The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

EMD Metric Learning

Zizhao Zhang, Yubo Zhang, Xibin Zhao*, Yue Gao*
Key Laboratory for Information System Security, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology
School of Software, Tsinghua University, China.
{zz-zh14,zhangyb17}@mails.tsinghua.edu.cn
{zxb,gaoyue}@tsinghua.edu.cn
* indicates corresponding authors

Abstract

Earth Mover's Distance (EMD), targeting at measuring the many-to-many distances, has shown its superiority and been widely applied in computer vision tasks, such as object recognition, hyperspectral image classification and gesture recognition. However, there is still little effort concentrated on optimizing the EMD metric towards better matching performance. To tackle this issue, we propose an EMD metric learning algorithm in this paper. In our method, the objective is to learn a discriminative distance metric for EMD ground distance matrix generation which can better measure the similarity between compared subjects. More specifically, given a group of labeled data from different categories, we first select a subset of training data and then optimize the metric for ground distance matrix generation. Here, both the EMD metric and the EMD flow-network are alternatively optimized until a steady EMD value can be achieved. This method is able to generate a discriminative ground distance matrix which can further improve the EMD distance measurement. We then apply our EMD metric learning method on two tasks, *i.e.*, multi-view object classification and document classification. The experimental results have shown better performance of our proposed EMD metric learning method compared with the traditional EMD method and the state-of-the-art methods. It is noted that the proposed EMD metric learning method can be also used in other applications.

Introduction

Variable-size descriptions of distributions, such as Gaussian Mixture Models (GMM) (Li, Wang, and Zhang 2013) have been widely used to represent multidimensional distributions in a compact way. These descriptors can be regarded as signatures, which are a set of the main clusters of a distribution. Here, each cluster is represented by a single point (*i.e.*, the cluster center) in the underlying space, together with a weight that denotes the size or the importance of that cluster. This is a common-used data description method and accordingly there are many methods to measure the distance between two signatures, such as K-L divergence (Goldberger, Gordon, and Greenspan 2003), Jensen-Shannon divergence (Endres and Schindelin 2003), and maximum mean discrepancy (Borgwardt et al. 2006). Among these methods, the Earth Mover's Distance (EMD) (Rubner, Tomasi, and Guibas 2000) is a general and flexible way to measure the dissimilarity between two signatures, which is a special case of the Wasserstein distances (Villani 2009) in its continuous form. EMD has shown its superior performance on the many-to-many matching problem and been used in many computer vision tasks, such as object recognition (Zhang et al. 2007), hyperspectral image classfication (Sun et al. 2015) and gesture recognition (Wang and Chan 2014).

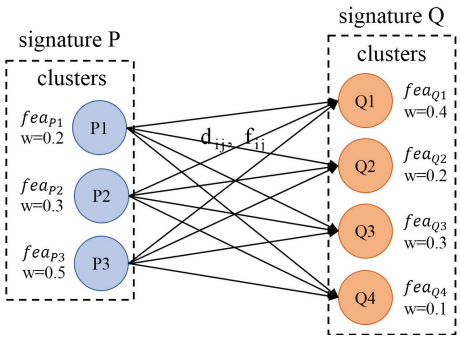


Figure 1: The illustration of Earth Mover's Distance, where fea_i is the feature for the center of P_i, and w_i is the corresponding weight of P_i. d_{ij} is the cost of shipping a unit of supply from P_i to Q_j, and f_{ij} is the corresponding flow.

The original EMD is based on a solution to the old transportation problem, as shown in Figure 1, which can be formalized as a linear programming problem: given a set of suppliers $\mathcal{P} = \{P_1, P_2, \ldots, P_n\}$, a set of consumers $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_m\}$, and a ground distance matrix $D = \{d_{ij}\}$ whose elements d_{ij} defines the cost of shipping a unit of supply from P_i to Q_j, the aim is to find an optimal set $\mathbf{F}$ of the flows f_{ij}, *i.e.* the amount of supply shipped from the i-th supplier P_i to the j-th consumer Q_j, to minimize the overall cost $\sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij}f_{ij}$, subjecting to the following constraints

$$\begin{array}{l} f_{ij} \geq 0 \quad (1 \leq i \leq n, 1 \leq j \leq m), \\ \sum_{i=1}^{n} f_{ij} \leq w_{Q_j} \quad (1 \leq j \leq m), \\ \sum_{j=1}^{m} f_{ij} \leq w_{P_i} \quad (1 \leq i \leq n), \end{array} \tag{1}$$

where w_{P_i} is the supply of the i-th supplier and w_{Q_j} is the capacity of the j-th consumer.

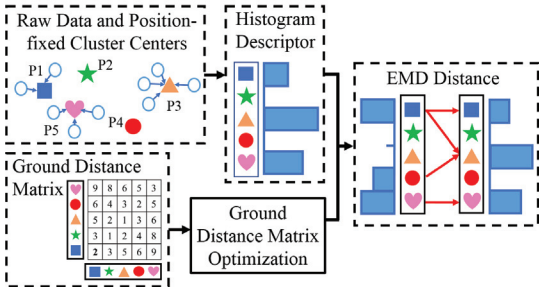


Figure 2: The illustration of ground distance metric optimization for EMD distance measurement in (Wang and Guibas 2012).

