Big Data Analytics

REVIEW Open Access

A review on multi-task metric learning

Peipei Yang[1]*, Kaizhu Huang[2] and Amir Hussain[3]

*Correspondence:
ppyang@nlpr.ia.ac.cn
[1]National Laboratory of Pattern Recognition, 95 East Zhongguancun Road, 100190 Beijing, China
Full list of author information is available at the end of the article

Abstract

Distance metric plays an important role in machine learning which is crucial to the performance of a range of algorithms. Metric learning, which refers to learning a proper distance metric for a particular task, has attracted much attention in machine learning. In particular, multi-task learning deals with the scenario where there are multiple related metric learning tasks. By jointly training these tasks, useful information is shared among the tasks, which significantly improves their performances. This paper reviews the literature on multi-task metric learning. Various methods are investigated systematically and categorized into four families. The central ideas of these methods are introduced in detail, followed by some representative applications. Finally, we conclude the review and propose a number of future work directions.

Keywords: Multi-task learning, Metric learning, Review

Background

In the area of machine learning, pattern recognition, and data mining, the concept of *distance metric* usually plays an important role. For many algorithms, a proper distance metric is critical to their performances. For example, the nearest neighbor classification relies on the metric to identify the nearest neighbor and determine their class, whilst k-means clustering uses the metric to determine which cluster a sample should belong to.

The metric is usually used as a measure of the similarity or dissimilarity, and there are various types of pre-defined distance metrics, such as Euclidean distance, cosine similarity, Hamming distance, etc. However, in practical applications, these general-purpose metrics are insufficient to catch the sundry particular properties of various tasks. Therefore, researchers propose learning a metric from data for particular tasks, to improve algorithm performance. This is termed metric learning [1–7].

With the advent of data science, challenging and evolving problems have arisen. Obtaining training data is a costly process, hence complex models are being trained on small datasets, resulting in poor generalization. Alongside this the number of tasks to be learnt has increased significantly. To overcome these problems, multi-task learning is proposed [8–13]. It aims to consider multiple tasks simultaneously at a higher level, whilst transferring useful information among different tasks to improve their performances.

After multi-task learning was proposed by Caruana [8] in 1997, various strategies have been designed based on different assumptions. There are also some closely related topics, such as transfer learning [14, 15], domain adaptation [16], meta-learning [17], life-long learning [18], learning to learn [19], etc. In spite of some minor discrepancies among them, they share the same basic idea that the performance is improved by considering multiple learning tasks jointly and sharing information with other tasks.

Under such a background, it is natural to consider the problem of multi-task metric learning. However, most multi-task learning algorithms designed for traditional models are difficult for applying to metric learning algorithms due to the obvious differences between the two kinds of models. To resolve this problem, a series of multi-task metric learning approaches are specifically designed for the metric learning models. By properly coupling multiple metric learning tasks, their performances are effectively improved.

Metric learning has the particularity that its effect on performance can be only given indirectly by the algorithm relying on the metric. This requires a different way to construct the multi-task learning framework from traditional models. As far as we know, there is no review at present on the multi-task metric learning, hence this paper will give a general overview of the existing works.

The rest of the paper is organized as follows. First we provide an overview of the basic concepts of metric learning and briefly introduce multi-task metric learning. Next, various strategies of multi-task metric learning approaches are reviewed. We then introduce some representative applications of multi-task metric learning, and conclude with a discussion on potential future issues.

Overview

In this section, we first provide an overview of metric learning, including its concept and several representative algorithms. Then a general description about multi-task metric learning is presented, leaving the details of the algorithms for the next section.

A brief review on metric learning

The notion of distance metric was originally a concept in mathematics, which refers to a function defined on $\mathcal{X}$ as $d : \mathcal{X} \times \mathcal{X} \to \mathbf{R}_+ =[0, +\infty)$ satisfying positiveness, symmetry, and triangle inequality [20]. In the community of machine learning, metric is unnecessary to keep its original definition from mathematics, but usually refers to a general measure of dissimilarity or similarity. A lot of machine learning algorithms use it to measure the dissimilarity between samples without explicitly referring its definition, such as nearest neighbor classification, k-means clustering, etc.

There have been various types of pre-defined metrics for general purposes. For example, assuming two points in the d-dimensional space $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X} = \mathbb{R}^d$, the most frequently used Euclidean distance is defined as $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$. Another example is the Mahalanobis metric [21] that is defined as $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)}$, where the symmetric positive semi-definite matrix $\mathbf{M}$ is the *Mahalanobis matrix* which determines the metric.

In spite of their widely-spread usage, the pre-defined metrics are incapable to capture the variety of real applications. Considering its importance to the performances of algorithms, researchers propose to learn a metric from the data instead of using the pre-defined metrics directly. By adapting the metric to the specific data for some algorithm, the performance is expected to be effectively improved. This is the central idea of the *metric learning*.

However, it is hardly practicable to learn a general metric by directly finding an optima in the functional space. A practical way is to define a family of metrics determined by some parameters, and transform the problem into the solving of the optimal parameters. Mahalanobis metric provides a perfect candidate for such a family of metrics, which has a simple formulation and is uniquely determined by the Mahalanobis matrix. In this case, the metric learning is equivalent to learning the Mahalanobis matrix.

Eric Xing et al. [1] proposes the idea of metric learning with the first algorithm in 2002. Since then, various metric learning methods have been proposed based on different strategies. Since metrics can be categorized into several families according to their properties, such as global vs. local, or linear vs. non-linear, the metric learning approaches can also be categorized accordingly. Mahalanobis metric is a typical global linear metric. Because existing multi-task learning approaches are almost based on global metrics, we focus on this type in this review, especially for the global linear metrics. Please refer to [22, 23] and their references for other types.

Most metric learning algorithms are formulated as a constrained optimization problem and the metric is obtained by solving this optimization. Since the distance is defined on two points, the supervised information to determine the metric, which is also called *side-information* in metric learning, is usually given by constraints on pairs or triplets as follows [22].

- Must-link / cannot-link constraints (positive/negative pairs)

$$\begin{aligned}\mathcal{S} =&\{(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ should be similar}\},\\ \mathcal{D} =&\{(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ should be dissimilar}\}.\end{aligned}$$

- Relative constraints (training triplets)

$$\mathcal{R} = \{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) : \mathbf{x}_i \text{ should be more similar to } \mathbf{x}_j \text{ than to } \mathbf{x}_k\}.$$

Using these constraints, we briefly introduce the strategies of some metric learning approaches. Xing's method [1] aims to maximize the sum of distances between dissimilar pairs while keeping the sum of squared distances between similar pairs to be small. It is an example of learning with positive/negative pairs. Large Margin Nearest Neighbors (LMNN) [2, 24] requires the k nearest neighbors to belong to the same class and pushes out all the imposters (instances of other classes existing in the neighborhood). The side-information is provided by the relative constraints. Information-theoretic metric learning (ITML) [3], which is also built with positive/negative pairs, models the problem with log-determinant. Sparse Metric Learning [6] uses the mixed $L_{2,1}$ norm to obtain a joint feature selection during metric learning, and Huang et al. [4, 5] proposes a unified framework for Generalized Sparse Metric Learning (GSML). Robust Metric Learning (RML) [25] deals with the noisy training constraints based on robust optimization.

It is notable that learning a Mahalanobis matrix can also be regarded as learning a linear transformation. For any symmetric positive semi-definite Mahalanobis matrix $\mathbf{M}$, there is a symmetric decomposition $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$ and the distance can be then reformulated as

$$
\begin{aligned}
d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) =& \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)} \\
=& \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{L}^\top \mathbf{L} (\mathbf{x}_i - \mathbf{x}_j)} \\
=& \sqrt{(\mathbf{L}\mathbf{x}_i - \mathbf{L}\mathbf{x}_j)^\top (\mathbf{L}\mathbf{x}_i - \mathbf{L}\mathbf{x}_j)} \\
=& \|\mathbf{L}\mathbf{x}_i - \mathbf{L}\mathbf{x}_j\|_2 .
\end{aligned}
\tag{1}
$$

By (1), the Mahalanobis metric defined by **M** is equivalent to the Euclidean distance after performing the linear transformation **L**, and thus metric learning can be also performed by learning such a linear transformation. Neighbourhood Component Analysis (NCA) [26] is an example of this class that optimizes the expected leave-one-out error of a stochastic nearest neighbor classifier by learning a linear transformation. Furthermore, the linear metric can be easily extended to the non-linear metric by replacing the linear transformation **L** with a non-linear transformation **f**, which is defined as

$$
d_{\mathbf{f}}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{f}(\mathbf{x}_i) - \mathbf{f}(\mathbf{x}_j)\|_2 .
$$

Then, the metric is obtained by learning an appropriate non-linear transformation **f**. Since the deep learning has achieved remarkable successes in computer vision and machine learning [27], some researchers proposed the deep metric learning recently [28, 29]. These methods resort to the deep neural network to learn a non-linear transformation, which are different from a traditional neural network in that their learning objective are given by constraints on distances.

There are a lot of metric learning methods because the metric plays an important role in many applications. We cannot introduce them in detail due to the limit of the space. Readers can refer to the paper [22] for a systematic review on metric learning.

An overview of multi-task metric learning

Since the concept of multi-task learning was proposed by Caruana [8] in 1997, this topic has attracted much attention from researchers in machine learning. Multiple different methods are proposed to construct a framework for simultaneously learning multiple tasks for conventional models, such as linear classifier or support vector machines. The performances of the original models are effectively improved by learning simultaneously.

