)=\hat{d}} p(d) \left[\alpha \sum_{\hat{w}} \sum_{w:\mathcal{C}_W^{(t+1)}(w)=\hat{w}} p(w|d) \log \frac{p(w|d)}{q^{(t+1)}(w|\hat{w})q^{(t+1)}(\hat{w}|\hat{d})} + \right. \\
&\quad \left. (1-\alpha) \sum_{\hat{l}} \sum_{l:\mathcal{C}_L^{(t+1)}(l)=\hat{l}} p(l|d) \log \frac{p(l|d)}{q^{(t+1)}(l|\hat{l})q^{(t+1)}(\hat{l}|\hat{d})}\right] \\
&= \sum_{\hat{d}} \sum_{d:\mathcal{C}_D^{(t+1)}(d)=\hat{d}} p(d) \left[\alpha \sum_{\hat{w}} \sum_{w:\mathcal{C}_W^{(t+1)}(w)=\hat{w}} p(w|d) \log \frac{p(w|d)}{q^{(t+1)}(w|\hat{d})} + (1-\alpha) \sum_{\hat{l}} \sum_{l:\mathcal{C}_L^{(t+1)}(l)=\hat{l}} p(l|d) \log \frac{p(l|d)}{q^{(t+1)}(l|\hat{d})}\right] \\
&= \sum_{\hat{d}} \sum_{d:\mathcal{C}_D^{(t+1)}(d)=\hat{d}} p(d) \left[\alpha \mathcal{D}(p(w|d)||q^{(t+1)}(w|\hat{d})) + (1-\alpha)\mathcal{D}(p(l|d)||q^{(t+1)}(l|\hat{d}))\right] \\
&\overset{(g)}{=} \Theta^{(t+1)} - \phi_C^{(t+1)}(\hat{W}, \hat{L})
\end{aligned}
$$

where (a) follows from Lemma 4.1, (b) follows from Step 4 of the IMAM algorithm, (c) follows by rearranging the summation, (d) and (f) follow since we hold the word and link clusters fixed in Step 4, (e) follows by non-negativity of the KL-divergence, and (g) follows from Lemma 4.1. Since the word and link clusters remain unchanged during the document clustering, i.e., $\phi_C^{(t)}(\hat{W},\hat{L}) = \phi_C^{(t+1)}(\hat{W},\hat{L})$, we can prove that $\Theta^{(t)} \geq \Theta^{(t+1)}$.

(ii) For word and link clustering: Note that the document clusterings remains unchanged in this phase. By using the properties of Step 6 and the similar technique as above, we can prove that

$$\begin{aligned}
\Theta^{(t+1)} &= \alpha \sum_{\hat{w}} \sum_{w:\mathcal{C}_W^{(t+1)}(w)=\hat{w}} p(w)\phi_W^{(t+1)}(w,\hat{w}) + (1-\alpha)\sum_{\hat{l}} \sum_{l:\hat{C}_L^{(t+1)}(l)=\hat{l}} p(l)\phi_L^{(t+1)}(l,\hat{l}) \\
&\geq \alpha \sum_{\hat{w}} \sum_{w:\mathcal{C}_W^{(t+2)}(w)=\hat{w}} p(w)\phi_W^{(t+2)}(w,\hat{w}) + (1-\alpha)\sum_{\hat{l}} \sum_{l:\mathcal{C}_L^{(t+2)}(l)=\hat{l}} p(l)\phi_L^{(t+2)}(l,\hat{l}) \\
&= \Theta^{(t+2)}
\end{aligned}$$

By combining steps (i) and (ii), it follows that in every iteration the algorithm IMAM monotonically decreases the objective function. □

Appendix C. Proof of Lemma 5.1

Proof. For any document $d \in D$ and any document cluster $\hat{d} \in \hat{D}$,
(i) If $\mathcal{C}_{D_W}(d) = \mathcal{C}_{D_L}(d)$

