# Manifold Regularized Transfer Distance Metric Learning

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

- <sup>1</sup> Key Laboratory of Machine Perception (MOE), Cooperative Medianet Innovation Center, School of EECS, Peking University, Beijing, China
- <sup>2</sup> School of Computer Engineering, Nanyang Technological University, Singapore

#### Abstract

The performance of many computer vision and machine learning algorithms are heavily depend on the distance metric between samples. It is necessary to e xploit abundant of side information like pairwise constraints to learn a robust and reliable distance metric. While in real world application, large quantities of labeled data is unavailable due to the high labeling cost. Transfer distance metric learning (TDML) can be utilized to tackle this problem by leveraging different but certain related source tasks to learn a target metric. The recently proposed decomposition based TDML (DTDML) is superior to other TDML methods in that much fewer variables need to be learned. In spite of this success, the learning of the combination coefficients in DTDML still relies on the limited labeled data in the target task, and the large amounts of unlabeled data that are typically available are discarded. To utilize both the information contained in the source tasks, as well as the unlabeled data in the target task, we introduce manifold regularization in DTDML and develop the manifold regularized transfer distance metric learning (MTDML). In particular, the target metric in MTDML is learned to be close to an integration of the source metrics under the manifold regularization theme. That is, the target metric is smoothed along each data manifold that is approximated by all the labeled and unlabeled data in the target task and each source metric. In this way, more reliable target metric could be obtained given the limited labeled data in the target task. Extensive experiments on the NUS-WIDE and USPS dataset demonstrate the effectiveness of the proposed method.

# **1** Introduction

Distance metric is critical in a number of machine learning and pattern recognition algorithms [4, 4, 11, 12, 17]. For instance, the simple k-nearest neighbor (kNN) classifier is sometimes superior to other well designed classifier with a suitable distance metric [4, 14]. Also in unsupervised settings such as clustering, a proper distance metric will help find a plausible patterns that meaningful to users.

In order to learn a robust distance metric to reveal data similarity, we need a large amount of side information [15, 16, 17]. These side information is usually cast in a form of pairwise constraint that indicates weather samples are similar or not. However, abundant labeled samples is unavailable in practice due to the high labeling cost. A natural solution to tackle this problem is to utilize the large amount of unlabeled data [5, 8, 9]. For example, Hoi et al. [5] proposed a Laplacian regularized metric learning algorithm for image retrieval. Recently, many works focus on transfer distance metric learning (TDML), which aims to derive a reliable target metric by leveraging auxiliary knowledge from some different but related tasks [10, 18, 19]. For example, the target metric is learned in [18] by minimizing the log-determinant divergence between the source metrics and target metric. Zhang and Yeung [19] proposed to learn the task relationships in transfer metric learning, and therefore, allow modeling of negative and zero transfer. DTDML [10] is superior to both of these two methods in that the number of variables to be learned is reduced significantly, by representing the target metric as a combination of some "base metrics". However, in the learning of the combination coefficients, only the source metrics pre-trained on the labeled data in the source tasks and limited labeled data in the target task are utilized. But the large amount of unlabeled data in the target task are discarded.

Therefore, in this paper, we propose manifold regularized transfer distance metric learning (MTDML), which could make full use of both the label information in the source tasks and all the available (labeled and unlabeled) data in the target task. In particular, a source metric is learned by utilizing the abundant labeled data in a certain source task. Then a data adjacency graph is constructed by applying a learned source metric on all the data in the target task. This leads to multiple graphs, and we integrate them to reveal the geometric structure of the target data distribution, which is assumed to be supported on a low-dimensional manifold. Finally, the Laplacian of the integrated graph is formulated as a regularization term to smooth the target metric. In this way, both the large amount of label information and the abundant unlabeled data in the target task are utilized in the learning of the target metric. The learned target metric not only is close to an integration of the source metrics, but also respects the data distribution of the target task. We therefore obtain more reliable solutions given the limited side information. In the optimization, the "base metric" combination coefficients and the source graph Laplacian integration weights are learned alternatively until converge. We conduct experiments 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.

