Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Multi-View Discriminant Transfer Learning

Pei Yang[1] **and Wei Gao**[2]
[1]South China University of Technology, Guangzhou, China
yangpei@scut.edu.cn
[2]Qatar Computing Research Institute, Doha, Qatar
wgao@qf.org.qa

Abstract

We study to incorporate multiple views of data in a *perceptive* transfer learning framework and propose a Multi-view Discriminant Transfer (MDT) learning approach for domain adaptation. The main idea is to find the optimal discriminant weight vectors for each view such that the correlation between the two-view projected data is maximized, while both the domain discrepancy and the view disagreement are minimized simultaneously. Furthermore, we analyze MDT theoretically from discriminant analysis perspective to explain the condition and reason, under which the proposed method is not applicable. The analytical results allow us to investigate whether there exist within-view and/or between-view conflicts, and thus provides a deep insight into whether the transfer learning algorithm work properly or not in the view-based problems and the combined learning problem. Experiments show that MDT significantly outperforms the state-of-the-art baselines including some typical multi-view learning approaches in single- or cross-domain.

1 Introduction

Transfer learning allows the domains, distributions, and feature spaces used in training being different from those in testing [Pan and Yang, 2010]. It utilizes labeled data available from some related (or source) domain in order to achieve effective knowledge transformation from it to the target domain. It is of great importance in many data mining applications, such as document classification [Sarinnapakorn and Kubat, 2007], sentiment classification [Blitzer *et al.*, 2011], collaborative filtering [Pan *et al.*, 2010], and Web search ranking [Gao *et al.*, 2010].

Many types of data are described with multiple views or perspectives. Multi-view learning aims to improve classifiers by leveraging the redundancy and consistency among distinct views [Blum and Mitchell, 1998; Rüping and Scheffer, 2005; Abney, 2002]. Most existing multi-view algorithms were designed for single domain, assuming that either view alone is sufficient for the prediction of target class. However, this view-consistency assumption is largely violated in the setting of transfer learning where training and test data are drawn from different distributions and/or even from distinct feature space. Little research has been done on multi-view transfer learning in the literature.

A fundamental problem in machine learning is to determine when and why a given technique is applicable [Martínez and Zhu, 2005]. However, for most existing transfer learning methods, the conditions regarding when the algorithms work properly are yet unclear. This paper is motivated to incorporate the multiple views of data across different domains in a *perceptive* transfer learning framework. Here "*perceptive*" means it is known when the proposed method works properly prior to its deployment. We proposed the Multi-view Discriminant Transfer (MDT) learning approach. Its objective is to find the optimal discriminant weight vectors for each view such that the correlation between the two-view projected data is maximized, while both the domain discrepancy and the view disagreement are minimized simultaneously. MDT incorporates the domain discrepancy and the view disagreement by taking a discriminant analysis approach, which can be transformed into a generalized eigenvalue problem. Then, we investigate theoretical conditions regarding when the proposed multi-view transfer method works properly from discriminant analysis perspective. The theoretical results allow us to measure the balance between the view-based discriminant power, and investigate whether there exist within-view and/or between-view conflicts. Under such conflicts, the learning algorithm may not work properly. Obviously, knowing when the proposed multi-view transfer learning work beforehand is crucial to many real-world applications especially when either domains or views are too "dissimilar".

The major contributions of this paper can be highlighted as follows: (1) We propose a novel MDT approach to incorporate the multi-view information across different domains for transfer learning. It incorporates the domain discrepancy and the view disagreement by taking a discriminant analysis approach, which leads to a compact and efficient solution. It addresses the questions of what and how to transfer. (2) We present a theoretical study on the MDT model to illustrate when and why the proposed method is not applicable, which answers the the question of when to transfer. To the best of our knowledge, there is no existing work focusing on the theory regarding when a multi-view transfer learning method works properly. (3) Experiments show that MDT significantly outperforms the state-of-the-art baselines.

2 Related Work

Transfer learning models data that are from related but not identically distributed sources. As pointed out by [Pan and Yang, 2010], there are three fundamental issues in transfer learning, i.e., what to transfer, how to transfer, and when to transfer. Despite the importance of avoiding negative transfer, little research has been done for "when to transfer" [Cao *et al.*, 2010; Yao and Doretto, 2010].

How to measure domain distance is also important to transfer learning. Pan et al. [2011] proposed transfer component analysis (TCA) for reducing distance between domains in a latent space for domain adaptation. Huang et al. [2006] proposed Kernel Mean Matching (KMM) approach to reweight the instances in source domain so as to minimize the marginal probability difference between two domains. Quanz and Huan [2009] defined the projected maximum mean discrepancy (MMD) to estimate the distribution distance under a given projection. We use the projected MMD to estimate the domain distance in both views because it is very effective and easy to be incorporated into our framework.

Multi-view learning has been studied extensively under single-domain setting, such as Co-Training [Blum and Mitchell, 1998] and its extensions [Collins and Singer, 1999; Dasgupta *et al.*, 2001]. Abney [2002] relaxed the view independence assumption and suggested that the disagreement rate of two independent hypotheses upper bounds the error rate of either hypothesis. Nevertheless, multi-view learning is not effective for transfer since they treat distinct domains indiscriminately.

