Information-theoretic Multi-view Domain Adaptation

Pei Yang[1,3]**, Wei Gao**[2]**, Qi Tan**[1]**, Kam-Fai Wong**[3]
[1]South China University of Technology, Guangzhou, China
{yangpei,tanqi}@scut.edu.cn
[2]Qatar Computing Research Institute, Qatar Foundation, Doha, Qatar
wgao@qf.org.qa
[3]The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
kfwong@se.cuhk.edu.hk

Abstract

We use multiple views for cross-domain document classification. The main idea is to strengthen the views' consistency for target data with source training data by identifying the correlations of domain-specific features from different domains. We present an Information-theoretic Multi-view Adaptation Model (IMAM) based on a multi-way clustering scheme, where word and link clusters can draw together seemingly unrelated domain-specific features from both sides and iteratively boost the consistency between document clusterings based on word and link views. Experiments show that IMAM significantly outperforms state-of-the-art baselines.

1 Introduction

Domain adaptation has been shown useful to many natural language processing applications including document classification (Sarinnapakorn and Kubat, 2007), sentiment classification (Blitzer et al., 2007), part-of-speech tagging (Jiang and Zhai, 2007) and entity mention detection (Daumé III and Marcu, 2006).

Documents can be represented by multiple independent sets of features such as words and link structures of the documents. Multi-view learning aims to improve classifiers by leveraging the redundancy and consistency among these multiple views (Blum and Mitchell, 1998; Rüping and Scheffer, 2005; Abney, 2002). Existing methods were designed for data from single domain, assuming that either view alone is sufficient to predict the target class 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.

Little research was done for multi-view domain adaptation. In this work, we present an Information-theoretical Multi-view Adaptation Model (IMAM) based on co-clustering framework (Dhillon et al., 2003) that combines the two learning paradigms to transfer class information across domains in multiple transformed feature spaces. IMAM exploits a multi-way-clustering-based classification scheme to simultaneously cluster documents, words and links into their respective clusters. In particular, the word and link clusterings can automatically associate the correlated features from different domains. Such correlations bridge the domain gap and enhance the consistency of views for clustering (i.e., classifying) the target data. Results show that IMAM significantly outperforms the state-of-the-art baselines.

2 Related Work

The work closely related to ours was done by Dai et al. (2007), where they proposed co-clustering-based classification (CoCC) for adaptation learning. CoCC was extended from information-theoretic co-clustering (Dhillon et al., 2003), where in-domain constraints were added to word clusters to provide the class structure and partial categorization knowledge. However, CoCC is a single-view algorithm.

Although multi-view learning (Blum and Mitchell, 1998; Dasgupta et al., 2001; Abney, 2002; Sridharan and Kakade, 2008) is common within a single domain, it is not well studied under cross-domain settings. Chen et al. (2011) proposed

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 270–274,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

CODA for adaptation based on co-training (Blum and Mitchell, 1998), which is however a pseudo multi-view algorithm where original data has only one view. Therefore, it is not suitable for the true multi-view case as ours. 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. However, instance-based approach is generally poor since new target features lack support from source data (Blitzer et al., 2011). We focus on feature-level multi-view adaptation.

3 Our Model

Intuitively, source-specific and target-specific features can be drawn together by mining their co-occurrence with domain-independent (common) features, which helps bridge the distribution gap. Meanwhile, the view consistency on target data can be strengthened if target-specific features are appropriately bundled with source-specific features. Our model leverages the complementary cooperation between different views to yield better adaptation performance.

3.1 Representation

Let D_S be the source training documents and D_T be the unlabeled target documents. Let C be the set of class labels. Each source document $d_s \in D_S$ is labeled with a unique class label $c \in C$. Our goal is to assign each target document $d_t \in D_T$ to an appropriate class as accurately as possible.

Let W be the vocabulary of the entire document collection $D = D_S \cup D_T$. Let L be the set of all links (hyperlinks or citations) among documents. Each $d \in D$ can be represented by two views, i.e., a bag-of-words set $\{w\}$ and a bag-of-links set $\{l\}$.

