Guided Learning: A New Paradigm for Multi-task Classification

Jingru Fu[1], Lei Zhang[1](✉), Bob Zhang[2], and Wei Jia[3]

[1] College of Communication Engineering, Chongqing University, Chongqing, China
{jrfu,leizhang}@cqu.edu.cn

[2] Department of Computer and Information Science, University of Macau,
Macau, China
bobzhang@umac.mo

[3] School of Computer and Information, Hefei University of Technology, Hefei, China
china.jiawei@139.com

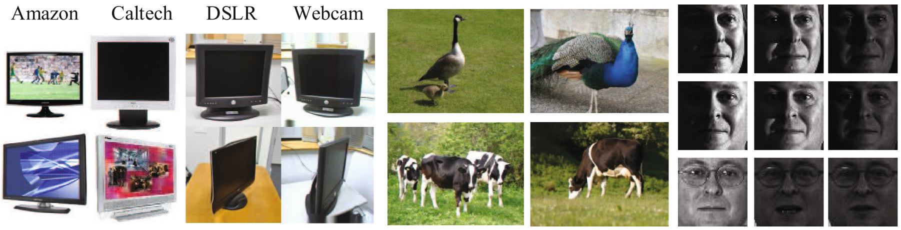
Abstract. A prevailing problem in many machine learning tasks is that the training and test data have different distribution (non i.i.d). Previous methods to solve this problem are called Transfer Learning (TL) or Domain Adaptation (DA), which belong to one stage models. In this paper, we propose a new, simple but effective paradigm, Guided Learning (GL), for multi-stage progressive training. This new paradigm is motivated by the "tutor guides student" learning mode in human world. Further, under the framework of GL, a Guided Subspace Learning (GSL) method is proposed for domain disparity reduction, which aims to learn an optimal, invariant and discriminative subspace through the guided learning strategy. Extensive experiments on various databases show that our method outperforms many state-of-the-art TL/DA methods.

Keywords: Guided Learning · Subspace Learning · Domain disparity

1 Introduction

Conventional machine learning algorithms are based on the assumption that the training and test data lie in the same feature space with the same distribution. However, this assumption may not hold in many real-world scenarios. Especially in the field of computer vision owing to various factors such as different camera devices, illuminations, background, etc. Fig. 1 shows some images of different distributions. When the disparity exists between the training and test data, the classification accuracy dropped dramatically [5]. However, retraining a new classifier often requires a large amount of labeled training data of the same distribution (i.i.d), which consumes a lot of human resources and is not realistic with the explosive growth of unlabeled data. TL/DA methods have been used to solve this problem [9]. They aim to transfer well-learned knowledge from the source domain (training set) to the target domain (test set). In this paper, we introduce a new paradigm, Guided Learning (GL), for solving such domain mismatch problem.

J. Zhou et al. (Eds.): CCBR 2018, LNCS 10996, pp. 239–246, 2018.
https://doi.org/10.1007/978-3-319-97909-0_26



a) 4DA b) MSRC and VOC2007 c) CMU PIE

Fig. 1. Some examples from different domains. (a) 4DA: Each column represents the data of Amazon, DSLR, Caltech and Webcam, respectively. (b) MSRC (left) and VOC2007 (right). (c) CMU PIE: The first two rows indicate different illuminations and poses, the last row indicates different expressions and glass occlusion.

Conventional TL/DA methods can be divided into classifier-based methods and representation-based methods [9,10]. The classifier-based methods tend to solve the domain disparity problem by adapting the existing classifiers to the data with different distributions, such as A-SVM [16]. However, they may not utilize the intrinsic information of the data, and it strongly depends on the specific classifier. Further, the representation-based methods tend to learn a better representation for classification, such as RDALR [4], TSL [11], LTSL [10], LSDT [17] and DTSL [15]. However, most of them only consider the domain adaptation at the data level, which ignore the global information of domains. SA [2] and CORAL [12] stand in another perspective, which tend to align the first-order and second-order statistical global features (e.g. PCA subspace and feature for domain discrepancy reduction at subspace level). Additionally, the TL/DA tends to find a classifier or transformation in one stage, which may not work when domain disparity is large.