# 2 Manifold Regularized Transfer Distance Metric Learning

In this section, we propose our manifold regularized transfer distance metric learning (MT-DML) model. The following are some notations that used throughout this paper: 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  $x_i, x_j \in \mathbb{R}^d$  and  $y_{ij} = \pm 1$  indicates  $x_i^l$  and  $x_i^l$  are similar/dissimilar to each other. The number of labeled samples  $n^l$  is very small, and thus 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  $\mathcal{D}_p = \{(x_{pi}, x_{pj}, y_{pij})\}_{i,j=1}^{n_p}, p = 1, \dots, m$ .

#### 2.1 Background and Overview of MTDML

In distance metric learning (DML), 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*:

$$\Phi(A) = \sum_{y_{ij}=1} \|x_i - x_j\|_A^2 - \mu \sum_{y_{ij}=-1} \|x_i - x_j\|_A^2$$
  
=  $\mathbf{tr}(\mathbf{S} \cdot A) - \mu \mathbf{tr}(\mathbf{D} \cdot A)$  (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,  $\mu$  is a positive trade-off parameter. Here, *S* and *D* are given by,

$$\mathbf{S} = \sum_{(x_i, x_j) \in S} (x_i - x_j) (x_i - x_j)^T,$$
  
$$\mathbf{D} = \sum_{(x_i, x_j) \in D} (x_i - x_j) (x_i - x_j)^T$$
(2)

The loss function (1) above is widely used in [5, 13, 18]. 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, especially less than the number of variables to be estimated in metric matrix A, such a simple regularization is often insufficient to control the model complexity. Transfer distance metric learning (TDML) methods [10, 18, 19] tackle this problem by leveraging auxiliary knowledge from some different but related source tasks. The recently proposed decomposition based TDML (DTDML) [10] algorithm is superior to the previous TMDL approaches in that much fewer variables are needed to be learned.

Given the *m* source tasks, we learn *m* corresponding metrics  $A_p \in \mathbb{R}^{d \times d}$ , p = 1, ..., mindependently. Considering that any metric *A* can be decomposed as  $A = U \operatorname{diag}(\theta) U^T = \sum_{i=1}^{d} \theta_i u_i u_i^T$ , 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. For example, we can apply singular value decomposition (SVD) to  $A_p$  and obtain a set of source eigenvectors  $U_p = [u_{p1} \dots u_{pd}]$ . Then the base metric is calculated as  $B_{pj} = u_{pj}u_{pj}^T$  and this leads to  $m \times d$  base metrics for *m* source tasks. Based on this idea, the formulation of DTDML is given by

$$\arg\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \Phi(\boldsymbol{\beta},\boldsymbol{\theta}) + \frac{\gamma_A}{2} \|\boldsymbol{A} - \boldsymbol{A}_{\boldsymbol{S}}\|_F^2 + \frac{\gamma_B}{2} \|\boldsymbol{\beta}\|_2^2 + \frac{\gamma_C}{2} \|\boldsymbol{\theta}\|_1^2$$
(3)

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 (3) 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 [2] to take advantage of all the given labeled and unlabeled information in a unified metric learning framework. The diagram of the proposed MTDML is shown in Figure. 1. Each source task learns a corresponding source metric independently. These metrics are used to construct several graphs by using the labeled and unlabeled data in target task. The different graph Laplacians are



Figure 1: Diagram of the proposed MTDML algorithm.

weightedly combined as an integrated Laplacian matrix L, which is used to preserve locality in the feature space and approximate the data distribution. Meanwhile, the target metric  $A = U \operatorname{diag}(\theta) U^T$  is represented by a group of base metrics that generated from decomposition of source metrics. By incorporating the limited label information, and smoothing the target metric A along the data manifold that is approximated with the integrated L, we learn the base metric combination coefficients and the graph Laplacian integration weights simultaneously, and finally, obtain the result metric.