Once the transportation problem is solved, the EMD distance is defined as the cost normalized by the total flow:

$$\mathrm{EMD}(\mathcal{P},\mathcal{Q}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij} f_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}}. \quad (2)$$

Compared with other distance measurements, EMD is able to account for cross-bin information and can be applied to the signatures with different sizes. Moreover, EMD can generate a flow-network representing how the mass in the clusters is transported between two signatures and thus calculate an optimal distance between two signatures.

As shown in the EMD definition, the ground distance matrix between two signatures plays an important role in the EMD flow optimization procedure. A well generated ground distance matrix can lead to a better flow, and yet achieve even better EMD distance measurement. In a simple case, if the cluster centers of all the signatures are the same (i.e. position-fixed), a signature can be degenerated into a histogram descriptor. In this scenario, the optimization task for a better EMD distance measurement is to directly learn an optimal ground distance matrix, which defines the distance between each pair of the cluster centers, as shown in Figure 2. To this end, a supervised EMD distance learning method (Wang and Guibas 2012) was introduced to learn an optimal ground distance matrix among the representing histograms. In particular, the triplet constraint was used in this method and the trained ground distance matrix can better reflect the cross-bin relationships, hence producing more accurate EMD distances and better performance in the application of face verification.

However, this method can only work on the situation with fixed-size histograms representing the subjects. In practice, it cannot guarantee that all the subjects could be described by fixed-size histogram, which limits the application of the ground distance matrix learning method. Moreover, the learned ground distance matrix in (Wang and Guibas 2012) may contain negative elements, which could lead to negative EMD distances. We note that it is more general in practice that each signature can be described by different types of data and a common-used ground distance matrix does not exist in these scenarios. Under such circumstances, it is not feasible to optimize the ground distance matrix for EMD in many applications. Different from optimizing the ground distance matrix directly, it is possible to learn an optimal EMD distance metric, *i.e.*, the metric to generate the ground distance matrix in EMD, which has not been investigated yet, to the best of our knowledge.

To tackle this issue, we propose an EMD metric learning algorithm that can work in a more general setting. In this method, the objective is to optimize the distance metric for ground distance matrix generation, which can be more discriminative on data categories. We first select a group of data pairs of homogeneous neighbors and inhomogeneous neighbors from the labeled training dataset, and then learn the metric to minimize the EMD between the signatures from the same category and maximize the EMD between the signatures from different categories by encoding the triplet constraints via a hinge-loss functions. Instead of the gradient descent method, we use the intrinsic algorithm for optimizing the ground distance metric to retain the positive definiteness of the metric. We note that with the updated EMD metric, the ground distance matrix and the corresponding flow for EMD will change accordingly. Under such circumstance, we propose to optimize the EMD metric and the flow-networks of EMD alternatively until convergence. The learned EMD metric can be used to calculate new ground distance matrix and update the flow-networks for EMD between each two compared signatures. Then, with the new flow-networks, the EMD metric can be optimized again. This process repeats until convergence. The merit of our proposed method lies in twofold. First, the EMD metric can update the ground distance matrix for each two signatures separately, which does not require all signatures should have the same histogram description. Second, the ground distance matrix is guaranteed to be non-negative. We have applied our EMD metric learning method on two tasks, *i.e.*, object classification and document classification, and experiments are conducted on two public benchmarks, including the National Taiwan University (NTU) 3D model dataset (Chen et al. 2003) and the Twitter Sentiment Corpus dataset (Sanders 2011). We have also compared the proposed method with state-of-the-art methods and the traditional EMD method, and the experimental results show better performance of the proposed method.

The main contributions of this work can be summarized as follows:

1. We propose an EMD metric learning algorithm targeting on a more general setting, which can dynamically optimize the ground distance matrix and yet lead to better EMD distance measurement.

2. We have applied the EMD metric learning method on two tasks, *i.e.*, multi-view object classification and text classification, and evaluated the performance.

The rest of this paper is organized as follows. We first introduce the related work on EMD applications. We then provide our proposed EMD metric learning method and its applications. Experiments and discussions are further provided and we finally conclude this paper.

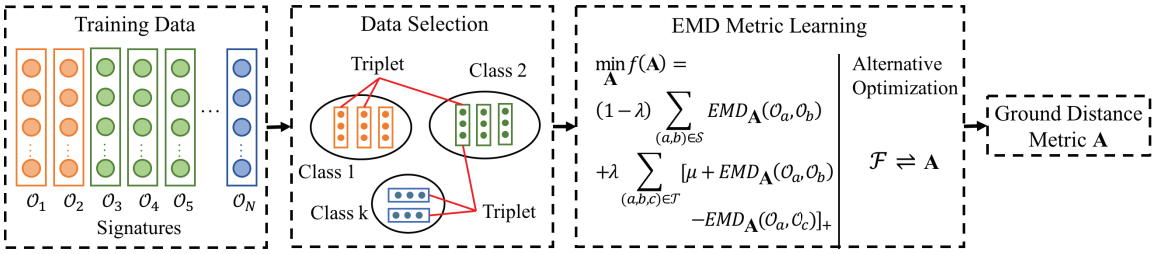


Figure 3: The framework of our proposed EMD metric learning algorithm.

Related work

In this section, we briefly review existing works on EMD and its applications, such as gesture recognition, hyperspectral image classification, document classification and face recognition.

