Learning Expected Hitting Time Distance*

De-Chuan Zhan and **Peng Hu** and **Zui Chu** and **Zhi-Hua Zhou**
National Key Laboratory for Novel Software Technology
Collaborative Innovation Center of Novel Software Technology and Industrialization
Nanjing University, Nanjing 210023, China
{zhandc, hup, chuz, zhouzh}@lamda.nju.edu.cn

Abstract

Most distance metric learning (DML) approaches focus on learning a Mahalanobis metric for measuring distances between examples. However, for particular feature representations, e.g., histogram features like BOW and SPM, Mahalanobis metric could not model the correlations between these features well. In this work, we define a non-Mahalanobis distance for histogram features, via Expected Hitting Time (EHT) of Markov Chain, which implicitly considers the high-order feature relationships between different histogram features. The EHT based distance is parameterized by transition probabilities of Markov Chain, we consequently propose a novel type of distance learning approach (LED, Learning Expected hitting time Distance) to learn appropriate transition probabilities for EHT based distance. We validate the effectiveness of LED on a series of real-world datasets. Moreover, experiments show that the learned transition probabilities are with good comprehensibility.

Introduction

Effectiveness of learning methods like k-means, k-NN substantially rely on the distance metric invoked. Most Distance Metric Learning (DML) methods, e.g., (Davis et al. 2007; Guillaumin, Verbeek, and Schmid 2009), focus on learning a (squared) Mahalanobis distance which is defined as: $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)^2 = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two instances in $\mathbb{R}^d$, $\mathbf{M} \in \mathbb{R}^{d\times d}$ is a symmetric positive semi-definite (PSD) matrix. $\mathbf{M}$ in above equation gives the definition of Mahalanobis distance metric, and it implicitly measures the second-order correlations between features.

It is noteworthy, however, that the second-order correlations could not express higher-order feature interactions in Mahalanobis metric; this may cause some problems on distance measurements. Fig. 1 gives a concrete example: The left two photos contain a same dog with different poses, while the right two subplots are the gray channel histograms respectively. The significant differences between histograms will lead to a large distance between these two images of dog, and may hurt the generalization ability no matter what

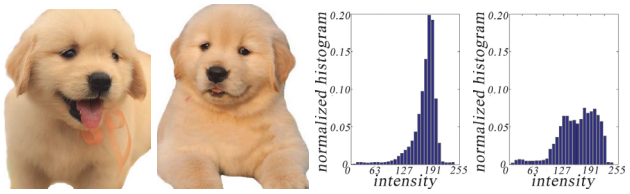


Figure 1: Illustration on the reason of classical Mahalanobis distance metric learning failure on histogram features: histogram changing tremendously on images of the same dog. The x-axis is the intensities, and the y-axis is the normalized frequencies (histogram), i.e., these histograms can be regarded as the intensities distribution.

kinds of Mahalanobis distances are learned and used. This phenomenon owes to the fact that traditional Mahalanobis distance metric only handle second-order correlations of individual features and leaves the high-order feature interactions unconsidered. While in reality, e.g., the case of Fig. 1, adjacent intensities are very similar on visual sense, and features in these photos should be strongly related. Similar phenomena occur on other histogram style features like BOW (Sivic and Zisserman 2003) and SPM (Lazebnik, Schmid, and Ponce 2006), etc. Mahalanobis distance metric could be inapplicable for these histogram feature representations where high-order feature interactions exist.

Non-Mahalanobis distances are raised in very recent. Kedem et al. (2012) proposed GB-LMNN, an extension of LMNN (Weinberger and Saul 2009), which utilizes gradient boosting regression trees to obtain a non-linear distance function; DeepML (Hu, Lu, and Tan 2014) learns nonlinear mappings for different views with deep brief networks for face verification; EMD (Rubner, Tomasi, and Guibas 2000) and Kullback-Leibler divergence investigate dissimilarities between two distributions yet itself a pre-defined distance measurement rather than performing distance learning; Shi, Bellet, and Sha (2014) proposed a sparse distance metric learning approach which can learn a group of local Mahalanobis distances based on fixed metric bases. These approaches are designed according to the characteristics of real problems, and have achieved better performance than ordinary Mahalanobis metrics in particular tasks. Nevertheless, those fixed non-linear distance functions or fixed distance

*This research was supported by NSFC (61333014) and CCF-Tencent Open Research Fund (RAGR20150117).

bases in above approaches have less flexibilities than learning distance metrics/measurements.

Expected Hitting Time (EHT) is one of the most critical concept in Markov Chains (Chen and Zhang 2008). A discrete-time Markov Chain is a kind of random process which involves dynamic systems represented as changing states along with time or steps. In Markov Chain, $\mathbf{T}_{pq}$ defines the direct transition probability between any of two states p and q, thus describes the second-order relationships between states. All these second-order transition probabilities form a matrix, i.e., the probability transition matrix of Markov Chain. Nevertheless, EHT implicitly makes use of the time-series relationships which is a kind of high-order interactions between state transitions. The EHT is defined as the average time or number of steps needed in all transitional paths from state p to q. For example, a transitional path from p to q can be with directly transiting p to q with probability $\mathbf{T}_{pq}$ and the corresponding hitting time equals to 1, or using other states as transients: state p can be transited to r and then to q with probability $\sum_r \mathbf{T}_{pr}\mathbf{T}_{rq}$ and the hitting time equals to 2, etc. $EHT_{p\to q}$ is the expected least efforts of transiting from state p to q, and can be further considered as the dissimilarity between these two states. When two states are with similar properties, the EHT can be hopefully smaller, and vice versa. Detailed definitions of EHT distance are described in Section 3.

