Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)

Constrained Metric Learning via Distance Gap Maximization

Wei Liu, Xinmei Tian, Dacheng Tao
School of Computer Engineering
Nanyang Technological University, Singapore
wliu.cu@gmail.com, xinmeitian@gmail.com, and
dctao@ntu.edu.sg

Jianzhuang Liu
Department of Information Engineering
The Chinese University of Hong Kong, Hong Kong
Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences, China
jzliu@ie.cuhk.edu.hk

Abstract

Vectored data frequently occur in a variety of fields, which are easy to handle since they can be mathematically abstracted as points residing in a Euclidean space. An appropriate distance metric in the data space is quite demanding for a great number of applications. In this paper, we pose robust and tractable metric learning under pairwise constraints that are expressed as similarity judgements between data pairs. The major features of our approach include: 1) it maximizes the gap between the average squared distance among dissimilar pairs and the average squared distance among similar pairs; 2) it is capable of propagating similar constraints to all data pairs; and 3) it is easy to implement in contrast to the existing approaches using expensive optimization such as semidefinite programming. Our constrained metric learning approach has widespread applicability without being limited to particular backgrounds. Quantitative experiments are performed for classification and retrieval tasks, uncovering the effectiveness of the proposed approach.

Introduction

In the field of content-based image retrieval (CBIR) and image categorization, choosing appropriate distance metrics plays a key role in establishing effective systems. Regular CBIR systems usually adopt Euclidean metrics for distance measure on images represented into a vector form. Unfortunately, the Euclidean distance is generally not effective enough in retrieving relevant images. A main reason stems from the well-known semantic gap between low-level visual features and high-level semantic concepts (Smeulders et al. 2000).

The commonly used relevance feedback scheme may remedy the semantic gap issue, which produces, aided by users, a set of constraints about relevance (similarity) or irrelevance (dissimilarity). These constraints along with involved image examples are called *log data*. Then the key to CBIR is to find an effective way of utilizing the log data in relevance feedback so that the semantic gap can be successfully reduced. A lot of ways could be studied to use the log data to boost the retrieval performance. In this paper, we explore to learn distance metrics from the log data toward image retrieval tasks or any related applications involving distance metric learning. Recently, learning distance metrics from pairwise constraints (or called *side information* (Xing et al. 2003)) has been proposed in the machine learning community. Different from previous distance metric learning approaches, we address some technical and practical issues found in applying distance metric techniques to real applications.

Particularly, we are aware that routine metric learning techniques may fail to learn reliable metrics when handling a small amount of log data. In this paper, we present a novel weakly supervised distance metric learning algorithm so as to incorporate the abundant unlabeled data in a learning task. Specifically, we develop an intuitive learning framework to integrate synergistic information from both the log data and the unlabeled data for the goal of coherently learning a distance metric. The proposed Constrained Metric Learning (CML) algorithm is elegantly formulated, resulting in a simple solution which can be solved efficiently.

Related Work

The major group of related work is the hot distance metric learning research in the machine learning community, which can be divided into three main categories. One is unsupervised learning approaches most of which attempt to find low-dimensional embeddings given high-dimensional input data. The well-known linear embedding techniques include Principal Component Analysis (PCA), Multidimensional Scaling (MDS) (Hastie, Tibshirani, and Friedman 2009), and Locality Preserving Projections (LPP) (He and Niyogi 2004). Some manifold based approaches study nonlinear embedding techniques such as Locally Linear Embedding (LLE) (Roweis and Saul 2000), Isomap (Tenenbaum, de Silva, and Langford 2000), etc.

The second category is supervised learning approaches for classification, where distance metrics are usually learned from the training data associated with explicit class labels. The representative techniques include Linear Discriminant Analysis (LDA) (Hastie, Tibshirani, and Friedman 2009) and some recently proposed methods such as Neighbourhood Components Analysis (NCA) (Goldberger, Roweis, and Salakhutdinov 2005), Maximally Collapsing Metric Learning (MCML) (Globerson and Roweis 2006), distance metric learning for Large Margin Nearest Neighbor classification (LMNN) (Weinberger, Blitzer, and Saul

2006)(Weinberger and Saul 2008), and local distance metric learning (Yang et al. 2006)(Frome et al. 2007).

