Distance Metric Facilitated Transportation between Heterogeneous Domains*

Han-Jia Ye[1], **Xiang-Rong Sheng**[1], **De-Chuan Zhan**[1], **Peng He**[2]
[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2] Tencent, China
{yehj,shengxr,zhandc}@lamda.nju.edu.cn, paulhe@tencent.com

Abstract

Lacking training examples is one of the main obstacles to learning systems. Transfer learning aims to extract and utilize useful information from related datasets and assists the current task effectively. Most existing methods restrict tasks connection on the same feature sets, or require aligned examples cross domains, even cannot take full advantage of the limited label information. In this paper, we focus on transferring between heterogeneous domains, i.e., those with different feature spaces, and propose the Metric Transporation on HEterogeneous REpresentations (MAPHERE) approach. In particular, an asymmetric transformation map is first learned to compensate the *cross-domain* feature difference based on linkage relationship between objects; then the *inner-domain* discrepancy is further reduced with learned optimal transportation. Note that both source domain and cross-domain relationship are fully utilized in MAPHERE, which helps improve target classification task a lot. Experiments on synthetic dataset validate the importance of the "metric facilitated" consideration, while results on real-world tasks show the superiority of the MAPHERE approach.

1 Introduction

A large amount of training examples leads to the effectiveness of a learning algorithm [Mohri *et al.*, 2012]. In real applications, however, this is not always the case due to the high instance/label collection cost [Li *et al.*, 2014; Li and Zhou, 2015]. The insufficiency of the training examples could be compensated by related tasks/data which are easier to obtain, so the knowledge from the source tasks can be extracted and adapted to help the learning on the target domain. Various methods and applications show that it is possible and necessary to deal with the divergence and do such kind of information transformation [Pan and Yang, 2010; Si *et al.*, 2010; Long *et al.*, 2014; Tommasi *et al.*, 2014].

*This research was supported by the National Key R&D Program of China (2018YFB1004300), the NSFC (61773198, 61632004), and the Tencent fund. Part of the work is implemented when the first author served as an intern in Tencent, China.

One common shift between tasks lies in the change of the statistics [Ben-David and Urner, 2012], including both conditional and marginal distributions. Thus, the consistency persistence followed by the distribution estimation lies the basis of the adaptation process. For example, instance weight keeps the statistical consistency between domains [Gretton *et al.*, 2012]; the Optimal Transport (OT) also show effectiveness to transfer source domain instances to the target space in an unsupervised way based on the least effort principle [Courty *et al.*, 2017a; 2017b]. Without restricting to this homogeneous case, the knowledge transfer could also be extended between heterogeneous feature sets. As in [Jiang and Li, 2017], paired inputs help learn better representation for a certain modality.

We consider a general supervised case on not-aligned heterogeneous domains with labeled examples, where source domain has relative more instances. There exist two apparent obstacles: the gap between two different feature sets, i.e., the two heterogeneous domains; and the change of distribution during the learning of the target task. In this paper, we solve the above two problems of heterogeneous transfer learning with the help of distance metric in a joint manner and propose the Metric Transporation on HEterogeneous REpresentations (MAPHERE) approach. Specifically, we treat all target domain instances as anchors, and an *asymmetric* transformation is learned to map instance from the source to the target domain, whose advantage is two-side. On the one hand, the projection implements a bridge to reduce the discrepancy between domains; besides, it finds a metric in the target domain for measuring a better distance between objects. Based on the metric facilitated pairwise cost, optimal transport depicts the distribution shift between tasks well. Since the similar and dissimilar relationship between both source domain and cross-domain examples direct the whole transfer, the label information is fully utilized to leverage and preserve the domain structure. Leveraging the distance metric and transportation, the nearest neighbor classifier can determine labels on the target task. Experiments on synthetic datasets show that by adjusting original distance measure between instances effectively, the utilization of metric with side information assists the target classification a lot. In addition, on real applications over images and texts, our MAPHERE approach appears better performance w.r.t. the state-of-the-art methods.

The rest of the paper starts with the related work. Then the MAPHERE approach is described in detail after a brief

introduction to the homogeneous optimal transport domain adaptation method. Last are experiments and conclusion.

2 Related Work

Owing to the ability to take advantage of related task knowledge to relieve the limited data burden of the current problem, transfer learning methods attract a lot of attention in the machine learning fields. Assuming there are only limited usable labeled examples in the target domain, how to extract as much information as possible from both current examples and from related domains in other tasks are stressed. Based on the type of source domain, methods of transfer learning could be partitioned to the homogeneous and heterogeneous cases [Pan and Yang, 2010]. For the former scenario, empirical estimation of distributions, for the marginal distribution [Ben-David *et al.*, 2010; Courty *et al.*, 2017b] or the joint distribution [Courty *et al.*, 2017a], are optimized to keep the consistency between two tasks. While in the heterogeneous case, the paired relationship between objects [Kulis *et al.*, 2011; Jiang and Li, 2017; Shi and Knoblock, 2017], the joint training over transformation and classifier [Hoffman *et al.*, 2013], or a shared subspace [Aljundi *et al.*, 2015] are learned to link two different domains together. Learned metric transformation in MAPHERE approach gives rise to the advantage transferring knowledge from source to the target domain without the feature set limit, and the leverage of side information persists the consistency between two distributions.