Therefore, we propose a new GL paradigm for domain disparity reduction through a progressive, guided, and multi-stage strategy. The GL paradigm is relevant but different from TL/DA methods that it is established upon the main idea of "tutor guides student" mode in human world. Tutor-students' teaching mode is general route in human learning process. In general, the tutor not only transfers expert knowledge to the students at a time, but to progressively guide the students achieving a certain learning purpose through the tutor's learning experience. Therefore, considering the domain difference between source and target domains, we propose a Guided Subspace Learning (GSL) method, which tends to progressively learn an optimal target subspace guided by source domain. The key contributions of this work are three-folds:

(1) Inspired by the "tutor guides student" learning mode in human world, we propose a new learning paradigm called Guided Learning (GL), which can achieve knowledge transfer in a progressive guided manner.
(2) Under the GL framework, we propose a Guided Subspace Learning (GSL) method for solving domain mismatch. Compared with the TL/DA methods, the concept of progressive guiding in GL makes the model more robust.

(3) The GSL method can simultaneously eliminate domain disparity at data level and subspace level. Finally, an optimal, invariant and discriminative target subspace can be achieved through subspace, data and label guidance.

2 Related Work

According to the objective of this paper, we present a overview of TL/DA methods from data and subspace level, respectively.

2.1 Data Level Approach

As mentioned before, this type of methods learn better feature representation from the data level. RDALR [4] presented a low-rank reconstruction constraint to reduce the domain shift, which can capture the intrinsic relationship in data. It assumes that the transformed source samples can be linearly reconstructed by target samples. TSL [11] solved the problem by minimizing Bregman divergence between the distribution of domains in a common subspace. LTSL [10] also used the reconstruction matrix and derived a generalized framework. LSDT [17] further presented sparse reconstruction constraint and generalized model into a kernel-based linear/nolinear framework. DTSL [15] imposed low-rank and sparse constraints on the reconstruction matrix to guarantee the global and local property. Then, it obtained a linear classifier by learning a non-negative label relaxation matrix. Obviously, those approaches heavily depend on the well-designed reconstruction matrices and sensitive to noise.

2.2 Subspace Level Approach

It is not enough to get robust representation for classification by only exploiting the data level information of two domains. Subspace level approach can align the statistical features of two domains. SA [2] seeks a domain invariant feature space by learning a linear mapping which aligns subspaces spanned by eigenvectors (obtained by PCA). This kind of statistical features have global domain information, so that the subspace level approaches are more robust to noise and outliers that are irrelevant to the target domain. It is worth mentioning that SA can be explained by the manifold learning perspective. SDA [13] considered the distribution difference in the subspace, and proved that SA can be extended to GFK [3] in the case of an infinite subspaces distribution alignment.

3 Proposed Method

3.1 Mathematical Notation

We first clarify the definition of terminologies. Given the source domain $\mathcal{S} = \{X_s, y_s\}$ and target domain $\mathcal{T} = \{X_t, y_t\}$, where $X_s \in \mathbb{R}^{D \times n_s}$ and $X_t \in \mathbb{R}^{D \times n_t}$ are samples, y_s and y_t are labels (note that y_t is only used during testing step). D is the dimensionality of the original samples, and n_s and n_t indicate the number of samples in source and target domain, respectively. Let $P_s \in \mathbb{R}^{D \times d}$ and $P_t \in \mathbb{R}^{D \times d}$ be the projection of the source domain and target domain, respectively, where d is the dimensionality of the invariant subspace. Define $Z \in \mathbb{R}^{n_s \times n_t}$ as the reconstruction matrix.

3.2 Problem Formulation

As mentioned above, GSL can reduce the distribution mismatch by learning a target subspace. To sum up, GSL can be composed of three parts: (1) subspace guidance; (2) data guidance; (3) label guidance.

(1) **Subspace Guidance:** We first guide the target subspace P_t by the source subspace P_s. Similar to SA, we expect that the subspaces of the two domains can be aligned to reduce the domain disparity. It can be easily achieved by minimizing the following Frobenius norm, instead of learning an additional mapping function:

$$\min_{P_s, P_t} \|P_s - P_t\|_F^2 \tag{1}$$

It treats two subspaces equally and may extremely preserve the useful information of the two data sets. Moreover, the subspaces of the two domains are adjusted at the same time, which encourages to seek a better P_t under the guidance of P_s.

(2) **Data Guidance:** Second, we expect to use the intrinsic information of data to guide the learning of P_t. For data guidance, we tend to seek an invariant subspace by forcing the target data linearly combined by source data. For revealing the underlying structure of source and target data, we constrain that each target data can be reconstructed by the neighbors of the source data. Mathematically, we can achieve this purpose by placing a low-rank constraint on the reconstruction matrix Z. Actually, this constraint has been extensively discussed in machine learning field due to its impact on subspace recovery [14]. This can be formulated as:

$$\min_{P_s, P_t, Z} \left\|P_t^T X_t - P_s^T X_s Z\right\|_F^2 + \alpha \|Z\|_* \tag{2}$$

By using term (2) together with (1), an invariant target subspace where the domain disparity has been largely reduced can be obtained.