However, these methods cannot be used directly for metric learning since there exist significant discrepancies between the conventional learning models and metric learning models. Taking the popular support vector machine (SVM) [30] as an example of conventional models, we can show the differences between it and metric learning. First, the training data of the two models are of different structures. For SVM, the training samples are given by points with a label for each one, while for metric learning they are given by pairs or triplets with a label for each one. Second, their models are of different types. The model of SVM is a single-input single-output function parameterized by a weight vector and a bias, while the model of metric learning is a double-input single-output function parameterized by a symmetric positive semi-definite matrix. Third, the algorithms take effect on the performance in different ways. For SVM, the classification accuracy is given by the algorithm directly, while for metric learning, the performance has to be evaluated indirectly by other algorithms working with the learned metric.

Due to the reasons mentioned above, strategies have to be specially designed to construct a multi-task metric learning model. They have to deal with two problem: (1) what type of useful information is shared among different metric learning tasks; (2) how such information is shared by the proposed model and algorithm. Parameswaran et al. [31] proposes the first multi-task metric learning approach in 2010, and in the following years a variety of strategies have been proposed for multi-task metric learning. We generally categorize them into the following families according to the way how the information is shared:

1 Assume that the Mahalanobis matrix of each metric is composed of several components and share some composition.
2 Pre-define the relationship among tasks or learn such relationship from data, and constrain the learning process with this relationship.
3 Use a common metric with proper regularization to couple the metrics.
4 Consider metric learning from the perspective of learning a transformation and share some parts of transformation.

There are some representative works in each family and we will introduce them in detail in next section. Figure 1 gives a summary of the multi-task metric learning approaches mentioned in this paper.

Review on multi-task metric learning approaches

In this section, we investigate the multi-task metric learning approaches published to-date and provide a detailed review on them. The methods are organized according to the type of strategies. We focus on only the models and algorithms in this section without mentioning their application backgrounds, which are left for the next section. The discussion about the relation between some closely related methods is also included.

Before diving into the details, we summarize the main features of these multi-task metric learning methods in Table 1. Besides, in this section, we always use **M** to represent

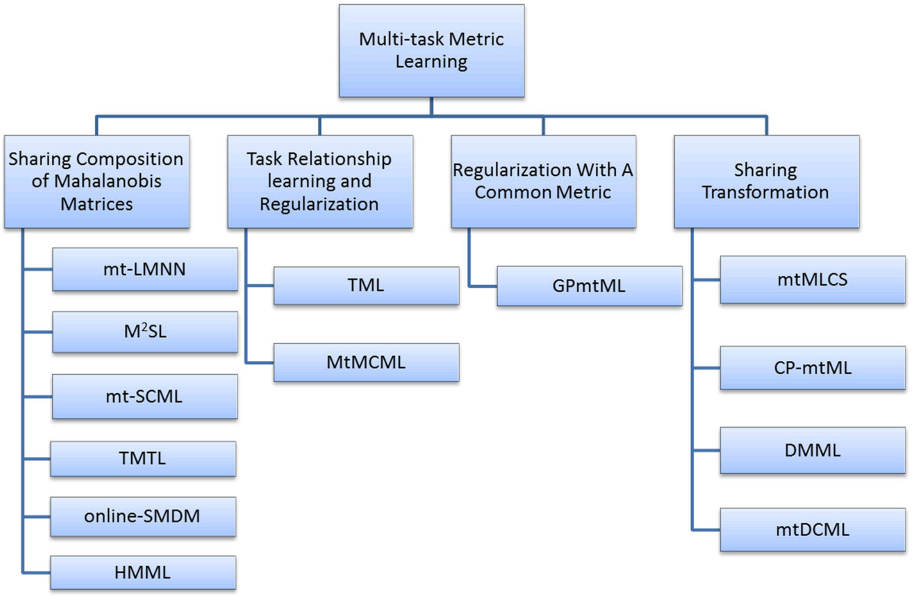
Multi-task Metric Learning
Sharing Composition of Mahalanobis Matrices
Task Relationship learning and Regularization
Regularization With A Common Metric
Sharing Transformation
mt-LMNN
M2SL
mt-SCML
TMTL
online-SMDM
HMML
TML
MtMCML
GPmtML
mtMLCS
CP-mtML
DMML
mtDCML

Fig. 1 A summary of multi-task metric learning approaches. This figure gives a summary of the approaches mentioned in this paper, where the name of each method is under the branch of its corresponding type

Table 1 Main features of multi-task metric learning methods

Name	Year	Multi-task Strategy	Solver	Dimension Reduction	Side-information	Regularizer
mt-LMNN	2010	Shared composition	Projected gradient descent	No	Triplets	Frobenius norm
TML	2010	Task relationship learning	Alternating Optimization	No	Pairs	Task covariance
mtMLCS	2011	Shared subspace	Gradient descent	Yes	Triplets	-
M^{2}SL	2012	Shared composition	Coordinate gradient descent	No	Pairs	Frobenius norm
GPmtML	2012	Geometry preserving	Alternating optimization	No	Triplets	Von Neumann divergence
mt-SCML	2014	Shared sparse representation	Regularized dual averaging	Yes	Triplets	ℓ_2/ℓ_1 norm
MtMCML	2014	Graph regularization	Alternating optimization	No	Pairs	Laplacian
TMTL	2015	Metric weighted sum	Direct calculation	No	Covariance	-
online-SMDM	2016	Shared composition	Online projected gradient descent	No	Pairs	Frobenius norm
CP-mtML	2016	Coupled projection	Stochastic gradient projection	Yes	Pairs	-
DMML	2016	Shared subnetwork	Sub-gradient descent	No	Pairs	-
HMML	2017	Shared composition	Not mentioned	No	Triplets	Trace norm
mtDCML	2017	Shared network	Gradient descent	No	Pairs	-

the Mahalanobis matrix to keep the notations uniform, which may be different from the original papers.

Sharing composition of Mahalanobis matrices

Since the Mahalanobis metric is uniquely determined by the Mahalanobis matrix, a natural way to couple multiple related metrics is to share some composition of their Mahalanobis matrices. Specifically, the Mahalanobis matrix of each task is assumed to be composed of both common composition shared by all tasks and task-specific composition preserving its specific properties. This strategy is the most popular way to construct a multi-task metric learning model and we introduce some representative ones below.

Large margin multi-task metric learning (mt-LMNN) Parameswaran et al. [31] proposes a multi-task learning model based on the idea to share a common composition of the Mahalanobis matrices. It is motivated by the regularized multi-task learning (RMTL) [9], and obtained by adapting RMTL to the large-margin nearest neighbor metric learning (LMNN) [2, 24]. To couple multiple tasks, each Mahalanobis matrix is decomposed into a common part $\mathbf{M}_0$ and a task-specific part $\mathbf{M}_t$. Thus the distance between two points $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$ defined by the metric of the t-th task is defined as

$$d_t(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{M}_0 + \mathbf{M}_t)(\mathbf{x}_i - \mathbf{x}_j), \tag{2}$$

By restricting that $\mathbf{M}_0 \succeq \mathbf{0}$ and $\mathbf{M}_t \succeq \mathbf{0}, \forall t$, the Mahalanobis matrix for each task is ensured to be positive semi-definite, which induces a positive semi-definite metric. In this model, $\mathbf{M}_0$ picks up the general trends across all tasks while $\mathbf{M}_t$ gathers the individual information for each task. The obtained regularization of mt-LMNN is

$$\gamma_0\|\mathbf{M}_0 - \mathbf{I}\|_{\mathrm{F}}^2 + \sum_{t=1}^{T} \gamma_t \|\mathbf{M}_t\|_{\mathrm{F}}^2. \tag{3}$$

In (3), the side-information is incorporated by constraints generated from triplets as LMNN [2]. The regularization on task-specific matrices $\mathbf{M}_t$'s represses the specialty of each task and encourages the shared part of all tasks, while the regularization on $\mathbf{M}_0$ restricts the common part to be close to the identity. They further make the learnt metric of each task not far from the Euclidean metric.

The hyper-parameters $\gamma_{t>0}$'s control the balance between the commonness and speciality, while γ_0 controls the regularization of the common part. As the increasing of $\gamma_{t>0}$, the task-specific parts become small and the learnt metrics of all tasks tend to be similar. When $\gamma_{t>0} \to \infty$, the algorithm learns a unique metric $\mathbf{M}_0$ for all tasks, while when $\gamma_{t>0} \to 0$, all tasks tend to be learnt individually. On the other hand, when $\gamma_0 \to \infty$, the common part $\mathbf{M}_0$ becomes identity and Euclidean metric is obtained. When $\gamma_0 \to 0$, there tends to be no regularization on the common part. This model is convex and can be solved effectively.

This is the first attempt to apply a multi-task approach to metric learning problem. It provides a simple yet effective way to improve the performance of metric learning by jointly learning multiple tasks. However, the idea of splitting each Mahalanobis matrix into a common part and an individual part is not easy to explain from the perspective of distance metric and can only deal with some simple cases.