Optimal Transport (OT) finds a coupling between two distributions based on the ground cost, which can be used as a map across two sets [Kolouri *et al.*, 2017]. Facilitated by the accelerated solver recently [Cuturi, 2013], the optimal transport has been successfully applied in various applications, such as multi-label cost computation [Frogner *et al.*, 2015], and domain adaptation [Courty *et al.*, 2017b; 2017a]. In the homogeneous case, the leaned transport plan is able to map all source domain instances together with labels to the target domain based on the Euclidean distance ground matrix. Supervision information can be incorporated into the group lasso regularizer on the transportation map. In MAPHERE, the advantage of OT is extended to the heterogeneous case. Better cost measurement with the learned transformation also assists the cross-task transportation estimation.

Distance metric learning aims to find better representations than the Euclidean one, considering the correlation between features and task property. Both pairwise and triplet side-information provide the direction for distance optimization [Davis *et al.*, 2007; Weinberger and Saul, 2009]. Transformations could be learned to fuse domain information together [Geng *et al.*, 2011; Wang *et al.*, 2014; Luo *et al.*, 2016]. Emphasizing domain linkages, MAPHERE transfers label and structure information across domain effectively.

3 Homogeneous Domain Adaptation with Optimal Transport

Consider two heterogeneous feature spaces, i.e., source domain with N_s examples $\{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{N_s}$, where instance $\mathbf{x}_i^s \in \mathbb{R}^{d_s}$ forms $X_s \in \mathbb{R}^{N_s \times d_s}$ and $y_i^s \in \{1, \ldots, C\}$ with totally C classes. While there are N_t examples $\{(\mathbf{x}_j^t, y_j^t)\}_{j=1}^{N_t}$ in the target domain. $\mathbf{x}_j^t \in \mathbb{R}^{d_t}$ forms $X_t \in \mathbb{R}^{N_t \times d_t}$ and $y_j^t \in \{1, \ldots, C\}$. The number of available instances in the target domain is much smaller than the one of the source domain, i.e., $N_t \ll N_s$. Denote $\mathbf{x}_i^s \sim \mathcal{X}_s$ with marginal distribution μ_s, and $\mathbf{x}_j^t \sim \mathcal{X}_t$ with marginal distribution μ_t, respectively. $\mathbf{1}$ is an all one-value vector, whose length could be determined by the context. In this supervised learning task, the goal is to utilize all examples from the source domain to help the classifier training of the target task.

First consider the homogeneous case when $d_s = d_t$, it is reasonable to assume that the change between source and target domains comes from an unknown, possibly non-linear transformation $\mathcal{T}(\cdot)$ on the input space. In addition, the label information can be preserved after this transformation, i.e., $p_s(y \mid \mathbf{x}_i^s) = p_t(y \mid \mathcal{T}(x_i^s))$, with p_s and p_t correspond to source and target conditional distributions. This transformation between domains could be discovered by the Kantorovitch relaxation of Optimal Transport (OT) [Courty *et al.*, 2017b], via optimizing a coupling Π over μ_s and μ_t:

$$\Pi = \arg\min_{\Pi} \int_{\mathcal{X}_s \times \mathcal{X}_t} c(\mathbf{x}^s, \mathbf{x}^t) \mathrm{d}\Pi(\mathbf{x}^s, \mathbf{x}^t) \ . \tag{1}$$

The coupling Π satisfies the constraint that the two marginals of Π are exactly equal to μ_s and μ_t, and $c(\mathbf{x}^s, \mathbf{x}^t)$ is the ground matrix measuring cost when moving from $\mathbf{x}^s$ to $\mathbf{x}^t$. Π provides a way to push-forward one distribution to another, and the value of Eq. 1 can be regarded as the divergence between two distributions, even there is no overlap between them. With limited instances, both source and target marginal distributions could be estimated in the empirical form, i.e., the uniform distribution over all examples. In this discrete case, Eq. 1 is transformed to:

$$\min_{T \geq 0} \sum_{ij} T_{ij} C_{ij}, \ T\mathbf{1} = \frac{1}{N_s}\mathbf{1}, \ T^\top \mathbf{1} = \frac{1}{N_t}\mathbf{1} \ . \tag{2}$$

$C \in \mathbb{R}^{N_s \times N_t}$ with $C_{ij} = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ is the pairwise cost matrix, and squared Euclidean distance is used in the following discussion. $T \in \mathbb{R}^{N_s \times N_t}$ is the transportation map from $\mathcal{X}_s$ to $\mathcal{X}_t$, which shows the proportion to move a unit mass of each source domain instance to the target. Two constraints over T require the transportation should be consistent with the marginal distributions. OT interpolates source and target distributions, and the transported source domain instance $\hat{\mathbf{x}}_i^t = \mathcal{T}(\mathbf{x}_i^s)$ could be represented as [Perrot *et al.*, 2016]:

$$\hat{\mathbf{x}}_i^t = \arg\min_{\mathbf{x}} \sum_{j=1}^{N_t} T_{ij} \|\mathbf{x} - \mathbf{x}_j^t\|_2^2 \ . \tag{3}$$

After taking derivatives w.r.t. $\mathbf{x}$ and set it to zero, we can get $\hat{\mathbf{x}}_i^t = N_s X_t^\top T_i^\top$, where T_i is the i-th row of T. Since the transformation is represented by the interpolation with all target domain instances, it is able to reflect a non-linear $\mathcal{T}$.

