Online Heterogeneous Transfer Metric Learning

Yong Luo 1 , Tongliang Liu 2 , Yonggang Wen 1 , Dacheng Tao 2
1 School of Computer Science and Engineering, Nanyang Technological University, Singapore
2 UBTECH Sydney AI Centre, SIT, FEIT, University of Sydney, Australia
yloo180@gmail.com, tongliang.liu@sydney.edu.au, ygwen@ntu.edu.sg, dacheng.tao@sydney.edu.au

Abstract
Distance metric learning (DML) has been demonstrated to be successful and essential in diverse applications. Transfer metric learning (TML) can help DML in the target domain with limited label information by utilizing knowledge from other related source domains. The heterogeneous TML (HTML), where the feature representations vary from the source to the target domain, is general and challenging. However, current HTML approaches are usually conducted in a batch manner and cannot handle sequential data. This motivates the proposed online HTML (OHTML) method. In particular, the distance metric in the source domain is pre-trained using some existing DML algorithms. To enable knowledge transfer, we assume there are large amounts of unlabeled corresponding data that have representations in both the source and target domains. By enforcing the distances (between these unlabeled samples) in the target domain to agree with those in the source domain under the manifold regularization theme, we learn an improved target metric. We formulate the problem in the online setting so that the optimization is efficient and the model can be adapted to new coming data. Experiments in diverse applications demonstrate both effectiveness and efficiency of the proposed method.

1 Introduction
Distance metric learning (DML) aims to learn a proper distance function to reveal the underlying data relationship [Xing et al., 2002]. It has been demonstrated to be successful in diverse applications, such as nearest-neighbor classification [Weinberger et al., 2005], clustering [Xing et al., 2002], content based image retrieval [Jain et al., 2008] and face verification [Chopra et al., 2005].

To learn an appropriate distance metric, we often need large amount of label information, such as class labels or pair (similar/dissimilar) or triplet (relative comparison) constraints. However, in real-world applications, the provided information is usually scarce due to the high labeling cost and DML is likely to fail in this scenario. Transfer metric learning (TML) [Zha et al., 2009] is able to alleviate this issue by transferring information or knowledge from other related source domains [Shao et al., 2016; Liu et al., 2017; Wang et al., 2017], where the distance metric is better. Directly applying the source metric to the target domain is infeasible when samples in the source and target domain lie in different feature spaces or have semantic gap. Such challenging heterogeneous TML (HTML) setting is popular in real-world applications. For example, we can use a large corpus of labeled English documents to help classify Spanish documents. The dimensions of the English and Spanish document representations are different due to the utilized different vocabularies. It is advantageous to use some existing expensive (high-performing but computational intensive) features (such as deep CNN [Chatfield et al., 2014]) to help learn a better metric for cheap features (such as LBP [Ojala et al., 2002]). We may also use the semantic tags to guide the metric learning of visual features.

Heterogeneous transfer learning (HTL) [Luo et al., 2017a; 2017b; Li et al., 2017] is able to handle the heterogeneous features [Xie et al., 2016; 2017], and some HTL approaches learn feature transformations to map the source and target samples into a common subspace [Wang and Mahadevan, 2011]. The learned transformation in the target domain can be used to derive an improved target distance metric. However, most of these approaches conduct the learning process on the entire training set. Hence, they are not applicable in the online setting, where training samples are provided sequentially (one by one). In addition, it is exhausted to retrain the model when new (labeled) training samples are available.