Multi-task multi-feature similarity learning learning $\left(M^2\mathbf{SL}\right)$ Wang et al. [32] proposes a multi-task multi-feature metric learning approach to adapt the metric learning to large scale visual applications. For each sample, M types of features are extracted and the metrics are learnt individually for each feature. For each feature channel, there are T tasks and each task learns a distance metric. To make the information shared among tasks, the Mahalanobis matrix of the t-th task in the m-th feature channel is defined to be a combination of a common part $\mathbf{M}_0^{(m)}$ and an individual part $\mathbf{M}_t^{(m)}$. Then the authors incorporate such a formulation into the idealized kernel learning [33] and obtain the multi-feature multi-task metric learning model as

$$\begin{aligned}
&\min_{\mathbf{b},\mathbf{M}} \frac{1}{2}\left(\gamma_0 \sum_{m=1}^{M} \frac{1}{b_0^{(m)}} \left\|\mathbf{M}_0^{(m)}\right\|_{\mathrm{F}} + \sum_{t=1}^{T}\sum_{m=1}^{M} \frac{\gamma_t}{b_t^{(m)}} \left\|\mathbf{M}_t^{(m)}\right\|_{\mathrm{F}}^2\right) \\
&\qquad + \frac{C}{N}\sum_{t=1}^{T}\sum_{ij\in S} \xi_t^{ij} + \frac{\eta}{2}\sum_{t=0}^{T} \|\mathbf{b}_t\|_p^2 \\
&\text{s.t. } \delta_t^{ij}\left(d_t^{ij} - \tilde{d}_t^{ij}\right) \ge \sigma_t^{ij} - \xi_t^{ij},\ \xi_t^{ij} \ge 0,\ b_t^{(m)} \ge 0,\ p > 1,\ \mathbf{M}_t^{(m)} \succeq \mathbf{0}
\end{aligned}$$

where the distance is defined as

$$\tilde{d}_t^{ij} = \sum_{m=1}^{M} \tilde{d}_t^{ij,m},\ \tilde{d}_t^{ij,m} = \left(\mathbf{x}_t^{i,m} - \mathbf{x}_t^{j,m}\right)^{\top}\left(\mathbf{M}_0^{(m)} + \mathbf{M}_t^{(m)}\right)\left(\mathbf{x}_t^{i,m} - \mathbf{x}_t^{j,m}\right).$$

The variable δ_t^{ij} denotes the label of similar/dissimilar labeled pairs, and σ_t^{ij} is a predefined threshold for hinge loss. The parameters $\mathbf{b}_0$ and $\mathbf{b}_t$ represent weights for the sharing part and discriminating parts respectively, and the last term is the regularization on these weights.

Using this approach, the information contained in different tasks is shared among them and the multiple features are used in a more effective way. It uses the same strategy as mt-LMNN to construct the multi-task metric learning model and thus has the similar advantages/disadvantages.

Multi-task sparse compositional metric learning (mt-SCML) Shi et al. [34] proposes a multi-task metric learning framework from the perspective of sparse combination. The authors first propose a sparse compositional metric learning (SCML) approach which regards a Mahalanobis matrix as a nonnegative weighted sum of K rank-1 positive semi-definite matrices:

$$\mathbf{M} = \sum_{i=1}^{K} w_i \mathbf{b}_i \mathbf{b}_i^{\top}, \text{ with } \mathbf{w} \geq 0, \tag{4}$$

where the $\mathbf{b}_i$'s are D-dimensional column vectors. Noting that the distance between any two points $(\mathbf{x}, \mathbf{y})$ determined by $\mathbf{M}$ is calculated by

$$d_{\mathbf{M}}^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^{\top} \mathbf{M} (\mathbf{x} - \mathbf{y}) = \sum_{i=1}^{K} w_i \left(\mathbf{b}_i^{\top} (\mathbf{x} - \mathbf{y}) \right)^2,$$

the vectors $\mathbf{b}_i$'s span the common low-dimensional subspace in which the metric is defined.

Using such a formulation, each rank-1 matrix is a basis and the metric can be reformulated as a sparse combination of these bases. Then the metric learning is a process of learning such weights, which is shown as

$$\min_{\mathbf{w}} \frac{1}{|C|} \sum_{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) \in C} L_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) + \beta \|\mathbf{w}\|_1,$$

where L defines the loss from side-information as

$$L_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = \left[1 + d_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j) - d_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_k) \right]_+$$

with $[\cdot]_+ = \max(0, \cdot)$, and the ℓ_1 regularization encourages a sparse solution of $\mathbf{w}$.

When there are T tasks to be learned together, the multi-task learning can be easily obtained by applying a structure regularization on these weights. To be specific, the authors assume that the different tasks share a common low-dimensional subspace for the reconstruction weights, and use a mixed norm to obtain the structure sparsity. The formulation of mt-SCML is shown as

$$\min_{\mathbf{W}} \sum_{t=1}^{T} \frac{1}{|C_t|} \sum_{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) \in C_t} L_{\mathbf{w}_t}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) + \beta \|\mathbf{W}\|_{2,1},$$

where $\mathbf{W}$ is a $T \times K$ nonnegative matrix where each row $\mathbf{w}_t$ defines the reconstruct weight vector for the t-th task, and $\|\mathbf{W}\|_{2,1}$ is the ℓ_2/ℓ_1 mixed norm. It equals to the ℓ_1 norm applied to the ℓ_2 norm of the columns of $\mathbf{W}$, which induces the group sparsity at the column level, i.e., it encourages some columns to be zero together and thus make the different tasks share the same reconstruction bases.

This method naturally introduces the idea of group sparse to construct multi-task metric learning, and the proposed approach is not difficult to be realized. However, this algorithm requires the set of rank-one metrics to be pre-trained and thus cannot be optimized simultaneously with the weights.

Two-level multi-task metric learning (TMTL) Liu, et al. [35] proposes a two-level multi-task metric learning approach that combines multiple metrics directly without an explicit optimization procedure. It is developed based on KISSME [36], which is a metric learning approach motivated by a statistical inference and defines the Mahalanobis matrix as

$$\mathbf{M} = \Sigma_S^{-1} - \Sigma_D^{-1}.$$

This model is extended to two-level multi-task learning paradigm in a rather simple way. The authors first learn a Mahalanobis matrix for each task respectively and a common metric for all samples. Then the final individual Mahalanobis matrix is given by a direct weighted composition

$$\hat{\mathbf{M}}_t^k = \mathbf{M}_0^k + \mu \mathbf{M}_t^k.$$

This method is so simple that no optimization procedure is needed. To be strict, it is not a typical metric learning and can deal with only some special problems.

Online semi-supervised multi-task distance metric learning (online-SMDM) Li et al. [37] proposes a semi-supervised metric learning approach that is capable to utilize the unlabeled data to learn the metric. The method is designed based on the regularized distance metric learning [38] and extended to a multi-task metric learning model called online semi-supervised multi-task distance metric learning. It assumes each Mahalanobis matrix to be composed of a common part $\mathbf{M}_0$ and a task-specific part $\mathbf{M}_t$ as [31] does, and proposes an online algorithm to solve the model effectively.

To utilize the unlabeled data during training process, the authors assign labels for the unlabeled pairs:

$$y_{ij} = \begin{cases} 1, \text{ if } \mathbf{x}_i \in N(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N(\mathbf{x}_i); \\ 0, \text{ otherwise}, \end{cases} \tag{5}$$

where $N(\mathbf{x}_i)$ indicates the nearest neighbor set of $\mathbf{x}_i$ calculated by Euclidean distance. The Eq. (5) indeed assumes that if a point is one of the nearest neighbors of the other point, they should have the same label. Then the model of unlabeled model can be formulated as

$$\begin{aligned} \min_{\mathbf{M}} \sum_{t=1}^{T} \Bigg\{ & \frac{2}{N_{tl}(N_{tl}-1)} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D_{tl}} g_l\left(y_{ij}\left[1 - \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{M}_t+\mathbf{M}_0}\right]\right) \\ & + \frac{2\beta}{N_{tu}(N_{tu}-1)} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D_{tu}} g_u\left(y_{ij}\left[1 - \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{M}_t+\mathbf{M}_0}\right]\right) \\ & + \frac{\lambda}{2}\|\mathbf{M}_t\|_{\mathrm{F}}^2 \Bigg\} + \gamma T \|\mathbf{M}_0\|_{\mathrm{F}}^2, \\ \text{s.t. } & \mathbf{M} \succeq \mathbf{0}, \end{aligned}$$

where D_{tl} and D_{tu} represent the sets of labeled data pairs and unlabeled data pairs respectively, N_{tl} and N_{tu} are the numbers of the labeled and unlabeled training data, λ and γ

are both hyper-parameters to control the regularization on the individual parts and the common part, and **M** represents all the $\mathbf{M}_t$'s and $\mathbf{M}_0$ for brevity.

This method utilizes the unlabeled data by assigning labels for them according to the original distances. The strategy of constructing the multi-task learning is same as the previous ones.

Hierarchical multi-task metric learning (HMML) Zheng et al. [39] proposes an approach to learn a hierarchical tree of multiple sparse metrics hierarchically over a visual tree. In this work, a visual tree is first constructed to organize the categories in a coarse-to-fine fashion. Then a top-down approach is used to learn multiple metrics along the visual tree, where the model is expected to benefit from leveraging both the inter-node visual correlation and the inter-level visual correlations.