Therefore, based on OT, domain adaptation could be progressed in two stages: first the optimal transportation map T is evaluated based on the cross-domain pairwise Euclidean distance cost; then all source domain instances, together with their label information, are transferred to the target domain, which augments the target task training set a lot. Last, a standard classifier is applied over this augmented set. The effectiveness of this strategy has been validated in [Perrot *et al.*, 2016; Courty *et al.*, 2017a; 2017b].

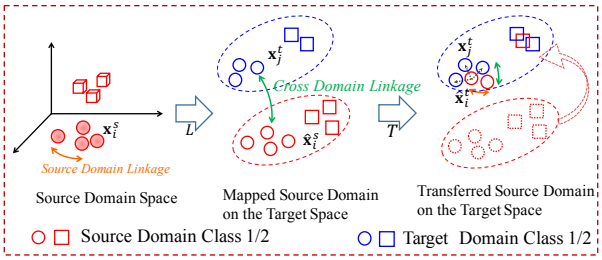


Figure 1: The illustration of the MAPHERE approach. Instances $\mathbf{x}_i^s$ in the source and $\mathbf{x}_j^t$ in the target domain could be with different feature spaces. An asymmetric map L is first learned to get rid of the space differences between two domains, then the optimal transport map T are learned on the mapped space to further transfers the mapped source instance $\hat{\mathbf{x}}_i^s$ to $\hat{\mathbf{x}}_i^t$, considering the change of distribution between these two tasks. With the help of source domain information, the training number burden of target task could be relieved. It is also notable that the relationship between source domain and cross domain instances are utilized during training.

4 The MAPHERE Approach

There are two main drawbacks of the previous OT transfer strategy. First, it is hard to estimate the right instance relationship with the Euclidean distance, since it neglects the differences between domains [Bellet *et al.*, 2015]. The provided label information, although limited, can help improve the distance estimation. Besides, since Euclidean distance computation requires features with the same form of representation, the OT based approaches are hard to be applied over heterogeneous feature sets. In this paper, we propose the Metric Transportation on heterogeneous representations (MAPHERE) approach, taking advantage of distance metric learning to solve above two considerations simultaneously.

In MAPHERE, an asymmetric transformation acts as a distance metric to fill the gap between diverse features, and facilitates the distance computation cross tasks. Based on the new distance measure, optimal transport furthermore reduces the differences between two domain distributions, and finds correspondences between the mapped source domain instances and the target set. Weakly supervised side information, i.e., the pairwise similar and dissimilar relationship between objects, is also incorporated in to direct the whole transfer process. Thus, same class instances will have small distances, while different class instances have larger ones. The main flow of MAPHERE is illustrated in Fig. 1.

The location of the target task examples are anchored during the whole training process, and the source domain instances are mapped to the target feature space with a transformation $L \in \mathbb{R}^{d_t \times d_s}$. This asymmetric movement avoids the inter-domain collapse and helps the final discriminative process [Kulis *et al.*, 2011]. The mapped source domain instance $\mathbf{x}_i^s$ can be represented as $\hat{\mathbf{x}}_i^s = L\mathbf{x}_i^s \in \mathbb{R}^{d_t}$. With the help of L, the mapped source domain instance $\hat{\mathbf{x}}_i^s$ has the same type of representation with $\mathbf{x}_j^t$ in the target domain.

Considering types of differences between two domains, there may still exist a discrepancy between two sets even in the same feature space. Therefore, the transfer process is continued by aligning the remaining distribution differences by the optimal transport over the target domain. The ground cost matrix between mapped source domain instances $\hat{\mathbf{x}}_i^s$ and target domain instances $\mathbf{x}_i^t$ can be computed in the squared form, $C_{ij} = \|\hat{\mathbf{x}}_i^s - \mathbf{x}_j^t\|_2^2$. The better the metric L revealing the relationship between instances, the more exact the cost estimation between two sets. Based on the new distance measurement, MAPHERE seeks a transportation plan $T \in \mathbb{R}^{N_s \times N_t}$ in the form of Eq. 2, and gets a *nonlinear* map between transformed source instances and target representations.

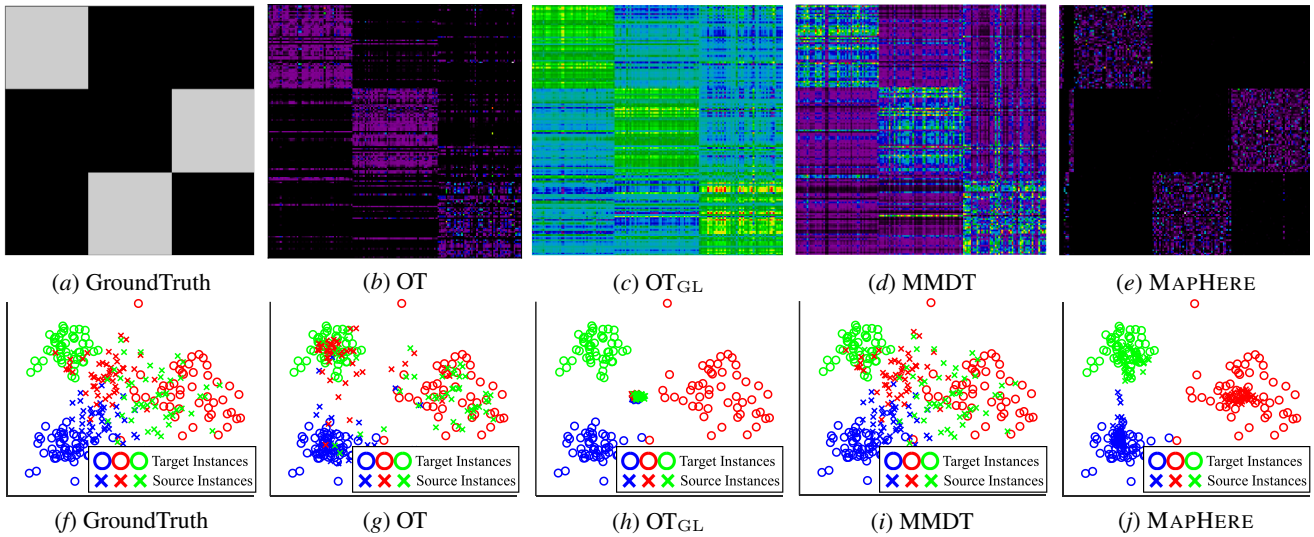
Up to now, the relationship between two tasks are linked with transformation L and transportation T in an unsupervised way. MAPHERE is also able to make use of the limited label information in both domains by considering the pairwise linkages between objects. It is the *relative* comparison between objects that reveals more useful information with limited instances. Formally, we construct P pairs, $\mathcal{P} = \{(\mathbf{x}_m^p, \mathbf{x}_n^p, y_{mn}^p)\}_{p=1}^P$, and in each pair, $y_{mn}^p \in \{-1, 1\}$ denotes whether two instances are similar or not. Following formulation reveals the satisfaction of similarity relationship:

