Manifold Regularized Transfer Distance Metric Learning

Haibo Shi[1]
shh@pku.edu.cn
Yong Luo[2]
yluo180@gmail.com
Chao Xu[1]
xuchao@cis.pku.edu.cn
Yonggang Wen[2]
ygwen@ntu.edu.sg

[1] Key Laboratory of Machine Perception (MOE), Cooperative Medianet Innovation Center, School of EECS, Peking University, Beijing, China
[2] School of Computer Engineering, Nanyang Technological University, Singapore

The performance of many computer vision and machine learning algorithms are heavily depend on the distance metric between samples. It is necessary to exploit abundant of side information like pairwise constraints to learn a robust and reliable distance metric[2, 3]. Let $\mathcal{D} = \{(x_i^l, x_j^l, y_{ij})\}_{i,j=1}^{n^l}$ denotes the labeled training set for the target task, wherein x_i, $x_j \in R^d$ and $y_{ij} = \pm 1$ indicates x_i^l and x_i^l are similar/dissimilar to each other.

Then, a metric is usually learned to minimize the distance between the data from the same class and maximize their distance otherwise. This leads to the following loss function for learning the metric A:

$$\begin{aligned}\Phi(A) &= \sum_{y_{ij}=1} \|x_i - x_j\|_A^2 - \mu \sum_{y_{ij}=-1} \|x_i - x_j\|_A^2 \\ &= \mathbf{tr}(\mathbf{S}\cdot A) - \mu \mathbf{tr}(\mathbf{D}\cdot A)\end{aligned} \tag{1}$$

where $d_A(x_i, x_j) = \|x_i - x_j\|_A = \sqrt{(x_i - x_j)^T A (x_i - x_j)}$ is the distance between two data points x_i and x_j, $\mathbf{tr}(\cdot)$ is the trace of matrix, μ is a positive trade-off parameter. Here, S and D are given by $\mathbf{S} = \sum (x_i - x_j)(x_i - x_j)^T$, $(x_i, x_j) \in \mathcal{S}$, $\mathbf{D} = \sum (x_i - x_j)(x_i - x_j)^T$, $(x_i, x_j) \in \mathcal{D}$.

The loss function (1) above is widely used in distance metric learning (DML) method. A regularization term $\Omega(A) = \|A\|_F^2$ can be added in (1) to control the model complexity. However, when the number of labeled data n^l is small, such a simple regularization is often insufficient to control the model complexity. The recently proposed decomposition based TDML (DTDML) [2] algorithm is superior to the previous TMDL approaches in that much fewer variables are needed to be learned. Given the m source tasks, we assume there are large amount of n^u unlabeled data $\{x_i^u, x_j^u\}$, as well as m different but related source tasks with abundant labeled training data $\mathcal{D}_p = \{(x_{pi}, x_{pj}, y_{pij})\}_{i,j=1}^{n_p}$, $p = 1, \ldots, m$. Then we learn m corresponding metrics $A_p \in \mathbb{R}^{d\times d}, p = 1, \ldots, m$ independently. Considering that any metric A can be decomposed as $A = U\mathrm{diag}(\theta)U^T = \sum_{i=1}^d \theta_i u_i u_i^T$, DTDML proposed to learn a combination of some base metrics to approximate the optimal target metric. The base metrics can be derived from the source metrics or some randomly generated base vectors. Based on this idea, the formulation of DTDML is given by

$$\arg\min_{\beta,\theta} \Phi(\beta, \theta) + \frac{\gamma_A}{2}\|A - A_S\|_F^2 + \frac{\gamma_B}{2}\|\beta\|_2^2 + \frac{\gamma_C}{2}\|\theta\|_1^2 \tag{2}$$

where $\Phi(\cdot)$ is some pre-defined convex loss, $A = \sum_{r=1}^{m\times d} \theta_r u_r u_r^T$, and $A_S = \sum_{p=1}^m \beta_p A_p$ is an integration of the source metrics.

Although the limited labeled samples in the target task and the auxiliary source metrics are effectively utilized in problem (2) by simultaneously minimizing the losses $\Phi(\beta, \theta)$ and the divergence between A_S and A, the large amount of unlabeled data in the target task are discarded. Therefore, we propose to utilize manifold regularization [1] to take advantage of all the given labeled and unlabeled information in a unified metric learning framework.