EMD was first introduced in (Rubner, Tomasi, and Guibas 2000), where it was used as a metric for image retrieval. More specifically, each image implied a distribution of points in the three-dimensional color space. Based on the clustering results, the data distribution was transformed into the signature description. In (Carley and Tomasi 2015), EMD was employed for gesture recognition. For each hand gesture, a silhouette signature containing segment information was extracted. The optional and additional ground distances were defined between segments as a convex function of the flows. Then, the signature dissimilarity was calculated using EMD. In the task of high-resolution remote sensing image classification (Zhang et al. 2013), each image was translated to a set of visual codes. By sampling and clustering the visual codes, a codebook was then generated and arranged via a strategy named local nearest neighbor codebook arrangement. For each image, the occurrence number of each code was counted to form a histogram representation. Then, the comparison between each image pair can be transformed to the EMD distance measurement, which was used for hyperspectral image classification. In (Li, Wang, and Zhang 2013), Li *et al.* proposed an EMD methodology using Gaussian Mixture Models for image matching. In this method, each image was modeled by a GMM, and the image matching task was conducted by comparing the corresponding GMMs via sparse representation-based EMD. Moreover, two kinds of ground distances between GMMs were defined and learned based on theory of information geometry. In the task of document classification (Kusner et al. 2015), each nonstop word in a document was first embedded into a feature vector using the *word2vec* network. Then, EMD was employed to measure the dissimilarity between two document by calculating the minimal cost that the embedded words of one document need to travel to the embedded words of another documents.

In (Wang and Guibas 2012), EMD was used to measure the distance between two face representations. In this work, the ground distance matrix was optimized using a supervised EMD learning method with triple-constraints. In the experiments of face verification, face descriptors were extracted based on reference identities which were pre-selected from the test faces. Although this work has shown better performance with the updated ground distance matrix, it requires that all the data are represented by fixed-size histograms, which is not always feasible in practice.

EMD Metric Learning

In this section, we detailed introduce the proposed EMD metric learning algorithm. Figure 3 illustrates the general framework of our proposed method. Given a set of signatures with labels, our objective is to learn an optimal metric for ground distance matrix generation, which could be more discriminative for classification. First, we select a group of data pairs from homogeneous neighbors and inhomogeneous neighbors, which are used as the training data for metric learning. Then, we conduct metric learning to minimize the EMD between the signatures from the same category and maximize the EMD between the signatures from different categories. As the learned EMD metric will lead to new flow-networks and yet new EMD distances, we propose an alternating optimization algorithm here to jointly update the EMD metric and the flow-networks until convergence. The final distance metric is used for EMD distance measurement. In this section, we first introduce the formulation of the EMD metric learning task, and then provide the training data selection method and the solution to the optimization task.

The Formulation of EMD Metric Learning

Given N signatures $\mathcal{O} = \{\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_N\}$, with $\mathcal{O}_i = \{\boldsymbol{X}_i, \boldsymbol{w}_i\}$, $i = 1, \ldots, N$, let $\boldsymbol{X}_i = (\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \ldots, \boldsymbol{x}_i^{n_i})$ denote the n_i feature vectors of the i-th signature and $\boldsymbol{w}_i = (w_i^1, w_i^2, \ldots, w_i^{n_i})$ denote the corresponding weight vector. $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$ is the set of labels with respect to $\mathcal{O}$.

To measure the distance between two signatures $\mathcal{O}_i$ and $\mathcal{O}_j$ by EMD, the ground distance matrix is required to be generated first. Here we use the Mahalanobis distance as the ground distance measurement, and the squared distance between the s-th feature vector of $\mathcal{O}_i$ and the t-th feature vector of $\mathcal{O}_j$ can be written as

$$d(\boldsymbol{x}_i^s, \boldsymbol{x}_j^t) = (\boldsymbol{x}_i^s - \boldsymbol{x}_j^t)^T \mathbf{A} (\boldsymbol{x}_i^s - \boldsymbol{x}_j^t), \tag{3}$$

where $\mathbf{A}$ is a global linear transformation of the underlying space, *i.e.*, the ground distance metric, which can be initialed as $\mathbf{I}$.

Given a set of training data with labels, the task here is to learn an optimal distance metric $\mathbf{A}$ for ground distance matrix generation, which should be more discriminative for the

data categories. More specifically, the objective is to minimize the EMD distances between the signatures from the same category and maximize the EMD distances between the signatures from different categories. Given a set of signature triplets $\mathcal{T} = \{(a, b, c) \mid y_a = y_b, y_a \neq y_c\}$, usually we need to guarantee that the distance between the signatures from the same category should be smaller than that from different categories. Here $y_a = y_b$ indicates that a and b are with the same labels and $y_a \neq y_c$ means that a and c are with different labels. Here, the triplet constraint for EMD metric can be written as

$$EMD_{\mathbf{A}}(\mathcal{O}_a, \mathcal{O}_b) < EMD_{\mathbf{A}}(\mathcal{O}_a, \mathcal{O}_c), (a, b, c) \in \mathcal{T}. \quad (4)$$

By encoding the triplet constraints via a hinge-loss function, the supervised EMD metric learning can be formulated as

$$\min_{\mathbf{A}} f(\mathbf{A}) = \sum_{(a,b,c)\in\mathcal{T}} [\mu + EMD_{\mathbf{A}}(\mathcal{O}_a, \mathcal{O}_b) - EMD_{\mathbf{A}}(\mathcal{O}_a, \mathcal{O}_c)]_+, \quad (5)$$

where μ is the non-negative real number which specifies the lower bound for the margin to fix the scale of the matrix A.