In this work, we focus on the classification and information retrieval problems where features are represented as *histograms*, and propose a novel non-Mahalanobis distance function based on the Expected Hitting Time (EHT). It is notable that we de-emphasize the original physical meaning on time-series of EHT but only model the feature-feature interactions with the transition probabilities in Markov Chains, and then the feature dissimilarities or distances with high-order feature interactions can be naturally modeled by EHT. Note that histogram representing an instance can be regarded as a type of *distribution* on features, and then the distances between instances are also able to be represented with averaged EHT under certain distributions. Since EHT strongly depends on transition matrix $\mathbf{T}$, in order to accurately model the distances between instances with EHT, $\mathbf{T}$ should be learned with supervised information, by minimizing the EHT based dissimilarities/distance between instances within the same class and maximizing that with different labels. We present an effective approach named Learning Expected hitting time Distance (LED) for the transition matrix $\mathbf{T}$ learning based on first order optimization techniques. LED is further compared with existing state-of-the-art metric learning algorithms on real-world datasets. Experimental results reveal the effectiveness of LED. Our main contribution includes:

- A novel EHT based distance for histogram representations: EHT can measure the distances between instances represented by *histograms*.
- An effective distance learning approach: LED is brought forward to learn an appropriate transition matrix and achieve better generalization ability of EHT distances.

The paper is organized as follows: Section 2 is related work. The novel EHT distance and LED are described in Section 3 and Section 4 respectively. Section 5 contains the experiments and finally Section 6 concludes.

Related work

The basic idea of distance metric learning is to find a distance metric with which the distance between data points in the same class is smaller than that from different classes. Representative algorithms for distance metric learning include: maximally collapsing metric learning (Globerson and Roweis 2006), information-theoretic metric learning (Davis et al. 2007), large-margin nearest neighbors (Weinberger and Saul 2009), logistic discriminant metric learning (Guillaumin, Verbeek, and Schmid 2009), optimization equivalence of divergences improves neighbor embedding (Yang, Peltonen, and Kaski 2014). More researches on Mahalanobis metric learning can be found in surveys (Bellet, Habrard, and Sebban 2013; Kulis 2013).

In recent years, some Non-Mahalanobis based distances have also been proposed aiming at the nonlinearities of problems. Typical works include distance defined with CNN (Chopra, Hadsell, and LeCun 2005); instance specific DML (Frome et al. 2007; Zhan et al. 2009; Ye, Zhan, and Jiang 2016); class specific DML (Weinberger and Saul 2008). However, most of these methods are variants to Mahalanobis distances or Mahalanobis distances in localities, which cannot fundamentally get out of the barrier of problem in Fig 1.

The Expected Hitting Time (EHT), which depends on the transition probabilities between states, is an important concept in Markov chain. It is usually used for stochastic differential equations (Yamada 1983). It is also noteworthy that the EHT of reaching the optimal solution for the first time, i.e., the Expected First Hitting Time (EFHT), is a fundamental theoretical issue of evolutionary algorithms, in analogous to time complexity of deterministic algorithms. It has been shown that EFHT has close relation to the convergence rate, and a powerful theoretical analysis approach has been developed (Yu and Zhou 2008).

In this paper, we adopt EHT to measure the distance between instances represented by histogram-like features, and the EHT-based distance is able to capture higher-order feature interactions to some extent.

EHT based Distance

Histograms are constituted by frequencies, and as a matter of fact normalized histograms are feature distributions. Mahalanobis distance, however, cannot capture the divergences between distributions, which may lead to the dilemma illustrated in Fig. 1: neglecting the high-order feature interactions may consequently degenerate the final performance, and specifically designed histogram similarities/dissimilarities (Cha 2007; Kedem et al. 2012; Ma, Gu, and Wang 2010) including HIK/EMD arbitrarily assign distances with pre-defined functions without comprehensive reasons. We try to solve this problem with a novel proposed Learning Expected hitting time Distance (LED) approach

in this paper, and will first introduce the distance measure based on EHT followed by the concrete LED approach.

Histogram features as Markov Chain states

Without any loss of generality, suppose we have the set of labeled training examples denoted as $\{(\mathbf{x}_i, y_i)|i = 1, 2, \cdots, n\}$, where $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{1, 2, \cdots, c\}$, n is the number of training instances and c is the number of categories. Note that $\mathbf{x}_i$ is represented with d histogram features, and it is reasonable to assume the feature space $\mathbb{R}^d$ is a finite discrete set. $\mathbf{x}_i^s$ is the quantity/frequencies of the s-th feature of $\mathbf{x}_i$. It is obvious that $\mathbf{x}_i^s \geq 0$ for any s. We can also denote the must-link pairs and cannot-link pairs as in literatures: $\mathcal{S} : \{\mathbf{x}_i, \mathbf{x}_j\}$ if $y_i = y_j$, $\mathcal{D} : \{\mathbf{x}_i, \mathbf{x}_k\}$ if $y_i \neq y_k$.