Our model explores multi-way clustering that simultaneously clusters documents, words and links. Let $\hat{D}$, $\hat{W}$ and $\hat{L}$ be the respective clustering of documents, words and links. The clustering functions are defined as $\mathcal{C}_D(d) = \hat{d}$ for document, $\mathcal{C}_W(w) = \hat{w}$ for word and $\mathcal{C}_L(l) = \hat{l}$ for link, where $\hat{d}$, $\hat{w}$ and $\hat{l}$ represent the corresponding clusters.

3.2 Objectives

We extend the information-theoretic co-clustering framework (Dhillon et al., 2003) to incorporate the loss from multiple views. Let $\mathcal{I}(X, Y)$ be mutual information (MI) of variables X and Y, our objective is to minimize the MI loss of two different views:

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

where

$$\Theta_W = \mathcal{I}(D_T, W) - \mathcal{I}(\hat{D}_T, \hat{W}) + \lambda \cdot \Big[\mathcal{I}(C, W) - \mathcal{I}(C, \hat{W})\Big]$$

$$\Theta_L = \mathcal{I}(D_T, L) - \mathcal{I}(\hat{D}_T, \hat{L}) + \lambda \cdot \Big[\mathcal{I}(C, L) - \mathcal{I}(C, \hat{L})\Big]$$

Θ_W and Θ_L are the loss terms based on word view and link view, respectively, traded off by α. λ balances the effect of word or link clusters from co-clustering. When $\alpha = 1$, the function relies on text only that reduces to CoCC (Dai et al., 2007).

For any $x \in \hat{x}$, we define conditional distribution $q(x|\hat{y}) = p(x|\hat{x})p(\hat{x}|\hat{y})$ under co-clustering $(\hat{X}, \hat{Y})$ based on Dhillon et al. (2003). 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: $q(w|\hat{d})$, $q(d|\hat{w})$, $q(l|\hat{d})$, $q(d|\hat{l})$, $q(c|\hat{w})$, $q(c|\hat{l})$.

Eq. 1 is hard to optimize due to its combinatorial nature. We transform it to the equivalent form based on Kullback-Leibler (KL) divergence between two conditional distributions $p(x|y)$ and $q(x|\hat{y})$, where $\mathcal{D}(p(x|y)||q(x|\hat{y})) = \sum_x p(x|y) log \frac{p(x|y)}{q(x|\hat{y})}$.

Lemma 1 (Objective functions) *Equation 1 can be turned into the form of alternate minimization: (i) For document clustering, 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*

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

(ii) For word and link clustering, we minimize

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

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

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

[1] We can obtain that $\phi_C(\hat{W}, \hat{L}) = \lambda \Big[\alpha(\mathcal{I}(C, W) - \mathcal{I}(C, \hat{W})) + (1 - \alpha)(\mathcal{I}(C, L) - \mathcal{I}(C, \hat{L}))\Big]$, which is constant since word/link clusters keep fixed during the document clustering step.

Lemma 1[2] allows us to *alternately* reorder either documents or both words and links by fixing the other in such a way that the MI loss in Eq. 1 decreases monotonically.

4 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.

According to Lemma 1, minimizing $\phi_D(d, \hat{d})$ for each d can reduce the objective function value iteratively (t denotes round id):

$$\mathcal{C}_D^{(t+1)}(d) = \arg\min_{\hat{d}} \Big[\alpha \cdot \mathcal{D}(p(w|d)||q^{(t)}(w|\hat{d})) + (1-\alpha) \cdot \mathcal{D}(p(l|d)||q^{(t)}(l|\hat{d}))\Big] \quad (2)$$

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

$$\mathcal{C}_{D_W}^{(t+1)}(d) = \arg\min_{\hat{d}} \mathcal{D}(p(w|d)||q^{(t)}(w|\hat{d})) \quad (3)$$

$$\mathcal{C}_{D_L}^{(t+1)}(d) = \arg\min_{\hat{d}} \mathcal{D}(p(l|d)||q^{(t)}(l|\hat{d})) \quad (4)$$

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 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 remains consistent with the view-based clusterings as much as possible. The more consistent the views, the better the document clustering, and then the better the word and link clustering, which creates a positive cycle.