$$\frac{1}{P}\sum_{p=1}^{P} \ell(y_{mn}^p(\gamma_{y_{mn}} - \|\mathbf{x}_m^p - \mathbf{x}_n^p\|_2^2)) \ .$$

$\ell(\cdot)$ is a convex loss function, which is usually an upper bound of the 0-1 loss. $\gamma_{y_{mn}}$ is the pre-defined threshold value for similar and dissimilar instances. For example, for square loss, $\ell(x) = (x-1)^2$, it requires the similar (resp. dissimilar) objects should have small distances near $\gamma_1 - 1$ (resp. $\gamma_{-1} + 1$). Thus, similar instances are pulled together, while dissimilar ones are pushed far away. This relationship should be kept both in source and target domains, so in MAPHERE, two types of pairs are extracted. The first type focuses on the relationship of the source domain. With similar and dissimilar instances generated based on nearest neighbors for each example, pairs indicate the class information of the source task. It is notable that this type of constraint preserves the source domain structure in the target space, and the concrete form of the pair could be $(N_s X_t^\top T_m^\top, N_s X_t^\top T_n^\top)$, which are the last transferred representations of source domain examples for $\mathbf{x}_m$ and $\mathbf{x}_n$. The second type of pairs includes one target instance and one source mapped instance, i.e., the cross-domain pairs. Since there are only limited target labeled examples, we can sample same class and different class instances from the mapped source set to generate pairs. Thus, for source domain instance $\mathbf{x}_m$ and target domain instance $\mathbf{x}_n$, we use the pair $(N_s X_t^\top T_m^\top, \mathbf{x}_n^t)$ in the objective. In summary, we utilize the pairwise linkage information from both source and cross domains, which directs the learning process of target transportation T, and indirectly influences the training of cross-domain projection L. The whole objective of MAPHERE is

$$\min_{L,T\geq 0} \frac{1}{P}\sum_{p=1}^{P} \ell(y_{mn}^p(\gamma_{y_{mn}} - d_{mn}^p)) + \lambda_1 \sum_{ij} T_{ij}C_{ij} + \lambda_2 \Omega(L)$$
$$s.t. \quad d_{mn}^p = \|\mathbf{x}_m^p - \mathbf{x}_n^p\|_2^2, \ T\mathbf{1} = \frac{1}{N_s}, \ T^\top \mathbf{1} = \frac{1}{N_t} \ . \quad (4)$$

The feature space transformation L and target domain transportation map T are learned in a joint manner. $\Omega(L)$ is the regularizer over L. λ_1 and λ_2 are two non-negative parameters. With learned transportation plan T, all source task instances $\{\mathbf{x}_i^s\}_{i=1}^{N_s}$ could be transferred to the target domain, together with their labels. With which, the final classification is conducted with 1 Nearest Neighbor on the target domain.



(*a*) GroundTruth (*b*) OT (*c*) OT$_{\text{GL}}$ (*d*) MMDT (*e*) MAPHERE

(*f*) GroundTruth (*g*) OT (*h*) OT$_{\text{GL}}$ (*i*) MMDT (*j*) MAPHERE

Figure 2: The effectiveness of MAPHERE approach on the synthetic dataset. We use different colors to differentiate the 3 classes, and use crosses/circles to denote source and target instances, respectively. The first row of plots show the transportation maps output by different methods, the darker the color, the smaller the transportation plan value. Different columns correspond to different methods. The first column is the ground truth result, and (*f*) is the original scatter of source and target instances. Plots (*g*) - (*j*) show the transformed source domain instances in the target domain by different methods, and original target domain instances. Best viewed in color.

5 Optimization for MAPHERE

The optimization for MAPHERE in Eq. 4 could be solved in an alternative manner. When the cross domain mapping L is fixed, the OT problem uses the distance between mapped source instances and the target instances as the ground metric. With the adjusted cost matrix, the transportation plan is updated under the supervision of types of linkages. Given the transportation T, the source domain instances are mapped to align with target anchors using the transportation weights.