Little was done for multi-view transfer. Chen et al. [2011] proposed 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 and may not be effective for the true multi-view case. Zhang et al. [2011] presented an instance-level multi-view transfer algorithm (MVTL-LM) that integrates classification loss and view consistency terms in a large-margin framework. Unlike MVTL-LM, MDT is of feature level which mines the correlations between views together with the domain distance measure to improve the transfer, and a theoretical analysis shows that the model is *perceptive*.

Linear discriminant analysis (LDA), which is also called Fisher discriminant analysis (FDA) [Fisher, 1938], searches for those vectors in the underlying feature space that can best discriminate classes. Its goal is to maximize the between-class distance while minimizing the within-class distance. LDA has played a major role in the areas of machine learning and pattern recognition, such as feature extraction, classification and clustering [Belhumeur *et al.*, 1997]. The idea of Kernel Fisher Discriminant (KFD) [Mika *et al.*, 2001] is to solve the problem of FDA in a kernel feature space, thereby yielding a nonlinear discriminant given the input space.

FDA2 [Diethe *et al.*, 2008], a two-view extension of FDA, was proposed to incorporate multi-view data with labels into the Canonical Correlation Analysis (CCA) [Melzer *et al.*, 2003] framework. Our proposed mothed extends FDA2 by taking into account the domain discrepancy and enhancing view consistency, thus leads to better adaptation performance.

Martínez and Zhu [2005] reported on a theoretical study demonstrating the condition the LDA-based methods do not work. They showed that the discriminant power is related to the eigensystems of the matrices that define the measure to be maximized and minimized. We further extend [Martínez and Zhu, 2005] to the multi-view scenario which could provide a deep insight into when the algorithm work properly on the view-based problems and the combined problem.

3 Multi-view Discriminant Transfer Model

3.1 Problem Statement

Suppose we are given a set of labeled source-domain data $D_s = \{(x_i, z_i, y_i)\}_{i=1}^{n}$ and unlabeled target-domain data $D_t = \{(x_i, z_i, ?)\}_{i=n+1}^{n+m}$ consisting of two views, where x_i and z_i are *column* vectors of the ith instance from the first and second views respectively, and $y_i \in \{-1, 1\}$ is its class label. The source and target domain data follow different distributions. Our goal is to assign the appropriate class label to the instance in the target domain.

Let $\phi(\cdot)$ be the kernel function of mapping the instances from the original feature space to a reproducing kernel Hilbert space (RKHS). Let w_x and w_z be the weights vectors in the mapped feature space for the first and the second views, respectively. Define the data matrix for the first view, $X = (X_s^T, X_t^T)^T$ where $X_s = (\phi(x_1), \cdots, \phi(x_n))^T$ and $X_t = (\phi(x_{n+1}), \cdots, \phi(x_{n+m}))^T$. Define Z, Z_s, and Z_t for the second view respectively. The class label vector of the source data is denoted by $y = (y_1, \cdots, y_n)^T$. Let n^+ and n^- be the number of positive and negative instances in the source domain.

3.2 Two-view Fisher Discriminant Analysis

Diethe et al. [2008] extended Fisher Discriminant Analysis (FDA) into FDA2 by incorporating the labeled two-view data into the Canonical Correlation Analysis (CCA) [Melzer *et al.*, 2003] framework as follows:

$$\max_{(w_x, w_z)} \frac{w_x^{\mathrm{T}} M_w w_z}{\sqrt{w_x^{\mathrm{T}} M_x w_x} \cdot \sqrt{w_z^{\mathrm{T}} M_z w_z}} \tag{1}$$

where

$$\begin{aligned} M_w &= X_s^{\mathrm{T}} y y^{\mathrm{T}} Z_s \\ M_x &= \frac{1}{n} \sum_{i=1}^{n} (\phi(x_i) - \mu_x)(\phi(x_i) - \mu_x)^{\mathrm{T}} \\ M_z &= \frac{1}{n} \sum_{i=1}^{n} (\phi(z_i) - \mu_z)(\phi(z_i) - \mu_z)^{\mathrm{T}} \end{aligned}$$

where μ_x and μ_z are the means of the source data from the two views such as $\mu_x = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$, respectively. The numerator in Eq.(1) reflects the between-class distance which needs to be maximized, while the denominator reflects the within-class distance which should be minimized. The above optimization problem is equivalent to selecting those vectors which maximize the Rayleigh quotient [Melzer *et al.*, 2003]

$$r = \frac{\xi^{\mathrm{T}} Q_w \xi}{\xi^{\mathrm{T}} P \xi} \tag{2}$$

where

$$Q_w = \begin{pmatrix} \mathbf{0} & M_w \\ M_w^{\mathrm{T}} & \mathbf{0} \end{pmatrix},\ P = \begin{pmatrix} M_x & \mathbf{0} \\ \mathbf{0} & M_z \end{pmatrix},\ \xi = \begin{pmatrix} w_x \\ w_z \end{pmatrix} \quad (3)$$

Note that Q_w encodes the between-class distance, while P encodes the compound information about the view-based within-class distances. ξ is the eigenvector. Such an optimization is different from [Diethe *et al.*, 2008] and facilitates the extension of FDA2 to cross-domain scenario, which will be presented in following sub-section.