Besides the triplet constraint, in order to preserve the topological structure of the data, we also need to guarantee that the homogeneous neighbours from the same category should be close. By jointly considering the topological regularizer and the triplet constraint, the formulation for EMD metric learning can be rewritten as

$$\begin{aligned}\min_{\mathbf{A}} f(\mathbf{A}) =& (1-\lambda) \sum_{(a,b)\in\mathcal{S}} EMD_{\mathbf{A}}(\mathcal{O}_a, \mathcal{O}_b) \\ &+ \lambda \sum_{(a,b,c)\in\mathcal{T}} [\mu + EMD_{\mathbf{A}}(\mathcal{O}_a, \mathcal{O}_b) \\ &\qquad - EMD_{\mathbf{A}}(\mathcal{O}_a, \mathcal{O}_c)]_+, \end{aligned} \quad (6)$$

where the set S is defined as all pairs of homogeneous neighbors from the same category and $0 \leq \lambda \leq 1$ is a tradeoff parameter to balance the triplet loss term and the topological regularizer.

Training Data Selection

We note that the number of all possible triplets in the training dataset could be very large. In such case, the number of relative distance constraints will grow cubically with respect to the size of the training set and thus leads to very high computational cost. Therefore, it is important to select a small set of training data for metric learning to reduce the computational cost. We just select k_g nearest neighbors with the same label (namely target neighbors) and k_i nearest neighbors with different labels (namely imposters) for each signature. Then, the overall amount of triplet constraints could be reduced to $k_i k_g N$.

However, we need to calculate the EMD distances between all pairs of signatures to find these target neighbors and imposters. Consider that the time complexity of solving the standard EMD problem is $O(n^3 \log n)$, it is very expensive for training data selection. To circumvent this issue, we use a cheap lower bound of the standard EMD to approximately select the nearest neighbors without computing the exact EMD distance (Kusner et al. 2015), which is much more efficient than the traditional EMD. Precisely, given two signatures $\mathcal{P} = \{P_1, P_2, \ldots, P_n\}$ and $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_m\}$, let $D = \{d_{ij}\}$ and $F = \{f_{ij}\}$ denote the ground distance matrix and the EMD flow-networks respectively. The relaxed EMD (REMD) is defined as

$$REMD(\mathcal{P}, \mathcal{Q}) = \begin{cases} \frac{\sum_{i=1}^{n} d_{ij^*} w_{P_i}}{\sum_{i=1}^{n} w_{P_i}} & \sum_{i=1}^{n} w_{P_i} \leq \sum_{j=1}^{m} w_{Q_j} \\ \frac{\sum_{j=1}^{m} d_{i^*j} w_{Q_j}}{\sum_{j=1}^{m} w_{Q_j}} & \sum_{i=1}^{n} w_{P_i} > \sum_{j=1}^{m} w_{Q_j}, \end{cases} \quad (7)$$

where $j^* = \arg\min_j d_{ij}$, $i^* = \arg\min_i d_{ij}$.

The intuition behind Eq. (7) is that if the total weight of $\mathcal{P}$ is less than that of $\mathcal{Q}$, then for each cluster in $\mathcal{P}$, we move all its mass to the closest cluster in $\mathcal{Q}$, and vice versa.

It is straight-forward to show that the $REMD(\mathcal{P}, \mathcal{Q})$ must lower bound $EMD(\mathcal{P}, \mathcal{Q})$. Now we assume that $\sum_{i=1}^{n} w_{P_i} \leq \sum_{j=1}^{m} w_{Q_j}$, according the definition of EMD, it follows $\sum_{j=1}^{m} f_{ij} = w_{P_i}$. Hence

$$\begin{aligned} EMD(\mathcal{P}, \mathcal{Q}) &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij} f_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}} \geq \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij^*} f_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}} \\ &= \frac{\sum_{i=1}^{n} d_{ij^*} \sum_{j=1}^{m} f_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}} = \frac{\sum_{i=1}^{n} d_{ij^*} w_{P_i}}{\sum_{i=1}^{n} w_{P_i}} \\ &= REMD(\mathcal{P}, \mathcal{Q}). \end{aligned} \quad (8)$$

Evidently, the conclusion remains true in the other case. To calculate the bound, we just need to calculate the pairwise distance between clusters and conduct a nearest neighbor search, which has a time complexity of $O(n^2)$ and allows us to speed up the data selection process considerably.

With the pairwise REMD distances among all training data, we can select a subset of triplets from the original training dataset for metric learning.

Optimization

With the selected training data, here we optimize the objective function in Eq. (6) to learn the EMD metric $\mathbf{A}$. We note that once the EMD metric $\mathbf{A}$ is updated, the ground distance matrix and the corresponding flow-networks $\mathcal{F}$ for EMD will change accordingly. Then, we need to re-calculate the EMD distances with respect to the new flow-networks. Under such circumstance, the optimization should target on both the EMD metric $\mathbf{A}$ and the associated flow-networks $\mathcal{F}$. By substituting Eq. (2) into Eq. (6), the objective function is further written as