To tackle this issue, we develop a novel online HTML (OHTML) method, which updates the target metric using the source knowledge and only one labeled target sample at each step. In particular, we first learn the distance metric in the source domain using some existing DML algorithms. The source metric learning can be performed offline, and only the obtained metric is needed in the target metric learning. Alternatively, if the source feature (such as deep CNN) is much more expressive than target feature, we can directly employ the simple Euclidean metric (or some other pre-defined metrics) in the source domain. To build a connection between the source and target domains, we assume there are abundant unlabeled training samples that have representations in both the source and target domains. For each pair of such
unlabeled samples, if they are close to each other in the source domain, their distances in the target domain should also be small. By formulating this as a manifold regularization term [Belkin et al., 2006], and simultaneously minimizing the empirical loss w.r.t. the current labeled training sample, we learn an improved distance metric for the target domain. Besides, LogDet divergence [Davis et al., 2007; Jain et al., 2008] is enforced to minimize the differences between the new obtained target metric and the previous one. This ensures that the updated metric parameter automatically satisfy the positive semi-definite (PSD) constraint. We do not require the costly PSD projection, and thus the updating algorithm is quite efficient.

There is a recent work of metric imitation (MI) [Dai et al., 2015] that aims to utilize the expensive source feature to learn an improved metric for the cheap target feature. Their formulation is also based on manifold learning, but it discards the valuable label information, and the target metric (parameterized by a linear transformation) is learned in a batch manner. Eigenvalue decomposition is involved in the optimization and thus their training complexity is high. The proposed OHTML is more advantageous than MI in that: 1) the target metric can be updated dynamically and adapt to patterns in the new coming data; 2) the optimization is much more efficient. We compare the proposed method with representative online DML algorithms [Jain et al., 2008; Jin et al., 2009] and competitive HTML approaches [Wang and Mahadevan, 2011; Dai et al., 2015] in various applications including object categorization, scene clustering, face verification, as well as image retrieval. The results demonstrate both effectiveness and efficiency of our method.

2 Online Heterogeneous Transfer Metric Learning

Problem setting: given a source and target domains with heterogeneous feature representations. The training set with side information for the target domain is given by \(D^t_M = \{(x^1_{MK}, x^2_{MK}, y_{MK})\}_{k=1}^{N_M} \), where \(x^1_{MK}, x^2_{MK} \in \mathbb{R}^{d_M} \), and \(y_{MK} = \pm 1 \) indicates \(x^1_{MK} = \text{donor} \) and \(x^2_{MK} = \text{source} \). In the target domain, the number of sample pairs \(N_M \) is small and the utilized feature is cheap. Hence the resulting distance metric obtained by applying existing DML algorithms may perform poorly. To improve the target metric, we assume there is a relevant source domain with training set \(D^s_S = \{(x^i_{Sk}, y_{MK})\}_{k=1}^{N_S} \) in the source domain, either the samples with side information are abundant (i.e., \(N_S \) is large), or the feature is more expressive or interpretable than that in the target domain. Therefore, a better distance metric can be obtained. To help the target metric learning use the source domain knowledge, we assume there are large amounts of unlabeled data that have representations in both the source and target domains, i.e., \(D^u = \{(x^U_{Sm}, x^U_{Mn})\}_{n=1}^{N_U} \). Such data are usually easy to collect in practice.

2.1 Problem Formulation

Our ultimate goal is to learn an appropriate Mahalanobis distance metric for the target domain by transferring knowledge from the source domain. The Mahalanobis distance is often defined as

\[
d_A(x^1_k, x^2_k) = (x^1_k - x^2_k)^T A(x^1_k - x^2_k),
\]

where \(A \) is the metric (parameter of the distance function), which is an positive semi-definite (PSD) matrix. To facilitate knowledge transfer [Du et al., 2013; Shao et al., 2014], we learn distance metric in the source domain beforehand using some existing DML algorithms, such as LMNN [Weinberger et al., 2005] and ITML [Davis et al., 2007]. This can be conducted offline and does not have impact on the computational complexity of target metric learning. In the target domain, the labeled training pairs are provided sequentially, i.e., only one labeled pair is available at each step. Suppose the pre-trained source metric is \(A_S \), then we have the following general formulation for updating the target metric \(A_M \):

\[
A_M^{k+1} = \arg\min_{A_M \succeq 0} F(A_M) = \Psi(A_M) + \gamma A \text{Div}(A_M, A^k_S) + \gamma I R_I(d_{A_M}, d_{A_S}, D^U),
\]