For an unlabeled instance, $(x_i, z_i, ?) \in D_t$, the classification decision function is given as follows:

$$f(x_i, z_i) = \left[w_x^{\mathrm{T}}\phi(x_i) + w_z^{\mathrm{T}}\phi(z_i) - b\right] \quad (4)$$

where the threshold $b = b_x + b_z$, b_x and b_z are chosen to bisect the two centers of mass of the source data from each view such as $w_x^T\mu_x^+ - b_x = b_x - w_x^T\mu_x^-$ where μ_x^+ and μ_x^- are the means of source positive and negative instances, respectively.

3.3 The Proposed MDT Model

Our goal is to incorporate FDA2, domain distance and view consistency into a unified discriminant analysis framework. The main idea is to find the optimal discriminant weight vectors for each view such that the correlation between the projections of the two-view data onto these weight vectors is maximized, while both the domain discrepancy and view disagreement are minimized simultaneously.

Domain Distance

Quanz and Huan [Quanz and Huan, 2009] defined the projected maximum mean discrepancy (MMD) to estimate the distribution distance under a given projection. Here we adopt projected MMD [Quanz and Huan, 2009] to estimate the domain distance for each view such as:

$$||\frac{1}{n}\sum_{i=1}^{n} w_x^T\phi(x_i) - \frac{1}{m}\sum_{i=n+1}^{n+m} w_x^T\phi(x_i)||^2 = w_x^T X^T \mathbf{L} X w_x$$

where

$$\mathbf{L} = \begin{pmatrix} \frac{\mathbf{1}_{n\times n}}{n^2} & -\frac{\mathbf{1}_{n\times m}}{nm} \\ -\frac{\mathbf{1}_{m\times n}}{nm} & \frac{\mathbf{1}_{m\times m}}{m^2} \end{pmatrix}$$

The domain distance for both views can be summed up as follows:

$$w_x^T X^T \mathbf{L} X w_x + w_z^T Z^T \mathbf{L} Z w_z = \xi^T Q_d \xi \quad (5)$$

where

$$Q_d = \begin{pmatrix} X^T\mathbf{L}X & \mathbf{0} \\ \mathbf{0} & Z^T\mathbf{L}Z \end{pmatrix}$$

View Consistency

Maximizing view consistency is equivalent to minimizing the disagreement of view-specific classifiers. We use both labeled source data and unlabeled target data to estimate the difference of predictions resulting from distinct views as follows:

$$\sum_{i=1}^{n+m} ||w_x^T\phi(x_i) - w_z^T\phi(z_i)||^2 = \xi^T Q_c\, \xi \quad (6)$$

where

$$Q_c = \begin{pmatrix} X^TX & -X^TZ \\ -Z^TX & Z^TZ \end{pmatrix}$$

Algorithm 1 Co-Train based MDT Algorithm

Input:
The source dataset $D_s = \{(x_i, z_i, y_i)\}_{i=1}^{n}$
The target dataset $D_t = \{(x_i, z_i, ?)\}_{i=n+1}^{n+m}$
Output:
Class label assigned to each instance in D_t;

1: **repeat**
2: Solve the generalized eigenvalue problem defined in Eq.(8), and then obtain the eigenvector ξ with the largest eigenvalue;
3: Use Eq.(4) to predict the target instance $(x_i, z_i) \in D_t$, which is labeled as $sign[f(x_i, z_i)]$;
4: Move κ most confident positive and negative instances with top absolute predicted scores $|f(x_i, z_i)|$ from D_t to D_s separately;
5: **until** Convergence is reached;

Overall Objective

In summary, Eq.(5) is to minimize domain distance $\xi^T Q_d \xi$, and Eq.(6) is to minimize view disagreement $\xi^T Q_c \xi$. The particular forms of both domain distance and view disagreement make them easier to be incorporated into the FDA2 framework. We define $Q = Q_w - c_1 Q_d - c_2 Q_c$ where c_1, c_2 are trade-off coefficients. By integrating the domain distance and view disagreement into Eq.(2), the overall objective of MDT is to maximize

$$r = \frac{\xi^{\mathrm{T}} Q \xi}{\xi^{\mathrm{T}} P \xi} \quad (7)$$

which is equivalent to solving the following generalized eigenvalue problem [Duda *et al.*, 2001]:

$$Q\xi = \lambda P \xi \quad (8)$$

where λ is the eigenvalue, and ξ is the eigenvector.

The eigenvectors corresponding to the largest eigenvalues represent the maximally correlated directions in feature space. It is straightforward to resolve this eigenvalue problem and obtain w_x and w_z.

Our Co-Train [Blum and Mitchell, 1998] based algorithm is given in Algorithm 1. In each iteration, it moves the most confident target instances to the source training set so that the performance can be gradually boosted. For the free parameter κ, we empirically set $\kappa = 5\%$.