(3) **Label Guidance:** Although an invariant subspace has been found, the discrimination of such invariant subspace is not enough for classification problems. Additionally, a large amount of label information of source domain is neglected. So, we further introduce label guidance strategy in both domains to improve the subspace discriminability. Firstly, we expect that the learned projections can serve as classifier, which can be achieved by forcing $P_t^T X_t$ close to the pseudo label matrix $\hat{Y}_t \in \mathbb{R}^{d \times n_t}$ ($d \geq c$, and c indicates the number of classes) with category information. Unfortunately, the pseudo label information of the target

domain is not accurate. Therefore, we propose to use the existing classifiers (e.g. SVM) to generate pseudo labels and then learn a discriminative target subspace alternatively, under the label guidance. Inspired by EDA [6], we introduce a relaxation matrix M to alleviate this effect while increasing the robustness of the framework. Secondly, to make full use of the known labels in the source domain and improve the accuracy of this strategy, we define the constructed label matrix $Y = \left[Y_s, \hat{Y_t}\right] \in \mathbb{R}^{d \times n}$ ($n = n_s + n_t$ indicates the total number of samples in both domains) as:

$$Y\{i,j\} = \begin{cases} 1, \ if \ x_j \in c_i \\ -1, \ otherwise \end{cases} \tag{3}$$

The purpose of label guidance strategy is to seek a discriminative P_t, which also approximates the common subspace between domains, formulated as:

$$\min_{P_t, M} \left\| P_t^T X - Y \circ M \right\|_F^2 \ \ s.t. \ M \succ 0 \tag{4}$$

where $X = [X_s, X_t] \in \mathbb{R}^{D \times n}$. $M \in \mathbb{R}^{D \times n}$ represents the relaxation matrix. $\circ$ is a hadamard product operator.

We can obtain the following ultimate objection function by incorporating the above three Eqs. (1), (2) and (4) as:

$$\begin{aligned} &\min_{P_s, P_t, M, Z} \beta \left\| P_s - P_t \right\|_F^2 + \left\| P_t^T X_t - P_s^T X_s Z \right\|_F^2 + \alpha \left\| Z \right\|_* + \tfrac{1}{2} \left\| P_t^T X - Y \circ M \right\|_F^2 \\ &s.t. \ M \succ 0 \end{aligned} \tag{5}$$

where β and α are trade-off parameters to balance the constraints. We iteratively update the pseudo labels of target domain data using the learned invariant and discriminative target subspace. Finally, an optimal, invariant, and discriminative target subspace P_t can be achieved in a progressive manner.

3.3 Optimization

It can be seen from problem (5) that four variables are involved when Y is fixed. To solve the problem, an inexact augmented Lagrange multiplier method (IALM) [14] is used. With an auxiliary variable L, the problem (5) can be converted into:

$$\begin{aligned} &\min_{P_s, P_t, M, Z, L} \beta \left\| P_s - P_t \right\|_F^2 + \left\| P_t^T X_t - P_s^T X_s Z \right\|_F^2 + \alpha \left\| L \right\|_* + \tfrac{1}{2} \left\| P_t^T X - Y \circ M \right\|_F^2 \\ &s.t. \ M \succ 0, \ Z = L \end{aligned} \tag{6}$$

Then, by using variables alternating strategy, we can derive the solution of each variable in IALM algorithm:

$$P_t = (2\beta I + 2X_t X_t^T + X X^T)^{-1} (2\beta P_s + 2X_t Z^T X_s^T P_s + X (Y \circ M)^T) \tag{7}$$

$$P_s = (2\beta I + 2X_s Z Z^T X_s^T)^{-1} (2\beta P_t + 2X_s Z X_t^T P_t) \tag{8}$$

$$Z = (2Xs^T P_s P_s^T X_s + \mu I)^{-1} (2Xs^T P_s P_t^T X_t + \mu (L - Y_1/\mu)) \tag{9}$$

$$L = arg\min_{L} \alpha \left\| L \right\|_{*} + \frac{\mu}{2} \left\| Z - L + Y_1/\mu \right\|_F^2 \tag{10}$$

$$M = arg\min_{M} \frac{1}{2} \left\| P_t^T X - Y \circ M \right\|_F^2 \tag{11}$$

where Y_1 is a Langrange multiplier, $\mu > 0$ is a penalty parameter and I is identity matrix. The optimal solution of formula (10) can be computed via the singular value thresholding (SVT) algorithm [1]. Problem (11) can be similarly solved by [15]. Then multiplier Y_1 and iteration step-size ρ ($\rho > 1$) are updated by:

$$\begin{cases} Y_1 = Y_1 + \mu(Z - L) \\ \mu = min(\rho\mu, \mu_{max}) \end{cases} \tag{12}$$