### 2.2 Problem Formalization

*Manifold regularization framework* [2] 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[1] 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 [2, 18], 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.,

$$\Omega(A) = \frac{1}{2} \sum_{p=1}^{m} \beta_p (\sum_{i,j} ||Px_i - Px_j||^2 W_p(i,j))$$

$$= \sum_{p=1}^{m} \beta_p \operatorname{tr}(XL_p X^T P P^T)$$

$$= \operatorname{tr}(XL X^T A)$$
(4)

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 (4) in (1), and adopting the decomposition based metric learning strategy in [10], we obtain the following optimization problem for our MTDML:

$$\arg\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \mathbf{tr}(\mathbf{S} \cdot A) - \mu \mathbf{tr}(\mathbf{D} \cdot A) + \gamma_A \mathbf{tr}(XLX^T A) + \frac{\gamma_B}{2} \|\boldsymbol{\beta}\|_2^2 + \frac{\gamma_C}{2} \|\boldsymbol{\theta}\|_2^2 \text{s.t.} \sum_{i=1}^m \beta_i = 1, \beta_i \ge 0, i = 1, \dots, m$$
(5)

where  $\gamma_A$ ,  $\gamma_B$ ,  $\gamma_C$  are positive trade-off parameters. By the use of the learned  $\theta^*$ , 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.

#### 2.3 Optimization

The solution of problem (5) can be obtained by alternately solving two sub-problem, which correspond to the minimization w.r.t  $\beta = [\beta_1, ..., \beta_m]^T$  and  $\theta = [\theta_1, ..., \theta_{m \times d}]^T$  respectively until convergence.

For fixed  $\beta$ , the optimization problem with respect to  $\theta$  is formulated as:

$$\arg\min_{\theta} \sum_{i=1}^{n} \theta_{i} \mathbf{tr}((\mathbf{S} - \mu \mathbf{D} + \gamma_{A} X L X^{T}) \cdot (u_{i} u_{i}^{T})) + \frac{\gamma_{C}}{2} \|\theta\|_{2}^{2}$$
(6)

Equation above can be further written in a compact form:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \boldsymbol{\theta}^T h + \frac{\gamma_C}{2} \|\boldsymbol{\theta}\|_2^2 \tag{7}$$

where  $h = [h_1, ..., h_n]$  with each  $h_i = \mathbf{tr}((\mathbf{S} - \mu \mathbf{D} + \gamma_A X L X^T)(u_i u_i^T))$ . Since (6) is a convex problem and has close form solution, we can get  $\theta^*$  efficiently.

With fixed  $\theta$ , the optimization problem with respect to  $\beta$  is formulated as:

$$\arg\min_{\beta} \gamma_{A} \sum_{i=1}^{m} \beta_{i} \mathbf{tr}(XL_{i}X^{T}A) + \frac{\gamma_{B}}{2} \|\beta\|_{2}^{2}$$
  
s.t. 
$$\sum_{i=1}^{m} \beta_{i} = 1, \beta_{i} \ge 0, i = 1, \dots, m$$
(8)

where A is derived from optimized  $\theta^*$  obtained last step. Similarly, (8) can be written as:

$$\beta^* = \arg\min_{\beta} \beta^T g + \frac{\gamma_B}{2} \|\beta\|_2^2$$
  
s.t.  $\sum_{i=1}^m \beta_i = 1, \beta_i \ge 0, i = 1, \dots, m$  (9)

where the constant term has been omitted,  $g = [g_1, ..., g_m]$  with each  $g_i = \gamma_A \operatorname{tr}(XL_i X^T A)$ . (9) is a standard quadratic programming problem with linear constraints. Typically, it can be solved by adopting coordinate descent algorithm. In each iteration, we select two elements  $\beta_i$  and  $\beta_j$  for updating while the others fixed. Due to the constraint  $\sum_{i=1}^{m} \beta_i = 1$ , the summation of  $\beta_i$  and  $\beta_j$  will not change after current iteration. Therefore, we obtain the following updating rule:

$$\beta_i^* = \frac{\gamma_B(\beta_i + \beta_j) + (h_j - h_i)}{2\gamma_B}$$

$$\beta_j^* = \beta_i + \beta_j - \beta_j^*$$
(10)

To satisfied  $\beta_i \ge 0$ , we set  $\beta_i^* = 0$  and  $\beta_j^* = \beta_i + \beta_j$  if  $\gamma_B(\beta_i + \beta_j) + (h_j - h_i) < 0$ . We iteratively traversal over all pairs of elements in  $\beta$  with (10) until the object function in (5) does not decrease.