Construction of the visual tree is composed of two key steps: (a) Active Sampling for Category Representation, which utilizes active sampling to find multiple representative samples for each image category. (b) Hierarchical Affinity Propagation Clustering for Node Partitioning, which is a top-down approach to hierarchical affinity propagation (AP) clustering. It starts from the root node containing all the image categories and ends at the leaf nodes containing only one single image category. Figure 2 gives an example

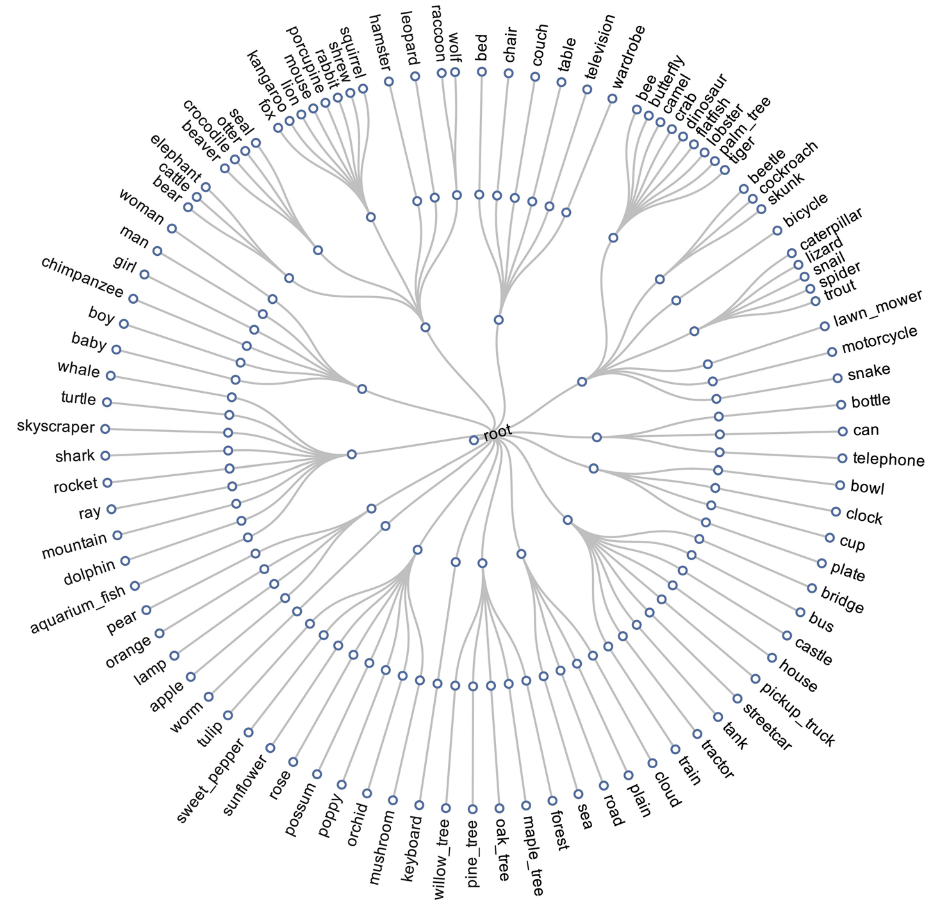

Fig. 2 An example of enhanced visual tree. The visual tree is constructed on the CIFAR-100 image set with 100 categories and its depth is 4. This figure is from the original paper [39]

of the enhanced visual tree for CIFAR-100. In this tree, the categories are organized in a hierarchical structure according to their similarities.

According to the construction procedure of the visual tree, categories on the same branch are more similar to each other than the ones on other branches. Thus, it is reasonable to perform multi-task metric learning over the sibling child nodes under the same parent node to utilize the inter-node visual correlation among them. The authors exploit the same strategy as mtLMNN [31] which decomposes the metric into a common part and an individual part as

$$d_t(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{M}_0 + \mathbf{M}_t)(\mathbf{x}_i - \mathbf{x}_j)},$$

where $\mathbf{M}_0$ defines the common metric shared among all sibling child nodes and $\mathbf{M}_t$ defines the node-specific metric.

For root node, the joint objective function is then defined as

$$\begin{aligned} \min_{\mathbf{M}_0,\ldots,\mathbf{M}_T} & \gamma_0 \|\mathbf{M}_0 - \mathbf{I}\|_{\mathrm{F}}^2 + \sum_{t=1}^{T} \alpha_t \mathrm{tr}[\,\mathbf{M}_0 + \mathbf{M}_t] \\ & + \sum_{t=1}^{T} \left[\gamma_t \|\mathbf{M}_t\|_{\mathrm{F}}^2 + \sum_{i,j} d_t^2(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i,j,k} \xi_{i,j,k} \right], \\ \text{s.t. } & d_t^2(\mathbf{x}_i, \mathbf{x}_k) - d_t^2(\mathbf{x}_i, \mathbf{x}_j) \geq 1 - \xi_{i,j,k}, \\ & \xi_{i,j,k} \geq 0, \\ & \mathbf{M}, \mathbf{M}_1, \ldots, \mathbf{M}_T \succeq \mathbf{0}. \end{aligned} \tag{6}$$

where the parameters γ_0 and γ_t's control the regularization on the common part and individual part respectively.

For non-root nodes at the mid-level of the visual tree, besides the inter-node correlations, the inter-level visual correlations between the parent node and its sibling child nodes at the next level should be also exploited. Since all nodes on the same branch are similar, any node p characterizes the common visual properties of its sibling child nodes. On the other hand, the task-specific metric $\mathbf{M}_p$ for node p contains the task-specific composition. Thus, it is reasonable to utilize the task-specific metric of node p to help the learning of its sibling child nodes. Based on this idea, the regularization $\beta\|\mathbf{M}_0 - \mathbf{M}_p\|^2$ is added into the objective of (6) for non-root nodes, where $\mathbf{M}_0$ is the common metric shared among the sibling child nodes under parent node p and $\mathbf{M}_p$ is the task-specific metric for node p at the upper level.

This method introduces the hierarchical visual tree into multi-task metric learning, which is used to guide the multi-task learning and thus provides a more powerful capability of describing the relationship among tasks.

Task relationship learning and regularization

Transfer metric learning by learning task relationship (TML) Zhang et al. [40, 41] proposes a multi-task metric learning by learning task relationship. This model is also a direct adaptation of a traditional multi-task learning approach to the metric learning task. The authors proposes a multi-task relationship learning (MTRL) [13] in their previous work which assumes all the parameter vectors to follow a matrix-variant normal distribution [42] and automatically learns the relationships between tasks by a regularization. Since the parameter to be learned in metric learning is a matrix rather than a vector, the authors concatenate all columns of the Mahalanobis matrix to form a vector for each task $\tilde{\mathbf{M}}_t = \text{vec}(\mathbf{M}_t)$ and then apply the regularization of MTRL to it: $\tilde{\mathbf{M}}\Omega^{-1}\tilde{\mathbf{M}}^\top$ where $\tilde{\mathbf{M}} = [\text{vec}(\mathbf{M}_1), \ldots, \text{vec}(\mathbf{M}_T)]$. It is equivalent to apply the following matrix-variant normal prior distribution to $\tilde{\mathbf{M}}_t$'s.

$$q(\tilde{\mathbf{M}}) = \mathcal{MN}_{d^2 \times T}(\tilde{\mathbf{M}}|\mathbf{0}_{d^2 \times T}, \mathbf{I}_{d^2} \otimes \Omega)$$

In this definition, the row covariance matrix $\mathbf{I}_{d^2}$ models the relationships between features and the column covariance matrix Ω models the relationships between different vectorized Mahalanobis matrices $\tilde{\mathbf{M}}$'s. Thus, Ω indeed determines the relationships between tasks. Since it cannot be given a priori in most cases, the authors propose to estimate it from data automatically.

The obtained model is shown in (7) and can be solved by alternating optimization.

$$\begin{aligned}
&\min_{\{\mathbf{M}_t\},\Omega} \sum_{t=1}^{T} \frac{2}{n_t(n_t-1)} \sum_{i<j} g\left(y_{i,j}^t \left[1 - \left\|\mathbf{x}_i^t - \mathbf{x}_j^t\right\|_{\Omega_t}^2\right]\right) \\
&\qquad + \frac{\lambda_1}{2} \sum_{t=1}^{T} \|\mathbf{M}_t\|_{\text{F}}^2 + \frac{\lambda_2}{2} \text{tr}(\tilde{\mathbf{M}}\Omega^{-1}\tilde{\mathbf{M}}^\top) \\
&\quad \text{s.t. } \mathbf{M}_t \succeq \mathbf{0}, \forall t \\
&\qquad \tilde{\mathbf{M}} = (\text{vec}(\mathbf{M}_1), \ldots, \text{vec}(\mathbf{M}_T)) \\
&\qquad \Omega \succeq \mathbf{0}, \ \text{tr}(\Omega) = 1.
\end{aligned} \tag{7}$$

In that paper, the authors further propose a transfer metric learning based on this model by training the concatenated Mahalanobis matrix of only target task while leaving other matrices fixed as source tasks. The idea of learning the relationship between tasks is interesting, but the covariance between the vectorized Mahalanobis matrices does not explain well from the perspective of distance metric.

Multi-task maximally collapsing metric learning (MtMCML) Ma et al. [43] proposes a multi-task metric learning approach using the graph-based regularization. To be specific, a graph is constructed to describe the relations between the tasks, where each node corresponds to a Mahalanobis matrix of one task, and an edge connecting two nodes represents the similarity between the two tasks. Thus an adjacency matrix $\mathbf{W}(0 \leq \mathbf{W}(i,j) \leq 1)$ is obtained where a higher $\mathbf{W}(i,j)$ indicates that metrics i and j are more related. The regularization is

$$\begin{aligned}
J(\mathbf{M}_1, \ldots, \mathbf{M}_T) =& \sum_{i=1}^{T} \sum_{j=1}^{T} \mathbf{W}(i,j) \|\tilde{\mathbf{M}}_i - \tilde{\mathbf{M}}_j\|_2^2 \\
=&\text{trace}\left(\tilde{\mathbf{M}}(\mathbf{DIA} - \mathbf{W})\tilde{\mathbf{M}}^\top\right) \\
=&\text{trace}\left(\tilde{\mathbf{M}}\mathbf{L}\tilde{\mathbf{M}}^\top\right),
\end{aligned} \tag{8}$$

where $\tilde{\mathbf{M}}_i = \text{vec}(\mathbf{M}_i)$ converts the Mahalanobis matrix of the i-task into a vector in a column-wise manner, $\mathbf{DIA}$ is a diagonal matrix where $\mathbf{DIA}(i,i) = \sum_{j=1}^{T} \mathbf{W}(i,j)$, and thus the matrix $\mathbf{L} = \mathbf{DIA} - \mathbf{W}$ indeed defines the graph Laplacian matrix. The model can be optimized by alternating method.