Once the guided P_t is obtained through the IALM algorithm, then an existing classifier can be used to get better pseudo-target-labels (also a better $\hat{Y}_t$) based on the optimal representation. To check the convergence, we define $\triangle P_t = \left\| P_t^{(t+1)} - P_t^{(t)} \right\|_F / \left\| P_t^{(t)} \right\|_F$, where t indicates iteration. Convergence is achieved when $\triangle P_t < \varepsilon$, where ε indicates a very small positive number.

4 Experiment

In this section, extensive experiments are conducted to justify the effectiveness of our method. The experiments on three different benchmark DA tasks, including 4DA object data set [3], MSRC-VOC2007 data set [8] and CMU PIE face data set [7]. Some examples are illustrated in Fig. 1.

Experimental Setting: In all experiments, we use SVM to progressively generate pseudo target labels. The dimensionality d of the invariant subspace is set as c (the number of classes) in each data set.

(1) **4DA Data Set:** 4DA consists of Office and Caltech-256. Office contains three real-world object domains, Amazon, Webcam and DSLR. 4DA is formulated with 10 shared categories of the two data sets. We use the same SURF features as [3]. Therefore, 4 domains: A (Amazon), C (Caltech-256), D (DSLR) and W (Webcam) are exploited. By deploying two different domains as the source domain and target domain alternatively, we construct 12 cross-domain tasks.

(2) **MSRC and VOC2007 Data Set:** MSRC contains 4,323 images of 18 classes, which was released by Microsoft Research Cambridge. VOC2007 contains 5011 images of 20 classes. 6 shared semantic classes: aeroplane, bicycle, bird, car, cow, sheep are formulated. Following the experimental setting as [15], two cross-domain tasks are constructed: MSRC vs VOC2007 and VOC2007 vs MSRC.

(3) **CMU PIE Face Data Set:** PIE contains 68 individuals with 41,368 face images of size 32 × 32. PIE1 (C05, left pose), PIE2 (C07, upward pose), PIE3 (C09, downward pose), PIE4 (C27, frontal pose), PIE5 (C29, right pose). The face images were captured by 13 different poses and 21 different illuminations

Table 1. Accuracy (%) On 3 types data sets. NA denotes no adaptation.

Data Set	Compared transfer learning methods							
	NA	SA	JDA [7]	TSL	RDALR	LTSL	DTSL	GSL
C→A(1)	50.09	48.02	51.46	52.30	52.51	24.11	53.34	**56.68**
C→W(2)	43.05	31.86	41.36	40.34	40.68	22.93	45.76	**47.12**
C→D(3)	47.77	42.68	46.50	49.04	45.22	14.58	**50.96**	49.04
A→C(4)	42.79	34.37	43.90	43.28	43.63	21.36	44.70	**45.24**
A→W(5)	37.03	33.90	33.90	34.58	35.93	18.17	38.31	**39.32**
A→D(6)	37.22	38.85	33.76	38.85	36.94	22.29	39.49	**43.95**
W→C(7)	29.47	30.01	31.17	31.43	28.05	**34.64**	30.28	32.24
W→A(8)	34.15	32.15	36.33	34.66	31.21	**39.46**	34.66	38.94
W→D(9)	80.62	83.44	77.71	79.62	83.44	72.61	82.80	**85.99**
D→C(10)	30.11	32.24	31.43	33.13	32.32	**35.35**	30.72	31.70
D→A(11)	32.05	33.40	38.41	32.57	33.72	**39.35**	33.19	36.95
D→W(12)	72.20	70.51	75.59	72.54	72.54	74.92	76.61	**79.32**
MSRC→VOC2007(1)	37.12	31.76	38.17	32.35	37.45	38.04	38.04	**41.76**
VOC2007→MSRC(2)	55.48	46.02	59.26	43.18	62.33	**67.06**	56.42	61.54
PIE1→PIE4(1)	51.76	42.75	25.14	46.68	41.66	20.01	81.29	**84.77**
PIE4→PIE4(2)	65.88	51.41	33.76	59.15	48.11	52.79	79.71	**83.85**
PIE4→PIE5(3)	51.96	47.92	29.47	45.22	48.84	47.00	71.02	**71.75**
PIE5→PIE4(4)	53.41	43.11	25.38	53.08	44.46	23.61	**66.09**	63.17
Average	47.33	43.02	41.82	45.67	45.50	37.13	52.96	**55.19**