4 Theoretical Analysis

We present the theoretical analysis on the proposed model to illustrate when the approach would not work properly.

Many machine learning problems can be formulated as an eigenvalue decomposition problem [Martínez and Zhu, 2005]. It is of great importance to analyse whether these algorithms work or not. Martínez and Zhu [2005] showed that when such approaches work properly is related to the eigensystems between Q and P. Specifically, the discriminant power $tr(P^{-1}Q)$ is related to the ratios between the eigenvalues of Q and P, as well as the angles between the their corresponding eigenvectors. The algorithm would become unstable if we cannot maximize $\xi^{\mathrm{T}} Q \xi$ and minimize $\xi^{\mathrm{T}} P \xi$ simultaneously, which is referred to as the conflict between the eigensystems of Q and P.

However, under the multi-view situation, these results can not be directly used to analyse the view-based discriminant power. It is blind to whether there exist between-view and/or within-view conflicts. Therefore, we further extend these results to a multi-view setting which is given in Lemma 1.

Lemma 1. *Suppose r_q, r_x, and r_z are the ranks of Q, M_x, and M_z, respectively. The discriminant power $tr(P^{-1}Q)$ is calculated as:*

$$tr(P^{-1}Q) = \sum_{i=1}^{r_q}\sum_{j=1}^{r_x}\frac{\lambda_{q_i}}{\lambda_{x_j}}\left[\begin{pmatrix}\xi_{x_j}\\ \mathbf{0}\end{pmatrix}^{\mathrm{T}}\xi_{q_i}\right]^2 + \sum_{i=1}^{r_q}\sum_{j=1}^{r_z}\frac{\lambda_{q_i}}{\lambda_{z_j}}\left[\begin{pmatrix}\mathbf{0}\\ \xi_{z_j}\end{pmatrix}^{\mathrm{T}}\xi_{q_i}\right]^2 \quad (9)$$

where $\lambda_{q_i}(1 \le i \le r_q)$ and ξ_{q_i} are the i-th largest eigenvalue and the corresponding eigenvector of $Q\xi = \lambda\xi$, $\lambda_{x_j}(1 \le j \le r_x)$ and ξ_{x_j} are the j-th largest eigenvalue and the corresponding eigenvector of $M_x\xi = \lambda\xi$, and $\lambda_{z_j}(1 \le j \le r_z)$ and ξ_{z_j} are the j-th largest eigenvalue and the corresponding eigenvector of $M_z\xi = \lambda\xi$.

The proof of the lemma is given in the Appendix. Lemma 1 shows that the total discriminant power $tr(P^{-1}Q)$ can be decomposed into view-base discriminant powers, i.e., the first and second items in the right hand side of Eq.(9). It indicates whether our proposed algorithm works properly or not is pertinent to the relationship among the eigensystems of Q, M_x, and M_z. Based on Lemma 1, each pair of eigensystems (x_j, q_i) (or (z_j, q_i)) will have a discriminant power such as $\frac{\lambda_{q_i}}{\lambda_{x_j}}\left[\begin{pmatrix}\xi_{x_j}\\ \mathbf{0}\end{pmatrix}^{\mathrm{T}}\xi_{q_i}\right]^2$. Those pairs with similar eigenvectors will have a higher weight $v(x_j, q_i) = \left[\begin{pmatrix}\xi_{x_j}\\ \mathbf{0}\end{pmatrix}^{\mathrm{T}}\xi_{q_i}\right]^2$ than those that differ. When the pair (x_j, q_i) that agree correspond to a small eigenvalue ratio $\frac{\lambda_{q_i}}{\lambda_{x_j}}$, the results are not guaranteed to be optimal. In this case, the results will be determined by the ratios between the eigenvalues of M_x and Q.

The power of Lemma 1 is that the results presented above allow us to measure the balance between the view-based discriminant power, and investigate whether there exist within-view and/or between-view conflicts. Specifically, within-view conflict means Q and M_x (or M_z) favor different solution directions, while between-view conflict means view-based classifiers favor different solution directions. A simplified illustrative example will be given in the next section. Therefore, it provides a deep insight into whether the algorithm work properly or not on the view-based problems and the combined learning problem, as well as their correlation.

Note that Q encodes the compound information about between-class distance, domain distance and view consistency that defines the measure to be maximized, while M_x and M_z encodes the information about the within-class distance that defines the measure to be minimized for the view-based learning problems, respectively. It is worth noting the interpretation here is applicable to both multi-view transfer learning ($c_1 \neq 0$) and general multi-view learning ($c_1 = 0$) since they share the same mathematical form as Eq.(8).

5 Experiments

5.1 Synthetic Dataset

First, we generate the synthetic dataset to provide an intuitive geometric interpretation to the theoretical analysis of the proposed model. Two three-class datasets with two views are generated. The datasets are detailed in Table 1. For each class, 100 instances are randomly drawn from a two-dimensional Gaussian distribution with the specified mean and covariance. The 2D scatter plots for the two synthetic datasets are shown in Figure 1 and 2. After the datasets are generated, we can obtain Q, M_x, and M_z for each dataset. Then the eigenvalues are given in Table 1, and the eigenvectors are shown as the dashed lines in Figure 1 and 2.