where \(\Psi(A_M) = V(A_M; x^1_{MK}, x^2_{MK}, y_{MK}) \) is the empirical loss w.r.t. \(A_M \) on the current training pair. We choose \(V(A_M; x^1_{MK}, x^2_{MK}, y_{MK}) = g(y_{MK}[1 - d_{A_M}(x^1_{MK}, x^2_{MK})]) \), where \(g(z) = \max\{0, b - z\} \) is the hinge loss, and we set \(b = 0 \) in this paper. For notational simplicity, we set \(\delta_k = x^1_k - x^2_k \) so that \(d_A(x^1_k, x^2_k) = \delta_k^T A \delta_k \). The regularization term \(\text{Div}(A_M, A^k_S) \) measures the difference between the new and previously obtained metric parameters. In this paper, we choose \(\text{Div}(\cdot, \cdot) \) to be the LogDet divergence [Davis et al., 2007], which is desirable in DML since it is scale-invariant and can make \(A_M \) automatically satisfy the PSD constraint \(A_M \succeq 0 \).

For any two unlabeled samples \((x^U_{Sm}, x^U_{Mn}) \) in \(D^U \), we calculate their distances \(d_{A_M}(x^U_{Sm}, x^U_{Mn}) \) and \(d_{A_S}(x^U_{Sm}, x^U_{Mn}) \) in the target and source domain respectively. Since the two distances correspond to the same unlabeled sample pair, they should agree with each other (\(d_{A_M} \) should be small if \(d_{A_S} \) is small). This intuition is formulated in the regularization term \(R_I \), which enables the source metric to guide the metric learning in the target domain. Different types of regularization can be employed for \(R_I \), such as the absolute difference between the distances if the source and target features have been properly normalized. In this paper, we design a manifold based regularizer since it can take the geometry of the data distribution into consideration. Because the matrix \(A_M \) is positive semi-definite, we decompose it as \(A_M = U_M U_M^T \). Hence the distance \(d_{A_M}(x_{Mi}, x_{Mj}) = (x_{Mi} - x_{Mj})^T U_M U_M^T (x_{Mi} - x_{Mj}) = \|U_M^T x_{Mi} - U_M^T x_{Mj}\|^2 \). Then we define a regularizer so that the transformation \(U_M \) is smooth over the source manifold, i.e.,

\[
R_I(\cdot) = \frac{1}{(N_U)^2} \sum_{i,j=1}^{N_U} W_{Si,j} \|U_M^T x_{M_i} - U_M^T x_{M_j}\|^2 = \frac{1}{(N_U)^2} \text{tr} \left( X_M^L S (X_M^U)^T A_M \right),
\]

where the source manifold is approximated by the data adjacency graph \(W_S \) calculated based on the distances \(d_{A_S} \).
in the source domain, $L_S$ is the graph Laplacian given by
$L_S = D_S - W_S$, $W_S$ is constructed using $k$ nearest-neighbor
(kNN) method, and $D_S$ is a diagonal matrix with the element
$D_{Si} = \sum_{j=1}^{N_i} W_{Si}$.
In this paper, we choose $W_{Si}$ to be a binary weight, i.e., $W_{Si} = 1$ if the j-th unlabeled sample is the neighbor of the $i$-th sample, and 0 otherwise. By substituting (3) into (2), we obtain the following specific optimization problem:

$$
\arg \min_{A_M \geq 0} F(A_M) = \xi_M + \gamma_A \text{tr}((A_M^{-1})^2 A_M) - \gamma_A \log \det(A_M) + \frac{(\gamma_1)}{(N_i)^2} \text{tr}(X_M^T L_S X_M)^T A_M),
$$

s.t. $\delta_{MK} A_M \delta_{MK} - 1 \leq \xi_M $, (4)

where we initialize $A_M^0$ as an identity matrix.