For brevity, we denote the cluster by $\hat{d}^*$, i.e., $\mathcal{C}_{D_W}(d) = \mathcal{C}_{D_L}(d) = \hat{d}^*$. By using Eq. 4, Eq. 8 and Eq. 9, we can obtain

$$\begin{cases} \mathcal{D}(p(w|d)||q(w|\hat{d}^*)) & \leq \quad \mathcal{D}(p(w|d)||q(w|\hat{d})) \\ \mathcal{D}(p(l|d)||q(l|\hat{d}^*)) & \leq \quad \mathcal{D}(p(l|d)||q(l|\hat{d})) \end{cases}$$

$$\Rightarrow \alpha\mathcal{D}(p(w|d)||q(w|\hat{d}^*)) + (1-\alpha)\mathcal{D}(p(l|d)||q(l|\hat{d}^*)) \leq \alpha\mathcal{D}(p(w|d)||q(w|\hat{d})) + (1-\alpha)\mathcal{D}(p(l|d)||q(l|\hat{d}))$$

$$\Rightarrow \mathcal{C}_D(d) = \mathcal{C}_{D_W}(d) = \mathcal{C}_{D_L}(d) = \hat{d}^*$$

$$\Rightarrow \delta_{\mathcal{C}_D,\mathcal{C}_{D_W}\cap\mathcal{C}_{D_L}}(d) = 1$$

(ii) If $\mathcal{C}_{D_W}(d) \neq \mathcal{C}_{D_L}(d)$

Obviously, we have $\delta_{\mathcal{C}_D,\mathcal{C}_{D_W}\cap\mathcal{C}_{D_L}}(d) = 0$. By combining (i) and (ii), we can obtain

$$\delta_{\mathcal{C}_D,\mathcal{C}_{D_W}\cap\mathcal{C}_{D_L}}(d) = \delta_{\mathcal{C}_{D_W},\mathcal{C}_{D_L}}(d) = \begin{cases} 1, & \text{if} \quad \mathcal{C}_{D_W}(d) = \mathcal{C}_{D_L}(d); \\ 0, & \text{otherwise} \end{cases}$$

For any $\mathcal{F}_i \in \Omega$, the indicator function $\delta_{\mathcal{F}_i,\mathcal{C}_{D_W}\cap\mathcal{C}_{D_L}}(d)$ can be rewritten as

$$\delta_{\mathcal{F}_i,\mathcal{C}_{D_W}\cap\mathcal{C}_{D_L}}(d) = \begin{cases} 1, & \text{if} \quad \mathcal{C}_{D_W}(d) = \mathcal{C}_{D_L}(d) = \mathcal{F}_i(d); \\ 0, & \text{if} \quad \mathcal{C}_{D_W}(d) = \mathcal{C}_{D_L}(d) \neq \mathcal{F}_i(d); \\ 0, & \text{otherwise} \end{cases}$$

Thus, we have $\delta_{\mathcal{C}_D,\mathcal{C}_{D_W}\cap\mathcal{C}_{D_L}}(d) \geq \delta_{\mathcal{F}_i,\mathcal{C}_{D_W}\cap\mathcal{C}_{D_L}}(d)$. This inequation holds true for any $\mathcal{F}_i \in \Omega$. Therefore, based on the definition of disagreement rate we can obtain

$$\eta(\mathcal{F}_i, \mathcal{C}_{D_W} \cap \mathcal{C}_{D_L}) = 1 - \frac{\sum_{d\in D}\delta_{\mathcal{F}_i,\mathcal{C}_{D_W}\cap\mathcal{C}_{D_L}}(d)}{|D|} \geq 1 - \frac{\sum_{d\in D}\delta_{\mathcal{C}_D,\mathcal{C}_{D_W}\cap\mathcal{C}_{D_L}}(d)}{|D|} = \eta(\mathcal{C}_D, \mathcal{C}_{D_W} \cap \mathcal{C}_{D_L})$$

Meanwhile, we can obtain $\eta(\mathcal{C}_D, \mathcal{C}_{D_W} \cap \mathcal{C}_{D_L}) = \eta(\mathcal{C}_{D_W}, \mathcal{C}_{D_L})$. □

References

Abney, S. (2002). Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, July 6-12, 2002, Philadelphia, PA, USA, pp. 360–367.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, June 23-30, 2007, Prague, Czech Republic, pp. 440–447.

Blitzer, J., Kakade, S., & Foster, D. P. (2011). Domain adaptation with coupled subspaces. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, April 11-13, 2011, Ft. Lauderdale, FL, USA, pp. 173–181.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, Madison, Wisconsin, USA, July 24-26, 1998, pp. 92–100.

Cai, P., Gao, W., Zhou, A. Y., & Wong, K. F. (2011a). Relevant knowledge helps in choosing right teacher: Active query selection for ranking adaptation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 24-28, 2011, Beijing, China, pp. 115–124.

Cai, P., Gao, W., Zhou, A. Y., & Wong, K. F. (2011b). Query weighting for ranking model adaptation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, June 19-24, Portland, Oregon, USA, pp. 112–122.