Figure 1 shows an example where the algorithm works well on the first synthetic dataset. According to Eq.(7), the objective is to maximize the measure given by Q, i.e., between-class distance from the two views, while minimizing those of M_x and M_z, i.e., within-class distance in the first and second view, respectively [1]. For the first view shown in Figure 1(a), Q would like to select ξ_{q_1} rather than ξ_{q_2} to maximize the between-class distance since $\lambda_{q_1} > \lambda_{q_2}$. Likewise, M_x prefers to select ξ_{x_2} rather than ξ_{x_1} to minimize the within-class distance in view 1 since $\lambda_{x_2} < \lambda_{x_1}$. Thus, both Q and M_x agree with each other on the same direction $\xi_1^* = \xi_{q_1} = \xi_{x_2}$. However, for the second view, Q would like to select ξ_{q_1} as a solution, whereas M_z prefers to select ξ_{z_2}. It indicates that there exists a within-view conflict. Based on Lemma 1, the model weights each pair of eigenvectors (ξ_{z_j}, ξ_{q_i}) according to their agreement. Here we have $v(q_1, z_1) = v(q_2, z_2) = 1 > v(q_1, z_2) = v(q_2, z_1) = 0$. In this case, whether the result is optimal or not will be determined by the eigenvalues ratio between Q and M_z. Since $\frac{\lambda_{q_1}}{\lambda_{z_1}} = 5.84 > \frac{\lambda_{q_2}}{\lambda_{z_2}} = 5.65$, the solution direction for view 2 will be $\xi_2^* = \xi_{q_1} = \xi_{z_1}$ with the corresponding larger eigenvalue ratio. In summary, since the two views agree on the same direction, the final solution direction is $\xi^* = \xi_1^* = \xi_2^*$, which is optimal though there are within-view conflict in the second view.

Figure 2 shows an example where the algorithm fails on the second synthetic dataset. Note that the parameters to generate the two datasets are nearly the same except for the highlighted means of the third class in the second view, as shown in Table 1. The similar analysis shows that there exists a conflict between the views. The algorithm selects $\xi^* = \xi_2^*(\perp\xi_1^*)$ as the final solution direction, which however is not correct.

5.2 Real Dataset

Data and Setup

Cora [McCallum *et al.*, 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

[1] To provide an intuitive interpretation, the example is simplified by considering within-class and between-class distances only.

Table 1: The description of the synthetic dataset.

Datasets	View 1		View 2		Eigenvalues					
	Covariance	Means for three classes	Covariance	Means for three classes	λ_{x_1}	λ_{x_2}	λ_{z_1}	λ_{z_2}	λ_{q_1}	λ_{q_2}
SynSet 1	diag(1,3)	[-5,0], [5,0], [0,5]	diag(3,1)	[-5,0], [5,0], **[0,5]**	7.19	2.57	8.24	2.65	48.16	14.96
SynSet 2	diag(1,3)	[-5,0], [5,0], [0,5]	diag(3,1)	[-5,0], [5,0], **[0,25]**	9.15	3.02	8.26	2.77	94.30	46.45

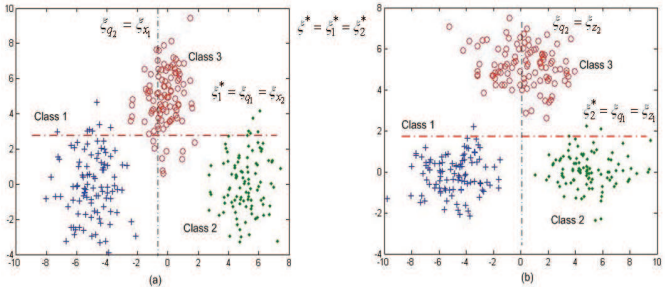


Figure 1: An example illustrating that the algorithm works properly on SynSet1. (a) In the first view, both Q and M_x agree with each other on the direction $\xi_1^* = \xi_{q_1} = \xi_{x_2}$. (b) In the second view, Q and M_z disagree with each other, and the solution direction for view 2 is $\xi_2^* = \xi_{q_1} = \xi_{z_1}$. Since the two views agree with each other, the final solution direction is $\xi^* = \xi_1^* = \xi_2^*$, which is optimal.