2.2 Solution

The solution of the optimization problem (4) is given as in the following theorem.

**Theorem 1.** The optimal solution of problem (4) is given by:

$$
A^{k+1}_M = \begin{cases} 
B_Mk, & \tau_M \leq 1; \\
(B_Mk - (s_{MK} - 1)B_Mk\delta_{MK}^T B_Mk)/\gamma_A + \gamma_M s_{MK}, & 0 \leq \tau_M < 1.
\end{cases}
$$

(5)

where $\tau_M = \gamma_M (1 - 1/s_{MK})$, $B_Mk = ((A_M^{-1})^2 + 2\gamma_M H_S)^{-1}$
with $H_S = (N_i)^{-1} X_M^T L_S X_M^T$ and $s_{MK} = \delta_{MK}^T B_Mk \delta_{MK}$.

**Proof.** For notational simplicity, we omit the subscript $M$ in the following derivation. We initialize $\xi_M$ and $\gamma_A$.

$$
\mathcal{L}(A, \xi_M, \gamma_A, \tau) = \xi_M + \gamma_A \text{tr}((A^{-1})^2 A) - \gamma_A \log \det(A) + \gamma_1 \text{tr}(H_S A) - \lambda \xi_M + \tau(y_k[\delta_{MK}^T A \delta_{MK} - 1] - \xi_k).
$$

where $H_S = X^T L_S (X^T)^T$. By taking the derivative of $\mathcal{L}$ with respect to $A$ and setting it to be zero, we have

$$
A^{-1} = B_k^{-1} + \frac{1}{\gamma_A} Z_k.
$$

Here $B_k = ((A^{-1})^2 + 2\gamma_A H_S)^{-1}$ and $Z_k = y_k \delta_{MK} \delta_{MK}^T$.

By using the Sherman-Morrison inverse formula, i.e., $(B + uv^T)^{-1} = B^{-1} - (B^{-1}uv^TB^{-1})^{-1}$, we obtain:

$$
A = B_k - \frac{\gamma_A}{\gamma_A + y_k s_k} B_k^T B_k,
$$

where $s_k = \delta_{MK}^T B_k \delta_{MK}$, and $\gamma_A + \gamma_1 y_k s_k \neq 0$. By setting $\frac{\partial \mathcal{L}(A, \xi_M, \lambda, \gamma_A)}{\partial \xi_M} = 0$, we have

$$
1 - \lambda = \tau = 0.
$$

**Algorithm 1** The proposed online heterogeneous transfer metric learning (OHTML) algorithm.

**Input:** Labeled training pairs in the source and target domains, i.e., $\{x_{SN}, y_{SN}\}$ and $\{x_{MN}, y_{MN}\}$: unlabeled corresponding data in both domains, i.e., $\{x_{SN}', y_{SN}'\}$.

**Pre-calculation:** Learn $A_S$ in using the labeled data in the source domain, and calculate $H_S$ using the unlabeled corresponding data based on the learned $A_S$.

**Hyper-parameters:** $\gamma_A$ and $\gamma_I$.

**Output:** $A_{M_{k+1}}$

1: Initialize $A_{M_0} = I$.
2: for $k = 0, 1, 2, \ldots \ do$
3: Receive a labeled training pair: $(x_{MK}, y_{MK})$.
4: Calculate the empirical loss $\Psi(A_M)$ based on $A_{MK}$.
5: If the loss is greater than zero:
6: Pre-compute $B_{MK}$ and $s_{MK}$;
7: Update $A_{M_{k+1}}$ using (5).
8: else $A_{M_{k+1}} \leftarrow A_{MK}$.
9: end for

Since $\lambda \geq 0$, we have $\tau \leq 1$. By substituting (8) and (9) into (6), we obtain a sub-problem $\tilde{\mathcal{L}}(\tau)$ w.r.t. $\tau$. By setting $\frac{\partial \tilde{\mathcal{L}}(\tau)}{\partial \tau} = 0$ and considering $B_k = ((A^{-1})^2 + 2\gamma_A H_S)^{-1}$, we have

$$
\tau = \frac{\gamma_A}{y_k}(1 - \frac{1}{s_k}).
$$

(10)

