Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Low-Rank Similarity Metric Learning in High Dimensions

Wei Liu[†] **Cun Mu**[‡] **Rongrong Ji**[♮] **Shiqian Ma**[§] **John R. Smith**[†] **Shih-Fu Chang**[‡]
[†]IBM T. J. Watson Research Center [‡]Columbia University [♮]Xiamen University [§]The Chinese University of Hong Kong
{weiliu,jsmith}@us.ibm.com cm3052@columbia.edu sfchang@ee.columbia.edu
rrji@xmu.edu.cn sqma@se.cuhk.edu.hk

Abstract

Metric learning has become a widespreadly used tool in machine learning. To reduce expensive costs brought in by increasing dimensionality, low-rank metric learning arises as it can be more economical in storage and computation. However, existing low-rank metric learning algorithms usually adopt nonconvex objectives, and are hence sensitive to the choice of a heuristic low-rank basis. In this paper, we propose a novel low-rank metric learning algorithm to yield bilinear similarity functions. This algorithm scales linearly with input dimensionality in both space and time, therefore applicable to high-dimensional data domains. A convex objective free of heuristics is formulated by leveraging trace norm regularization to promote low-rankness. Crucially, we prove that all globally optimal metric solutions must retain a certain low-rank structure, which enables our algorithm to decompose the high-dimensional learning task into two steps: an SVD-based projection and a metric learning problem with reduced dimensionality. The latter step can be tackled efficiently through employing a linearized Alternating Direction Method of Multipliers. The efficacy of the proposed algorithm is demonstrated through experiments performed on four benchmark datasets with tens of thousands of dimensions.

Introduction

During the past decade, distance metric learning, typically referring to learning a Mahalanobis distance metric, has become ubiquitous in a variety of applications stemming from machine learning, data mining, information retrieval, and computer vision. Many fundamental machine learning algorithms such as K-means clustering and k-nearest neighbor classification all need an appropriate distance function to measure pairwise affinity (or closeness) between data examples. Beyond the naive Euclidean distance, the distance metric learning problem intends to seek a better metric through learning with a training dataset subject to particular constraints arising from fully supervised or semi-supervised information, such that the learned metric can reflect domain-specific characteristics of data affinities or relationships. A vast number of distance metric learning approaches have been proposed, and the representatives include supervised metric learning which is the main focus in the literature, semi-supervised metric learning (Wu et al. 2009; Hoi, Liu, and Chang 2010; Liu et al. 2010a; 2010b; Niu et al. 2012), *etc*. The suite of supervised metric learning can further be divided into three categories in terms of forms of supervision: the methods supervised by instance-level labels (Goldberger, Roweis, and Salakhutdinov 2004; Globerson and Roweis 2005; Weinberger and Saul 2009), the methods supervised by pair-level labels (Xing et al. 2002; Bar-Hillel et al. 2005; Davis et al. 2007; Ying and Li 2012), and the methods supervised by triplet-level ranks (Ying, Huang, and Campbell 2009; McFee and Lanckriet 2010; Shen et al. 2012; Lim, McFee, and Lanckriet 2013). More methods can also be found in two surveys (Kulis 2012) and (Bellet, Habrard, and Sebban 2013).

Despite the large amount of literature, relatively little work has dealt with the problem where the dimensionality of input data can be extremely high. This is relevant to a wide variety of practical data collections, *e.g.*, images, videos, documents, time-series, genomes, and so on, which frequently involve from tens to hundreds of thousands of dimensions. Unfortunately, learning full-rank metrics in a high-dimensional input space $\mathbb{R}^d$ quickly becomes computationally prohibitive due to high computational complexities ranging from $O(d^2)$ to $O(d^{6.5})$. Additionally, limited memory makes storing a huge metric matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$ a heavy burden.

This paper concentrates on a more reasonable learning paradigm, *Low-Rank Metric Learning*, also known as structural metric learning. (Davis and Dhillon 2008) demonstrated that in the small sample setting $n \ll d$ (n is the number of training samples) learning a low-rank metric in a high-dimensional input space is tractable, achieving linear space and time complexities with respect to the dimensionality d. We follow this learning setting $n \ll d$ throughout the paper. Given a rank $r \ll d$, a distance metric can be expressed into a low-rank form $\mathbf{M} = \mathbf{L}\mathbf{L}^\top$ which is parameterized by a rectangular matrix $\mathbf{L} \in \mathbb{R}^{d\times r}$. $\mathbf{L}$ constitutes a low-rank basis in $\mathbb{R}^d$, and also acts as a dimensionality-reducing linear transformation since any distance between two inputs $\boldsymbol{x}$ and $\boldsymbol{x}'$ under such a low-rank metric $\mathbf{M}$ can be equivalently viewed as a Euclidean distance between the transformed inputs $\mathbf{L}^\top\boldsymbol{x}$ and $\mathbf{L}^\top\boldsymbol{x}'$. Therefore, the use of low-rank metrics allows reduced storage of metric matrices (only saving

$\mathbf{L}$) and simultaneously enables efficient distance computation ($O(dr)$ time). Despite the benefits of using low-rank metrics, the existing methods most adopt nonconvex optimization objectives like (Goldberger, Roweis, and Salakhutdinov 2004; Torresani and Lee 2006; Mensink et al. 2013; Lim and Lanckriet 2014), and are thus confined to the subspace spanned by a heuristic low-rank basis $\mathbf{L}^0$ even including the convex method (Davis and Dhillon 2008). In terms of the training costs, most low-rank metric learning approaches including (Goldberger, Roweis, and Salakhutdinov 2004; Torresani and Lee 2006; McFee and Lanckriet 2010; Shen et al. 2012; Kunapuli and Shavlik 2012; Lim, McFee, and Lanckriet 2013) yet incur quadratic and cubic time complexities in the dimensionality d. A few methods such as (Davis and Dhillon 2008) and (Mensink et al. 2013) are known to provide efficient $O(d)$ procedures, but they both rely on and are sensitive to an initialization of the heuristic low-rank basis $\mathbf{L}^0$.

This work intends to pursue a convex low-rank metric learning approach without resorting to any heuristics. Recently, one trend in similarity learning (Kar and Jain 2011; 2012; Chechik et al. 2010; Crammer and Chechik 2012; Cheng 2013) emerges, which attempts to learn similarity functions in the form of $\mathcal{S}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \mathbf{M} \boldsymbol{x}'$ and appears to be more flexible than distance metric learning. In light of this trend, we incorporate the benefits of both low-rank metrics and similarity functions to learn a low-rank metric $\mathbf{M}$ that yields a discriminative similarity function $\mathcal{S}(\cdot)$. Specifically, we propose a novel low-rank metric learning algorithm by designing a convex objective which consists of a hinge loss enforced over pairwise supervised information and a trace norm regularization penalty promoting low-rankness of the desired metric. Significantly, we show that minimizing such an objective guarantees to obtain an optimal solution of a certain low-rank structure as $\mathbf{M}^\star = \mathbf{U}\mathbf{W}^\star\mathbf{U}^\top$. The low-rank basis $\mathbf{U}$ included in any optimal metric $\mathbf{M}^\star$ turns out to lie in the column (or range) space of the input data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, and $\mathbf{W}^\star$ can thus be sought within a fixed subspace whose dimensionality is much lower than the original dimensionality.