In this work, the authors empirically set the adjacency matrix as $\mathbf{W}(i,j) = 1$, which indeed defines every pair of tasks are related. It is not difficult to prove that such a regularization is just a variant of the regularization of mt-LMNN. Therefore, these two methods are closely related in this special case.

This work naturally introduces the graph regularization into multi-task learning by applying a Laplacian to the vectorized Mahalanobis matrices. However, the relationship between two metrics is still vague, and the Laplacian matrix $\mathbf{L}$ is not easy to be reasonably determined.

Regularization with a common metric

A framework for approaches based on common metric Yang et al. [44] proposes a general framework for multi-task metric learning to solve the scenario that all metrics are similar to a common one. The optimization problem is shown in (9) where $\mathbf{M}_t$ is the Mahalanobis matrix of the t-th task where $\mathbf{M}_0$ is the common one.

$$\begin{aligned} \min_{\{\mathbf{M}_t\},\mathbf{M}_c} & \sum_t \left(L(\mathbf{M}_t, \mathcal{X}_t) + \gamma D(\mathbf{M}_t, \mathbf{M}_c)\right) + \gamma_0 D(\mathbf{M}_0, \mathbf{M}_c) \\ \text{s.t. } & \mathbf{M}_t \in \mathcal{C}_t(\mathcal{X}_t), \\ & \mathbf{M}_t \succeq \mathbf{0}, \end{aligned} \tag{9}$$

In this framework, the loss L and constraints $\mathcal{C}_t$ are used to incorporate the side-information from training samples into the learning process, while the regularization $D(\mathbf{M}_t, \mathbf{M}_c)$ encourages the metric of each task to be similar to a common one $\mathbf{M}_c$, and $D(\mathbf{M}_0, \mathbf{M}_c)$ further regularizes the common metric to be close to a pre-defined metric. Without more prior information available, $\mathbf{M}_0$ is set to the identity $\mathbf{I}$ to define a Euclidean metric.

The mt-LMNN can be easily included as a special case of this framework by $D(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|_F^2$. The only difference exists on the constraints: the Mahalanobis matrix of the t-th task in mt-LMNN is $\mathbf{M}_0 + \mathbf{M}_t$, where both the two parts are positive semi-definite; the Mahalanobis matrix of the t-th task in (9) with Frobenius norm is $\mathbf{M}_t$ and the positive semi-definiteness of only this matrix is required. The authors indicate that the latter actually provides a more reasonable model because the constraints in mt-LMNN is unnecessary to be so strict.

Geometry preserving multi-task metric learning (GPmtML) Yang et al. [44] proposes the geometry preserving multi-task metric learning approach based on the general framework (9). Different from most previous approaches, the GPmtML considers the multi-task metric learning problem from the perspective of propagating the relative distance. The authors indicate that learning of a metric is a process of embedding the supervised information from training samples into the learnt metric, and thus it is reasonable to couple the multiple tasks by sharing the embedded supervised information among metrics. As we have illustrated, it is an important class of metric learning approaches which learn the metric from relative distances given by triplets, and thus it is reasonable to propagate such relationships about the relative distance between metrics. Motivated by this, the authors propose the concept of *geometry preserving probabilistic* [44, 45] to measure such kind of characteristic between two metrics defined by Mahalanobis matrices **A** and **B**.

$$\begin{aligned}\mathrm{PG}_f(d_{\mathbf{A}}, d_{\mathbf{B}}) =& \mathrm{P}\left[d_{\mathbf{A}}(\mathbf{x}_1, \mathbf{y}_1) > d_{\mathbf{A}}(\mathbf{x}_2, \mathbf{y}_2) \ \wedge \ d_{\mathbf{B}}(\mathbf{x}_1, \mathbf{y}_1) > d_{\mathbf{B}}(\mathbf{x}_2, \mathbf{y}_2)\right] \\ &+ \mathrm{P}\left[d_{\mathbf{A}}(\mathbf{x}_1, \mathbf{y}_1) < d_{\mathbf{A}}(\mathbf{x}_2, \mathbf{y}_2) \ \wedge \ d_{\mathbf{B}}(\mathbf{x}_1, \mathbf{y}_1) < d_{\mathbf{B}}(\mathbf{x}_2, \mathbf{y}_2)\right] \\ &+ \mathrm{P}\left[d_{\mathbf{A}}(\mathbf{x}_1, \mathbf{y}_1) = d_{\mathbf{A}}(\mathbf{x}_2, \mathbf{y}_2) \ \wedge \ d_{\mathbf{B}}(\mathbf{x}_1, \mathbf{y}_1) = d_{\mathbf{B}}(\mathbf{x}_2, \mathbf{y}_2)\right],\end{aligned}$$

where $(\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2) \sim f$ and $\wedge$ denotes the logical *"and"* operator.

Then the geometry preserving multi-task metric learning is proposed which aims to increase the geometry preserving probabilistic. The method is obtained by using the von Neumann divergence [46, 47] (10) as regularization in (9).

$$D_{\mathrm{vN}}(\mathbf{M}, \mathbf{M}_c) = \mathrm{tr}\left(\mathbf{M} \log \mathbf{M} - \mathbf{M} \log \mathbf{M}_c - \mathbf{M} + \mathbf{M}_c\right) \tag{10}$$

By a series of theoretical analysis, this method is proved to be able to encourage a higher geometry preserving geometry, and thus more likely to keep the relative distances of samples across different metrics. The introduced regularization is jointly convex and thus the problem can be effectively solved by alternating optimization.

This is the first paper that attempts to construct the multi-task metric learning by sharing the supervised side-information among tasks. It provides a reasonable explanation from the perspective of metric learning. However, the macrostructure of the model is too simple and thus cannot describe more complex relationship among tasks.

Sharing transformation

According to (1). Learning a Mahalanobis distance is equivalent to learning a corresponding linear transformation. There are indeed some metric learning algorithms that aim to learn such a transformation directly, and it naturally provides a way to construct a multi-task metric learning by sharing some parts of transformation.

Multi-task metric learning based on common subspace (mtMLCS) Yang et.al [48] proposes a multi-task learning method based on the assumption of common subspace. The idea is motivated by multi-task feature learning [11] which learns a common sparse representations across multiple tasks. Based on the same assumption that all the tasks share a common low-dimensional subspace, the authors propose a multi-task framework for metric learning by transformation.

To couple multiple tasks with a common low-dimensional subspace, the authors notice that for any low-rank Mahalanobis matrix **M**, the corresponding linear transformation matrix **L** is of full row rank and has the size of $r \times d$, where $r = \mathrm{rank}(\mathbf{M})$ is the dimension of the subspace. Applying a compact SVD to **L**, there is $\mathbf{L} = \mathbf{U}\Lambda\mathbf{V}^\top$ where **V** is a $d \times r$ matrix defining a projection to the low-dimensional subspace, and $\mathbf{U}\Lambda$ defines a transformation in the subspace. This fact motivates a straightforward multi-task strategy with common subspace: to share a common projection matrix **V** and learn an individual transformation $\mathbf{R}_t \doteq \mathbf{U}_t\Lambda_t$ for each task.

However, it is computationally complex to apply an orthogonal constraint to **V**. On the other hand, it's notable that the orthogonality is not necessary for **V** to define a subspace. As well as **V** is of the size $r \times d$ and $r < d$, it indeed defines a subspace of dimensionality no more than r with some extra full-rank transformation in the subspace. Therefore, a common matrix **L** of size $r \times d$ is used to realize the common projection instead of $\mathbf{V}^\top$, and the extra transformation can be absorbed by $\mathbf{R}_t$. The obtained model for multi-task metric learning is to a transformation for each task $\mathbf{L}_t = \mathbf{R}_t\mathbf{L}_0$ where $\mathbf{L}_0$ defines the common subspace and $\mathbf{R}_t$ defines the task-specific metric. This strategy is then incorporated into the LMCA [49] which is a variant of LMNN [2] by learning the transformation.

This approach is simple to implement. Compared with the approaches that learn metrics by learning Mahalanobis matrices, mtMLCS does not require the symmetric positive-definite constraints, and thus is much easier to optimize. However, this model is not convex and thus the global optimum cannot be obtained.

Coupled projection multi-task metric learning (CP-mtML) Bhattarai et al. [50] proposes a multi-task metric learning approach which also focuses on the methods that learns a linear transformation. In this paper, the authors refer the transformation in (1) as "projection", and the idea to couple different tasks is to decompose it into a common projection and a task-specific projection. Different from mtMLCS in which the common projection and task-specific projection are concatenated, CP-mtML decomposes the projection in the manner of distance:

$$\begin{aligned} d_t^2(\mathbf{x}_i, \mathbf{x}_j) =& d_{\mathbf{L}_0}^2(\mathbf{x}_i, \mathbf{x}_j) + d_{\mathbf{L}_t}^2(\mathbf{x}_i, \mathbf{x}_j) \\ =& \|\mathbf{L}_0\mathbf{x}_i - \mathbf{L}_0\mathbf{x}_j\|_2^2 + \|\mathbf{L}_t\mathbf{x}_i - \mathbf{L}_t\mathbf{x}_j\|_2^2 \end{aligned}$$

It is easy to show that the relation among different tasks is the same as mt-LMNN where both of them obtain the distance by summing the squared distances of common and task-specific parts:

$$\begin{aligned} d_t^2(\mathbf{x}_i, \mathbf{x}_j) =& (\mathbf{x}_i - \mathbf{x}_j)^\top \left(\mathbf{L}_0^\top\mathbf{L}_0 + \mathbf{L}_t^\top\mathbf{L}_t\right)(\mathbf{x}_i - \mathbf{x}_j) \\ =& (\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{M}_0 + \mathbf{M}_t)(\mathbf{x}_i - \mathbf{x}_j) \end{aligned}$$

The authors pointed out that there are important differences between the two approaches. First, the side-information of mt-LMNN is based on triplets while CP-mtML is based on similar/dissimilar pairs. Second, using the formulation of projection, it is easy to obtain a low-rank metric. Third, the authors propose a scalable SGD based learning algorithm. Finally, it can work in online setting.