Considering that $0 \leq \tau \leq 1$, we obtain

$$
\tau = \text{median}\{\frac{\gamma_A}{y_k}(1 - \frac{1}{s_k}), 0, 1\}.
$$

(11)

By substituting (11) into (8), we have

$$
A = \begin{cases} 
B_k, & 0 < \frac{2\gamma_A}{y_k}(1 - \frac{1}{s_k}) \leq 1; \\
B_k - (s_{MK} - 1)B_k \delta_{MK} \delta_{MK}^T B_k)/\gamma_A + \gamma_M s_{MK}, & \frac{2\gamma_A}{y_k}(1 - \frac{1}{s_k}) < 1.
\end{cases}
$$

(12)

This completes the proof.

In practice, we can calculate $B_{MK}$ using the Sherman-Morrison inverse formula, i.e., $(A + UV)^{-1} = A^{-1} - A^{-1}U(I + VA^{-1}U)^{-1}A^{-1}$, and obtain

$$
B_{MK} = A_{MK} - \frac{\gamma_A}{\gamma_I A_{MK}^T H_S(I + \frac{\gamma_I}{\gamma_A} A_{MK}^T H_S)^{-1}} A_{MK}.
$$

(13)

We summarize the main procedure of the proposed OHTML algorithm in Algorithm 1. The following theorem guarantees that the obtained resulting matrix $A_{M_{k+1}}$ calculated using Algorithm 1 is positive definite.

**Theorem 2.** The resulting matrix $A_{M_{k+1}}$ in (5) is positive definite.

**Proof.** For notational simplicity, we omit the subscript $M$ in the following derivation. We use $S_{d+}$ to denote the...
sets of positive semi-definite and positive definite matrices respectively. We have that the matrix \( H_S \in S^d_+ \), since the integrated Laplacian matrix \( L_S \in S^d_+ \). Because the inverse of a positive definite matrix is also positive definite, we conclude that \( B_k \in S^d_+ \) and \( s_k = \delta_k B_k \delta_k > 0 \) for any non-zero vector \( \delta_k \).

If \( \gamma_A (1 - \frac{1}{s_k}) \leq 0 \), it is obvious that \( x^T A^{k+1} x = x^T B_k x > 0 \). This indicates that \( A^{k+1} \in S^d_+ \).

If \( 0 < \gamma_A (1 - \frac{1}{s_k}) < 1 \),

\[
x^T A^{k+1} x = x^T B_k x - \frac{(s_k - 1) x^T B_k \delta_k \delta_k^T B_k x}{s_k^2} + \frac{x^T B_k \delta_k \delta_k^T B_k x}{s_k^2}.
\]

(14)

It is easy to verify that \( (x^T B_k x)(\delta_k^T B_k \delta_k) - (x^T B_k \delta_k)^2 \geq 0 \) according to the Cauchy-Schwarz inequalities. Therefore, \( x^T A^{k+1} x > 0 \) for any non-zero vector \( x \), and \( A^{k+1} \in S^d_+ \).

If \( \gamma_A (1 - \frac{1}{s_k}) \geq 1 \), when \( y_k = 1 \), for any vector \( x \in \mathbb{R}^d \), we have

\[
x^T A^{k+1} x = x^T B_k x - \frac{x^T B_k \delta_k \delta_k^T B_k x}{\gamma_A + s_k}.
\]

(15)

Hence \( A^{k+1} \in S^d_+ \). When \( y_k = -1 \), we have \( \gamma_A > s_k \) since \( \gamma_A (\frac{1}{s_k} - 1) \geq 1 \). Therefore, for any vector \( x \in \mathbb{R}^d \), we also have

\[
x^T A^{k+1} x = x^T B_k x + \frac{x^T B_k \delta_k \delta_k^T B_k x}{\gamma_A - s_k} > 0.
\]