By virtue of this key finding, our algorithm intelligently decomposes the challenging high-dimensional learning task into two steps which are both computationally tractable and cost a total $O(d)$ time. The first step is dimensionality reduction via SVD-based projection, and the second step is lower dimensional metric learning by applying a linearized modification of the Alternating Direction Method of Multipliers (ADMM) (Boyd et al. 2011). The linearized ADMM maintains analytic-form updates across all iterations, works efficiently, and converges rapidly. We evaluate the proposed algorithm on four benchmark datasets with tens of thousands of dimensions, demonstrating that the low-rank metrics obtained by our algorithm accomplish prominent performance gains in terms of kNN classification.

Low-Rank Similarity Metrics

Different from most previous metric learning approaches that used the learned metric into a distance function like $\sqrt{(\boldsymbol{x}-\boldsymbol{x}')^\top \mathbf{M}(\boldsymbol{x}-\boldsymbol{x}')}$, in this paper we pursue a low-rank metric $\mathbf{M} = \mathbf{L}\mathbf{L}^\top$ for a *bilinear similarity* function to measure the similarity between two arbitrary inputs $\boldsymbol{x}$ and $\boldsymbol{x}'$ in $\mathbb{R}^d$:

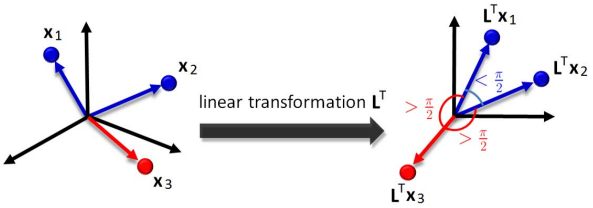

Figure 1: The effect of a desired low-rank similarity metric that induces a linear transformation $\mathbf{L}^\top$. Points with the same color denote similar data examples, while points with different colors denote dissimilar data examples.

$$\begin{aligned}\mathcal{S}_{\mathbf{M}}(\boldsymbol{x}, \boldsymbol{x}') &= \boldsymbol{x}^\top \mathbf{M} \boldsymbol{x}' = \boldsymbol{x}^\top \mathbf{L}\mathbf{L}^\top \boldsymbol{x}' \\ &= \left(\mathbf{L}^\top \boldsymbol{x}\right)^\top \left(\mathbf{L}^\top \boldsymbol{x}'\right), \end{aligned} \tag{1}$$

which is parameterized by a low-rank basis $\mathbf{L} \in \mathbb{R}^{d\times r}$ with the rank $r \ll d$. Naturally, such a similarity function characterized by a low-rank metric essentially measures the inner-product between two transformed inputs $\mathbf{L}^\top \boldsymbol{x}$ and $\mathbf{L}^\top \boldsymbol{x}'$ in a lower dimensional space $\mathbb{R}^r$.

Motivated by (Kriegel, Schubert, and Zimek 2008) which presented a statistical means based on *angles* rather than distances to identify outliers scattering in a high-dimensional space, we would think that angles own a stronger discriminating power than distances in capturing affinities among data points in high dimensions. Suppose that we are provided with a set of pairwise labels $\{y_{ij} \in \{1, -1\}\}$ of which $y_{ij} = 1$ stands for a similar data pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ while $y_{ij} = -1$ for a dissimilar pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The purpose of our supervised similarity metric learning is to make an *acute* angle between any similar pair whereas an *obtuse* angle between any dissimilar pair after applying the learned metric $\mathbf{M}$ or equivalently mapping data points from $\boldsymbol{x}$ to $\mathbf{L}^\top \boldsymbol{x}$. Formally, this purpose is expressed as

$$\mathcal{S}_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} > 0, & y_{ij} = 1, \\ < 0, & y_{ij} = -1. \end{cases} \tag{2}$$

Fig. 1 exemplifies our notion for those angles between similar and dissimilar data pairs.

Learning Framework

Let us inherit the conventional deployment of pairwise supervision adopted by most metric learning methods, *e.g.*, (Xing et al. 2002; Bar-Hillel et al. 2005; Davis et al. 2007; Davis and Dhillon 2008; Wu et al. 2009; Hoi, Liu, and Chang 2010; Liu et al. 2010a; 2010b; Niu et al. 2012; Ying and Li 2012). This paper tackles a supervised similarity metric learning problem provided with a set of labeled pairs $\mathcal{I} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j, y_{ij})\}$. Suppose that n ($\ll d$) data examples $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$ are gathered in the set $\mathcal{I}$. Inspired by the aforementioned angle-based learning principle in Eq. (2), we design a margin criterion over the similarity

outputs of the supervised pairs, that is, $\mathcal{S}_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 1$ for $y_{ij} = 1$ and $\mathcal{S}_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq -\epsilon$ (ϵ is a small positive value) for $y_{ij} = -1$. Note that setting the margin values to 1 and -1 for positive and negative pairs in the context of similarity learning may be infeasible for multi-class data. For example, there are data points $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$ from three different classes. Imposing $\mathcal{S}_{\mathbf{M}}(\boldsymbol{x}_1, \boldsymbol{x}_2) \leq -1$, $\mathcal{S}_{\mathbf{M}}(\boldsymbol{x}_1, \boldsymbol{x}_3) \leq -1$, and $\mathcal{S}_{\mathbf{M}}(\boldsymbol{x}_2, \boldsymbol{x}_3) \leq -1$ is problematic since the former two likely lead to $\mathcal{S}_{\mathbf{M}}(\boldsymbol{x}_2, \boldsymbol{x}_3) > 0$.

This similarity margin criterion ensures that after applying the linear transformation $\mathbf{L}^\top$ induced by a desired metric $\mathbf{M}$, the angle between a similar pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is acute since their inner-product $\left(\mathbf{L}^\top \boldsymbol{x}_i\right)^\top \left(\mathbf{L}^\top \boldsymbol{x}_j\right) = \mathcal{S}_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is large enough, and meanwhile the angle between a dissimilar pair is obtuse since their inner-product is negative.

In order to promote low-rankness of the target metric, we leverage nuclear norm $\|\mathbf{M}\|_*$, the sum of all singular values of $\mathbf{M}$, as a regularization penalty in our learning framework. Nuclear norm has been corroborated to be able to encourage low-rank matrix structure in the matrix completion literature (Fazel, Hindi, and Boyd 2001; Candès and Recht 2009; Recht, Fazel, and Parrilo 2010). Moreover, since our framework constrains $\mathbf{M}$ to be in the positive semidefinite (PSD) cone $\mathbb{S}_+^d$ to make $\mathbf{M}$ a valid metric, the nuclear norm $\|\mathbf{M}\|_*$ can be replaced by the trace norm $\mathrm{tr}(\mathbf{M})$.