Since this method learns the metric by optimizing on the transformation **L**, it has the similar merits and faults as mtMLCS. It is also designed for the simple case where the tasks are correlated by a common Mahalanobis matrix.

Deep multi-task metric learning (DMML) Soleimani et al. [51] proposes a multi-task learning version of deep metric learning. The method is constructed based on the discriminative deep metric learning (DDML) [29]. For any pair of points, the DDML transforms the two points with a neural network, and then the distance is defined to be the Euclidean distance of their transformations. Thus the process of metric learning is done by learning the parameters of the network.

The DMML uses a straightforward way to construct a multi-task version of DDML by sharing the same first layer. Assuming there are T tasks, the outputs for two points $\mathbf{x}_{i,t}, \mathbf{x}_{j,t}$ in the t-th task are $\mathbf{h}_{1,t}^{(1)} = s\left(\mathbf{W}^{(1)}\mathbf{x}_{i,t} + \mathbf{b}^{(1)}\right)$ and $\mathbf{h}_{2,t}^{(1)} = s\left(\mathbf{W}^{(1)}\mathbf{x}_{j,t} + \mathbf{b}^{(1)}\right)$, where all tasks share a common weights matrix $\mathbf{W}^{(1)}$ and a common bias vector $\mathbf{b}^{(1)}$, and s is a nonlinear operator such as tanh. Then the outputs the second layer is calculated separately for each task as $h_{1,t}^{(2)} = s\left(\mathbf{W}_t^{(2)}\mathbf{h}_{1,t}^{(1)} + \mathbf{b}_t^{(2)}\right)$ and $h_{2,t}^{(2)} = s\left(\mathbf{W}_t^{(2)}\mathbf{h}_{2,t}^{(1)} + \mathbf{b}_t^{(2)}\right)$, where each task use the task-specific weights matrix $\mathbf{W}_t^{(2)}$ and bias vector $\mathbf{b}_t^{(2)}$, and s is the non-linear operator again. The obtained distance now can be calculated by

$$d^2(\mathbf{x}_{i,t}, \mathbf{x}_{j,t}) = \|\mathbf{h}_{1,t}^{(2)} - \mathbf{h}_{1,t}^{(2)}\|_2^2.$$

Then the model is learned by the following optimization problem:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{b}} J =& \frac{1}{2}\sum_{t=1}^{T}\sum_{i,j} g\left(1 - l_{i,j}(\tau - d^2(\mathbf{x}_{i,t}, \mathbf{x}_{j,t}))\right) \\ &+ \frac{\lambda}{2}\left(\|\mathbf{W}^{(1)}\|_\mathrm{F}^2 + \|\mathbf{b}^{(1)}\|_2^2\right) + \frac{\lambda}{2}\sum_{t=1}^{T}\left(\|\mathbf{W}_t^{(2)}\|_\mathrm{F}^2 + \|\mathbf{b}_t^{(2)}\|_2^2\right),\end{aligned}$$

where $g(z) = \frac{1}{\beta}\log(1 + \exp(\beta z))$ is the smoothed approximation for $[z]_+ = \max(z, 0)$, and β controls its sharpness.

This method is based on a simple yet effective idea which a part of the network weights are shared across multiple tasks. It is not difficult to implement by slightly modify the original network architecture. However, only the first layer is shared across different tasks in this model, which may be not the optimal choice and it is not easy to determine how many layers should be shared.

Deep convernets metric learning with multi-task learning (mtDCML) McLaughlin et al. [52] proposes to introduce auxiliary tasks in the model to help the metric learning task. The central idea is also to learn the distance metric by learning a feature representation. Denoting the subnetwork to transform the sample to the feature representation as G, and the network parameters as w, the learned distance can be calculated by the Euclidean distance of their representations as

$$D(\mathbf{x}_1, \mathbf{x}_2; w) = \|G(\mathbf{x}_1; w) - G(\mathbf{x}_2; w)\|_2.$$

The network is trained using sample pairs in the training dataset. The cost for similar/dissimilar pairs are shown below:

$$\mathcal{V}_S(\mathbf{x}_1, \mathbf{x}_2; w) = \frac{1}{2}D(\mathbf{x}_1, \mathbf{x}_2; w)^2,\ \mathcal{V}_D(\mathbf{x}_1, \mathbf{x}_2; w) = \frac{1}{2}\left(\max(0, m - D(\mathbf{x}_1, \mathbf{x}_2; w))\right)^2.$$

Then the cost function is written as

$$\mathcal{V}(\mathbf{x}_1, \mathbf{x}_2|y; w) = (1 - y)\mathcal{V}_S(\mathbf{x}_1, \mathbf{x}_2; w) + y\mathcal{V}_D(\mathbf{x}_1, \mathbf{x}_2; w).$$

To improve the metric learning task, the authors include other related auxiliary tasks into the objective and obtain the multi-task version:

$$\mathcal{C}_m(X) = \sum_{(\mathbf{x}_1^i, \mathbf{x}_2^i)\in X} \mathcal{V}\left(\mathbf{x}_1^i, \mathbf{x}_2^i|y^i; w\right) + \sum_t \alpha_t \mathcal{T}_t\left(G\left(\mathbf{x}_1^i\right)|l_1^{i,t}; w\right) + \sum_t \alpha_t \mathcal{T}_t\left(G\left(\mathbf{x}_2^i\right)|l_2^{i,t}; w\right),$$

where $\mathcal{T}_t$ is an auxiliary task which helps to learn a better representation.

The selection of the auxiliary task depends on the problem of interest and there are a variety of choices. For the example in [52] where the main task is a metric learning for face verification, all the auxiliary tasks involve assigning one of several mutually exclusive labels to each training image. Thus the following softmax regression cost function is used

$$\mathcal{T}_t\left(\mathbf{z}|l^t; \mathbf{w}\right) = \sum_{j\in L_t} \mathbf{1}\left\{l^t = j\right\}\log\frac{e^{\mathbf{w}_j^\top \mathbf{z}}}{\sum_{q\in L_t} e^{\mathbf{w}_q^\top \mathbf{z}}},$$

where $\mathbf{z}$ is the feature representation of the input image $G\left(\mathbf{x}_l^i\right)$ or $G\left(\mathbf{x}_2^i\right)$, L_t is the label set for the t-th task, and $\mathbf{1}\{l^t = j\}$ is an indicator function that takes value one when j is equal to the ground truth l^t and zero otherwise. Using this framework, several auxiliary tasks can be included by using different label set L_t, such as identification, attributes, pose, etc. Please refer to [52] for more details.

The strategy to construct the multi-task metric learning used in this paper is common in the community of multi-task learning. It is a flexible model by using different auxiliary tasks. However, for some task, it is difficult to choose a proper auxiliary task, and a bad auxiliary task may induce deterioration of the performance.

Applications

Multi-task metric learning has been widely used in a variety of practical applications, and we would like to introduce some representative works in this section.

Semantic categorization and social tagging with knowledge transfer among tasks

Wang et al. [32] uses their proposed multi-task multi-feature similarity learning to solve the large scale visual applications. The metrics for visual categorization and automatic tagging are learned jointly based on the framework, which benefits from several perspectives. First, M^{2}SL learns a metric for each feature instead of concatenating the multiple features into one feature. This effectively reduces the computation complexity growth from $O\left(M^2d^2\right)$ to $O\left(Md^2\right)$ and also the risk of over-fitting. Second, the multi-task framework is more flexibility to explore the intrinsic model sharing and feature weighting relations on image data with large amount of classes. Third, the knowledge is transferred among semantic labels and social tagging information by the model. This combines the information fusion from both sides for effective image understanding.

The authors compare the performances of two versions of M^{2}SL (linear and kernelized) with some other methods and the experimental results are shown in Fig. 3. From the

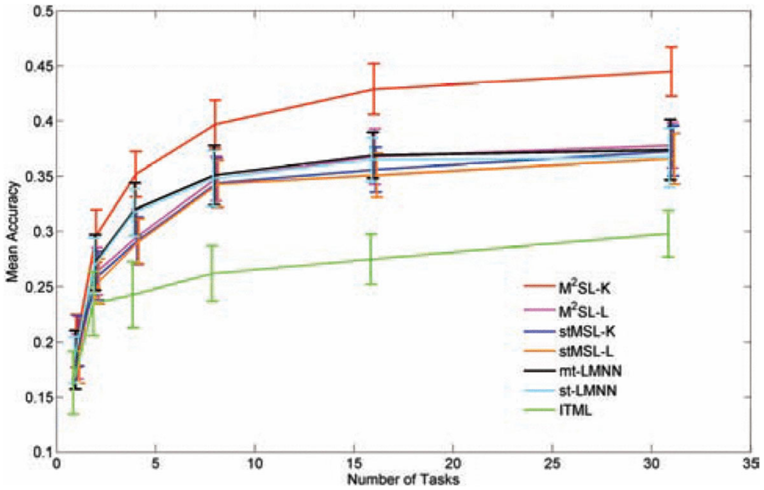

Fig. 3 Comparison of M^{2}SL with other methods. The performance curves of M^{2}SL and other methods on ImageNet-250 dataset are shown, where M^{2}SL-K and M^{2}SL-L indicate the kernelized and linear M^{2}SL respectively. The *x*-axis represents the number of tasks while *y*-axis the mean accuracy of all tasks. This figure is from the original paper [32]

results, the kernelized M^{2}SL always achieves the best performance, especially when the number of tasks are greater. For the linear M^{2}SL, it also outperforms the single-task MSL. Thus, the knowledge transfer by multi-task learning effectively improves the performance of metric learning.