(16)

This completes the proof. \( \square \)

The complexity of the proposed algorithm mainly depends on calculation of the matrix \( B_{MK} \), which involves multiplication and inversion of \( d_M \times d_M \) matrices. Therefore, the complexity of the proposed algorithm is \( O(N_M d_M^2) \), where \( N_M \) and \( d_M \) are the number of labeled training pairs and feature dimension in the target domain respectively. The constant \( c \leq 3 \) is determined by the utilized multiplication and inversion algorithms. It should be noted that the matrix \( H_S \) can be pre-computed and hence the complexity is independent on the source domain, as well as the number of unlabeled correspondences. Therefore, the proposed method is quite efficient as long as \( d_M \) is not very large.

3 Experiments

In this section, we evaluate the proposed OHTML algorithm in four different applications: object categorization, scene clustering, face verification and image retrieval. In the first three applications, we investigate how much the source domain with powerful but computationally expensive feature could help DML in the target domain with cheap feature. In the last application, we utilize the interpretable text feature to help DML with visual feature, which is often hard to be interpreted.

3.1 Experimental Setup

Specifically, we compare with the following methods:

- **EU**: calculating the distance between samples in the target domain by applying the simple Euclidean metric, which is served as the baseline.
- **LEGO** [Jain et al., 2008]: a competitive online DML algorithm based on LogDet regularization.
- **RDML** [Jin et al., 2009]: an efficient online DML algorithm that is robust for high dimensional data.
- **DAMA** [Wang and Mahadevan, 2011]: a competitive heterogeneous transfer learning (HTL) approach based on manifold alignment. It utilizes class labels to align the heterogeneous domains.
- **MI** [Dai et al., 2015]: a recently proposed metric imitation approach that utilizes the expensive feature to help learn a good metric for cheap feature. Large amounts of unlabeled correspondences are used for knowledge transfer via manifold approximation.
- **OHTML**: the proposed online HTML method. OHTML(EU) means that we set \( A_S \) as an identity matrix in term (3), i.e., do not learn the source metric and directly employ Euclidean metric in the source domain.

For the single DML algorithms (LEGO and RDML), only the limited labeled training pairs in the target domain are utilized, and no additional information from the source domain is leveraged. For DAMA and MI, a linear transformation \( U_M \) is learned for the target domain and we derive the metric parameter as \( A_M = U_M^T U_M \).

The candidate set for choosing the trade-off hyper-parameters is \{10^4, 10^5, 10^6\} if unspecified in the original papers. The hyper-parameters \( \gamma_A \) and \( \gamma_I \) are tuned on the set \{10^4, 10^5, 10^6\} and \{10^4, 10^5, 10^6\} respectively. Hyper-parameter determination is still an open issue in HTL due to the limited labeled data. To this end, best performance over the candidate sets are reported for all compared methods.

For all the different methods, kernel principal component analysis (KPCA) is adopted to explore some nonlinearity in the data, and also reduce the dimensionality. We run the experiments ten times by randomly choosing the labeled set.

The algorithms are implemented using Matlab and the experiments are conducted on a 3.4 GHz Intel Xeon E5-2687W (8 cores) computer.

3.2 Object Categorization

This set of experiments is conducted on the popular Caltech-101 [Fei-Fei et al., 2004] dataset, which contains 8,677 images that belong to 101 object categories. The expensive deep CNN [Chatfield et al., 2014] and cheap PHOG [Bosch et al., 2007] are adopted as the feature representations in the source and target domains respectively. The features are provided by [Dai et al., 2015], where the original feature dimensions...
are 4096 and 40 respectively. The resulting dimensions after KPCA are 512 and 40. Half of the dataset is used for training and the remaining is for test. In both the source and target domains, we randomly select \{10, 15, 20, 25, 30\} labeled training samples for each category. The labeled training pairs are constructed according to the strategy in [Weinberger et al., 2005], and k-NN is adopted as the classifier.