$$\begin{aligned} &\min_{\mathbf{A},\mathcal{F}} f(\mathbf{A}, \mathcal{F}) = (1-\lambda) \textstyle\sum_{(a,b)\in\mathcal{S}} EMD_{\mathbf{A}}(\mathcal{O}_a, \mathcal{O}_b) \\ &+\lambda \textstyle\sum_{(a,b,c)\in\mathcal{T}} [\mu + EMD_{\mathbf{A}}(\mathcal{O}_a, \mathcal{O}_b) - EMD_{\mathbf{A}}(\mathcal{O}_a, \mathcal{O}_c)]_+, \\ &= (1-\lambda) \textstyle\sum_{(a,b)\in\mathcal{S}} \left(\sum_{i=1}^{n_a} \sum_{j=1}^{n_b} F_{a,b}(i,j) d(\boldsymbol{x}_a^i, \boldsymbol{x}_b^j) \right) \\ &+\lambda \textstyle\sum_{(a,b,c)\in\mathcal{T}} \Big[\mu + \left(\sum_{i=1}^{n_a} \sum_{j=1}^{n_b} F_{a,b}(i,j) d(\boldsymbol{x}_a^i, \boldsymbol{x}_b^j) \right) \\ &- \left(\sum_{i=1}^{n_a} \sum_{j=1}^{n_c} F_{a,c}(i,j) d(\boldsymbol{x}_a^i, \boldsymbol{x}_c^j) \right) \Big]_+ . \end{aligned} \quad (9)$$

When $\mathcal{F}$ is fixed, the optimization task can be transformed into the problem of Mahalanobis distance metric learning. On the other hand, If $\mathbf{A}$ is fixed, the problem can be splited into $|\mathcal{S}| + 2|\mathcal{T}|$ independent traditional EMD sub-problems,

which can be solved by the Hungarian algorithm. Therefore, the formulation is bi-convex with respect to $\mathbf{A}$ and $\mathcal{F}$ jointly. To solve the above optimization task, we propose to optimize the EMD metric $\mathbf{A}$ and the flow-networks $\mathcal{F}$ of EMD alternatively until convergence.

Given the flow-network $\mathcal{F}$ for the selected training data, the learning task can be further rewritten in the matrix form as

$$\begin{aligned}\min_{\mathbf{A}} f(\mathbf{A}) = (1-\lambda) \sum_{(a,b)\in\mathcal{S}} tr(G_{a,b}^T \mathbf{A}) \\ +\lambda \sum_{(a,b,c)\in\mathcal{T}} [\mu + tr(G_{a,b}^T \mathbf{A}) - tr(G_{a,c}^T \mathbf{A})]_+,\end{aligned} \tag{10}$$

where

$$\begin{aligned} G_{a,b} &= G_{a,b}^1 - G_{a,b}^0, \\ G_{a,b}^1 &= X_a \text{diag}(F_{a,b}\boldsymbol{e}) X_a^T + X_b \text{diag}(\boldsymbol{e}^T F_{a,b}) X_b^T, \\ G_{a,b}^0 &= X_a F_{a,b} X_b^T + X_b F_{a,b}^T X_a^T. \end{aligned} \tag{11}$$

The gradient of Eq. (10) with respect to $\mathbf{A}$ is calculated by

$$\begin{aligned}\nabla f(\mathbf{A}) = (1-\lambda) \sum_{(a,b)\in\mathcal{S}} G_{a,b} + \lambda \sum_{(a,b,c)\in\mathcal{T}'} (G_{a,b} - G_{a,c}), \\ \mathcal{T}' = \mathcal{T} - \{(a,b,c) \mid tr(G_{a,c}^T \mathbf{A}) - tr(G_{a,b}^T \mathbf{A}) > \mu\}.\end{aligned} \tag{12}$$

To retain the positive definiteness of $\mathbf{A}$, a common used optimization techniques is the projected gradient method, which applies gradient descent followed by a projection onto the positive-definite cone (Han et al. 2017). However, in recent years, several intrinsical interactive methods for optimization on matrix manifolds have achieved better performance and convergence (Arsigny et al. 2007). This kind of methods preserves the manifold structure of positive-definite matrix group, *i.e.*, let the variable $\mathbf{A}$ still belong to the corresponding manifolds in each iteration. Therefore, we use the intrinsic algorithm (Ying et al. 2017) to optimize $\mathbf{A}$.

The set of symmetric positive-definite matrices is a smooth Riemannian manifold, denoted by $\mathcal{M}$. The tangent space of $\mathcal{M}$ at the point P is the vector space of derivations at point P, denoted by $T_P\mathcal{M}$. Given two points on the manifold, the locally distance-minimizing curve connecting them is termed the geodesic. The geodesic on the manifold going through P with tangent vector Z is given by

$$\Gamma_{(P,Z)}(\alpha) = P^{\frac{1}{2}} \exp(\alpha P^{-\frac{1}{2}} Z P^{-\frac{1}{2}}) P^{\frac{1}{2}}. \tag{13}$$

Therefore, we update $\mathbf{A}_t$ to $\mathbf{A}_{t+1}$ on the positive-definite matrix group by computing:

$$\mathbf{A}_{t+1} = \mathbf{A}_t^{\frac{1}{2}} \exp(\alpha \mathbf{A}_t^{-\frac{1}{2}} \nabla f(\mathbf{A}_t) \mathbf{A}_t^{-\frac{1}{2}}) \mathbf{A}_t^{\frac{1}{2}}, \tag{14}$$

where α is the optimal step size at each iteration.