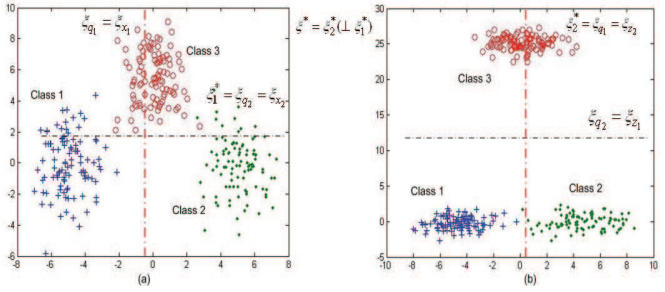


Figure 2: An example illustrating that the algorithm fails on SynSet2. (a) In the first view, Q and M_x disagree with each other, and the solution direction for view 1 is $\xi_1^* = \xi_{q_2} = \xi_{x_2}$. (b) In the second view, Q and M_z agree with each other on the direction $\xi_2^* = \xi_{q_1} = \xi_{z_2}$. In this case, the two views disagree with each other. The algorithm selects $\xi^* = \xi_2^*(\perp \xi_1^*)$ as the final solution direction, which however is not correct.

subset of Cora with 5 top categories and 10 sub-categories:
- 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/ (1102)
- ML_1: /machine_learning/probabilistic_methods/ (687)
- ML_2: /machine_learning/genetic_algorithms/ (670)

We used a similar way as [Pan and Yang, 2010] 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 method ensures the domains of labeled and unlabeled data related due to the same top categories, but the domains are different because they are drawn from different sub-categories.

We preprocessed the data for both text and link information. We removed words or links with frequency less than 5. Then the standard TF-IDF [Salton and Buckley, 1988] technique was applied to both the text and link datasets. Moreover, we generated the merged dataset by putting both the word and link features together. The MDT algorithm used the RBF kernel to map the data from the original feature space to the RKHS. The classification error rate on target data is used as evaluation metric, which is defined as the number ratio between the misclassified instances and the total instances in the target domain.

Performance Comparison

We compared MDT with a variety of the state-of-the-art algorithms such as Transductive SVM (TSVM) [Joachims, 1999] which is a semi-supervised classifier, traditional multi-view algorithm Co-Training [Blum and Mitchell, 1998], large-margin-based multi-view transfer learner MVTL-LM [Zhang *et al.*, 2011] and Co-Training based adaptation algorithm CODA[2] [Chen *et al.*, 2011].

For simplicity, we used the postfix -C, -L and -CL to denote that the classifier was fed with the text, link and merged dataset, respectively. Both the text and link datasets were fed to the multi-view classifiers Co-Training, MVTL-LM and MDT. Since CODA is a pseudo multi-view adaptation algorithm, to fit our scenario, the CODA was fed with the merged dataset. For each dataset, we repeated the algorithms five times and reported the average performance.

Table 2 shows the results. TSVM performed poorly for adaptation when using either content or link features. Simply merging the two sets of features make some improvements, implying that text and link can be complementary, but it may degrade the confidence of the classifier on some instances whose features become conflict because of merge. Co-Training can avoid this problem by boosting the confidence of classifiers built on the distinct views in a complementary way, thus performs a little better than TSVMs. Since both TSVM and Co-Training don't consider the distribution gap, they performed worse than the adaptation algorithms such as MVTL-LM, CODA and MDT.

Since FDA2 only utilized the labeled information, its generalization performance is not comparable with the semi-supervised methods such as TSVM-CL and Co-Training.

[2]http://www1.cse.wustl.edu/~mchen/code/coda.tar

Table 2: Comparision of adaptation error rate on different datasets.

Algorithms	DA-EC	DA-NT	DA-OS	DA-ML	EC-NT	EC-OS	EC-ML	NT-OS	NT-ML	OS-ML	Average
TSVM-C	0.293	0.175	0.276	0.217	0.305	0.355	0.333	0.364	0.205	0.202	0.272
TSVM-L	0.157	0.137	0.261	0.114	0.220	0.201	0.205	0.501	0.106	0.170	0.207
TSVM-CL	0.214	0.114	0.262	**0.107**	0.177	0.245	0.168	0.396	0.101	0.179	0.196
Co-Train	0.230	0.163	0.175	0.171	0.296	0.175	0.206	0.220	0.132	0.128	0.190
MVTL-LM	0.192	0.108	**0.068**	0.183	0.261	0.176	0.264	0.288	0.071	0.126	0.174
CODA	0.234	**0.076**	0.109	0.150	0.178	0.187	0.322	0.240	**0.025**	0.087	0.161
FDA2	0.407	0.159	0.267	0.212	0.324	**0.154**	0.277	0.255	0.088	0.152	0.229
MDT	**0.107**	0.082	0.102	0.118	**0.154**	0.167	**0.149**	**0.178**	0.072	**0.057**	**0.119**

MDT significantly outperformed FDA2 on most of the datasets. Note that FDA2 is a special case of our approach ($c_1 = c_2 = 0$). MDT outperformed FDA2 by taking the domain discrepancy into consideration and enhancing the view consistency.

It is shown that MDT outperformed MVTL-LM. This is because MDT fully leverages the correlation between views by projecting the two-view data onto the discriminant directions. Since the content and links may share some common topics, both views are correlated to each other at the semantic level. MDT utilizes two views of the same underlying semantic content to extract a shared representation, which helps improve the generalization prediction performance. Moreover, incorporating the projected domain distance measure into the optimization framework to minimize the domain discrepancy is another competency of MDT.

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, since its objective is non-convex, CODA may suffer from sub-optimal solution on view splitting. Furthermore, CODA cannot fully utilize both the text and link information since the pseudo views generated are essentially not as complementary as true multiple views in our case. It performed worse than MDT, indicating that pseudo views might be detrimental. In contrast, MDT incorporated view consistency and domain distance by taking a discriminant analysis approach and performed better.