Performance w.r.t. Different Number of Labeled Samples
The classification accuracies and training costs w.r.t. a varied number of labeled training samples are shown in Fig. 1. We do not show the curve of OHTML(EU) in the time cost figure since the source metric $A_S$ is pre-calculated and the training costs of OHTML and OHTML(EU) are the same. From the results, we observe that: 1) the accuracies of all different methods improve with an increasing number of labeled samples. The single domain DML algorithms (LEGO and RDML) are only slightly better than the EU baseline, while the HTL approaches (DAMA, MI and OHTML) outperform them significantly. This demonstrates that knowledge of the source domain can help the target metric learning and be successfully transferred to the target domain by the different HTL approaches; 2) MI is better than DAMA when the number of labeled data is small (such as 10), but DAMA is better when more labeled data are provided. This is because MI only utilizes the unlabeled corresponding data to facilitate knowledge transfer, while DAMA relies on labeled data; 3) the proposed OHTML outperforms MI consistently since we can leverage the label information in the target domain. OHTML(EU) is a bit worse than OHTML since we do not learn the source metric in OHTML(EU). The accuracy decrease is not significant since the utilized source feature is much more expressive than the target feature and the estimated data adjacency graph $W_S$ in the source domain can be used to guide the target metric learning even without learning the source metric; 4) OHTML is also superior to DAMA given limited labeled data since we use the unlabeled correspondences to build domain connection; 5) the training time of DAMA increases sharply when more labeled instances are given, while the proposed OHTML is much steady and the costs are only slightly higher than the single domain DML algorithms, and much lower than MI and DAMA. This demonstrates efficiency of our method.

Performance w.r.t. Different Number of Unlabeled Correspondences
A certain percentage of training data is used as the unlabeled corresponding data. The classification accuracies and training costs w.r.t. a varied percentage are shown in Fig. 2. We only report the performance of MI and OHTML since other approaches do not use such data. It can be seen from the results that: 1) accuracies of both MI and OHTML increase when more unlabeled correspondences are provided, and the proposed method outperforms MI consistently; 2) the training cost of MI increases dramatically while OHTML is much steady.
### 3.3 Scene Clustering

We conduct scene clustering on the Scene-15 [Lazebnik et al., 2006] dataset. It consists of 4,585 images from 15 natural scene categories. CNN [Chatfield et al., 2014] and LBP [Ojala et al., 2002] are used as the feature representations of the source and target domains respectively. The resulting dimensions after KPCA are 512 and 50 respectively. We randomly split the dataset into equal size for training and test. The number of labeled samples is 10 for each category. Following [Dai et al., 2015], spectral clustering is applied to group the data and the evaluation criterion is the purity of clustering. We report the performance in Table 1.

We can see from the results that the HTL approaches are much better than the single domain DML algorithms, which are only comparable to the EU baseline. MI and OHTML outperform DAMA significantly and the computational cost of DAMA is low. This is because the total number of labeled samples (15×10) is quite small in this set of experiments. The standard deviation of MI is zero since: 1) it is an unlabeled approach and the learned target metric does not depend on the varied labeled sets; 2) the labeled set is not utilized for test in clustering.

### 3.4 Face Verification

In this subsection, we employ the well-known labeled face in the wild (LFW) [Huang et al., 2007] dataset, where there are 13,233 face images of 5,749 individuals. The source and target features are CNN and LBP [Chen et al., 2013] respectively. The dimension of LBP is reduced to 400 suggested by [Chen et al., 2013]. We conduct experiments under the unrestricted protocol since DAMA needs the class label information, and no outside data are utilized. We adopt the standard 10-folds split of the dataset [Huang et al., 2007], and each fold is used for test in turn. The performance is reported in Table 2.