Integrating the angle principle and the trace norm regularizer, the objective of our proposed *Low-Rank Similarity Metric Learning* (LRSML) framework is formulated as

$$\min_{\mathbf{M} \in \mathbb{S}_+^d} f(\mathbf{M}) := \sum_{i,j=1}^{n} \left[\tilde{y}_{ij} - y_{ij} \mathcal{S}_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right]_+ + \alpha \mathrm{tr}(\mathbf{M}), \tag{3}$$

where $\tilde{y}_{ij} = \begin{cases} \epsilon, & y_{ij} = -1 \\ y_{ij}, \text{otherwise} \end{cases}$, $[x]_+ = \max(0, x)$, and $\alpha > 0$ is the regularization parameter. Above all, the objective in Eq. (3) is convex and does not introduce any heuristic low-rank basis which was required by the prior convex method (Davis and Dhillon 2008) and those nonconvex methods (Goldberger, Roweis, and Salakhutdinov 2004; Torresani and Lee 2006; Mensink et al. 2013; Lim and Lanckriet 2014) in low-rank distance metric learning. Second, minimizing the hinge loss in Eq. (3) enforces the margin criterion on the similarity function $\mathcal{S}_{\mathbf{M}}$ so that the dissimilar pairs can be obviously distinguished from the similar ones. It is worth clarifying that we do not pursue $\mathcal{S}_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0$ (*i.e.*, $\epsilon = 0$) in the case of $y_{ij} = -1$ in Eq. (2) because it may lead to the trivial solution $\mathbf{M} = \mathbf{0}$ to Eq. (3).[1]

Due to the space limit, all the proofs of lemmas and theorems presented in this section are placed in the supplemental material.

[1]We observed the trivial solution $\mathbf{M} = \mathbf{0}$ in the case of $\epsilon = 0$ when the training data examples are fewer and the regularization parameter α is relatively large.

Optimality Characterization

Despite being convex, the framework in Eq. (3) is not easy to directly tackle when the data points of $\mathcal{X}$ live in a space of high dimensions, *i.e.*, d is very large. Under the high-dimensional scenario, basic matrix manipulations upon $\mathbf{M}$ are too expensive to execute. For example, projecting $\mathbf{M}$ onto $\mathbb{S}_+^d$ costs $O(d^3)$. Nonetheless, by delving into problem (3) we can characterize a certain low-rank structure of its all optimal solutions. In particular, such a low-rank structure can be proven to be the form of $\mathbf{U}\mathbf{W}\mathbf{U}^\top$, in which $\mathbf{U}$ is a low-rank basis lying in the column space of the data matrix $\mathbf{X}$.

Let us do the singular value decomposition (SVD) over the data matrix $\mathbf{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$, obtaining $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \sum_{i=1}^{m} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, where m ($\leq n \ll d$) is the rank of $\mathbf{X}$, $\sigma_1, \cdots, \sigma_m$ are the positive singular values, $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \cdots, \sigma_m)$, and $\mathbf{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m] \in \mathbb{R}^{d \times m}$ and $\mathbf{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m] \in \mathbb{R}^{n \times m}$ are the matrices composed of left- and right-singular vectors, respectively. It is obvious that $\mathrm{range}(\mathbf{X}) \equiv \mathrm{range}(\mathbf{U})$, in which we write $\mathrm{range}(\mathbf{A}) = \mathrm{span}\big(\{\boldsymbol{a}_i\}_i\big)$ as the column space of a given matrix $\mathbf{A}$. Significantly, the optimality of problem (3) can be characterized as follows.

Lemma 1. *For any optimal solution $\mathbf{M}^\star$ to problem (3), we have $\mathbf{M}^\star \in \left\{\mathbf{U}\mathbf{W}\mathbf{U}^\top | \mathbf{W} \in \mathbb{S}_+^m\right\}$.*

Lemma 1 reveals that although we are seeking $\mathbf{M}^\star$ in $\mathbb{S}_+^d$, any optimal solution $\mathbf{M}^\star$ must obey the certain low-rank structure $\mathbf{U}\mathbf{W}\mathbf{U}^\top$ because of the nature of problem (3). This important discovery enables us to first project the data onto the low-rank basis $\mathbf{U}$ and then seek $\mathbf{W}$ via optimizing a new problem in Eq. (4) which is at a much smaller scale than the raw problem in Eq. (3).

Theorem 1. *For any $i \in [1:n]$, denote $\tilde{\boldsymbol{x}}_i = \mathbf{U}^\top \boldsymbol{x}_i \in \mathbb{R}^m$. Then $\mathbf{M}^\star$ is an optimal solution to problem (3) if and only if $\mathbf{M}^\star = \mathbf{U}\mathbf{W}^\star\mathbf{U}^\top$ in which $\mathbf{W}^\star$ belongs to the set*

$$\arg\min_{\mathbf{W} \in \mathbb{S}_+^m} \tilde{f}(\mathbf{W}) := \sum_{i,j=1}^{n} \left[\tilde{y}_{ij} - y_{ij} \tilde{\boldsymbol{x}}_i^\top \mathbf{W} \tilde{\boldsymbol{x}}_j\right]_+ + \alpha \mathrm{tr}(\mathbf{W}). \tag{4}$$

Theorem 1 indicates that through a simple projection step ($\boldsymbol{x}_i \to \mathbf{U}^\top \boldsymbol{x}_i$), we are able to reduce the high-dimensional metric learning task in Eq. (3) to a lower dimensional metric learning problem in Eq. (4) with guarantees, thereby greatly reducing the computational costs. However, the scale of problem (4) could still be not neglectable, which severely limits the use of the off-the-shell interior point based solvers such as SDPT3 (Toh, Todd, and Tütüncü 1999). In the next subsection, we will devise an efficient first-order method to solve problem (4).

Similar projection tricks have been recently exploited and proven for convex distance metric learning problems regularized by squared Frobenius norm (Chatpatanasiri et al. 2010) and several other strictly convex functions (Kulis 2012), whose proof strategies, however, cannot carry over to trace norm since trace norm is not strictly convex. To the best of our knowledge, the structural representation $\mathbf{U}\mathbf{W}^\star\mathbf{U}^\top$

characterizing optimal metric solutions is firstly explored here for trace norm regularized convex metric learning.

Linearized ADMM

Now we develop an efficient optimization method to exactly solve Eq. (4). To facilitate derivations, we define three matrix operators: the Hadamard product $\circ$, the inner-product $\langle\cdot,\cdot\rangle$, and the Frobenius norm $\|\cdot\|_{\mathrm{F}}$.

We first introduce a dummy variable matrix $\mathbf{Z}=[z_{ij}]_{ij}\in\mathbb{R}^{n\times n}$ to equivalently rewrite problem (4) as

$$\begin{aligned}\min_{\mathbf{W}\in\mathbb{S}_+^m,\mathbf{Z}}\ &\sum_{i,j=1}^{n}[z_{ij}]_+ + \alpha\mathrm{tr}(\mathbf{W}) \\ \text{s.t.}\quad &\mathbf{Z}-\tilde{\mathbf{Y}}+\mathbf{Y}\circ\big(\tilde{\mathbf{X}}^\top\mathbf{W}\tilde{\mathbf{X}}\big)=\mathbf{0},\end{aligned} \tag{5}$$

in which three constant matrices $\tilde{\mathbf{X}}=[\tilde{\boldsymbol{x}}_1,\cdots,\tilde{\boldsymbol{x}}_n]=\mathbf{U}^\top\mathbf{X}\in\mathbb{R}^{m\times n}$, $\mathbf{Y}=[y_{ij}]_{ij}\in\mathbb{R}^{n\times n}$, and $\tilde{\mathbf{Y}}=[\tilde{y}_{ij}]_{ij}\in\mathbb{R}^{n\times n}$ are involved. Either of the objective function and constraint in Eq. (5) is separable in terms of two variables $\mathbf{Z}$ and $\mathbf{W}$, which naturally suggests a use of alternating direction methods. To this end, we employ the increasingly popular Alternating Direction Method of Multipliers (ADMM) (Boyd et al. 2011) to cope with Eq. (5). ADMM works on the following augmented Lagrangian function:

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{Z},\mathbf{W};\mathbf{\Lambda}) :=\ &\sum_{i,j=1}^{n}[z_{ij}]_+ + \alpha\mathrm{tr}(\mathbf{W}) \\ &+\Big\langle\mathbf{\Lambda},\mathbf{Z}-\tilde{\mathbf{Y}}+\mathbf{Y}\circ\big(\tilde{\mathbf{X}}^\top\mathbf{W}\tilde{\mathbf{X}}\big)\Big\rangle \\ &+\frac{\rho}{2}\Big\|\mathbf{Z}-\tilde{\mathbf{Y}}+\mathbf{Y}\circ\big(\tilde{\mathbf{X}}^\top\mathbf{W}\tilde{\mathbf{X}}\big)\Big\|_{\mathrm{F}}^2,\end{aligned} \tag{6}$$

where $\mathbf{\Lambda}=[\lambda_{ij}]_{ij}\in\mathbb{R}^{n\times n}$ is the multiplier for the linear constraint in Eq. (5) (*a.k.a.* dual variable), and $\rho>0$ is the penalty parameter. ADMM will proceed through successively and alternatingly updating the three variable matrices $\mathbf{Z}^k$, $\mathbf{W}^k$ and $\mathbf{\Lambda}^k$ ($k=0,1,\cdots$).

$\mathbf{Z}$, $\mathbf{\Lambda}$-Updates

Here we first present the updates for $\mathbf{Z}$ and $\mathbf{\Lambda}$ which are easier to achieve than the update for $\mathbf{W}$.

The $\mathbf{Z}$-update is achieved by minimizing $\mathcal{L}_\rho(\mathbf{Z},\mathbf{W};\mathbf{\Lambda})$ with respect to $\mathbf{Z}$ while keeping $\mathbf{W}$ and $\mathbf{\Lambda}$ fixed, that is,

$$\begin{aligned}\mathbf{Z}^{k+1} :=\arg\min_{\mathbf{Z}}\mathcal{L}_\rho\big(\mathbf{Z},\mathbf{W}^k;\mathbf{\Lambda}^k\big) &= \arg\min_{\mathbf{Z}}\sum_{i,j=1}^{n}[z_{ij}]_+ \\ &+\frac{\rho}{2}\Big\|\mathbf{Z}-\tilde{\mathbf{Y}}+\mathbf{Y}\circ\big(\tilde{\mathbf{X}}^\top\mathbf{W}^k\tilde{\mathbf{X}}\big)+\mathbf{\Lambda}^k/\rho\Big\|_{\mathrm{F}}^2.\end{aligned}$$

To solve this $\mathbf{Z}$ subproblem, we utilize the proximal operator $\mathcal{T}_\theta(a)=\arg\min_{z\in\mathbb{R}}\theta[z]_+ +\frac{1}{2}(z-a)^2$ ($\theta>0$) to handle $[z_{ij}]_+$. $\mathcal{T}_\theta(a)$ is a proximal mapping of $\theta[z]_+$ and has shown in (Ye, Chen, and Xie 2009) to output as follows

$$\mathcal{T}_\theta(a)=\begin{cases}a-\theta, & a>\theta,\\ 0, & 0\le a\le\theta,\\ a, & a<0.\end{cases} \tag{7}$$

By applying $\mathcal{T}_{1/\rho}$ elementwisely, we obtain a closed-form solution to the $\mathbf{Z}$ subproblem as

$$\mathbf{Z}^{k+1} := \mathcal{T}_{1/\rho}\big(\tilde{\mathbf{Y}}-\mathbf{Y}\circ(\tilde{\mathbf{X}}^\top\mathbf{W}^k\tilde{\mathbf{X}})-\mathbf{\Lambda}^k/\rho\big). \tag{8}$$

The $\mathbf{\Lambda}$-update is prescribed to be

$$\mathbf{\Lambda}^{k+1} := \mathbf{\Lambda}^k+\rho\big(\mathbf{Z}^{k+1}-\tilde{\mathbf{Y}}+\mathbf{Y}\circ(\tilde{\mathbf{X}}^\top\mathbf{W}^{k+1}\tilde{\mathbf{X}})\big). \tag{9}$$

$\mathbf{W}$-Update

In the standard ADMM framework, $\mathbf{W}$ is updated by minimizing $\mathcal{L}_\rho(\mathbf{Z},\mathbf{W};\mathbf{\Lambda})$ with respect to $\mathbf{W}$ while $\mathbf{Z}$ and $\mathbf{\Lambda}$ are fixed. Noticing $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top=\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}=\mathbf{\Sigma}^2$, we derive the $\mathbf{W}$-update as follows

$$\begin{aligned}\mathbf{W}^{k+1} &:= \arg\min_{\mathbf{W}\in\mathbb{S}_+^m}\mathcal{L}_\rho\big(\mathbf{Z}^{k+1},\mathbf{W};\mathbf{\Lambda}^k\big) \\ &= \arg\min_{\mathbf{W}\in\mathbb{S}_+^m}\alpha\mathrm{tr}(\mathbf{W})+\Big\langle\mathbf{\Lambda}^k,\mathbf{Y}\circ\big(\tilde{\mathbf{X}}^\top\mathbf{W}\tilde{\mathbf{X}}\big)\Big\rangle \\ &\quad+\frac{\rho}{2}\Big\|\mathbf{Z}^{k+1}-\tilde{\mathbf{Y}}+\mathbf{Y}\circ\big(\tilde{\mathbf{X}}^\top\mathbf{W}\tilde{\mathbf{X}}\big)\Big\|_{\mathrm{F}}^2 \\ &= \arg\min_{\mathbf{W}\in\mathbb{S}_+^m}\frac{\rho}{2}\big\|\tilde{\mathbf{X}}^\top\mathbf{W}\tilde{\mathbf{X}}\big\|_{\mathrm{F}}^2+\langle\alpha\mathbf{I},\mathbf{W}\rangle \\ &\quad+\Big\langle\mathbf{\Lambda}^k+\rho\mathbf{Z}^{k+1}-\rho\tilde{\mathbf{Y}},\mathbf{Y}\circ\big(\tilde{\mathbf{X}}^\top\mathbf{W}\tilde{\mathbf{X}}\big)\Big\rangle \\ &= \arg\min_{\mathbf{W}\in\mathbb{S}_+^m}\frac{\rho}{2}\big\langle\mathbf{W}\mathbf{\Sigma}^2,\mathbf{\Sigma}^2\mathbf{W}\big\rangle+\langle\alpha\mathbf{I},\mathbf{W}\rangle \\ &\quad+\Big\langle\tilde{\mathbf{X}}\big(\mathbf{Y}\circ(\mathbf{\Lambda}^k+\rho\mathbf{Z}^{k+1}-\rho\tilde{\mathbf{Y}})\big)\tilde{\mathbf{X}}^\top,\mathbf{W}\Big\rangle \\ &= \arg\min_{\mathbf{W}\in\mathbb{S}_+^m}\frac{1}{2}\big\langle\mathbf{W}\mathbf{\Sigma}^2,\mathbf{\Sigma}^2\mathbf{W}\big\rangle \\ &\quad+\Big\langle\alpha\mathbf{I}/\rho+\tilde{\mathbf{X}}\big(\mathbf{Y}\circ(\mathbf{Z}^{k+1}-\tilde{\mathbf{Y}}+\mathbf{\Lambda}^k/\rho)\big)\tilde{\mathbf{X}}^\top,\mathbf{W}\Big\rangle \\ &= \arg\min_{\mathbf{W}\in\mathbb{S}_+^m}g\big(\mathbf{W};\mathbf{Z}^{k+1},\mathbf{\Lambda}^k\big),\end{aligned} \tag{10}$$