4.1 Disagreement Rate of Views

For any document, a consistency indicator function with respect to the two view-based clusterings can be defined as follows (t is omitted for simplicity):

Definition 1 (Indicator function) *For any $d \in D$,*

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

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

Definition 2 (Disagreement rate)

$$\eta(\mathcal{C}_{D_W}, \mathcal{C}_{D_L}) = 1 - \frac{\sum_{d \in D} \delta_{\mathcal{C}_{D_W},\mathcal{C}_{D_L}}(d)}{|D|} \quad (5)$$

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, Eq. 5 is not continuous nor convex, which is difficult to optimize directly. By using the optimization based on Lemma 1, we can show empirically that disagreement rate is monotonically decreased (see Section 5).

4.2 View Combination

In practice, view-based document clusterings in Eq. 3 and 4 are not computed explicitly. Instead, Eq. 2 directly optimizes view combination and produces the document clustering. Therefore, it is necessary to disclose how consistent it could be with the view-based clusterings.

Suppose $\Omega = \{\mathcal{F}_D | \mathcal{F}_D(d) = \hat{d}, \hat{d} \in \hat{D}\}$ is the set of all document clustering functions. For any $\mathcal{F}_D \in \Omega$, we obtain the disagreement rate $\eta(\mathcal{F}_D, \mathcal{C}_{D_W} \cap \mathcal{C}_{D_L})$, where $\mathcal{C}_{D_W} \cap \mathcal{C}_{D_L}$ denotes the clustering resulting from the overlap of the view-based clusterings.

Lemma 2 *$\mathcal{C}_D$ always minimizes the disagreement rate for any $\mathcal{F}_D \in \Omega$ such that*

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

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

Lemma 2 suggests that IMAM always finds the document clustering with the minimal disagreement rate to the overlap of view-based clusterings, and the minimal value of disagreement rate equals to the disagreement rate of the view-based clusterings.

[2] Due to space limit, the proof of all lemmas will be given in a long version of the paper.

Table 1: View disagreement rate η and error rate ϵ that decrease with iterations and their Pearson's correlation γ.

Round		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 2: Comparison of error rate with baselines.

Data	TSVM	Co-Train	CoCC	MVTL-LM	IMAM
DA-EC	0.214	0.230	0.149	0.192	**0.138**
DA-NT	0.114	0.163	0.106	0.108	**0.069**
DA-OS	0.262	0.175	0.075	0.068	**0.039**
DA-ML	0.107	0.171	0.109	0.183	**0.047**
EC-NT	**0.177**	0.296	0.225	0.261	0.192
EC-OS	0.245	0.175	0.137	0.176	**0.074**
EC-ML	**0.168**	0.206	0.203	0.264	0.173
NT-OS	0.396	0.220	0.107	0.288	**0.070**
NT-ML	0.101	0.132	0.054	0.071	**0.032**
OS-ML	0.179	0.128	0.051	0.126	**0.021**
Average	0.196	0.190	0.122	0.174	**0.085**

5 Experiments and Results

Data and Setup

Cora (McCallum et al., 2000) is an online archive of computer science articles. The documents in the archive are categorized into a hierarchical structure. We selected a subset of Cora, which contains 5 top categories and 10 sub-categories. We used a similar way as Dai et al. (2007) 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 dataset denoted as DA-EC consists of source domain: DA_1(+), EC_1(-); and target domain: DA_2(+), EC_2(-).

The classification error rate ϵ is measured as the proportion of misclassified target documents. In order to avoid the infinity values, we applied Laplacian smoothing when computing the KL-divergence. We tuned α, λ and the number of word/link clusters by cross-validation on the training data.