In order to model the interactions between histogram features, Markov Chain is employed. One of the most important parts of Markov Chain is the probability transition matrix $\mathbf{T}$, in which the (p, q)-th element of $\mathbf{T}$ (denoted as $\mathbf{T}_{pq}$) is the transition probability from state p to q, and $\sum_q \mathbf{T}_{pq} = 1$. In this paper, we treat the states of Markov Chain process as single histogram features and the physical meaning of $\mathbf{T}_{pq}$ is the transiting probability of "moving one unit from feature p to feature q", where $\mathbf{T}_{pp} = 0$. It is obvious that $0 < p, q \leq d$. As a consequence, $\mathbf{T}_{pq}$ can be turned into a type of criterion for measuring difference between feature p and q, i.e., the first Hitting Time of transiting one unit of feature p to q, which can be written according to (Chen and Zhang 2008): $HT_{p\to q} = \min\{t : v_0 = p, v_t = q\}$, where v_0 indicates the start state of Markov Chain, and $v_0 = p$ reveals at time 0, the concerned instance $\mathbf{x}$ have the property that $\mathbf{x}^p = 1$ and $\mathbf{x}^s = 0$ ($s \neq p$). Similarly, we have v_t reflects the hitting state. Then $HT_{p\to q}$ reflects "the minimized efforts for moving one unit of feature from p to q". As a consequence, the expectation of $HT_{p\to q}$, furthermore, figures out the average efforts which are reasonable for measuring the divergence between two different histogram features.

Matching distribution to feature

Following the Theorem 1 inspired by (Chen and Zhang 2008), we can straightly get the expected efforts for transiting a given distribution $\mathbf{x}_a \in \mathbb{R}^d$ to feature s, that is $EHT_{\mathbf{x}_a\to s} = \mathbf{x}_a^\top(\mathbf{I} - \mathbf{T}_{-s})^{-1}\mathbf{1}$, where $\mathbf{T}_{-s}$ denotes the probability transition matrix obtained by letting the s-th row ($\mathbf{T}_{s\cdot}$) and the s-th column ($\mathbf{T}_{\cdot s}$) equal 0.

Theorem 1. *Providing a finite irreducible Markov Chain on state set V, $\forall s \in V$, the expected hitting time of transiting from a certain distribution $\mathbf{x}_a$ to state/feature s is :*

$$EHT_{\mathbf{x}_a\to s} = \mathbf{x}_a^\top(\mathbf{I} - \mathbf{T}_{-s})^{-1}\mathbf{1}, \tag{1}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{1}$ is all one vector.

It is notable that $(\frac{\mathbf{1}}{d})^\top(\mathbf{I} - \mathbf{T}_{-s})^{-1}\mathbf{1}$ equals the expected value for transiting uniform distribution to state/feature s, while the physical meaning behind Theorem 1 is by setting the s-th feature as the absorbing state, and $EHT_{\mathbf{x}_a\to s}$ is the expected value of $(\mathbf{I} - \mathbf{T}_{-s})^{-1}\mathbf{1}$ over distribution $\mathbf{x}_a$, i.e., Eq. 1 is the efforts should be made for transiting all components (frequencies or counts) in $\mathbf{x}_a$ to a single histogram feature s.

Matching distribution to distribution

Based on expectation value of $EHT_{\mathbf{x}_a\to s}$ providing the transitional efforts between distribution (or instance) $\mathbf{x}_a$ and features s, we can further define $EHT_{\mathbf{x}_i\to\mathbf{x}_j}$ by treating each element $\mathbf{x}_j^s$ within the distribution of $\mathbf{x}_j$ as the absorbing state s and then taking average, consequently we have:

$$EHT_{\mathbf{x}_i\to\mathbf{x}_j} = \sum_{s=1}^{d} \mathbf{x}_j^s \cdot EHT_{\mathbf{x}_i\to s}. \tag{2}$$

EHT based Distance

Eq. 2 defines the transitional efforts of transiting distribution of instance $\mathbf{x}_i$ to $\mathbf{x}_j$ and can be directly used for measuring the dissimilarity between these two instances, however, considering the substantive transiting are made on the differences between this instances pair, we can further define $\mathbf{x}_{i,j} = (\mathbf{x}_i - \mathbf{x}_j)_+$, where $\mathbf{z}_+$ is a vector function and $\mathbf{z}_+ = \max(\mathbf{0}, \mathbf{z})$; we then consider the $\mathrm{Dis}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{T})$ related to the efforts for transiting from $\mathbf{x}_{i,j}$ to $\mathbf{x}_{j,i}$ and vice versa, i.e., we have the EHT based distance defined as

$$\mathrm{Dis}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{T}) = EHT_{\mathbf{x}_{i,j}\to\mathbf{x}_{j,i}} + EHT_{\mathbf{x}_{j,i}\to\mathbf{x}_{i,j}}. \tag{3}$$

Here the bidirectional transitions are considered for the symmetry property for constructing a distance function, and the $EHT_{\mathbf{x}_{i,j}\to\mathbf{x}_{j,i}}$ is a distribution to distribution EHT distance, and the physical meaning of $EHT_{\mathbf{x}_{i,j}\to\mathbf{x}_{j,i}}$ is the average efforts of moving units from $\mathbf{x}_{i,j}$ to the opposite direction differences $\mathbf{x}_{j,i}$. To better illustrate the formation of EHT distance in Eq. 3, a concrete example with fixed transition matrix $\mathbf{T}$ is shown in Fig. 2. Fig. 2a gives the detailed transition diagram together with the transition probabilities in $\mathbf{T}$ (diag elements of $\mathbf{T}$ are configured as zeros and not shown), and the EHT between single features which are calculated according to (Chen and Zhang 2008). It is noteworthy that in Fig. 2d, there are four components, i.e., $EHT_{p\to q}$, etc., constituted the overall EHT distance, and each of these components should be weighted by $\mathbf{x}_j^s$ and $\mathbf{x}_a$ according to Eq. 2 and Eq. 1 accordingly.