From the results we can see that: 1) although the single domain DML algorithms are very efficient, the improvements on accuracy are quite limited compared with the EU baseline; 2) DAMA is superior to MI since the provided label information in both domains is enough for it to achieve satisfactory performance. By making use of both the unlabeled corresponding data and label information in the target domain, we obtain the best accuracy, and is more efficient than MI and DAMA.

Table 2: Verification accuracy and training cost on the LFW dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
<th>CPU time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EU</td>
<td>0.831±0.009</td>
<td>NA</td>
</tr>
<tr>
<td>LEGO</td>
<td>0.843±0.012</td>
<td>3.283±0.224</td>
</tr>
<tr>
<td>RDML</td>
<td>0.853±0.008</td>
<td>1.533±0.216</td>
</tr>
<tr>
<td>DAMA</td>
<td>0.906±0.023</td>
<td>409.702±17.555</td>
</tr>
<tr>
<td>MI</td>
<td>0.885±0.014</td>
<td>945.203±1.780</td>
</tr>
<tr>
<td>OHTML(EU)</td>
<td>0.918±0.011</td>
<td>106.375±0.874</td>
</tr>
<tr>
<td>OHTML</td>
<td>0.924±0.010</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Retrieval MAP and training cost on the Corel5K dataset. The number of labeled instances for each concept is 10.

<table>
<thead>
<tr>
<th>Methods</th>
<th>MAP</th>
<th>CPU time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EU</td>
<td>0.265±0.000</td>
<td>NA</td>
</tr>
<tr>
<td>LEGO</td>
<td>0.274±0.007</td>
<td>0.673±0.030</td>
</tr>
<tr>
<td>RDML</td>
<td>0.266±0.001</td>
<td>0.375±0.029</td>
</tr>
<tr>
<td>DAMA</td>
<td>0.268±0.002</td>
<td>0.627±0.020</td>
</tr>
<tr>
<td>MI</td>
<td>0.291±0.000</td>
<td>4.176±0.067</td>
</tr>
<tr>
<td>OHTML(EU)</td>
<td>0.292±0.002</td>
<td>3.314±0.402</td>
</tr>
<tr>
<td>OHTML</td>
<td>0.296±0.003</td>
<td></td>
</tr>
</tbody>
</table>

### 3.5 Image Retrieval

We further apply the different methods to image retrieval and the Corel5K [Duygulu et al., 2002] dataset is used for evaluation. The dataset contains 5,000 images belonging to 50 concepts (100 images for each concept). The semantic tag is used as the source feature and the bag-of-visual word (BoVW) based on the local SIFT [Lowe, 2004] is adopted as the target feature. Dimensions of both the tag and BoVW features are reduced to 100, and half of the data are used for training. Following [Dai et al., 2015], mean average precision (MAP) is adopted as the evaluation criterion, and the results are shown in Table 3.

In this application, DAMA is only comparable to the EU baseline and single domain DML algorithms. This may be because there is semantic gap between the textual and visual features. Thus it is harder to build connections between the source and target domains using the limited label information. MI and the proposed OHTML are much better and our method outperforms MI in terms of both MAP score and training cost.

### 4 Conclusion

This paper presents a general online model for heterogeneous transfer metric learning (HTML). The model is based on the direct pairwise distance minimization between the source and target domain. By formulating it under the manifold regularization theme, we obtain an efficient online HTML (OHTML) algorithm. Both effectiveness and efficiency of the proposed method are verified in diverse applications. We mainly conclude from the results that: 1) it is advantageous to utilize the expensive or interpretable feature to help learning a relatively good metric for the cheap feature or the feature that is hard to interpret; 2) the developed online model can significantly accelerate HTML with little accuracy sacrifice in most cases, especially when the labeled data are limited in the target domain.

### Acknowledgments

This work is supported by Singapore NRF2015ENC-GDCR01001-003, administrated via IMDA, NRF2015ENC-GBICR001-012, administrated via BCA, and Australian Research Council Projects FL-170100117, DP-140102164, and LP-150100671.
References