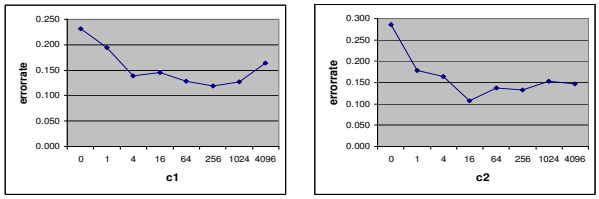
Parameter Sensitivity

Here we examine how our algorithm is influenced by the trade-off coefficients c_1 and c_2. The search range for c_1 and c_2 are $\{0, 1, 4, 16, 64, 256, 1024, 4076\}$. The results on DA-EC are shown in Figure 3. We observe that the best results can be achieved when $c_1 = 256$ and $c_2 = 16$. The algorithm performed worse when either domain distance ($c_1 = 0$) or view consistency ($c_2 = 0$) is not taken into consideration. However, when the magnitude of the value is very large given $c_1 = 4096$, the domain distance part will dominate the entire objective which would deteriorate the accuracy. We have the similar trend of error rate by increasing c_2.

As a result, we tune the trade-off parameters c_1 and c_2 for each dataset by cross-validation on the source data.

6 Conclusion

We present the MDT approach which incorporates the domain distance and view consistency into the FDA2 framework to improve the adaptation performance. Experiments



(a) Error rate varies with c_1 (c_2=16) (b) Error rate varies with c_2 (c_1=256)

Figure 3: The sensitivity of performance varies with c_1 and c_2.

show that MDT performed significantly better than the state-of-the-art baselines. Furthermore, we report on the theoretical analysis of the proposed approach and discuss the condition that the given technique is applicable.

Next we will extend our model to the scenario of multiple (>2) views/domains, which is not straightforward to implement. Though the conflicts between views/domains are not observed on the real dataset, it is more likely to occur in the situations of multiple views/domains and needs further investigation. Similar to [Martínez and Zhu, 2005], we will develop a robust algorithm in attempting to avoid the conflicts.

A Proof of Lemma 1

Proof. Suppose r_p, r_q, r_x, and r_z are the ranks of P, Q, M_x, and M_z, respectively. Since P, M_x and M_z are symmetric, there exist respective orthogonal matrices U_p, U_x and U_z to diagonalize them. Thus, P, M_x, and M_z can be written as the similar form such as $P = U_p \Lambda_p U_p^{\mathrm{T}} = \sum_{j=1}^{r_p} \lambda_{p_j} \xi_{p_j} \xi_{p_j}^{\mathrm{T}}$, where $U_p = (\xi_{p_1}, \cdots, \xi_{p_{r_p}})$, $\Lambda_p = \mathtt{diag}\{\lambda_{p_1}, \cdots, \lambda_{p_{r_p}}\}$ and $\lambda_{p_1} \geq \lambda_{p_2} \geq \cdots \geq \lambda_{p_{r_p}}$. On the other hand, P is a block matrix which can also be written as

$$\begin{aligned} P &= \begin{pmatrix} M_x & \mathbf{0} \\ \mathbf{0} & M_z \end{pmatrix} = \begin{pmatrix} U_x \Lambda_x U_x^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & U_z \Lambda_z U_z^{\mathrm{T}} \end{pmatrix} \\ &= \begin{pmatrix} U_x & \mathbf{0} \\ \mathbf{0} & U_z \end{pmatrix} \begin{pmatrix} \Lambda_x & \mathbf{0} \\ \mathbf{0} & \Lambda_z \end{pmatrix} \begin{pmatrix} U_x^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & U_z^{\mathrm{T}} \end{pmatrix} \end{aligned} \tag{10}$$

Then we can connect the eigensystem of P to those of M_x and M_z as follows

$$\{\lambda_{p_1}, \cdots, \lambda_{p_{r_p}}\} = \{\lambda_{x_1}, \cdots, \lambda_{x_{r_x}}\} \cup \{\lambda_{z_1}, \cdots, \lambda_{z_{r_z}}\} \tag{11}$$

$$\xi_{p_j} = \begin{cases} \begin{pmatrix} \xi_{x_j} \\ \mathbf{0} \end{pmatrix}, & if \ \ \lambda_{p_j} \in \{\lambda_{x_1}, \cdots, \lambda_{x_{r_x}}\} \\ \begin{pmatrix} \mathbf{0} \\ \xi_{z_j} \end{pmatrix}, & if \ \ \lambda_{p_j} \in \{\lambda_{z_1}, \cdots, \lambda_{z_{r_z}}\} \end{cases} \tag{12}$$

where $1 \leq j \leq r_p = r_x + r_z$. Hence, we could reach the final conclusion as follows