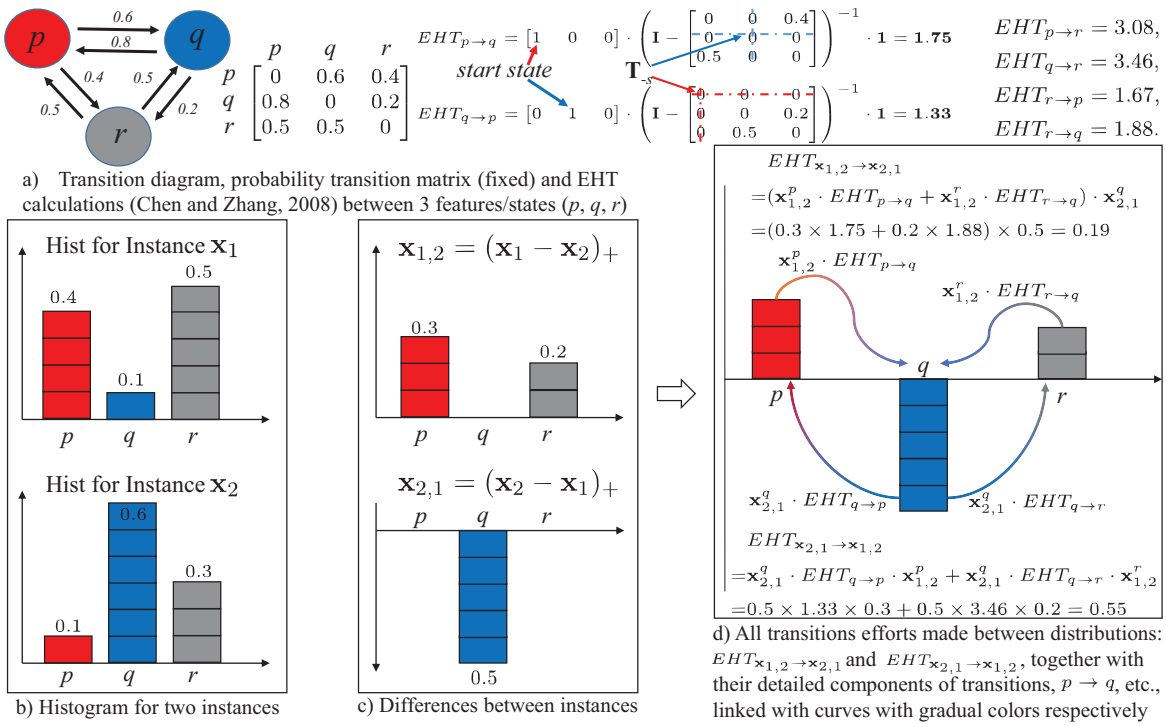
Note that the probability transition matrix $\mathbf{T}$ is the parameter and can strongly affect the EHT distance. Besides, in the transition matrix, $\mathbf{T}_{pq}$ indicates the "one step" feature interaction (transition), while the high-order interactions can be implicitly covered by the "expected multiple steps" in EHT distances/dissimilarities naturally.

Learning EHT Distance

In order to achieve better classification performance, classical distance metric learning approaches obtain the Mahalanobis distance metric with the help of side information. While for EHT based distance, it's a crucial problem how to get a good probability transition matrix $\mathbf{T}$ rather than the metric matrix $\mathbf{M}$. Instead of manually designing transition matrix according to feature similarities or correlations, in this work we try to "learn" the transition matrix automatically inspired by distance metric learning.

The formulation

We propose the Learning Expected hitting time Dissimilarity (LED) approach to update the transition probabilities by



a) Transition diagram, probability transition matrix (fixed) and EHT calculations (Chen and Zhang, 2008) between 3 features/states (p, q, r)

b) Histogram for two instances

c) Differences between instances

d) All transitions efforts made between distributions: $EHT_{\mathbf{x}_{1,2}\to\mathbf{x}_{2,1}}$ and $EHT_{\mathbf{x}_{2,1}\to\mathbf{x}_{1,2}}$, together with their detailed components of transitions, $p \to q$, etc., linked with curves with gradual colors respectively

Figure 2: A concrete example of EHT based distance calculation between two instances $\mathbf{x}_1$ and $\mathbf{x}_2$ with 3 features (p, q, r) given a fixed probability transitional matrix, and the $\text{Dis}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{T}) = EHT_{\mathbf{x}_{1,2}\to\mathbf{x}_{2,1}} + EHT_{\mathbf{x}_{2,1}\to\mathbf{x}_{1,2}} = 0.19{+}0.55 = 0.74$.

minimizing or maximizing the pairwised expected hitting time distance with side supervision information. Good distance metric always preserves the instances from the same class close and separates the instances from different classes apart. Inspired by this, we have a straight forward objective function for EHT distance based on probability transition matrix $\mathbf{T}$:

$$\arg\min_{\mathbf{T}} F(\mathbf{T}) = \sum_{\mathcal{S}} \text{Dis}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{T}) - \lambda \sum_{\mathcal{D}} \text{Dis}(\mathbf{x}_i, \mathbf{x}_k, \mathbf{T}),$$
$$\text{s.t.} \sum_q \mathbf{T}_{pq} = 1, \quad \mathbf{T}_{pq} \geq 0, \quad \mathbf{T}_{pp} = 0, \quad \forall p, q \leq d. \quad (4)$$

To simplify the consideration, we have $\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \epsilon, \|\mathbf{x}_i - \mathbf{x}_k\|_2^2 \leq \epsilon$, i.e., here we restrict on discussion in localities. Moreover, λ is a trade off parameter, which is fixed to 1 in our experiments and the constraints in Eq. 4 ensure $\mathbf{T}$ a valid probability transition matrix.