Manifold regularization implies the geometry of the intrinsic data probability distribution is supported on the low-dimensional manifold. The Laplacian of the adjacency graph computed in an unsupervised manner using Laplacian Eigenmap with both labeled and unlabeled samples. The data manifold can be approximated with the graph Laplacian. Moreover, the distance measure is a key point for graph Laplacian construction. Since both the integrated source metric A_S and target metric A are derived from the same feature space and related tasks, these two metrics should be similar. Rather than explaining this similarity by simply minimizing the least squares difference in DTDML, we formulate it as a smoothing penalty term. Based on the obtained source metric A_p, we construct an adjacency graph W_p by using all the labeled and unlabeled data in the target task. This leads to multiple graphs W_p, $p = 1, \ldots, m$. Considering the target metric A, distance between two samples can be further written as, $d_A(x_i, x_j) = \sqrt{(x_i - x_j)^T A (x_i - x_j)} = \sqrt{(x_i - x_j)^T PP^T (x_i - x_j)}$ with $P \in \mathbb{R}^{d\times d}$. As a consequence, it is equivalent to learn the target metric A and the linear mapping P. Following the manifold regularization principle, we can smooth P along the data manifold [1, 3], which is approximated by the Laplacian of the graph W_p. By summing over all the different graphs $\{W_p\}_{p=1}^m$, we obtain the following regularizer for the mapping P as well as the metric A, i.e.,

$$\begin{aligned}\Omega(A) &= \frac{1}{2}\sum_{p=1}^m \beta_p \Big(\sum_{i,j} \|Px_i - Px_j\|^2 W_p(i,j)\Big) \\ &= \mathbf{tr}(XLX^T A)\end{aligned} \tag{3}$$

where $L = \sum_{p=1}^m \beta_p L_p$, is the integrated graph Laplacian, and each $L_p = D_p - W_p$. Here, D_p is a diagonal matrix with the entity $D_{pii} = \sum_{j=1}^{n^l+n^u} W_{pij}$. In this way, target metric A is not only close to an integration of the source metrics, and also smooth along the data manifold. This leads to lower model complexity compared with DTDML, and thus better generalization ability for metric learning.

By introducing the regularizer (3) in (1), and adopting the decomposition based metric learning strategy in [2], we obtain the following optimization problem for our MTDML:

$$\begin{aligned}\arg\min_{\beta,\theta} &\mathbf{tr}(\mathbf{S}\cdot A) - \mu\mathbf{tr}(\mathbf{D}\cdot A) + \gamma_A \mathbf{tr}(XLX^T A) \\ &+ \frac{\gamma_B}{2}\|\beta\|_2^2 + \frac{\gamma_C}{2}\|\theta\|_2^2 \\ &\text{s.t.} \sum_{i=1}^m \beta_i = 1, \beta_i \geq 0, i = 1, \ldots, m\end{aligned} \tag{4}$$

where $\gamma_A, \gamma_B, \gamma_C$ are positive trade-off parameters selected empirically by grid search. With learned θ^*, we can easily construct $A^* = \sum_{r=1}^{m\times d} \theta_r^* u_r u_r^T$ as optimal distance metric for next step classification.

In the optimization, the "base metric" combination coefficients and the source graph Laplacian integration weights are learned alternatively until converge. We therefore obtain more reliable solutions given the limited side information. Experiments are conducted on NUS-WIDE, which is a challenge web image annotation dataset and USPS, a handwritten digit classification dataset. The results confirm the effectiveness of the proposed MTDML.

Acknowledgements
This research was partially supported by grants from NBRPC 2011CB302400, NSFC 61375026, 2015BAF15B00.

[1] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 7: 2399–2434, 2006.

[2] Y Luo, T Liu, D Tao, and C Xu. Decomposition-based transfer distance metric learning for image classification. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 23(9):3789–3801, 2014.

[3] Zheng-Jun Zha, Tao Mei, Meng Wang, Zengfu Wang, and Xian-Sheng Hua. Robust distance metric learning with auxiliary knowledge. In *IJCAI*, pages 1327–1332, 2009.