Results and Discussions

Table 1 shows the monotonic decrease of view disagreement rate η and error rate ϵ with the iterations and their Pearson's correlation γ is nearly perfectly positive. This indicates that IMAM gradually improves adaptation by strengthening the view consistency. This is achieved by the reinforcement of word and link clusterings that draw together target- and source-specific features that are originally unrelated but co-occur with the common features.

We compared IMAM with (1) Transductive SVM (TSVM) (Joachims, 1999) using both words and links features; (2) Co-Training (Blum and Mitchell, 1998); (3) CoCC (Dai et al., 2007): Co-clustering-based single-view transfer learner (with text view only); and (4) MVTL-LM (Zhang et al., 2011): Large-margin-based multi-view transfer learner.

Table 2 shows the results. Co-Training performed a little better than TSVM by boosting the confidence of classifiers built on the distinct views in a complementary way. But since Co-Training doesn't consider the distribution gap, it performed clearly worse than CoCC even though CoCC has only one view.

IMAM significantly outperformed CoCC on all the datasets. In average, the error rate of IMAM is 30.3% lower than that of CoCC. 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.

MVTL-LM performed worse than CoCC. It suggests that instance-based approach is not effective when the data of different domains are drawn from different feature spaces. Although MVTL-LM regulates view consistency, it cannot identify the associations between target- and source-specific features that is the key to the success of adaptation especially when domain gap is large and less commonality could be found. In contrast, CoCC and IMAM uses multi-way clustering to find such correlations.

6 Conclusion

We presented a novel feature-level multi-view domain adaptation approach. The thrust is to incorporate distinct views of document features into the information-theoretic co-clustering framework and strengthen the consistency of views on clustering (i.e., classifying) target documents. The improvements over the state-of-the-arts are significant.

References

Steven Abney. 2002. Bootstrapping. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 360-367.

John Blitzer, Mark Dredze and Fernado Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440-447.

John Blitzer, Sham Kakade and Dean P. Foster. 2011. Domain Adaptation with Coupled Subspaces. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 173-181.

Avrim Blum and Tom Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92-100.

Minmin Chen, Killian Q. Weinberger and John Blitzer. 2011. Co-Training for Domain Adaptation. In *Proceedings of NIPS*, pages 1-9.

Wenyuan Dai, Gui-Rong Xue, Qiang Yang and Yong Yu. 2007. Co-clustering Based Classification for Out-of-domain Documents. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 210-219.

Sanjoy Dasgupta, Michael L. Littman and David McAllester. 2001. PAC Generalization Bounds for Co-Training. In *Proceeding of NIPS*, pages 375-382.

Hal Daumé III and Daniel Marcu. 2006. Domain Adaptation for Statistical Classifiers. *Journal of Artificial Intelligence Research*, 26(2006):101-126.

Inderjit S. Dhillon, Subramanyam Mallela and Dharmendra S. Modha. 2003. Information-Theoretic Co-clustering. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 210-219.

Thorsten Joachims. 1999. Transductive Inference for Text Classification using Support Vector Machines. In *Proceedings of Sixteenth International Conference on Machine Learning*, pages 200-209.

Jing Jiang and Chengxiang Zhai. 2007. Instance Weighting for Domain Adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264-271.

Andrew K. McCallum, Kamal Nigam, Jason Rennie and Kristie Seymore. 2000. Automating the Construction of Internet Portals with Machine Learning. *Information Retrieval*, 3(2):127-163.

Stephan Rüping and Tobias Scheffer. 2005. Learning with Multiple Views. In *Proceedings of ICML Workshop on Learning with Multiple Views*.

Kanoksri Sarinnapakorn and Miroslav Kubat. 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.

Karthik Sridharan and Sham M. Kakade. 2008. An Information Theoretic Framework for Multi-view Learning. In *Proceedings of the 21st Annual Conference on Computational Learning Theory*, pages 403-414.

Dan Zhang, Jingrui He, Yan Liu, Luo Si and Richard D. Lawrence. 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*, pages 1208-1216.