The optimization problem of Eq. 4 is non-convex, yet can be solved with gradient descent. However, according to Eq. 3, the objective goal in Eq. 4 contains $\mathbf{T}_{-s}$ $(s \leq d)$ rather than the whole transition matrix $\mathbf{T}$, and this becomes a barrier for optimizating the matrix $\mathbf{T}$. Considering in gradient descent, the main target focuses on the update of $\mathbf{T}^t$ (the superscript t indicates the iteration steps), we can update $\mathbf{T}_{-s}$ for each s and represent $\mathbf{T}^t$ with those $\mathbf{T}_{-s}$. For simplifying the discussion, we can denote $\mathbf{Y}_s = \mathbf{I} - \mathbf{T}_{-s}$, and update $\mathbf{Y}_s$ in each iteration first. Once we collect all updated $\mathbf{Y}_s^{t+1}$, we can obtain the updated $\hat{\mathbf{T}}^{t+1} = \frac{1}{d}\sum_{s=1}^{d}(\mathbf{I} - \mathbf{Y}_s^{t+1})$. By taking differential of the objective goal on $\mathbf{Y}_s$ for instance pairs $\mathbf{x}_i, \mathbf{x}_j$, we have the gradient $\nabla F(\mathbf{Y}_s)$ equals:

$$-{\mathbf{Y}_s}^{-1}\left(\sum_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{S}} \mathbf{v}(i,j) - \lambda \cdot \sum_{\mathbf{x}_i,\mathbf{x}_k\in\mathcal{D}} \mathbf{v}(i,k)\right)\mathbf{1}^\top \cdot \mathbf{Y_s}^{-1}, \quad (5)$$

where $\mathbf{v}(i,j)$ is defined as: $\mathbf{x}_{i,j}\mathbf{x}^s_{j,i} + \mathbf{x}_{j,i}\mathbf{x}^s_{i,j}$,

By line search on the gradient step size τ, we can first get the updated $\mathbf{Y}^{t+1}$ according to the derivative of objective function, and we project each row of the updated transition matrix $\hat{\mathbf{T}}$ into a simplex for obtaining the final $\mathbf{T}$ at the last iteration, which ensures the constraints on transition matrix $\mathbf{T}$, i.e., condition of $\sum_q \mathbf{T}_{pq} = 1, \mathbf{T}_{pq} \geq 0, \mathbf{T}_{pp} = 0$ hold.

SGD Acceleration

From Eq. 5, it can be found that the matrix inversion of $\mathbf{Y}_s$, $s \in \{1, \cdots, d\}$, will consume the computational resources most. In this section, we focus on this issue for reducing the time complexity with stochastic gradient optimization techniques in incremental learning settings.

Different from batch updating, in incremental learning configuration, instances come one by one and only a few of instance pairs can be used at t. We can therefore rewrite the stochastic gradient version of LED by constructing the stochastic gradient approximation $\gamma F(\mathbf{Y}_s)$ of $\nabla F(\mathbf{Y}_s)$. We sample a training example pair $\{\mathbf{x}_i, \mathbf{x}_j\}$ according to a Multinomial distribution Multi $(\gamma_{1,1}, \ldots, \gamma_{n,k})$, where k is the fixed size of neighborhoods, and compute $\gamma_{i,j}$ as:

$$\gamma_{i,j}(F) = -\theta(i,j){\mathbf{Y}_s}^{-1}\mathbf{v}(i,j)\mathbf{1}^\top\mathbf{Y_s}^{-1}, \quad (6)$$

Algorithm 1 LED-SGD pseudo code

Input: $\mathbf{x}_i$: Instances with histogram features, $i = 1, 2, \cdots, n$;
τ: Step size;
ε: Neighborhood threshold;
λ: Trade off;
$t_{\max}$: the maximum iteration number
Output: updated transition matrix $\mathbf{T}^{t+1}$

Initialize $\mathbf{T}^1$ with random matrix, setting $\mathbf{T}^1_{pp} = 0$
while true do
  $t = t + 1$
  for each $s \in [1, d]$ **do**
    $\mathbf{Y}_s^{t+1} = \mathbf{I} - \mathbf{T}^t_{-s}$
    for each $\| x_i - x_j \| \leq \varepsilon$ **do**
      Update $(\mathbf{Y}_s^{t+1})^{-1}$ with Eq. 8
    end for
    $\hat{\mathbf{T}}^{t+1} = \frac{1}{d} \sum_{s=1}^{d} (\mathbf{I} - \mathbf{Y}_s^{t+1})$
  end for
  Project each row of $\hat{\mathbf{T}}^{t+1}$ to be a simplex to obtain the transition matrix $\mathbf{T}^{t+1}$
  if $t > t_{\max}$ **then**
    break
  end if
end while
return $\mathbf{T}^{t+1}$;