With fixed transportation map T revealing a type of cross-domain pairwise similarity, we tune the feature space map L:

$$\min_{L} \lambda_1 \sum_{ij} T_{ij} \|L\mathbf{x}_i^s - \mathbf{x}_j^t\|_2^2 + \lambda_2 \Omega(L) \tag{5}$$

Since $T_{ij} \geq 0$, the subproblem actually finds a map between source and target examples, weighted by the transportation plan between them. The pair with larger weights T_{ij} shows more representation similarity considering the distribution change, thus it requires more strength to reduce their differences. We can get the closed form solution of L, i.e.,

$$L = (\lambda_1 \sum_{ij} T_{ij} \mathbf{x}_j^t \mathbf{x}_i^{s\top} + \lambda_2 L_0)(\lambda_1 \sum_{ij} T_{ij} \mathbf{x}_i^s \mathbf{x}_i^{s\top} + \lambda_2 L_0)^{-1} , \tag{6}$$

when the biased regularizer $\Omega(L) = \|L - L_0\|_F^2$ is used. L_0 is a prior for the transformation L.

With optimized L, the cost matrix C could be computed between transformed source and target instances. The optimization subproblem for T is difficult since there are equality constraints over T in the objective, and T also exists in the metric loss function, as shown in the following formulation:

$$\min_{T \geq 0} \frac{1}{P} \sum_{p=1}^{P} (y_{mn}^p (\gamma_{y_{mn}} - \|\mathbf{x}_m^p - \mathbf{x}_n^p\|_2^2) - 1)^2 + \lambda_1 \sum_{ij} T_{ij} C_{ij}$$
$$T\mathbf{1} = \frac{1}{N_s}\mathbf{1},\ T^\top \mathbf{1} = \frac{1}{N_t}\mathbf{1} . \tag{7}$$

Considering the possible nonlinear term over T, the objective can be solved with conditional gradient descent method [Courty *et al.*, 2017b], a.k.a. Frank-Wolfe optimizer. It *linearizes* the objective by first order approximation, which could be solved in an easier and efficient way. The convergence property of the conditional gradient descent could be found in [Jaggi, 2013]. Square loss is used in the remaining part of the paper. Since there are two types of pairs, we consider their derivatives separately. When two instances in a pair both come from the source domain, we have the loss term $o_{il} = (y_{il}(\gamma_{y_{il}} - \|N_s X_t^\top T_i^\top - N_s X_t^\top T_l^\top\|_2^2) - 1)^2$, which has the derivative for T_i (similar for T_l)

$$\frac{\partial o_{il}}{\partial T_i} = -4 o_{il}^{\frac{1}{2}} y_{il} N_s^2 (T_i X_t X_t^\top - T_l X_t X_t^\top) .$$

While for a cross domain pair, i.e., $o_{ij} = (y_{ij}(\gamma_{y_{ij}} - \|N_s X_t^\top T_i^\top - \mathbf{x}_j^t\|_2^2) - 1)^2$, we have

$$\frac{\partial o_{ij}}{\partial T_i} = -4 o_{ij}^{\frac{1}{2}} y_{ij} (N_s^2 T_i X_t X_t^\top - N_s \mathbf{x}_j^{t\top} X_t^\top) .$$

We initialize the transformation L first, then the optimal transport plan T could be solved based on the refined cost matrix. Since each sub-problem decreases the total loss function of the objective in Eq. 4, the final algorithm will converge at last. In the implementation, L_0 is an identity matrix in the homogeneous case; while in the heterogeneous scenario, L_0 comes from a least square problem mapping the source domain instances to the center of corresponding target domain classes. Two types of pairs are sampled based on Euclidean distance nearest neighbors of instances (5NN and 3NN are used to construct source and target domain same class similar pairs, while 1NN are used to generate both domains' impostors). $\lambda_1 = 1$ and $\lambda_2 = 10$ are default parameters.