Chen, M.M., Weinberger, K. Q., & Blitzer, J. (2011). Co-training for domain adaptation. In *Proceedings of Advances in Neural Information Processing Systems 24*, December 12-14, 2011, Granada, Spain, pp. 1–9.

Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. In *Proceedings of 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 100–110.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Wiley-Interscience.

Dai, W. Y., Xue, G. R., Yang, Q., & Yu, Y. (2007a). Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, California, USA, August 12-15, 2007, pp. 210–219.

Dai, W. Y., Yang, Q., Xue, G. R., & Yu, Y. (2007b). Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, Oregon, USA, June 20-24, 2007, pp. 193–200.

Dasgupta, S., Littman, M. L., & McAllester, D. (2001). PAC generalization bounds for co-training. In *Proceedings of Advances in Neural Information Processing Systems 14*, December 9-14, 2002, Vancouver, British Columbia, Canada, pp. 375–382.

Daumé III, H., & Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26(2006):101–126.

Daumé III, H., Kumar, A., & Saha, A. (2010). Co-regularization based semi-supervised domain adaptation. In *Proceedings of Advances in Neural Information Processing Systems 23*, December 6-9, 2010, Vancouver, Canada, pp. 478–496.

Dayanik, A. A., Lewis, D. D., Madigan, D., Menkov, V., & Genkin, A. (2006). Constructing informative prior distributions from domain knowledge in text classification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, August 6-11, 2006, pp. 493–500.

Dhillon, I. S., Mallela, S., & Modha, D. S. (2003). Information-theoretic co-clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, August 24 - 27, 2003, pp. 210–219.

Gao, W., Blitzer, J., Zhou, M., & Wong, K. F. (2009). Exploiting bilingual information to improve web search. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, August 2-7, 2009, Singapore, pp. 1075–1083.

Gao, W., Cai, P., Wong, K. F., & Zhou, A. Y. (2010). Learning to rank only using training data from related domain. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 19-23, 2010, Geneva, Switzerland, pp. 162–169.

Gao, W., & Yang, P. (2014). Democracy is good for ranking: Towards multi-view rank learning and adaptation in web search. In *Proceedings of the 7th International ACM Conference on Web Search and Data Mining*, Feburary 25-27, 2014, New York City, USA, pp. 63–72.

He, J. R., & Lawrence, R. (2011) A graph-based framework for multi-task multi-view learning. In *Proceedings of the 28th International Conference on Machine Learning*, Washington, Jun 28-Jul 2, 2011, pp. 25–32.

Lewis, D. D. (2004). Reuters-21578 test collection. *http://www.daviddlewis.com/*.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning*, Bled, Slovenia, June 27-30, 1999, pp. 200–209.

Jiang, J., & Zhai, C. X. (2007). Instance weighting for domain Adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, June 23-30, 2007, Prague, Czech Republic, pp. 264–271.

Ling, X., Dai, W. Y., Xue, G. R., Yang, Q., & Yu, Y. (2008). Spectral domain-transfer learning. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, Nevada, USA, August 24-27, 2008, pp. 488–496.

McCallum, A. K., Nigam, K., Rennie, J., & Seymore, K. (2000). Automating the construction of Internet portals with machine learning. *Information Retrieval*, 3(2):127–163.

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

Rüping, S., & Scheffer, T. (2005). Learning with multiple views. In *Proceedings of 2005 ICML Workshop on Learning with Multiple Views*.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Sarinnapakorn, K., & Kubat, M. (2007). Combining sub-classifiers in text categorization: A DST-based solution and a case study. *IEEE Transactions Knowledge and Data Engineering*, 19(12):1638–1651.

Sridharan, K., & Kakade, S. M. (2008). An information theoretic framework for multi-view learning. In *Proceedings of the 21st Annual Conference on Learning Theory*, Helsinki, Finland, July 9-12, 2008, pp. 403–414.

Yang, P., Gao, W., Tan, Q., & Wong, K. F. (2012). Information-theoretic multi-view domain adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, July 8-11, 2012, Jeju Island, Korea, pp. 270–274.

Zhang, D., He, J. R., Liu, Y., Si, L., & Lawrence, R. D. (2011). Multi-view transfer learning with a large margin approach. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA, August 21-24, 2011, pp. 1208–1216.

Zhang, J. T., & Huan, J. (2012). Inductive multi-task learning with multiple view data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, China, August 12-16, 2012, pp. 543–551.