where $\theta(i, j)$ is a scale function which equals 1 if $\{\mathbf{x}_i, \mathbf{x}_j\} \in \mathcal{S}$, and $-\lambda$ if $\{\mathbf{x}_i, \mathbf{x}_j\} \in \mathcal{D}$. As a consequence, the update rule for $\mathbf{Y}_s$ becomes:

$$\begin{aligned} \mathbf{Y}_s^{t+1} &= \mathbf{Y}_s^t - \tau \gamma_{i,j}(F) \\ &= \mathbf{Y}_s^t + \tau \theta(i,j) (\mathbf{Y}_s^t)^{-1} \mathbf{v}(i,j) \mathbf{1} (\mathbf{Y}_\mathbf{s}^\mathbf{t})^{-\mathbf{1}} \\ &= \mathbf{Y}_s^t + \tau \theta(i,j) (\mathbf{Y}_s^t)^{-1} \left[\mathbf{x}_{i,j} \ \mathbf{x}_{j,i} \right] \\ &\quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x}_{j,i}^s \mathbf{1}^\top \\ \mathbf{x}_{i,j}^s \mathbf{1}^\top \end{bmatrix} (\mathbf{Y}_s^t)^{-1}. \end{aligned} \tag{7}$$

For simplification of the discussion, we can substitute some variables in Eq. 7 as:

$$\mathbf{U}^t = (\mathbf{Y}_s^t)^{-1} \left[\mathbf{x}_{i,j} \ \ \mathbf{x}_{j,i} \right],$$

$$\mathbf{V}^t = \begin{bmatrix} \mathbf{x}_{j,i}^s \mathbf{1}^\top \\ \mathbf{x}_{i,j}^s \mathbf{1}^\top \end{bmatrix} (\mathbf{Y}_s^t)^{-1},$$

where $\mathbf{U} \in \mathbb{R}^{d \times 2}$ while $\mathbf{V} \in \mathbb{R}^{2 \times d}$. From Eq. 7, it clearly shows that the stochastic gradient of $\mathbf{Y}_s$ on each instance pair $\mathbf{x}_i, \mathbf{x}_j$ in Eq. 6 is a matrix of rank 2. More important, we observe that only the inverse of $\mathbf{Y}_s$ are actually required in calculating the stochastic gradient approximation, therefore following the Sherman-Morrison-Woodbury formula, we have the update rule rewritten as:

$$(\mathbf{Y}_s^{t+1})^{-1} = (\mathbf{Y}_s^t)^{-1} - (\mathbf{Y}_s^t)^{-1} \mathbf{U}^t \mathbf{Q}^{-1} \mathbf{V}^t (\mathbf{Y}_s^t)^{-1}, \tag{8}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{Q} = (\mathbf{I} + \mathbf{V}^t (\mathbf{Y}_s^t)^{-1} \mathbf{U}^t)$.

With the low rank update rule described in Eq. 8, the stochastic gradient updating can be carried out efficiently. Procedures for stochastic version of LED are listed in Algorithm 1. With the transition matrix $\mathbf{T}$ obtained by LED or LED-SGD, we can predict new instances with distance based classifiers, such as k-NN.

Table 1: Brief dataset description including dataset type, number of instances and number of classes.

Name	Type	#instances	#classes
PASCAL-VOC2009	Image	7818	20
CalTech101	Image	9144	102
sub-ImageNet	Image	119236	193
R8 of Reuters21578	Text	21578	8
20Newsgroups	Text	16926	20

Experiments

In this section, we evaluate the proposed LED approaches by comparing them with 5 state-of-the-art distance metric learning methods. Datasets used in our empirical investigations are from image and text classification problems. In detail, 3 image and 2 text datasets are used in our experiments, and the brief datesets descriptions are summarized in Table 1. For PASCAL-VOC2009, the goal is to recognize objects from 20 visual object classes in realistic scenes. CalTech101 collects pictures of objects from 102 categories and contains 40 to 800 images per category. Reuters-21578 is a collection of documents that published on Reuters Newswire in 1987, from which a subset of 8 classes according to (Cachopo 2007) is used in our experiments. 20Newsgroups is a collection of newsgroup documents from 20 different newsgroups, each of which corresponds to a different topic. Histogram features are extracted for both image and text datasets, i.e., for PASCAL-VOC2009, 1000 BOW features are extracted; while 1000 BOW features and 1500 SPM features are extracted from CalTech-101; 105 HOG+HSV features for subset of imageNet and 1500 TF (term frequency) features for both of the text datasets. BOW and SPM are extracted based on sparse sampling SIFT descriptors (Lowe 1999). All histogram features are normalized. For facilitating the discussion, we denote the dataset in specific configurations with the notation of "dataset name-feature type", e.g., 'CalTech101-BOW' in result tables and figures.

In order to validate the effectiveness of LED, the 5 state-of-the-art distance metric learning methods compared are:

- Information Theoretic Metric Learning (ITML) (Davis et al. 2007) which minimizes differential relative entropy on the distance functions;
- Logistic Discriminant Metric Learning (LDML) (Guillaumin, Verbeek, and Schmid 2009) which uses logistic discriminant to learn a metric from labeled instance pairs;
- Laplacian Regularized Metric Learning (LRML) (Hoi, Liu, and Chang 2010) which learns robust distance metrics in an effective graph regularization framework.
- Parametric Local Metric Learning (PLML) (Wang, Kalousis, and Woznica 2012) which learns a smooth metric matrix function over the data manifolds defined by local metric;
- Divergences Neighbor Embedding (DNE) (Yang, Peltonen, and Kaski 2014) which minimizes information divergences in generalized stochastic neighbor embedding.