where $\mathbf{I}$ denotes the $m\times m$ identity matrix. Unfortunately, there is no analytical solution available for this $\mathbf{W}$ subproblem in Eq. (10). Trying iterative optimization methods to solve Eq. (10) would, however, limit the efficiency of ADMM. In what follows, we develop a linearized modification of the original ADMM framework in Eqs. (8)(10)(9) by rectifying the $\mathbf{W}$-update in Eq. (10) to an easier one.

Linearization. Rather than dealing with intrinsically quadratic $g\big(\mathbf{W};\mathbf{Z}^{k+1},\mathbf{\Lambda}^k\big)$, we resort to optimizing a surrogate function $\widehat{g}_\tau\big(\mathbf{W};\mathbf{Z}^{k+1},\mathbf{\Lambda}^k\big)$ which is the linear approximation to g and augmented by a quadratic proximal term.

Concretely, the rectified $\mathbf{W}$-update is shown as

$$\begin{aligned}
\mathbf{W}^{k+1} &:= \arg\min_{\mathbf{W}\in\mathbb{S}_+^m} \widehat{g}_\tau\big(\mathbf{W};\mathbf{Z}^{k+1},\mathbf{\Lambda}^k\big) \\
&= \arg\min_{\mathbf{W}\in\mathbb{S}_+^m} \frac{1}{2\tau}\big\|\mathbf{W}-\mathbf{W}^k\big\|_{\mathrm{F}}^2 + g\big(\mathbf{W}^k;\mathbf{Z}^{k+1},\mathbf{\Lambda}^k\big) + \\
&\quad + \left\langle \nabla_{\mathbf{W}} g\big(\mathbf{W};\mathbf{Z}^{k+1},\mathbf{\Lambda}^k\big)\Big|_{\mathbf{W}=\mathbf{W}^k}, \mathbf{W}-\mathbf{W}^k \right\rangle \\
&= \arg\min_{\mathbf{W}\in\mathbb{S}_+^m} \big\langle \mathbf{G}^k, \mathbf{W}-\mathbf{W}^k\big\rangle + \frac{1}{2\tau}\big\|\mathbf{W}-\mathbf{W}^k\big\|_{\mathrm{F}}^2 \\
&= \arg\min_{\mathbf{W}\in\mathbb{S}_+^m} \frac{1}{2\tau}\big\|\mathbf{W}-\big(\mathbf{W}^k-\tau\mathbf{G}^k\big)\big\|_{\mathrm{F}}^2 \\
&= \mathcal{P}_{\mathbb{S}_+^m}\big(\mathbf{W}^k-\tau\mathbf{G}^k\big). \qquad (11)
\end{aligned}$$

In Eq. (11) we introduce the matrix

$$\begin{aligned}
\mathbf{G}^k &= \nabla_{\mathbf{W}} g\big(\mathbf{W};\mathbf{Z}^{k+1},\mathbf{\Lambda}^k\big)\Big|_{\mathbf{W}=\mathbf{W}^k} \qquad (12)\\
&= \alpha\mathbf{I}/\rho + \tilde{\mathbf{X}}\big(\mathbf{Y}\circ(\mathbf{Z}^{k+1}-\tilde{\mathbf{Y}}+\mathbf{\Lambda}^k/\rho)\big)\tilde{\mathbf{X}}^\top + \mathbf{\Sigma}^2\mathbf{W}^k\mathbf{\Sigma}^2.
\end{aligned}$$

The introduced operator $\mathcal{P}_{\mathbb{S}_+^m}(\cdot)$ represents the projection operation onto the PSD cone $\mathbb{S}_+^m$; for example, $\mathcal{P}_{\mathbb{S}_+^m}(\mathbf{A}) = \sum_{i=1}^m [\gamma_i]_+ \boldsymbol{p}_i\boldsymbol{p}_i^\top$ for any symmetric matrix $\mathbf{A}\in\mathbb{R}^{m\times m}$ whose eigenvector-eigenvalue pairs are $(\boldsymbol{p}_i,\gamma_i)_{i=1}^m$. The introduced parameter $\tau>0$ is the step size which is to be specified later.

The advantage of linearizing $g\big(\mathbf{W};\mathbf{Z}^{k+1},\mathbf{\Lambda}^k\big)$ is that the resulting $\mathbf{W}$-update in Eq. (11) enjoys an analytical expression and consequently works very efficiently. This modification of the standard ADMM, *i.e.*, optimizing $\widehat{g}_\tau$ instead of g, is the core ingredient of our developed linearized ADMM framework, namely *linearized ADMM*. The following theorem certifies that the developed linearized ADMM is theoretically sound, producing a globally optimal solution to the raw problem in Eq. (5).

Theorem 2. *Given* $0<\tau<\frac{1}{\|\tilde{\mathbf{X}}\|_{op}^4}=\frac{1}{\|\mathbf{X}\|_{op}^4}$, *the sequence* $\big\{\big(\mathbf{Z}^k,\mathbf{W}^k,\mathbf{\Lambda}^k\big)\big\}_k$ *generated by the linearized ADMM in Eqs. (8)(11)(9) starting with any symmetric* $\big(\mathbf{Z}^0,\mathbf{W}^0,\mathbf{\Lambda}^0\big)$ *converges to an optimal solution of problem (5).*

Note that $\|\cdot\|_{op}$ denotes the operator norm of matrices. For simplicity, we initialize $\mathbf{W}^0=\mathbf{I}_{m\times m}$ and $\mathbf{\Lambda}^0=\mathbf{0}_{n\times n}$ to launch the linearized ADMM.

Algorithm

We summarize the proposed *Low-Rank Similarity Metric Learning* (LRSML) approach in Algorithm 1, which consists of two steps: 1) SVD-based projection, and 2) lower dimensional metric learning via the linearized ADMM. In the projection step, the SVD of $\mathbf{X}$ can be efficiently performed in $O(dn^2)$ time by running the eigen-decomposition over the matrix $\mathbf{X}^\top\mathbf{X}$. If the data vectors in $\mathbf{X}$ are zero-centered, such a decomposition is exactly PCA. In the linearized ADMM step, we use the scaled dual variable matrix $\tilde{\mathbf{\Lambda}}^k=\mathbf{\Lambda}^k/\rho$. Importantly, we claim that there is no need to explicitly compute and even store the large $d\times d$ metric matrix $\mathbf{M}$; instead, it only needs to store the low-rank basis $\mathbf{U}$ and update the much smaller $m\times m$ metric matrix $\mathbf{W}^k$, as indicated by Theorem 1. In doing so, LRSML circumvents expensive matrix manipulations such as projecting $\mathbf{M}$ onto $\mathbb{S}_+^d$, therefore accomplishing considerable savings in both storage and computation. The total time complexity of LRSML is bounded by $O(dn^2+Tn^3)$ which scales linearly with the input dimensionality d. In practice, we find that the linearized ADMM converges rapidly within $T=1,000$ iterations under the setting of $\rho=1,\tau=0.01$ (see the convergence curves in the supplemental material).