In sum, the two group of coefficients are learning alternately by this two sub-problem until convergence.

## **3** Experiment Evaluation

In this section, experiments are conducted to validate the effectiveness of the proposed **MT-DML** on two popular datasets, a challenge real-world web image annotation dataset and a handwritten image dataset. For comparison purpose, we also evaluate the following methods:

- **RDML[6]** A regularized distance metric learning method. The regularization term,  $||A||_F^2$ , used to control model complexity. An online algorithm is demonstrated to be efficient and effective for solving the problem. In this paper, we conduct an aggregation strategy which is simply applying RDML on the training set that consists of data from both source and target tasks. This method also serves as a baseline for regularized metric learning.
- LRML[5] A semi-supervised distance metric learning method by utilizing unlabeled data. The LRML algorithm is formulated as Semidefinite Program (SDP) and solved by convex optimization techniques. This method serves as a baseline for semi-supervised metric learning.

- L-DML & M-DML[18] A transfer metric learning method that utilizes the auxiliary knowledge by adopting Bregman divergence as well also manifold regularization. The entries in target metric are estimated also by solving a SDP problem and we reimplement the algorithm by the toolbox, SDPT3 solver.
- **TML[19]** A transfer metric learning algorithm proposed recently. Relationship between the source and target tasks is learned for transfer by solving a second-order cone programming (SOCP). An online algorithm is developed for target metric's optimization.
- **STML[19]** A metric learning model based on the TML. It is provided as semi-supervised extension in the same paper with [19].
- **DTDML**[10] A transfer metric learning method learned by decomposition based method. Our method is built on top of DTDML and is further improved by exploiting useful information contained in the unlabeled data.

With metric learned by above algorithms, we apply 1-Nearest Neighbor classifier on the test set and report the average classification accuracy.

## 3.1 Web Image Annotation

We first utilize the well-known natural image dataset NUS-WIDE[3] to justify the significance of the proposed method for learning a reliable distance metric, especially under the data deficiency condition. This dataset contains 269,648 images and corresponding features. Our experiments use features extracted by SIFT [7] descriptor with 500-D bag of visual words, which is available on the homepage<sup>1</sup> of the dataset.

To perform transfer meaningfully, we select 12 animal concepts: bear, bird, cat, cow, dog, elk, fish, fox, horse, tiger, whale and zebra. For each concept, 100 samples are randomly picked from the dataset. In this setting of experiment, we randomly select 6 concepts as source tasks and target task requires annotation of all others. It forms as a multi-class problem. The pair constraints are labeled as similar if they are from the same class, otherwise dissimilar.

Method	Split 1	Split 2	Split 3
RDML	30.8	27.7	32.7
L-DML	28.6	28.3	31.7
M-DML	28.9	29.9	33.0
TML	32.3	31.7	36.6
DTDML	33.6	33.7	37.3
MTDML	35.3	35.3	38.7

Table 1: Average Annotation Performance on NUS-WIDE Dataset

The source metrics are trained by using the RDML method with all available data in the source tasks. For the target task, we divide the dataset by 6 labeled training samples for each class and 300 for testing, others as unlabeled data for semi-supervised setting. All trade-off parameters are determined empirically by adopting grid search. The overall classification

<sup>&</sup>lt;sup>1</sup>http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

result is reported in Table. 3.1. We perform 3 different random splits of the concept set and run the program for 5 iterations and calculate the average accuracy.

We then show that the performance of different semi-supervised distance metric learning method w.r.t the number of labeled samples in each class. The result shows in Figure. 2. **MTDML** always performs the best compared with other semi-supervised metric learning methods. In particular, our method harvests a 6.5% improvement on the average over M-DML when only 4 labeled samples are used in each class.



Figure 2: Comparison with other Semi-supervised Distance Metric Learning method.