Table 2: Accuracy (Avg.± Std.) comparisons with 6 other approaches on datasets with histogram feature representations. The best performance is marked with bold.

Datasets	k-NN	DNE	ITML	PLML	LDML	LRML	LED	LED-SGD
CalTech101-BOW	.153±.005	.132±.007	.384±.008	.411±.012	.424± .017	.086± .002	**.448±.022**	.444±.020
CalTech101-SPM	.501±.009	.587±.015	.688±.009	.645±.011	.604± .003	.281± .007	**.732±.005**	.731±.008
ImageNet-HOG	.315±.005	.071±.060	.242±.008	.138±.009	.275± .004	.184± .003	.322±.006	**.325±.005**
VOC2009-BOW	.529±.012	.155±.002	.597±.013	.627±.009	.602± .016	.337± .027	**.629±.011**	.628±.010
20NewsGroups-TF	.481±.021	.136±.014	.663±.034	.649±.014	.632± .031	.203± .042	**.671±.007**	.670±.007
Reuters-TF	.920±.004	.795±.028	.940±.004	.938±.004	.944± .004	.815± .008	**.946±.003**	.940±.004

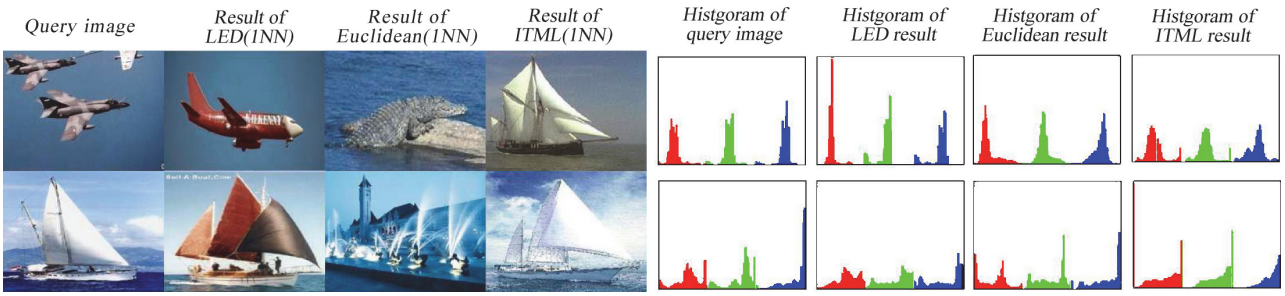


Figure 3: CBIR retrieved results (LED, Euclidean and ITML) and histograms. The first column is two random picked up query images, column 2–4 are top-1 retrieved results for compared methods, remain columns are RGB histograms for query or retrieved images. LED learned EHT distance can retrieve accurate objects (plane, sailboat) with different dominant hue.

Experimental results on k-NN with Euclidean distance are also listed as baseline in our experiments. All compared implementations are with the form as their respective literatures reported. 5-NN is utilized as the final classifier, while all other parameters of compared methods are tuned according to the original reports respectively. For LED and LED-SGD, the probability transition matrix $\mathbf{T}$ is initialized randomly and projected to a simplex described by the constraints in Eq. 4. Step size is tuned with line search, and λ is simply configured to 1 for default.

2/3 instances are used as training examples while remains are used for testing. Experiments on each dataset are performed 30 times with random splits. Average classification accuracies together with the standard derivations are recorded in Table 2. The best result is marked with bold.

Table 2 clearly shows, both of the proposed methods, i.e., LED and LED-SGD, are superior to those compared approaches obviously. In detail, LED outperforms other 7 compared methods, including LED-SGD, on 5 datasets, i.e., *CalTech101* with BOW features and SPM features, *VOC2009* with BOW features, and both of the text datasets (*20NewsGroups* and *Reuters*); LED-SGD outperforms other 7 compared approaches on *sub-ImageNet* with HOG features. Moreover, LEG-SGD is superior to other 5 state-of-the-art distance metric learning methods and k-NN on 4 datasets, including *CalTech101* with BOW and SPM, *VOC2009* with BOW and *20Newsgroups*. Experimental results have validated the effectiveness of LED/LED-SGD.

Table 3: Words corresponding to the top $\mathbf{T}_{pq}$ in transition matrix learned by LED on text data.

Word (q)	Words (p) with top transition probability
time	speed, scale, earlier, quality
people	company, official, responsibility, hurt
government	illegal, shooting, nsa, vote
phone	Hd, game, location, voice
windows	computing, network, electronics, systems

In order to reveal the superiority of LED comprehensively, we conduct more investigations on both image and text datasets. On *CalTech101* with RGB color histogram, we perform random queries, and due to the page limits, Fig. 3 only gives the top-1 (1-NN) images retrieved by LED, Euclidean and ITML metric on two of the random queries, each of which is plotted in a row. The x-axis of RGB color histograms in Fig. 3 are ordered by color Red, Green and Blue. From the visualization effects, we can find LED can pick up the CBIR results with totally different dominant hues nonetheless including similar semantical meanings, while other distance metrics lean to choose images with similar RGB histograms.