Algorithm 1 Low-Rank Similarity Metric Learning

Input: the data matrix $\mathbf{X}\in\mathbb{R}^{d\times n}$ ($d\gg n$), the pairwise label matrix $\mathbf{Y}\in\{1,-1\}^{n\times n}$, the regularization parameter $\alpha>0$, three positive constants ϵ,ρ,τ, the budget iteration number T.

SVD Projection: perform the SVD of $\mathbf{X}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ to obtain $\mathbf{U}\in\mathbb{R}^{d\times m}$, $\mathbf{\Sigma}\in\mathbb{R}_+^{m\times m}$, and compute the projected data $\tilde{\mathbf{X}}=\mathbf{U}^\top\mathbf{X}\in\mathbb{R}^{m\times n}$.

Linearized ADMM:

form the matrix $\tilde{\mathbf{Y}}\in\{1,\epsilon\}^{n\times n}$ according to $\mathbf{Y}$;

initialize $\mathbf{W}^0=\mathbf{I}_{m\times m}$, $\tilde{\mathbf{\Lambda}}^0=\mathbf{0}_{n\times n}$, $\mathbf{S}^0=\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$;

for $k=1,\cdots,T$ **do**

$\mathbf{Z}^k \longleftarrow \mathcal{T}_{1/\rho}(\tilde{\mathbf{Y}}-\mathbf{Y}\circ\mathbf{S}^{k-1}-\tilde{\mathbf{\Lambda}}^{k-1})$,

$\mathbf{G}^{k-1} \longleftarrow \frac{\alpha}{\rho}\mathbf{I}_{m\times m}+\tilde{\mathbf{X}}(\mathbf{Y}\circ(\mathbf{Z}^k-\tilde{\mathbf{Y}}+\tilde{\mathbf{\Lambda}}^{k-1}))\tilde{\mathbf{X}}^\top+\mathbf{\Sigma}^2\mathbf{W}^{k-1}\mathbf{\Sigma}^2$,

$\mathbf{W}^k \longleftarrow \mathcal{P}_{\mathbb{S}_+^m}(\mathbf{W}^{k-1}-\tau\mathbf{G}^{k-1})$, $\mathbf{S}^k \longleftarrow \tilde{\mathbf{X}}^\top\mathbf{W}^k\tilde{\mathbf{X}}$,

$\tilde{\mathbf{\Lambda}}^k \longleftarrow \tilde{\mathbf{\Lambda}}^{k-1}+\mathbf{Z}^k-\tilde{\mathbf{Y}}+\mathbf{Y}\circ\mathbf{S}^k$,

if $\|\mathbf{Z}^k-\tilde{\mathbf{Y}}+\mathbf{Y}\circ\mathbf{S}^k\|_{\mathrm{F}}^2/n^2\le 10^{-12}$ **then**

break,

end if

end for.

Output: run the eigen-decomposition of $\mathbf{W}^k=\mathbf{H}\mathbf{E}\mathbf{H}^\top$ in which $\mathbf{E}\in\mathbb{R}^{r\times r}$ retains the positive eigenvalues, and then output the optimal low-rank basis $\mathbf{L}^\star=\mathbf{U}\mathbf{H}\mathbf{E}^{1/2}\in\mathbb{R}^{d\times r}$ (the corresponding low-rank metric is $\mathbf{M}^\star=\mathbf{L}^\star\mathbf{L}^{\star\top}$).

Discussion

A primary advantage of our analysis in Lemma 1 and Theorem 1 is avoiding the expensive projection onto the high-dimensional PSD cone $\mathbb{S}_+^d$, which was required by the previous trace norm regularized metric learning methods such as (McFee and Lanckriet 2010; Lim, McFee, and Lanckriet 2013). In the supplemental material, we further provide in-depth theoretic analysis (see Lemma 3 and Theorem 3) to comprehensively justify the low-rank solution structure $\mathbf{M}^\star=\mathbf{U}\mathbf{W}^\star\mathbf{U}^\top$ for any convex loss function in terms of $\boldsymbol{x}_i^\top\mathbf{M}\boldsymbol{x}_j$ regularized by trace norm $\mathrm{tr}(\mathbf{M})$ or squared Frobenius norm $\|\mathbf{M}\|_{\mathrm{F}}^2$. As a result, our analysis would directly lead to scalable $O(d)$ algorithms for a task of low-rank distance or similarity metric learning supervised by instance-level, pairwise, or listwise label information. For example, our analysis would give an $O(d)$-time algorithm for optimizing the low-rank distance metric learning objective (a hinge loss based on listwise supervision plus a trace norm regularizer) in (McFee and Lanckriet 2010) through following our proposed two-step scheme, SVD projection + lower dimensional metric learning.

We remark here that linearized ADMM techniques have been investigated in the optimization community, and applied to the problems arising from compressed sensing (Yang and Zhang 2011), image processing (Zhang et al.

Table 1: Basic descriptions of four datasets used in the experiments.

Dataset	# Classes	# Samples	# Dimensions
Reuters-28	28	8,030	18,933
TDT2-30	30	9,394	36,771
UIUC-Sports	8	1,579	87,040
UIUC-Scene	15	4,485	21,504

2010), and nuclear norm regularized convex optimization (Yang and Yuan 2013). Nevertheless, the linearized ADMM method that we developed in this paper is self-motivated in the sense that it was derived for solving a particular optimization problem of low-rank metric learning. To the best of our knowledge, linearized ADMM has not been applied to metric learning before, and the analytic-form updates produced by our linearized ADMM method are quite advantageous. Note that (Lim, McFee, and Lanckriet 2013) applied the standard ADMM to solve a subproblem in every iteration of their optimization method, but executing ADMM at only one time costs the cubic time complexity $O(T'd^3)$ (T' is the iteration number of ADMM) which is computationally infeasible in high dimensions. By contrast, our linearized ADMM tackles the entire metric optimization in $O(Tn^3)$ time, performing at least one order of magnitude faster than the standard ADMM by virtue of the elegant $\mathbf{W}$-update.

Experiments

We carry out the experiments on four benchmark datasets including two document datasets **Reuters-28** and **TDT2-30** (Cai, He, and Han 2011), and two image datasets **UIUC-Sports** (Li and Fei-Fei 2007) and **UIUC-Scene** (Lazebnik, Schmid, and Ponce 2006). Note that we choose 28 categories in the raw Reuters-21578 corpus (Cai, He, and Han 2011) such that each category has no less than 20 examples. In **Reuters-28** and **TDT2-30**, each document is represented by an ℓ_2 normalized TF-IDF feature vector; in **UIUC-Sports** and **UIUC-Scene**, each image is represented by an ℓ_2 normalized sparse-coding feature vector (Wang et al. 2010). The basic information about these four datasets is shown in Table 1.