	MAPHERE	OT_{IT}	OT_{GL}	OT_{MT}	JDOT	MMDT	ARC	GFK	$1NN_S$	$1NN_T$	$LMNN_T$	$LMNN_H$
A→C	**43.5±2.1**	34.6±3.0	38.7±2.2	39.6±1.4	41.3±3.7	39.8±2.3	34.3±2.1	37.8±1.9	36.0±1.3	31.9±3.2	32.4±3.0	34.7±3.7
A→D	56.6±4.7	59.3±4.0	**59.6±4.7**	45.6±3.5	38.1±3.6	54.3±4.3	32.1±3.8	51.5±3.6	33.6±4.4	53.3±4.3	50.0±3.5	54.7±5.0
A→W	69.9±4.8	69.6±5.0	**70.2±4.2**	51.5±4.8	39.6±3.7	64.9±5.7	34.0±7.0	59.4±4.3	33.7±3.6	66.3±3.9	62.6±4.5	67.6±5.2
C→A	**55.1±3.3**	49.2±3.8	50.4±2.8	44.6±1.7	45.2±2.7	51.1±3.4	39.4±2.8	46.4±2.9	37.4±3.0	47.3±4.2	43.0±3.8	50.4±4.7
C→D	58.3±5.7	59.2±5.2	**61.6±4.8**	47.5±4.8	39.9±4.6	52.8±4.8	33.4±5.5	58.1±3.9	31.9±5.8	54.2±4.8	46.0±6.5	57.4±4.5
C→W	**74.2±5.4**	68.5±5.6	69.4±5.5	53.8±6.7	36.9±7.4	62.8±5.2	31.5±6.9	63.3±5.9	28.6±6.1	65.1±6.3	55.8±5.1	65.1±5.3
D→A	**54.7±3.0**	49.4±2.9	47.3±1.8	40.8±1.5	41.3±2.0	50.4±3.4	34.8±2.0	40.8±2.6	33.6±1.8	47.8±3.6	40.6±3.8	49.7±4.0
D→C	**40.0±2.0**	34.1±2.6	36.8±1.6	36.5±1.5	37.4±1.2	35.7±3.3	35.0±1.3	30.6±2.0	31.2±1.2	32.2±3.0	28.0±3.0	33.8±3.0
D→W	82.2±2.0	71.1±4.1	80.5±2.5	**84.6±2.2**	80.2±2.1	74.4±3.1	74.0±6.2	75.0±2.9	76.9±2.2	66.2±4.6	65.4±3.8	69.7±3.8
W→A	**54.9±3.1**	49.9±3.3	48.4±2.4	43.0±1.9	40.5±1.6	50.6±3.7	36.1±3.0	43.3±2.3	32.2±3.0	48.3±3.5	41.7±3.7	50.9±4.0
W→C	**38.0±2.0**	33.0±3.0	36.1±1.6	35.4±2.0	35.4±2.6	34.9±3.6	32.7±2.4	30.0±3.1	27.7±2.6	30.7±3.9	28.6±3.4	32.6±3.5
W→D	69.7±4.1	63.1±5.1	72.4±3.5	**75.4±3.2**	69.8±3.5	62.5±4.4	68.2±3.4	71.9±4.1	64.6±4.3	54.8±5.2	56.9±5.1	61.1±5.8
Mean	**58.1±3.5**	53.4±4.0	56.0±3.10	49.9±2.9	45.5±3.2	52.8±3.9	40.5±3.9	50.7±3.3	38.9±3.3	49.8±4.2	45.9±4.1	52.3±4.4

Table 1: Classification comparison of MAPHERE with others on Office-Caltech datasets. The mean accuracy±std. are shown in the table, where the values with the highest performance are presented in bold. Each row corresponds to a transfer task, while each column is a method.

6 Experiments

We validate the effectiveness of our MAPHERE approach from various perspectives. Based on a synthetic example, the importance of the "metric facilitated" consideration is stressed; then MAPHERE is extensively compared with other methods in both homogeneous and heterogeneous case in both image and text classifications.

6.1 Synthetic Illustration

We first illustrate the necessity of "metric facilitated" consideration in the domain adaptation task on a synthetic dataset. A 2D three-class dataset is generated based on Gaussian distribution, where the source and target domains are with different centers but the same variances. There are 50 instances in each class. The corresponding classes in source and target domains are first generated in near locations, and then we exchange the label of the latter two classes in the target domain. Since the direct space location information is not correct for two of the three classes, the transfer learning methods should utilize the label information to do a good transfer.

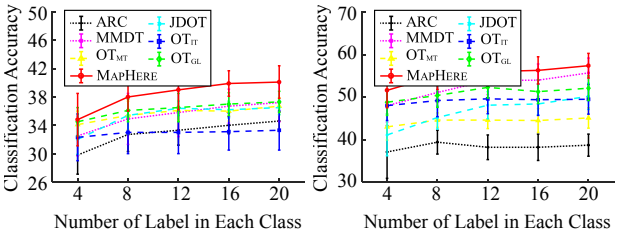
We compare with unsupervised Optimal Transport based Domain Adaptation (OT), its supervised extension with Group Lasso (OT_{GL}) [Courty *et al.*, 2017b], and Maximum Margin Domain Adaptation (MMDT) [Hoffman *et al.*, 2013]. OT_{GL} considers the class information with the group lasso regularizer on transportation plan. MMDT learns a transformation over the source domain hypothesis, and trains classifier and cross-domain mapping jointly. All methods are investigated with default parameters. The learned transportation map and mapped source domain results are shown in Fig. 2.

The first row of plots in Fig. 2 are the plots of true and learned transportation maps, i.e., a type of similarity matrix between source and target domain instances. The darker the color, the smaller the value. Plot (a) shows the ground truth transportation, where the permutation between last two classes can be clearly found. The block diagonal structure exists in (b)-(d). OT finds the cross-domain mapping with the Euclidean distance, so source domain instances are uniformly mapped to the nearby target instances, neglecting their labels. OT_{GL} and MMDT consider the class information. For MMDT, we plot the scaled inner product similarity matrix between domains as its transportation plan. Since missing a good metric between instances, it is hard for them to find a good match between distance computation and class correspondence. In (c)-(d), the relationship between classes is disturbed. For MAPHERE, the class mapping is correctly learned, which is similar to the ground-truth. The second row in Fig. 2 demonstrates the mapped source domain instances in the target space. Similarly, (f) is the original source and target plot. From results in (g)-(j), only our MAPHERE approach maps source domain instances to corresponding target classes correctly. The performance of MAPHERE validates the importance to utilize the specific distance computation and the label information in transfer learning.

6.2 Homogeneous Image Classification

We first test the performance of MAPHERE approach on homogeneous domain adaptation task on the Office-Caltech dataset. There are totally four domains, namely Amazon (A), Caltech (C), DSLR (D) and Webcam (W), and 10 classes for each. Each domain is used as source and target alternatively, which generates 12 different tasks. We use the same protocol (including the splits) as [Perrot and Habrard, 2015]. For each task, we repeat the investigations 20 trials, and in each trial, there are 8 labeled source examples (20 if the source is Amazon) and 3 labeled target examples are selected. All instances are normalized thanks to the zscore and the dimensionality is reduced to 20 using a PCA.