and/or expressions. Alternatively, we constructed 4 cross-domain tasks: PIE1 vs PIE4, PIE4 vs PIE1, PIE4 vs PIE5, and PIE5 vs PIE4.

Specifically, the experimental results on the three datasets are shown in Table 1, from which we can observe that our GSL method outperforms other TL/DA methods in most tasks. The average classification performance of GSL shows significant improvement than others.

5 Conclusion

We firstly propose a new learning paradigm called Guided Learning (GL), which is inspired by the "tutor guides student" learning mode in human world. In order to solve the problem of domain mismatch in multi-task classification, we further proposed a Guided Subspace Learning (GSL) method, which aims to progressively seek an optimal target subspace through the GL paradigm. The proposed GSL is imposed the optimality, invariance, and discrimination by proposing three strategies, including subspace guidance, data guidance and label guidance. Notably, the label guidance strategy is constructed by formulating label relaxation and progressive target pseudo target label pre-computing method. The proposed GL provides a new learning mechanism for multi-task classification as TL/DA methods do. Experimental results demonstrate that our method outperforms many state-of-the-art TL/DA methods.

Acknowledgements. This work was supported by the National Science Fund of China under Grants (61771079).

References

1. Jian Feng Cai, C., Emmanuel, J.S., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**(4), 1956–1982 (2008)
2. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: IEEE ICCV, pp. 2960–2967 (2014)
3. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: IEEE CVPR, pp. 2066–2073 (2012)
4. Hong Jhuo, I, Liu, D., Lee, D.T., Chang, S.F.: Robust visual domain adaptation with low-rank reconstruction. In: CVPR, pp. 2168–2175 (2012)
5. Kan, M., Junting, W., Shan, S., Chen, X.: Domain adaptation for face recognition: targetize source domain bridged by common subspace. IJCV **109**(1–2), 94–109 (2014)
6. Lei, Z., Zhang, D.: Robust visual knowledge transfer via extreme learning machine-based domain adaptation. IEEE Trans. IP **25**(10), 4959–4973 (2016)
7. Long, M., Wang, J., Ding, G., Sun, J., Yu P.S.: Transfer feature learning with joint distribution adaptation. In: IEEE ICCV, pp. 2200–2207 (2014)
8. Long, M., Wang, J., Ding, G., Sun, J., Yu, P.S.: Transfer joint matching for unsupervised domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1410–1417 (2014)
9. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
10. Shao, M., Kit, D., Yun, F.: Generalized transfer subspace learning through low-rank constraint. IJCV **109**(1–2), 74–93 (2014)
11. Si, S., Tao, D., Geng, B.: Bregman divergence-based regularization for transfer subspace learning. IEEE Trans. Knowl. Data Eng. **22**(7), 929–942 (2010)
12. Sun, B., Feng, J., Saenko, K.: Correlation alignment for unsupervised domain adaptation. In: Csurka, G. (ed.) Domain Adaptation in Computer Vision Applications. ACVPR, pp. 153–171. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58347-1_8
13. Sun, B., Saenko, K.: Subspace distribution alignment for unsupervised domain adaptation. In: BMVC, pp. 24.1–24.10 (2015)
14. Wright, J., Ganesh, A., Rao, S., Ma, Y.: Robust principal component analysis: exact recovery of corrupted low-rank matrices. J. ACM **87**(4), 20:3–20:56 (2009)
15. Xu, Y., Fang, X., Wu, J., Li, X., Zhang, D.: Discriminative transfer subspace learning via low-rank and sparse representation. IEEE Trans. IP **25**(2), 850–863 (2016)
16. Yang, J., Yan, R., Hauptmann, A.G.: Cross-domain video concept detection using adaptive SVMS. In: ACM International Conference on Multimedia, pp. 188–197 (2007)
17. Zhang, L., Zuo, W., Zhang, D.: LSDT: latent sparse domain transfer learning for visual adaptation. IEEE Trans. IP **25**(3), 1177–1191 (2016)