The baseline method is named as "Original", which takes original feature vectors for 1NN classification, and the linear SVM is also included in comparison. Here we evaluate and compare eight metric and similarity learning methods including: two classical linear dimensionality reduction methods Latent Semantic Analysis (LSA) (Deerwester et al. 1990) and Fisher Linear Discriminant Analysis (FLDA) (Hastie, Tibshirani, and Friedman 2009), two low-rank variants LSA-ITML and CM-ITML (Davis and Dhillon 2008) of a representative distance metric learning method Information-Theoretic Metric Learning (ITML) (Davis et al. 2007) (LSA-ITML uses LSA's output as its heuristic low-rank basis, while CM-ITML uses the class means as its heuristic low-rank basis), two recent low-rank distance metric learning methods Metric Learning to Rank (MLR) (McFee and Lanckriet 2010) and Metric Learning for the Nearest Class Mean classifier (MLNCM) (Mensink et al. 2013) (MLNCM also uses LSA's output as its heuristic low-rank basis), a similarity learning method Adaptive Regularization Of MAtrix models (AROMA) (Crammer and Chechik 2012), and the Low-Rank Similarity Metric Learning (LRSML) method proposed in this paper. Except MLNCM which requires to use the Nearest Class Mean classifier, all the compared metric/similarity learning methods collaborate with the 1NN classifier. Note that we choose to run LSA instead of PCA because of the nonnegative and sparse nature of the used feature vectors. LSA-ITML, CM-ITML, MLR and LRSML are convex methods, while MLNCM is a nonconvex method. We choose the most efficient version of AROMA, which restricts the matrix $\mathbf{M}$ used in the similarity function to be diagonal. To efficiently implement MLR, we follow the two-step scheme, which has been justified in this paper, SVD projection followed by MLR with lower dimensionality.

Except AROMA which does not impose the PSD constraint on $\mathbf{M}$, each of these referred methods can yield a low-rank basis $\mathbf{L}$. For them, we try three measures between two inputs $\boldsymbol{x}$ and $\boldsymbol{x}'$: (i) *distance* $\left\|\mathbf{L}^\top\boldsymbol{x}-\mathbf{L}^\top\boldsymbol{x}'\right\|$, (ii) *inner-product* $\left(\mathbf{L}^\top\boldsymbol{x}\right)^\top\left(\mathbf{L}^\top\boldsymbol{x}'\right)$, and (iii) *cosine* $\frac{(\mathbf{L}^\top\boldsymbol{x})^\top(\mathbf{L}^\top\boldsymbol{x}')}{\|\mathbf{L}^\top\boldsymbol{x}\|\|\mathbf{L}^\top\boldsymbol{x}'\|}$. For LSA, FLDA, LSA-ITML, CM-ITML, MLR and MLNCM, distance and cosine measures are tried; while for LRSML, inner-product and cosine measures are tried since the inner-product measure (ii) is exactly the bilinear similarity function shown in Eq. (1). The baseline "Original" gives the same results under the three measures as the feature vectors are ℓ_2 normalized already.

Let C be the number of classes in every dataset. On **Reuters-28** and **TDT2-30**, we select $5\times C$ up to $30\times C$ samples for training such that each category covers at least one sample; we pick up the same number of samples for cross-validation; the rest of samples are for testing. We repeat 20 times of random training/validation/test splits, and then report the average classification error rates and training time for all the competing methods in Tables 2 and 3. On the two image datasets, we follow the commonly used evaluation protocols like (Lazebnik, Schmid, and Ponce 2006; Wang et al. 2010). On **UIUC-Sports**, we select 10 to 70 samples per class for training, a half number of samples for validation, and the remaining ones for testing; on **UIUC-Scene**, we obey the similar setting and the training samples range from 10 to 100 per class. By running 20 times, the average of per-class recognition rates as well as the training time are reported in the tables of the supplemental material. To run our proposed method LRSML, we fix $\epsilon=0.1,\rho=1$, and find that $\tau=0.01$ makes the linearized ADMM converge within $T=1,000$ iterations on all datasets. The results shown in all the referred tables demonstrate that LRSML consistently leads to the highest accuracy in terms of 1NN classification. Across all datasets, the best two results are achieved by LRSML with two measures; the cosine measure almost results in slightly higher classification accuracy than the inner-product measure. More experimental results are presented in the supplemental material.

Conclusions

The proposed LRSML formulated a trace norm regularized convex objective which was proven to yield a globally optimal solution with a certain low-rank structure. With the optimality guarantee, the challenging high-dimensional metric learning task can be reduced to a lower dimensional metric learning problem after a simple SVD projection. The linearized ADMM was further developed to efficiently solve the reduced problem. Our approach bears out to maintain linear space and time complexities in the input dimensionality, consequently scalable to high-dimensional data domains. Through extensive experiments carried out on four benchmark datasets with up to 87,000 dimensions, we certify that our learned low-rank similarity metrics are well suited to high-dimensional problems and exhibit prominent performance gains in kNN classification. In future work, we would like to investigate generalization properties of the proposed LRSML like (Cao, Guo, and Ying 2012), and aim to accomplish a tighter generalization bound for trace norm regularized convex metric learning.

Acknowledgments

When doing an earlier version of this work, Wei Liu was supported by Josef Raviv Memorial Postdoctoral Fellowship. The authors would like to thank Prof. Gang Hua and Dr. Fuxin Li for constructive suggestions which are helpful to making the final version.

References

Bar-Hillel, A.; Hertz, T.; Shental, N.; and Weinshall, D. 2005. Learning a mahalanobis metric from equivalence constraints. *JMLR* 6:937–965.

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

Boyd, S. P.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* 3(1):1–122.

Cai, D.; He, X.; and Han, J. 2011. Locally consistent concept factorization for document clustering. *IEEE Transactions on Knowledge and Data Engineering* 23(6):902–913.

Candès, E. J., and Recht, B. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics* 9(6):717–772.

Cao, Q.; Guo, Z.-C.; and Ying, Y. 2012. Generalization bounds for metric and similarity learning. Technical report.

Chatpatanasiri, R.; Korsrilabutr, T.; Tangchanachaianan, P.; and Kijsirikul, B. 2010. A new kernelization framework for mahalanobis distance learning algorithms. *Neurocomputing* 73(10):1570–1579.

Chechik, G.; Sharma, V.; Shalit, U.; and Bengio, S. 2010. Large scale online learning of image similarity through ranking. *JMLR* 11:1109–1135.

Cheng, L. 2013. Riemannian similarity learning. In *Proc. ICML*.

Crammer, K., and Chechik, G. 2012. Adaptive regularization for weight matrices. In *Proc. ICML*.

Davis, J. V., and Dhillon, I. S. 2008. Structured metric learning for high dimensional problems. In *Proc. KDD*.

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

Deerwester, S. C.; Dumais, S. T.; Landauer, T. K.; Furnas, G. W.; and Harshman, R. A. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41(6):391–407.

Fazel, M.; Hindi, H.; and Boyd, S. P. 2001. A rank minimization heuristic with application to minimum order system approximation. In *Proc. the American Control Conference*, volume 6, 4734–4739.

Globerson, A., and Roweis, S. 2005. Metric learning by collapsing classes. In *NIPS 18*.

Goldberger, J.; Roweis, S.; and Salakhutdinov, R. 2004. Neighbourhood components analysis. In *NIPS 17*.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition, Springer.