We compare with optimal transport for domain adaptation with two different class regularizers OT_{IT} and OT_{GL} [Courty *et al.*, 2017b], mapping estimation for optimal transport OT_{MT} [Perrot *et al.*, 2016], joint domain adaptation method JDOT [Courty *et al.*, 2017a], Maximum Margin learning for domain invariant representation (MMDT) [Hoffman *et al.*, 2013], Asymmetric Regularized Cross-domain transformation (ARC) [Kulis *et al.*, 2011], Geodistic Flow Kernel (GFK) [Gong *et al.*, 2012] and Hypothesis biased Large Margin Nearest Neighbor ($LMNN_H$) [Perrot and Habrard, 2015]. All methods do classification with 1NN on the target domain, with the help of transformed instances or classifier.



(*a*) *W→C* (*b*) *C→A*

Figure 3: Changes of accuracy for transfer learning methods as the number of source task labels in each class increases. Results over two tasks, i.e., W→C and C→A, are shown.

	MAPHERE	MMDT	HFA	$1NN_T$	SVM_T
A→C	36.1±3.7	**38.7±2.1**	35.1±2.9	31.9±3.2	35.2±2.9
A→D	**57.6±4.1**	51.1±4.6	56.7±4.0	53.3±4.3	57.0±5.5
A→W	**71.9±5.8**	62.8±4.8	67.9±5.0	66.3±3.9	70.0±5.6
C→A	**52.2±3.9**	48.9±3.5	51.6±3.9	47.3±4.2	51.0±4.2
C→D	**58.9±4.7**	51.4±4.6	57.1±4.4	54.2±4.8	57.6±4.4
C→W	**71.2±4.6**	59.7±4.4	66.6±4.7	65.1±6.3	66.5±5.4
D→A	**52.4±4.0**	47.0±3.1	50.7±3.7	47.8±3.6	50.4±3.8
D→C	**36.3±2.9**	33.7±3.1	35.4±3.2	32.2±3.0	34.9±2.6
D→W	**73.5_4.2**	68.7±3.5	67.4±3.5	66.2±4.6	68.4±4.4
W→A	**53.9±4.0**	47.9±3.3	52.0±4.4	48.3±3.5	51.5±4.1
W→C	**34.0±3.7**	32.8±3.6	33.9±3.9	30.7±3.9	33.2±4.1
W→D	61.0±4.3	**63.2±4.8**	59.8±6.0	54.8±5.2	59.9±6.8
Mean	**54.9±4.2**	53.0±4.5	50.5±3.8	49.8+4.2	52.9±4.1

Table 2: Classification comparison of MAPHERE with others on Office-Caltech datasets for heterogeneous transfer. The mean accuracy±std. are shown in the table, where the values with the highest performance are presented in bold. Each row corresponds to a transfer task, while each column is a method.

The LMNN and 1NN results on the source and target domains are listed as baselines ($LMNN_T$, $1NN_S$, and $1NN_T$). Test accuracy values (mean±std.) are listed in Table 1. From the results, OT methods can achieve good performance, which shows that joint consideration of source and target distributions is useful, and OT is able to learn the transportation map in some cases. Our MAPHERE approach can achieve the best performance in most tasks, which attributes to the feature space transformation and effective utilization pairwise linkage. Compared with ARC, which only learns a domains transformation, the superiority of MAPHERE validates the necessity of the further distribution consideration with OT.

The changes of performance given different amount of source labeled data are also investigated, as shown in Fig. 3. The number of labeled examples in each source task class increases from 4 to 20. MAPHERE is compared with ARC, MMDT, OT_{MT}, JDOT, OT_{IT}, and OT_{GL}. Due to the page limit, 2 tasks, i.e., W→C and C→A, are listed. From the results, the performance of all methods increases when there are more source domain labels, which shows the helpfulness of enough source task labeled data. In addition, our MAPHERE approach achieves best performance with all the label ratios. Therefore, our MAPHERE approach is effective even there are only limited number of labels from the source domain.

	MAPHERE	MMDT	HFA	$1NN_T$	SVM_T
Cornell	**35.1±7.0**	20.2±1.9	28.6±6.6	31.4±4.9	24.5±5.0
Wisconsin	**44.0±3.5**	21.3±1.5	30.8±6.2	28.2±6.7	31.2±8.8
Washington	**46.9±3.4**	16.5±1.9	30.3±9.8	28.0±7.7	32.8±11.9
Mean	**42.0±4.6**	19.3±1.8	29.9±7.5	29.2±6.4	29.5±8.6

Table 3: Classification comparison of MAPHERE with others on WebKB datasets. The mean accuracy±std. are shown in the table. Values with the highest performance are presented in bold.

6.3 Heterogeneous Image Classification

The extension to heterogeneous cases of MAPHERE is also considered on the Office-Caltech dataset, with the same experimental settings as before. But in this case, the source and target tasks have different dimensions, i.e., PCA to 30 for source domain, and 20 for target tasks. Results are shown in Table 2, where SVM on the target task is listed as a baseline. Besides, we compare with HFA [Li *et al.*, 2014], which can learn with augmented features. MAPHERE also performs best in this heterogeneous case. Compared with MMDT, the superiority of MAPHERE validate the assistance of OT based distribution alignment after transformation, and the effectiveness of using pairwise information between objects.