Note that the tangent vector should be symmetric in Eq.(13), but actually the gradient of the objective function in Eq.(12) is not symmetric, which further leads $\mathbf{A}_{t+1}$ to be nonsymmetric. To circumvent this issue, we add a symmetrization operator on the gradient. That is, we have

$$\mathbf{A}_{t+1} = \mathbf{A}_t^{\frac{1}{2}} \exp(\alpha \mathbf{A}_t^{-\frac{1}{2}} \mathbf{Sym}[\nabla f(\mathbf{A}_t)] \mathbf{A}_t^{-\frac{1}{2}}) \mathbf{A}_t^{\frac{1}{2}}, \tag{15}$$

Algorithm 1 EMD Metric Learning

Input: Training data set $\mathcal{O}$, maximal iteration of EMD metric learning k_{out}, and parameters k_g, k_i, λ and μ
Output: EMD metric $\mathbf{A}$

```
Initialize A = I.
Calculate the REMD between each pair of signatures.
Construct the set S and T by selecting k_i target neighbors and k_g imposters for each training instance.
Initialize the learning rate α and maximal iteration k_in of EMD metric learning.
for each s ∈ [1, k_out] do
    Fix A, solve for the EMD flow-networks F.
    Fix F.
    for each t ∈ [1, k_in] do
        Compute the gradient ∇f(A_t) by Eq.(12).
        Update A by Eq.(15).
return A
```

where $\mathbf{Sym}[\nabla f(\mathbf{A}_t)] = \frac{\nabla f(\mathbf{A}_t) + \nabla f(\mathbf{A}_t)^T}{2}$.

Then, we can update the EMD metric $\mathbf{A}$ using the gradient descent method. With the new $\mathbf{A}$, we can further calculate the flow-networks $\mathcal{F}$ for all the selected training data. In this way, we can alternatively update $\mathbf{A}$ and the flow $\mathcal{F}$ until convergence. The newly generated EMD metric $\mathbf{A}$ can be used for further EMD distance measurement. The overall workflow of the proposed EMD metric learning method is shown in Algorithm 1.

Applications of EMD Metric Learning

In this section, we apply our proposed EMD metric learning method to two tasks, *i.e.*, multi-view object classification and document classification.

Multi-View Object Classification. In this application, each object is represented by a group of views taken from the cameras with different angles. For each view of the object, a Convolutional Neural Network (CNN) feature (Nie et al. 2017) is extracted for description due to its superior performance on the task of object retrieval and recognition. It is noteworthy that our framework has no specific requirements on the feature extraction, which indicates that other features can be also used here and the CNN feature is just an example. After feature extraction, considering that multiple views may contain redundancy, we conduct view clustering following the settings in (Gao et al. 2012) to generate view clusters. For each object, a group of representative views are selected from these clusters, associated with corresponding weights based on the cluster size. In this way, each object can be represented by a signature of representative views and EMD can be used to measure the distance between objects for classification. Our EMD metric learning method can be employed to optimize the EMD distance between objects.

Document Classification. In this application, each document is represented by a bag-of-words (BOW) feature and can be modeled as a signature (Kusner et al. 2015). To describe the document, the *word2vec* (Mikolov et al. 2013) model, a three-layered neural network, is used to generate a feature representation for each nonstop word. Then, the nor-

malized bag-of-words (nBOW) vector is extracted from each document, reflecting the frequency (weight) of each nonstop word in this document. In this way, we construct the signature description for each document and the EMD metric learning method can be used to learn an optimal EMD distance for document classification.

Experiments

Testing Datasets and Experimental Settings

To validate the proposed EMD metric learning method, we have conducted experiments on two datasets, *i.e.*., the National Taiwan University 3D model dataset (NTU) (Chen et al. 2003) and the Twitter Sentiment Corpus dataset (TWITTER) (Sanders 2011). The NTU dataset contains 401 objects from 16 categories, including bomb, bottle, car, chair, cup, door, guitar, gun, map, plane, starship, stick, sword, table, tank, truck. In this dataset, each object contains 60 views, and the 4096-d CNN feature is extracted for each view. To reduce the computational cost during the metric learning process, we further apply PCA to reduce the feature dimension to 300. The TWITTER dataset contains 2176 tweets labeled with three types of sentiments, including 'positive', 'negative', and 'neutral'. In this dataset, each document is represented by a 300-d feature, and the average number of unique words per document is 9.9.

We randomly select 20%, 30%, 40% and 50% of all data per each category as labeled training data and all the rest are used for testing. The dataset splitting process repeats 10 times and the average accuracy of classification results is used for evaluation. We empirically set the parameter μ in Eq. (6) to 0.1 on both datasets. The parameter λ is set to 0.5 on the NTU dataset, and 0.2 on the TWITTER dataset.

In experiments, the following methods are used for comparison:

1. Manifold Discriminant Analysis (MDA) (Wang and Chen 2009). In MDA, an embedding space is learned to maximize manifold margin between two compared manifolds by modeling each image set as a manifold.
2. Covariance Discriminative Learning (CDL) (Wang et al. 2012). In CDL, the natural second-order statistic is employed to model the image set and the distance is measured using the linear discriminant analysis.
3. Covariance Discriminative Learning with Partial Least Squares (CDL_PLS) (Wang et al. 2012). CDL_PLS is improved from the CDL to conduct the partial least squares through means of latent variables.
4. Log-Euclidean Metric Learning (LEML) (Huang et al. 2015b). In LEML, a Mahalanobis metric is learned to make the matrix logarithms changed to a discriminative tangent space from original space.
5. Projection Metric Learning (PML) (Huang et al. 2015a). In PML, a Fisher LDA-like framework is introduced to learn a Mahalanobis-like matrix directly on Grassmann manifold, which maps the original Grassmann manifold to a lower-dimensional, more discriminative one.
6. EMD, *i.e.*, the traditional EMD method.