Hoi, S. C.; Liu, W.; and Chang, S.-F. 2010. Semi-supervised distance metric learning for collaborative image retrieval and clustering. *ACM Transactions on Multimedia Computing, Communications and Applications* 6(3).

Kar, P., and Jain, P. 2011. Similarity-based learning via data driven embeddings. In *NIPS 24*.

Kar, P., and Jain, P. 2012. Supervised learning with similarity functions. In *NIPS 25*.

Kriegel, H.-P.; Schubert, M.; and Zimek, A. 2008. Angle-based outlier detection in high-dimensional data. In *Proc. KDD*.

Kulis, B. 2012. Metric learning: a survey. *Foundations and Trends® in Machine Learning* 5(4):287–364.

Kunapuli, G., and Shavlik, J. 2012. Mirror descent for metric learning: A unified approach. In *Proc. ECML*.

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

Li, L.-J., and Fei-Fei, L. 2007. What, where and who? classifying event by scene and object recognition. In *Proc. ICCV*.

Lim, D. K. H., and Lanckriet, G. 2014. Efficient learning of mahalanobis metrics for ranking. In *Proc. ICML*.

Lim, D. K. H.; McFee, B.; and Lanckriet, G. 2013. Robust structural metric learning. In *Proc. ICML*.

Liu, W.; Ma, S.; Tao, D.; Liu, J.; and Liu, P. 2010a. Semi-supervised sparse metric learning using alternating linearization optimization. In *Proc. KDD*.

Liu, W.; Tian, X.; Tao, D.; and Liu, J. 2010b. Constrained metric learning via distance gap maximization. In *Proc. AAAI*.

McFee, B., and Lanckriet, G. 2010. Metric learning to rank. In *Proc. ICML*.

Mensink, T.; Verbeek, J.; Perronnin, F.; and Csurka, G. 2013. Distance-based image classification: Generalizing to new classes at near zero cost. *TPAMI* 35(11):2624–2637.

Niu, G.; Dai, B.; Yamada, M.; and Sugiyama, M. 2012. Information-theoretic semi-supervised metric learning via entropy regularization. In *Proc. ICML*.

Recht, B.; Fazel, M.; and Parrilo, P. A. 2010. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review* 52(3):471–501.

Shen, C.; Kim, J.; Wang, L.; and van den Hengel, A. 2012. Positive semidefinite metric learning using boosting-like algorithms. *JMLR* 13:1007–1036.

Table 2: **Reuters-28** dataset: classification error rates and training time of eight competing metric/similarity learning methods.

	5×28 training samples			30×28 training samples		
Method	**Measure**	**Error Rate** (%)	**Train Time** (sec)	**Measure**	**Error Rate** (%)	**Train Time** (sec)
Original	–	23.94±1.63	–	–	16.56±0.55	–
Linear SVM	–	21.25±1.46	0.03	–	9.16±0.43	0.12
LSA	distance	42.21 ±4.48	0.02	distance	24.56±5.30	0.55
	cosine	21.99±1.89		cosine	11.85±0.48	
FLDA	distance	39.46±6.10	0.03	distance	15.19±1.38	0.75
	cosine	21.34±3.07		cosine	10.05±0.65	
LSA-ITML	distance	20.17±0.95	47.2	distance	10.06±0.99	6158.0
	cosine	21.29±1.02		cosine	9.06±0.45	
CM-ITML	distance	33.38±3.57	7.3	distance	18.18±0.74	219.3
	cosine	23.49±1.42		cosine	13.28±0.76	
MLR	distance	27.08±2.75	17.8	distance	13.29±1.14	1805.7
	cosine	20.22±1.44		cosine	9.86±0.53	
MLNCM	distance	21.07±1.62	1.1	distance	12.62±0.55	138.5
	cosine	22.88±2.52		cosine	13.51±0.79	
AROMA	similarity	21.25±1.53	33.1	similarity	13.49±0.63	2328.1
LRSML	inner-product	**15.56±1.55**	9.1	inner-product	**8.84±0.61**	676.0
	cosine	**14.83±1.20**		cosine	**7.29±0.40**	

Table 3: **TDT2-30** dataset: classification error rates and training time of eight competing metric/similarity learning methods.

	5×30 training samples			30×30 training samples		
Method	**Measure**	**Error Rate** (%)	**Train Time** (sec)	**Measure**	**Error Rate** (%)	**Train Time** (sec)
Original	–	10.12±1.24	–	–	6.75±0.39	–
Linear SVM	–	9.80±2.46	0.06	–	2.59±0.24	0.37
LSA	distance	30.01±10.14	0.02	distance	8.52±1.19	0.68
	cosine	8.19±1.49		cosine	4.26±0.44	
FLDA	distance	36.57±4.80	0.04	distance	7.50±0.62	1.08
	cosine	8.77±1.96		cosine	3.56±0.34	
LSA-ITML	distance	11.35±2.99	35.7	distance	5.68±0.76	1027.4
	cosine	5.85±1.14		cosine	2.84±0.33	
CM-ITML	distance	32.56±4.87	7.8	distance	7.67±0.71	297.7
	cosine	8.75±2.48		cosine	3.42±0.36	
MLR	distance	18.50±3.51	30.1	distance	7.50±0.75	1036.4
	cosine	8.04±1.11		cosine	3.84±0.29	
MLNCM	distance	8.48±1.38	1.5	distance	4.69±0.32	231.5
	cosine	8.58±1.64		cosine	4.39±0.38	
AROMA	similarity	11.31±2.55	35.6	similarity	4.27±0.41	2655.6
LRSML	inner-product	**4.74±0.83**	10.2	inner-product	**2.37±0.20**	813.9
	cosine	**4.95±1.06**		cosine	**2.27±0.15**	

Toh, K. C.; Todd, M. J.; and Tütüncü, R. H. 1999. Sdpt3 - a matlab software package for semidefinite programming, version 1.3. *Optimization Methods and Software* 11(1-4):545–581.

Torresani, L., and Lee, K.-C. 2006. Large margin component analysis. In *NIPS 19*.

Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; and Gong, Y. 2010. Locality-constrained linear coding for image classification. In *Proc. CVPR*.

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

Wu, L.; Jin, R.; Hoi, S. C. H.; Zhu, J.; and Yu, N. 2009. Learning bregman distance functions and its application for semi-supervised clustering. In *NIPS 22*.

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; and Russell, S. 2002. Distance metric learning with application to clustering with side-information. In *NIPS 15*.

Yang, J., and Yuan, X. 2013. Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation* 82(281):301–329.

Yang, J., and Zhang, Y. 2011. Alternating direction algorithms for ℓ_1 problems in compressive sensing. *SIAM Journal on Scientific Computing* 33(1):250–278.

Ye, G.-B.; Chen, Y.; and Xie, X. 2009. Efficient variable selection in support vector machines via the alternating direction method of multipliers. In *Proc. AISTATS*.

Ying, Y., and Li, P. 2012. Distance metric learning with eigenvalue optimization. *JMLR* 13:1–26.

Ying, Y.; Huang, K.; and Campbell, C. 2009. Sparse metric learning via smooth optimization. In *NIPS 22*.

Zhang, X.; Burger, M.; Bresson, X.; and Osher, S. 2010. Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM Journal on Imaging Sciences* 3(3):253–276.