7 Investigation on Web Page Classification

In this section, we investigate MAPHERE over a web page classification problem.[1] Web pages are described by their web texts and linkage information with others. In this experiment, source domain is the web texts and the target domain is the linkage information. Dimensionality reduction is applied to each domain and keeps 50% energy. Source and target dimensionality are 40, 17 (Cornell), 41, 24 (Wisconsin); and 40, 27 (Washington), respectively. Each task has 5 classes. There are only 5 instances for each class in the target domain. Results are shown in Table 3. MMDT and HFA cannot perform well in this task. MAPHERE also gets good results, thanks to the learned transformation and the distribution change consideration.

8 Conclusion

How to utilize limited labeled examples from the related task is important in machine learning fields, and the main obstacles lie in the fact both feature space and distribution will change. We propose the Metric Transportation on HEterogeneous REpresentations (MAPHERE) approach in this paper, which learns the transformation between feature space and the transportation plan handling the non-stationary distribution jointly. Based on the learned transformation, a better distance metric is evaluated on the target task. The limited label information from source task is also organized as pairwise linkages between both source and cross-domain objects. Experiments on both homogeneous and heterogeneous cases show that the MAPHERE approach deals with the transfer learning task effectively. Future directions include the consideration of novel classes in both source and target tasks.

[1]Data from http://www.cs.cmu.edu/ webkb/

References

[Aljundi *et al.*, 2015] R. Aljundi, R. Emonet, D. Muselet, and M. Sebban. Landmarks-based kernelized subspace alignment for unsupervised domain adaptation. In *The 28th IEEE Conference on Computer Vision and Pattern Recognition*, pages 56–63, Boston, MA., 2015.

[Bellet *et al.*, 2015] A. Bellet, A. Habrard, and M. Sebban. Metric learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 9(1):1–151, 2015.

[Ben-David and Urner, 2012] S. Ben-David and R. Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Algorithmic Learning Theory - 23rd International Conference*, pages 139–153, Lyon, France, 2012.

[Ben-David *et al.*, 2010] S. Ben-David, T. Lu, T. Luu, and D. Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136, Sardinia, Italy, 2010.

[Courty *et al.*, 2017a] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems 30*, pages 3733–3742. Cambridge, MA.: MIT Press, 2017.

[Courty *et al.*, 2017b] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

[Cuturi, 2013] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Cambridge, MA.: MIT Press, 2013.

[Davis *et al.*, 2007] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 209–216, Corvalis, OR., 2007.

[Frogner *et al.*, 2015] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. A. Poggio. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems 28*, pages 2053–2061. Cambridge, MA.: MIT Press, 2015.

[Geng *et al.*, 2011] B. Geng, D. Tao, and C. Xu. DAML: domain adaptation metric learning. *IEEE Transactions on Image Processing*, 20(10):2980–2989, 2011.

[Gong *et al.*, 2012] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, Providence, RI., 2012.

[Gretton *et al.*, 2012] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[Hoffman *et al.*, 2013] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell. Efficient learning of domain-invariant image representations. *CoRR*, abs/1301.3224, 2013.

[Jaggi, 2013] M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 427–435, Atlanta, GA., 2013.

[Jiang and Li, 2017] Q.-Y. Jiang and W.-J. Li. Deep cross-modal hashing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3278, Honolulu, HI., 2017.

[Kolouri *et al.*, 2017] S. Kolouri, S. Rim Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[Kulis *et al.*, 2011] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition*, pages 1785–1792, Colorado Springs, CO., 2011.

[Li and Zhou, 2015] Y.-F. Li and Z.-H. Zhou. Towards making unlabeled data never hurt. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 37(1):175–188, 2015.

[Li *et al.*, 2014] W. Li, L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1134–1148, 2014.

[Long *et al.*, 2014] M. Long, J. Wang, G. Ding, D. Shen, and Q. Yang. Transfer learning with graph co-regularization. *IEEE Transactions on Knowledge and Data Engineering*, 26(7):1805–1818, 2014.

[Luo *et al.*, 2016] Y. Luo, Y. Wen, and D. Tao. On combining side information and unlabeled data for heterogeneous multitask metric learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 1809–1815, New York, NY., 2016.

[Mohri *et al.*, 2012] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2012.

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

[Perrot and Habrard, 2015] M. Perrot and A. Habrard. A theoretical analysis of metric hypothesis transfer learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1708–1717, Lille, France, 2015.

[Perrot *et al.*, 2016] M. Perrot, N. Courty, R. Flamary, and A. Habrard. Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems 29*, pages 4197–4205. Cambridge, MA.: MIT Press, 2016.

[Shi and Knoblock, 2017] Y. Shi and C. A. Knoblock. Learning with previously unseen features. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2722–2729, Melbourne, Australia, 2017.

[Si *et al.*, 2010] S. Si, D. Tao, and B. Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transaction on Knowledge and Data Engineering*, 22(7):929–942, 2010.

[Tommasi *et al.*, 2014] T. Tommasi, F. Orabona, and B. Caputo. Learning categories from few examples with multi model knowledge transfer. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 36(5):928–941, 2014.

[Wang *et al.*, 2014] H. Wang, W. Wang, C. Zhang, and F. Xu. Cross-domain metric learning based on information theory. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2099–2105, Québec, Canada, 2014.

[Weinberger and Saul, 2009] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.