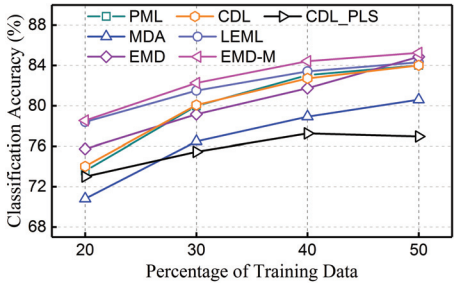


Figure 4: Experimental comparison on the NTU dataset.

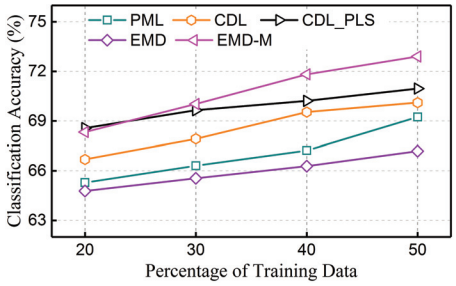


Figure 5: Experimental comparison on the TWITTER dataset.

7. EMD metric learning (EMD-M), *i.e.*, the proposed method.

We note that in the TWITTER dataset, MDA and LEML cannot be performed due to the number of words per document is too small and there may exists duplicate words in one document, which results in some mistakes in the training process. Therefore, we compare all 7 methods in the NTU datasets while only compare 5 methods in the TWIITER dataset.

Experimental results

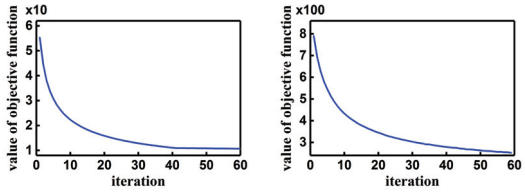
Experimental results on two datasets are demonstrated in Figure 4 and Figure 5. As shown in these results, we can have the following observations:

1. Compared with the state-of-the-art methods, the proposed EMD-M method achieves higher classification performance on both datasets. For example, EMD-M achieves the classification accuracy of 78.54%, 82.26%, 84.42% and 85.25% when 20%, 30%, 40% and 50% data are used for training on the NTU dataset. On the TWITTER dataset, EMD-M achieves gains of 5.29%, 3.98% and 2.74% compared with PML, CDL, and CDL_LDA when 50% data are used for training. Similar results can be observed in other experiments.
2. Compared with traditional EMD method, the proposed EMD-M achieves better performance. For example, EMD-M obtains gains of 5.48%, 6.83%, 8.32% and 8.5% on the TWITTER dataset when 20%, 30%, 40%, and 50% data are used for training, compared with EMD.

The better performance of our proposed method can be dedicated to the following reasons. We note that although

EMD is effective on many-to-many matching, it is limited on the robustness of data comparison. We can observe that EMD works better on the NTU dataset while performs worse on the TWITTER dataset. To overcome this limitation, our proposed method is able to learn an optimal EMD metric and yet lead to better ground distance matrix for EMD distance calculation. In this way, the EMD-M method can be more discriminative using the training data and thus achieve much better performance compared with EMD.

On convergence



(a) the NTU Dataset (b) the TWITTER Dataset

Figure 6: The variation of objective function with respect to the iteration of our learning method.

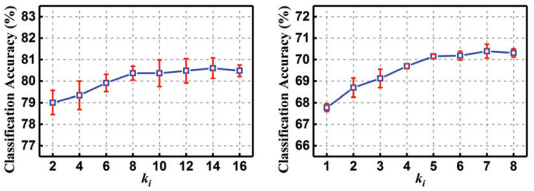
In our proposed EMD-M method, the EMD metric $\mathbf{A}$ and the flow-network $\mathcal{F}$ are alternatively updated. Therefore, the convergence speed is important for the learning process. Here we further investigate the change of the cost values of the objective function with respect to the iterations. The data is shown in Figure 6. As shown in these results, the objective function can reduce very fast and it will converge after about 20 iterations. These results can demonstrate that our proposed method can achieve the optimal performance efficiently.

On parameters

In our framework, there are two types of key parameters, including the selected number of training data and λ in Eq. (6). Here we investigate the influence of these parameters on the performance.

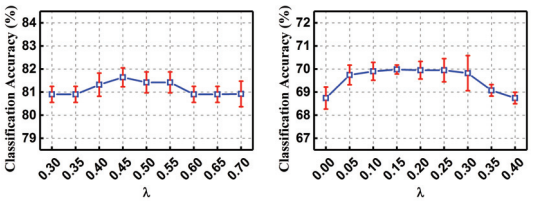
For the training data selection, there are parameters for the number of selected target neighbors and imposters for each signature, *i.e.*, k_g and k_i. These parameters control the size of training data. Usually, the more training data and the corresponding constraints we use, the better the performance of resulting metric is. Here we fix k_g as 3 and vary k_i in the range of $[2, 16]$ on the NTU dataset, and $[1, 8]$ on the TWITTER dataset, with 30% training data. Experimental results are shown in Figure 7. As shown in these results, we can notice that the performance is steady when k_i is large enough, such as above 10 on the NTU dataset. An interesting observation is that on the TWITTER dataset, we need fewer imposters to achieve the stable performance than on the NTU dataset. It illustrates that the more categories there are, the more imposters we need to select for each training instance.