Person re-identification over camera networks Ma et al. [43] uses their proposed multi-task maximally collapsing metric learning to solve the person re-identification over camera networks. Person re-identification in a camera network is a challenging problem because the data are collected from different cameras. The method to use a common metric overlooks the differences between cameras, and thus the authors propose to use a multi-task learning approach for this problem. With the MtMCML, an particular metric is learned for each pair of cameras, while the common information can be shared among them. The experimental results show that the multi-task approach works substantially better than other state-of-the-art methods as shown in Fig. 4.

Large-scale face retrieval Bhattarai et al. [50] uses their proposed coupled projection multi-task metric learning to solve the large-scale face retrieval. They use the multi-task framework to learn different tasks on heterogeneous datasets simultaneously, where a common projection is used to share information among these tasks. The tasks include face identity, age recognition, and expression recognition. By jointly learning these tasks, the authors get an improved performance as shown in Fig. 5.

Offline signature verification Soleimani et al. [51] aims to deal with the offline signature verification problem using the deep multi-task metric learning. For offline signature verification, there are writer-dependent (WD) approaches and writer-independent (WI) approaches. These two approaches benefits from their particular advantages respectively. These two approaches are well integrated in this model where the shared layer acts as a WI approach while the separated layers learn WD factors. In the experiments, the DMML

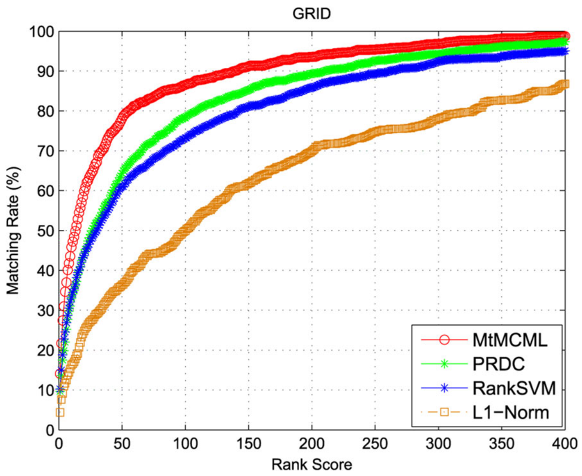

Fig. 4 Comparison of MtMCML with other methods. The performance of MtMCML and other methods on GRID datasets are presented, where the *x*-axis and *y*-axis represent the rank score and matching rate respectively. From the results, the multi-task learning approach evidently improves the performance of matching rate. This figure is from the original paper [43]

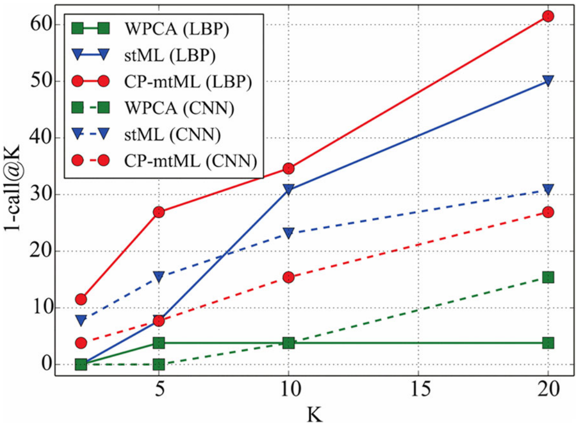

Fig. 5 Comparison of CP-mtML with other methods. Age retrieval performance (1-call@*K*) for different *K* with auxiliary task of identity matching. This figure is from the original paper [50]

achieves better performance than other methods. For example, on the UTSig dataset and using the HOG feature, the DMML achieves equal error rate (ERR) of 17.45% while the SVM achieves ERR of 20.63%; using the DRT feature, the DMML achieves ERR of 20.28% while the SVM achieves ERR of 27.64%.

Hierarchical large-scale image classification Zheng et al. [39] uses their proposed hierarchical multi-task metric learning to solve the large-scale image classification problem. To deal with the large-scale problem, the authors first learn a visual tree to organize large number of image categories hierarchically in a coarse-to-fine fashion. Then a series metrics are learnt hierarchically. Using the HMML, both the inter-node visual correlations and the inter-level visual correlations are utilized. The inter-node correlation is obtained directly from the multi-task framework, while the inter-level correlation is obtained by passing the task-specific part into the next level. The experimental results shown in Fig. 6 demonstrate that the multi-task model obtain better performance on large-scale classification.

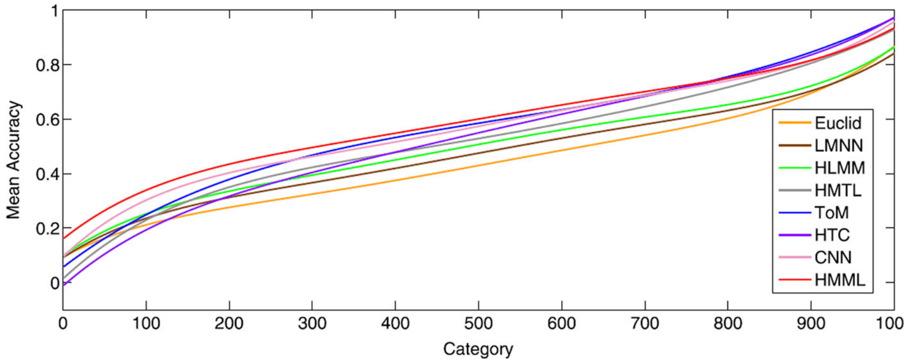

Fig. 6 Comparison of HMML with other methods. Accuracy comparison on the ILSVRC-2012 image set with 1000 image categories. This figure is from the original paper [39]

Person re-identification with auxiliary tasks McLaughlin et al. [52] uses the multi-task learning to improve the performance of person re-identification. Using their proposed deep convernets metric learning with multi-task learning, the authors train the network to jointly perform verification and identification and to recognize attributes related to the clothing and pose of the person in each image. The main job of the network is to learn a metric using similar and dissimilar pairs. With the help of auxiliary tasks (attribute recognition), the network learn a metric to give a satisfactory performance. Figure 7 shows the experimental results. It is obvious that the accuracy is effectively improved by introducing auxiliary tasks.

Conclusion

In this paper, we have systematically reviewed multi-task metric learning. Following a brief overview of metric learning, various multi-task learning approaches are categorized into four families and introduced respectively. We then review the motivations, models, and algorithms of them, and also discuss and compare some closely related approaches. Finally some representative applications of multi-task metric learning are illustrated.

For future work, we suggest potential issues for exploration. First, the theoretical analysis of multi-task metric learning should be addressed. There has long been an important issue yielding multiple results [53–56], with most studies focusing on how multi-task learning improves the generalization [57] of a conventional algorithm. However, as mentioned earlier, the metric learning improves the performances of the algorithms who use the metric indirectly. This makes these results difficult for application to metric learning algorithms. There has also been some research [58–61] on the theoretical analysis of metric learning, however it has been to difficult to explain these in the context of multi-task learning, Whilst Yang et al. [44] has attempted to provide an intuitive explanation, the issue pertaining to multi-task learning remains unresolved. Second, how to avoid the negative transfer among tasks. Existing approaches are designed to couple multiple metrics without considering the problem of negative transfer, and thus it is likely to deteriorate the performances when the tasks are not related. Third, most existing multi-task metric learning approaches are designed for global linear metrics. Thus it should be extended to

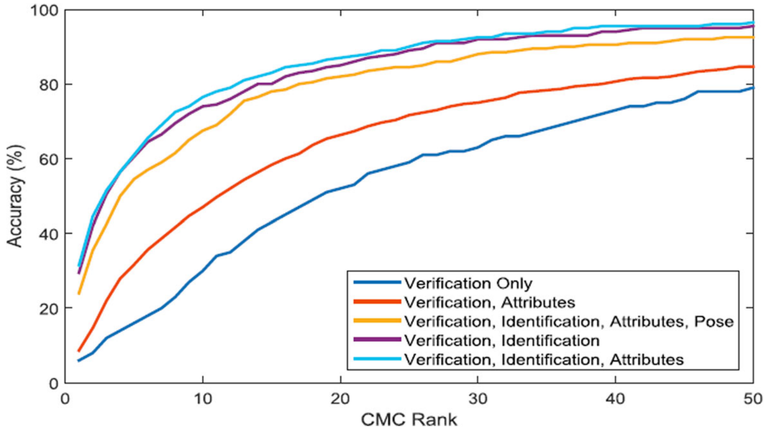

Fig. 7 Comparison of mtDCML with other methods. CMC curve after multitask training on VIPeR dataset. This figure is from the original paper [52]

more types of metric learning approaches, including local metric learning and non-linear metric learning. Finally, increased applications of multi-task metric learning are expected to be discovered.

Funding
The paper was partially supported by National Natural Science Foundation of China (NSFC) under grant no.61403388, no.61473236, Natural science fund for colleges and universities in Jiangsu Province under grant no. 17KJD520010, Suzhou Science and Technology Program under grant no. SYG201712, SZS201613, Key Program Special Fund in XJTLU (KSF-A-01), and UK Engineering and Physical Sciences Research Council (EPSRC) grant numbers EP/I009310/1, EP/M026981/1.