#### 3.2 Handwritten Image Classification

USPS<sup>2</sup> is a handwritten digit dataset, which contains 7,291 samples. Each sample is constructed by an image of size  $16 \times 16$  in raw pixels. These raw images are treated as features of d = 256 dimension. Here, we conduct the same experiment settings with DTDML for comparison, i.e. nine tasks: 0/6, 0/8, 1/4, 2/7, 3/5, 4/7, 4/9, 5/8 and 6/8, each corresponding to a binary classification task. While training, we choose one of the nine tasks as target task and others as source tasks. Also, the source metrics for our method are obtained by RDML and all trade-off parameters are set empirically with grid search.

We present the average performance over all settings of MTDML with some related distance metric learning method. The result shows in Figure. (3).

From the result, we observe that: 1) Comparing to RDML, our method achieves a better classification performance which shows that source metrics and unlabeled data greatly help with learning target metric. 2) Our algorithm is superior than LRML which may due to exploiting information in source metrics as well as the finding of a bad local minima in LRML optimization steps. 3) The performance of all compared methods tends to be improved while training samples increasing. MTDML outperforms other methods and especially stable under training data deficiency condition.



Figure 3: USPS dataset in comparison with some baseline method.

# 4 Conclusion

This paper proposed the manifold regularization transfer distance metric learning model, **MTDML**. By integrating source and target tasks with a data geometric constraint, MTDML learns a robust and reliable metric from both labeled samples and unlabeled samples. Moreover, our algorithm can be solved quite efficiently under an alternative optimization framework. From the experimental validation on the web image annotation and handwritten image classification task, MTDML outperforms other transfer metric learning and semi-supervised metric learning methods, especially under training data deficiency conditions.

#### Acknowledgements

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

## References

- [1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.
- [2] 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.
- [3] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, page 48. ACM, 2009.

- [4] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image autoannotation. In *International Conference on Computer Vision*, pages 309–316. IEEE, 2009.
- [5] Steven CH Hoi, Wei Liu, and Shih-Fu Chang. Semi-supervised distance metric learning for collaborative image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2008.
- [6] Rong Jin, Shijun Wang, and Yang Zhou. Regularized distance metric learning: Theory and algorithm. In *Advances in neural information processing systems*, pages 862–870, 2009.
- [7] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- [8] Yong Luo, Dacheng Tao, Bo Geng, Chao Xu, and Stephen J Maybank. Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE Transactions on Image Processing*, 22(2):523–536, 2013.
- [9] Yong Luo, Dacheng Tao, Chang Xu, Chao Xu, Hong Liu, and Yonggang Wen. Multiview vector-valued manifold regularization for multilabel image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 24(5):709–722, 2013.
- [10] Yong Luo, Tongliang Liu, Dacheng Tao, and Chao Xu. Decomposition-based transfer distance metric learning for image classification. *IEEE Transactions on Image Processing*, 23(9):3789–3801, 2014.
- [11] Yong Luo, Tongliang Liu, Dacheng Tao, and Chao Xu. Multi-view matrix completion for multi-label image classification. *IEEE Transactions on Image Processing*, 24(8): 2355–2368, 2015.
- [12] Yong Luo, Dacheng Tao, Kotagiri Ramamohanarao, Chao Xu, and Yonggang Wen. Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE transactions on Knowledge and Data Engineering*, page DOI: 10.1109, 2015.
- [13] Hua Wang, Feiping Nie, and Heng Huang. Robust distance metric learning via simultaneous 11-norm minimization and maximization. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1836–1844, 2014.
- [14] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In Advances in neural information processing systems, pages 1473–1480, 2005.
- [15] Eric P Xing, Michael I Jordan, Stuart Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 505–512, 2002.
- [16] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2, 2006.
- [17] Liu Yang, Rong Jin, and Rahul Sukthankar. Bayesian active distance metric learning. *arXiv preprint arXiv:1206.5283*, 2012.

- [18] 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.
- [19] Yu Zhang and Dit-Yan Yeung. Transfer metric learning with semi-supervised extension. ACM Transactions on Intelligent Systems and Technology (TIST), 3(3):54, 2012.