Another important parameter is λ, which controls the weights of the loss term and the regularizer. We vary λ from 0.3 to 0.7 on the NTU dataset, while from 0 to 0.4 on the TWITTER dataset. The experimental results are shown in



(a) the NTU Dataset (b) the TWITTER Dataset

Figure 7: The performance evaluation by varying the sampling rates with 30% data used for training.



(a) the NTU Dataset (b) the TWITTER Dataset

Figure 8: The performance evaluation by varying the parameter λ with 30% data used for training.

Figure 8. We can notice that the proposed method can also achieve steady performance when λ is round 0.5 and 0.2 on the two datasets respectively, and show relatively poor performance when λ is overly small or large, which is consistent with the impact of λ in the presented model. Too small values of λ eliminate the influence of local correlations, while overly large values overemphasize the topological structure of the data.

Conclusion

In this paper, we propose an EMD metric learning algorithm. The proposed method is able to learn an optimal distance metric and yet lead a better ground distance matrix for EMD distance calculation between two signatures. In the learning stage, we alternatively update the EMD metric and flow-network for EMD distance. We have applied it on the tasks of multi-view object classification and document classification. We have conducted experiments on two datasets, and the comparison with the state-of-the-art methods demonstrates the effectiveness of the proposed method. We also compare our proposed EMD-M method with the traditional EMD method. Experimental results show about 5% gains for our EMD-M method. The framework can also be further applied to many other tasks, such as face verification and object recognition.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2017YFC011300, in part by the National Natural Science Funds of China under Grant 61671267 and Grant 61527812, in part by National Science and Technology Major Project (No. 2016ZX01038101), MIIT IT funds (Research and application of TCN key tech-

nologies) of China, and The National Key Technology R&D Program (No. 2015BAG14B01-02).

References

Arsigny, V.; Fillard, P.; Pennec, X.; and Ayache, N. 2007. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications* 29(1):328–347.

Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H.-P.; Schölkopf, B.; and Smola, A. J. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22(14):e49–e57.

Carley, C., and Tomasi, C. 2015. Single-frame indexing for 3d hand pose estimation. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, 101–109.

Chen, D.-Y.; Tian, X.-P.; Shen, Y.-T.; and Ouhyoung, M. 2003. On visual similarity based 3d model retrieval. In *Computer Graphics Forum*, volume 22, 223–232. Wiley Online Library.

Endres, D. M., and Schindelin, J. E. 2003. A new metric for probability distributions. *IEEE Transactions on Information Theory* 49(7):1858–1860.

Gao, Y.; Tang, J.; Hong, R.; Yan, S.; Dai, Q.; Zhang, N.; and Chua, T.-S. 2012. Camera constraint-free view-based 3-d object retrieval. *IEEE Transactions on Image Processing* 21(4):2269–2281.

Goldberger, J.; Gordon, S.; and Greenspan, H. 2003. An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *Proceedings of IEEE International Conference on Computer Vision*, 487. IEEE.

Han, J.; Cheng, G.; Li, Z.; and Zhang, D. 2017. A unified metric learning-based framework for co-saliency detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Huang, Z.; Wang, R.; Shan, S.; and Chen, X. 2015a. Projection metric learning on grassmann manifold with application to video based face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 140–149.

Huang, Z.; Wang, R.; Shan, S.; Li, X.; and Chen, X. 2015b. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *Proceedings of International Conference on Machine Learning*, 720–729.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *Proceedings of International Conference on Machine Learning*, 957–966.

Li, P.; Wang, Q.; and Zhang, L. 2013. A novel earth mover's distance methodology for image matching with gaussian mixture models. In *Proceedings of IEEE International Conference on Computer Vision*, 1689–1696.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nie, W.; Cao, Q.; Liu, A.; and Su, Y. 2017. Convolutional deep learning for 3d object retrieval. *Multimedia Systems* 23(3):325–332.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision* 40(2):99–121.

Sanders, N. J. 2011. Sanders-twitter sentiment corpus.

Sun, W.; Li, W.; Li, J.; and Lai, Y. M. 2015. Band selection using sparse nonnegative matrix factorization with the thresholded earths mover distance for hyperspectral imagery classification. *Earth Science Informatics* 8(4):907–918.

Villani, C. 2009. *The Wasserstein distances*. Berlin, Heidelberg: Springer Berlin Heidelberg. 93–111.

Wang, C., and Chan, S. 2014. A new hand gesture recognition algorithm based on joint color-depth superpixel earth mover's distance. In *Proceedings of the 4th International Workshop on Cognitive Information Processing*, 1–6. IEEE.

Wang, R., and Chen, X. 2009. Manifold discriminant analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 429–436. IEEE.

Wang, F., and Guibas, L. J. 2012. Supervised earth movers distance learning and its computer vision applications. In *Proceedings of European Conference on Computer Vision*, 442–455. Springer.

Wang, R.; Guo, H.; Davis, L. S.; and Dai, Q. 2012. Covariance discriminative learning: A natural and efficient approach to image set classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2496–2503. IEEE.

Ying, S.; Wen, Z.; Shi, J.; Peng, Y.; Peng, J.; and Qiao, H. 2017. Manifold preserving: An intrinsic approach for semisupervised distance metric learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, J.; Marszalek, M.; Lazebnik, S.; and Schmid, C. 2007. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision* 73(2):213–238.

Zhang, Y.; Sun, X.; Wang, H.; and Fu, K. 2013. High-resolution remote-sensing image classification via an approximate earth mover's distance-based bag-of-features model. *IEEE Geoscience and Remote Sensing Letters* 10(5):1055–1059.