Availability of data and materials
Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Authors' contributions
PY carried out the whole structure of the idea and the mainly drafted the manuscript. KH provided the guidance of the whole manuscript and revised the draft. AH participated the discussion and gave valuable suggestion on the idea. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]National Laboratory of Pattern Recognition, 95 East Zhongguancun Road, 100190 Beijing, China. [2]Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, 215123 Suzhou, China. [3]University of Stirling, FK9 4LA Stirling, UK, Scotland.

Received: 10 August 2017 Accepted: 30 January 2018
Published online: 27 February 2018

References

1. Xing EP, Ng AY, Jordan MI, Russell SJ. Distance metric learning with application to clustering with side-information. In: Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]. 2002. p. 505–12. http://papers.nips.cc/paper/2164-distance-metric-learning-with-application-to-clustering-with-side-information.
2. Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res. 2009;10:207–44.
3. Davis JV, Kulis B, Jain P, Sra S, Dhillon IS. Information-theoretic metric learning. In: Proceedings of the 24th International Conference on Machine Learning. 2007. p. 209–16.
4. Huang K, Ying Y, Campbell C. Gsml: A unified framework for sparse metric learning. In: Ninth IEEE International Conference on Data Mining. 2009. p. 189–98.
5. Huang K, Ying Y, Campbell C. Generalized sparse metric learning with relative comparisons. Knowl Inf Syst. 2011;28(1):25–45.
6. Ying Y, Huang K, Campbell C. Sparse metric learning via smooth optimization. In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A, editors. Advances in Neural Information Processing Systems 22. 2009. p. 2214–222.
7. Ying Y, Li P. Distance metric learning with eigenvalue optimization. J Mach Learn Res. 2012;13:1–26.
8. Caruana R. Multitask learning. Mach Learn. 1997;28(1):41–75.
9. Evgeniou T, Pontil M. Regularized multi-task learning. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004. p. 109–17.
10. Argyriou A, Micchelli CA, Pontil M, Ying Y. A spectral regularization framework for multi-task structure learning. In: Advances in Neural Information Processing Systems 20. 2008. p. 25–32.
11. Argyriou A, Evgeniou T. Convex multi-task feature learning. Mach Learn. 2008;73(3):243–72.
12. Zhang J, Ghahramani Z, Yang Y. Flexible latent variable models for multi-task learning. Mach Learn. 2008;73(3): 221–42.
13. Zhang Y, Yeung DY. A convex formulation for learning task relationships in multi-task learning. In: Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence. 2010. p. 733–442.
14. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59.
15. Dai W, Yang Q, Xue GR, Yu Y. Boosting for transfer learning. In: Proceedings of the 24th International Conference on Machine Learning, ICML '07. New York: ACM; 2007. p. 193–200.
16. Gopalan R, Li R, Chellappa R. Domain adaptation for object recognition: An unsupervised approach. In: Proceedings of IEEE International Conference on Computer Vision, ICCV 2011. p. 999–1006.

17. Vilalta R, Drissi Y. A perspective view and survey of meta-learning. Artif Intell Rev. 2002;18(2):77–95.
18. Thrun S. Lifelong learning algorithms. In: Learning to Learn. USA: Springer; 1998. p. 181–209.
19. Thrun S, Pratt L. Learning to Learn. USA: Springer; 2012.
20. Burago D, Burago Y, Ivanov S. A Course in Metric Geometry. USA: American Mathematical Society; 2001. Chap. Ch 1.1.
21. Mahalanobis PC. On the generalised distance in statistics. In: Proceedings National Institute of Science, vol. 2. India; 1936. p. 49–55.
22. Bellet A, Habrard A, Sebban M. A survey on metric learning for feature vectors and structured data. arXiv preprint arXiv:1306.6709v4, 2014.
23. Kulis B. Metric learning: A survey. Found Trends Mach Learn. 2013;5(4):287–364.
24. Weinberger KQ, Blitzer J, Saul L. Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems 18. 2006.
25. Huang K, Jin R, Xu Z, Liu CL. Robust metric learning by smooth optimization. In: The 26th Conference on Uncertainty in Artificial Intelligence. 2010. p. 244–51.
26. Goldberger J, Roweis S, Hinton G, Salakhutdinov R. Neighbourhood components analysis. In: Advances in Neural Information Processing Systems. 2004. p. 513–20.
27. Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw. 2015;61:85–117.
28. Salakhutdinov R, Hinton G. Learning a nonlinear embedding by preserving class neighbourhood structure. In: Artificial Intelligence and Statistics. 2007. p. 412–9.
29. Hu J, Lu J, Tan Y. Discriminative deep metric learning for face verification in the wild. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. 2014. p. 1875–82.
30. Vapnik VN. Statistical Learning Theory, 1st ed. USA: Wiley; 1998.
31. Parameswaran S, Weinberger K. Large margin multi-task metric learning. In: Advances in Neural Information Processing Systems 23. 2010. p. 1867–75.
32. Wang S, Jiang S, Huang Q, Tian Q. Multi-feature metric learning with knowledge transfer among semantics and social tagging. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012. 2012. p. 2240–7.
33. Kwok JT, Tsang IW. Learning with idealized kernels. In: Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA; 2003. p. 400–7. http://www.aaai.org/Library/ICML/2003/icml03-054.php.
34. Shi Y, Bellet A, Sha F. Sparse compositional metric learning. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada; 2014. p. 2078–084. http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8224.
35. Liu H, Zhang X, Wu P. Two-level multi-task metric learning with application to multi-classification. In: 2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015; 2015. p. 2756–60.
36. Köstinger M, Hirzer M, Wohlhart P, Roth PM, Bischof H. Large scale metric learning from equivalence constraints. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012; 2012. p. 2288–95.
37. Li Y, Tao D. Online semi-supervised multi-task distance metric learning. In: IEEE International Conference on Data Mining Workshops, ICDM Workshops 2016, December 12-15, 2016, Barcelona, Spain; 2016. p. 474–9.
38. Jin R, Wang S, Zhou Y. Regularized distance metric learning: Theory and algorithm. In: Advances in Neural Information Processing Systems, vol. 22. 2009. p. 862–70.
39. Zheng Y, Fan J, Zhang J, Gao X. Hierarchical learning of multi-task sparse metrics for large-scale image classification. Pattern Recogn. 2017;67:97–109.
40. Zhang Y, Yeung DY. Transfer metric learning by learning task relationships. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010.
41. Zhang Y, Yeung DY. Transfer metric learning with semi-supervised extension. ACM Trans Intell Syst Tech (TIST). 2012;3(3):54–15428.
42. Gupta AK, Nagar DK. Matrix Variate Distributions. Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics, vol. 104. London: Chapman & Hall; 2000.
43. Ma L, Yang X, Tao D. Person re-identification over camera networks using multi-task distance metric learning. IEEE Trans Image Process. 2014;23(8):3656–70.
44. Yang P, Huang K, Liu CL. Geometry preserving multi-task metric learning. Mach Learn. 2013;92(1):133–75.
45. Yang P, Huang K, Liu CL. Geometry preserving multi-task metric learning. In: European Conference on Machine Learning and Knowledge Discovery in Databases, vol. 7523. 2012. p. 648–64.
46. Dhillon IS, Tropp JA. Matrix nearness problems with bregman divergences. SIAM J Matrix Anal Appl. 2008;29:1120–46.
47. Kulis B, Sustik MA, Dhillon IS. Low-rank kernel learning with bregman matrix divergences. J Mach Learn Res. 2009;10:341–76.
48. Yang P, Huang K, Liu C. A multi-task framework for metric learning with common subspace. Neural Comput Applic. 2013;22(7-8):1337–47.
49. Torresani L, Lee K. Large margin component analysis. In: Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006; 2006. p. 1385–92. http://papers.nips.cc/paper/3088-large-margin-component-analysis.
50. Bhattarai B, Sharma G, Jurie F. Cp-mtml: Coupled projection multi-task metric learning for large scale face retrieval. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016; 2016. p. 4226–35.
51. Soleimani A, Araabi BN, Fouladi K. Deep multitask metric learning for offline signature verification. Pattern Recogn Lett. 2016;80:84–90.
52. McLaughlin N, del Rincón JM, Miller PC. Person reidentification using deep convnets with multitask learning. IEEE Trans Circ Syst Video Techn. 2017;27(3):525–39.

53. Baxter J. A bayesian/information theoretic model of learning to learn via multiple task sampling. Mach Learn. 1997;28(1):7–39.
54. Baxter J. A model of inductive bias learning. J Artif Intell Res. 2000;12:149–98.
55. Blitzer J, Crammer K, Kulesza A, Pereira F, Wortman J. Learning bounds for domain adaptation. In: Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007; 2007. p. 129–36. http://papers.nips.cc/paper/3212-learning-bounds-for-domain-adaptation.
56. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. Mach Learn. 2010;79(1-2):151–75.
57. Bousquet O, Elisseeff A. Stability and generalization. J Mach Learn Res. 2002;2:499–526.
58. Balcan MF, Blum A, Srebro N. A theory of learning with similarity functions. Mach Learn. 2008;72(1-2):89–112.
59. Wang L, Sugiyama M, Yang C, Hatano K, Feng J. Theory and algorithm for learning with dissimilarity functions. Neural Comput. 2009;21(5):1459–84.
60. Perrot M, Habrard A. A theoretical analysis of metric hypothesis transfer learning. In: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015; 2015. p. 1708–17. http://jmlr.org/proceedings/papers/v37/perrot15.html.
61. Bellet A, Habrard A. Robustness and generalization for metric learning. Neurocomputing. 2015;151:259–67.