$$\begin{aligned} tr(P^{-1}Q) &= \sum_{i=1}^{r_q} \sum_{j=1}^{r_p} \frac{\lambda_{q_i}}{\lambda_{p_j}} (\xi_{p_j}^{\mathrm{T}} \xi_{q_i})^2 \\ &= \sum_{i=1}^{r_q} \sum_{j=1}^{r_x} \frac{\lambda_{q_i}}{\lambda_{x_j}} \left[\begin{pmatrix} \xi_{x_j} \\ \mathbf{0} \end{pmatrix}^{\mathrm{T}} \xi_{q_i} \right]^2 + \sum_{i=1}^{r_q} \sum_{j=1}^{r_z} \frac{\lambda_{q_i}}{\lambda_{z_j}} \left[\begin{pmatrix} \mathbf{0} \\ \xi_{z_j} \end{pmatrix}^{\mathrm{T}} \xi_{q_i} \right]^2 \end{aligned}$$

where the first term follows from [Martínez and Zhu, 2005] and the second follows from Eq.(11), and Eq.(12). □

References

[Abney, 2002] Steven Abney. Bootstrapping. In *Proceedings of ACL*, pages 360-367, 2002.

[Belhumeur *et al.*, 1997] P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):711-720, 1997.

[Blitzer *et al.*, 2011] John Blitzer, Sham Kakade and Dean P. Foster. Domain Adaptation with Coupled Subspaces. In *Proceedings of AISTATS*, pages 173-181, 2011.

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of COLT*, pages 92-100, 1998.

[Cao *et al.*, 2010] Bin Cao, Sinno Jialin Pan, Yu Zhang, Dit-Yan Yeung, and Qiang Yang. Adaptive transfer learning. In *Proceedings of AAAI*, 2010.

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

[Collins and Singer, 1999] M. Collins and Y. Singer. Unsupervised Models for Named Entity Classification. In *Proceedings of EMNLP*, pages 100-110, 1999.

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

[Diethe *et al.*, 2008] Tom Diethe, David R. Hardoon and John Shawe-Taylor. Multiview Fisher Discriminant Analysis. In *Proceedings of NIPS Workshop on Learning from Multiple Sources*, 2008.

[Duda *et al.*, 2001] R.O. Duda, P.E. Hart, and D.G. Stock. Pattern Classification, second ed. Wiley, 2001.

[Fisher, 1938] R.A. Fisher. The Statistical Utilization of Multiple Measurement. *Annals of Eugenics*, vol.8, pages 376-386, 1938.

[Gao *et al.*, 2010] Wei Gao, Peng Cai, Kam-Fai Wong, Aoying Zhou. Learning to Rank only using Training Data from related Domain. In *Proceedings of SIGIR*, pages 162-169, 2010.

[Huang *et al.*, 2006] Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, Bernhard Schölkopf. Correcting Sample Selection Bias by Unlabeled Data. In *Proceedings of NIPS*, pages 601-608, 2006.

[Joachims, 1999] Thorsten Joachims. Transductive Inference for Text Classification using Support Vector Machines. In *Proceedings of ICML*, pages 200-209, 1999.

[Martínez and Zhu, 2005] Aleix M. Martínez, Manli Zhu. Where Are Linear Feature Extraction Methods Applicable? *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 27(12):1934-1944, 2005.

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

[Melzer *et al.*, 2003] Thomas Melzer, Michael Reiter, Horst Bischof. Appearance Models based on Kernel Canonical Correlation Analysis. *Pattern Recognition (PR)*, 36(9):1961-1971, 2003.

[Mika *et al.*, 2001] Sebastian Mika, Alexander Smola, Bernhard Schölkopf. An Improved Training Algorithm for Kernel Fisher Discriminants. In *Proceedings of AISTATS*, 2001.

[Pan and Yang, 2010] Sinno J. Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345-1359, 2010.

[Pan *et al.*, 2011] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, Qiang Yang. Domain Adaptation via Transfer Component Analysis. *IEEE Transactions on Neural Networks (TNN)*, 22(2):199-210, 2011.

[Pan *et al.*, 2010] Weike Pan, Evan Wei Xiang, Nathan Nan Liu, Qiang Yang. Transfer Learning in Collaborative Filtering for Sparsity Reduction. In *Proceedings of AAAI*, pages 230-235, 2010.

[Quanz and Huan, 2009] Brian Quanz and Jun Huan. Large Margin Transductive Transfer Learning. In *Proceedings of CIKM*, pages 1327-1336, 2009.

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

[Salton and Buckley, 1988] G. Salton and C. Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513-523, 1988.

[Sarinnapakorn and Kubat, 2007] Kanoksri Sarinnapakorn and Miroslav Kubat. Combining Sub-classifiers in Text Categorization: A DST-Based Solution and a Case Study. *IEEE Transactions Knowledge and Data Engineering*, 19(12):1638-1651, 2007.

[Yao and Doretto, 2010] Yi Yao and Gianfranco Doretto. Boosting for transfer learning with multiple sources. In *Proceedings of CVPR*, pages 1855-1862, 2010.

[Zhang *et al.*, 2011] Dan Zhang, Jingrui He, Yan Liu, Luo Si and Richard D. Lawrence. Multi-view Transfer Learning with a Large Margin Approach. In *Proceedings of KDD*, pages 1208-1216, 2011.