Since the physical meanings of color histograms for images is not intuitively expressed, to reveal the under-

lying reasons why LED and EHT distance can perform well, we carry out a further investigation on text dataset (20Newsgroups-TF). The transition probabilities learned by LED are recorded. We randomly pick up 5 different words (features q), and then list 4 words (features p) which are with top-4 $\mathbf{T}_{pq}$ for each q. In this way, the pairs of feature p and q are strongly related with high-order interactions. The results are shown in Table 3. Considering the comprehensibility of texts, the inner relationships between word p and q in Table 3 are apparent: taking the word "government" for example, the words with top transition probabilities are "illegal", "shooting", "nsa" and "vote", which often appear in the category of "talk.politics.guns" and belong to the topic focused by "gevernment". This phenomenon indicates that the transition probabilities learned by LED are meaningful.

Conclusion

In this paper, we figure out the natural weakness of Mahalanobis metric when instances are represented by histogram features, and then propose a novel non-Mahalanobis distance learning approach, LED, based on the expected hitting time (EHT) distances which can utilize the high-order interactions between histogram features. In EHT distance, feature relationships are modeled by Markov Chain and only depend on the probability transition matrix, therefore the main target of LED is seeking for an appropriate transition matrix for better EHT distance to improve the generalization ability of successive classifiers.

It is noteworthy that the LED can be solved with stochastic gradient descend and the efficiency is guaranteed by the low-rank update in each iteration. Our experiments on image and text datasets reveal the effectiveness of LED comparing to 5 state-of-the-art distance metric learning methods. Additional experiments on discovering the comprehensibility and underlying reasons for superiorities of LED are also performed and the results demonstrate that the transition matrix learned by LED can provide meaningful results.

The drawback of EHT based distance is it could be non-metric and may consequently lead to loss of potential good properties. How to leverage the advantages of EHT and Mahalanobis metric in a learning framework should be investigated in future. Besides, EHT itself has the abilities on modeling time-series information which are neglected in this work, further discussions on how to explore these abilities for distributional feature measuring can be an interesting work as well.

References

Bellet, A.; Habrard, A.; and Sebban, M. 2013. A survey on metric learning for feature vectors and structured data. *Preprint ArXiv:1306.6709*.

Cachopo, A. M. 2007. *Improving Methods for Single-label Text Categorization*. Ph.D. Dissertation, Univ. Técnica de Lisboa.

Cha, S.-H. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *Int. Joul. Math. Models and Methods in Applied Sciences* 1(4):300–307.

Chen, H., and Zhang, F. 2008. The expected hitting times for finite markov chains. *Linea. Alge. and its Appl.* 428(11):2730–2749.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proc. CVPR*, 539–546.

Davis, J. V.; Kulis, B.; Jain, P.; Sra, S.; and Dhillon, I. S. 2007. Information-theoretic metric learning. In *Proc. ICML*, 209–216.

Frome, A.; Singer, Y.; Sha, F.; and Malik, J. 2007. Learning globally consistent local distance functions for shape-based image retrieval and classification. In *Proc. ICCV*, 1–8.

Globerson, A., and Roweis, S. T. 2006. Metric learning by collapsing classes. In *NIPS 19*. Cambridge: MIT Press. 451–458.

Guillaumin, M.; Verbeek, J.; and Schmid, C. 2009. Is that you? metric learning approaches for face identification. In *Proc. ICCV*, 498–505.

Hoi, S. C.; Liu, W.; and Chang, S.-F. 2010. Semi-supervised distance metric learning for collaborative image retrieval and clustering. *ACM Trans. Multimedia Comp., Comm., Appl.* 6(3):18–44.

Hu, J.; Lu, J.; and Tan, Y.-P. 2014. Discriminative deep metric learning for face verification in the wild. In *Proc. CVPR*, 1875–1882.

Kedem, D.; Tyree, S.; Sha, F.; Lanckriet, G. R.; and Weinberger, K. Q. 2012. Non-linear metric learning. In *NIPS 25*. Cambridge: MIT Press. 2573–2581.

Kulis, B. 2013. Metric learning: A survey. *Foundations and Trends in Mach. Learn.* 5(4):287–364.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, 2169–2178.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *Proc. ICCV*, 1150–1157.

Ma, Y.; Gu, X.; and Wang, Y. 2010. Histogram similarity measure using variable bin size distance. *CVIU* 114(8):981–989.

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

Shi, Y.; Bellet, A.; and Sha, F. 2014. Sparse compositional metric learning. *Preprint ArXiv:1404.4105*.

Sivic, J., and Zisserman, A. 2003. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 1470–1477.

Wang, J.; Kalousis, A.; and Woznica, A. 2012. Parametric local metric learning for nearest neighbor classification. In *NIPS 25*. Cambridge: MIT Press. 1610–1618.

Weinberger, K. Q., and Saul, L. K. 2008. Fast solvers and efficient implementations for distance metric learning. In *Proc. ICML*, 1160–1167.

Weinberger, K. Q., and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *JMLR* 10:207–244.

Yamada, K. 1983. A bound for the expected hitting time of storage processes. *Stochas. Proce. and their Appl.* 14(1):93–105.

Yang, Z.; Peltonen, J.; and Kaski, S. 2014. Optimization equivalence of divergences improves neighbor embedding. In *Proc. ICML*, 460–468.

Ye, H.-J.; Zhan, D.-C.; and Jiang, Y. 2016. Instance specific metric subspace learning: A bayesian approach. In *Proc. AAAI*.

Yu, Y., and Zhou, Z.-H. 2008. A new approach to estimating the expected first hitting time of evolutionary algorithms. *AIJ* 172(15):1809–1832.

Zhan, D.-C.; Li, M.; Li, Y.-F.; and Zhou, Z.-H. 2009. Learning instance specific distances using metric propagation. In *Proc. ICML*, 1225–1232.