The third category is weakly supervised learning approaches which try to learn distance metrics with pairwise constraints, or known as side information (Xing et al. 2003). Each constraint indicates whether two data points are relevant (similar) or irrelevant (dissimilar) in a particular learning task. A well-known metric learning method with these constraints was proposed by Xing et al. (Xing et al. 2003) who cast the learning task into a convex optimization problem and applied the generated solution to data clustering. Following their work, there are several emerging metric techniques in this "weakly supervised" direction. For instance, Relevant Component Analysis (RCA) learns a global linear transformation by exploiting only the equivalent (relevant) constraints (Bar-Hillel et al. 2005). Discriminative Component Analysis (DCA) improves RCA via incorporating the inequivalent (irrelevant) constraints (Hoi et al. 2006). Lately, an Information-Theoretic Metric Learning (ITML) approach is presented to express the weakly supervised metric learning problem as a Bregman optimization problem (Davis et al. 2007)(Davis and Dhillon 2008).

Our previous work Output Regularized Metric Learning (ORML) (Liu, Hoi, and Liu 2008) falls into the third category. However, ORML was proposed merely for image retrieval and limited to the particular querying-feedback background. To remedy this limitation, this paper studies the general weakly supervised metric learning scenario and suggests an approach without being limited to particular backgrounds. Consequently, our approach can accommodate itself to broad applications including semi-supervised classification, relevance-feedback based image retrieval, and constrained clustering with background knowledge (Bennett, Bradley, and Demiriz 2000)(Wagstaff et al. 2001).

Constrained Metric Learning

In this section, we propose a novel weakly supervised metric learning technique, i.e., Constrained Metric Learning (CML), to produce metrics with high fidelity.

Problem Statement

Assume that we are given a set of n data points $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n \subseteq \mathbb{R}^m$, and two sets of pairwise constraints among these data points:

$$\begin{aligned} \mathcal{S} &= \{(i,j) \mid \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are judged to be similar}\} \\ \mathcal{D} &= \{(i,j) \mid \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are judged to be dissimilar}\}, \end{aligned} \quad (1)$$

where $\mathcal{S}$ is the set of similar pairwise constraints, and $\mathcal{D}$ is the set of dissimilar pairwise constraints. Each pairwise constraint (i, j) indicates if two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are relevant or irrelevant judged by users under some application context. Note that it is not necessary for all the points in $\mathcal{X}$ to be involved in $\mathcal{S}$ or $\mathcal{D}$.

For any pair of points $\mathbf{x}_i$ and $\mathbf{x}_j$, let $d(\mathbf{x}_i, \mathbf{x}_j)$ denote the distance between them. To compute this distance, let $M \in \mathbb{R}^{m\times m}$ be a symmetric metric matrix. Then we can express the distance measure as follows:

$$d_M(\mathbf{x}_i,\mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_M = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top M(\mathbf{x}_i - \mathbf{x}_j)}. \quad (2)$$

In practice, the symmetric matrix M is a valid metric if and only if it satisfies the non-negativity and triangle inequality conditions. In other words, M must be positive semidefinite, i.e., $M \succeq 0$. Generally, the matrix M parameterizes a family of Mahalanobis distances on the vector space $\mathbb{R}^m$. As an extreme case, when setting M to be the identity matrix $I \in \mathbb{R}^{m\times m}$, the distance in eq. (2) becomes the common Euclidean distance.

CML Prototype. *The constrained distance metric learning problem is to learn a symmetric matrix $M \in \mathbb{R}^{m\times m}$ from a collection of data points $\mathcal{X}$ on a vector space $\mathbb{R}^m$ together with a set of similar pairwise constraints $\mathcal{S}$ and a set of dissimilar pairwise constraints $\mathcal{D}$. This problem can be formulated as the following optimization prototype:*

$$\max_{M\succeq 0} \frac{g(M,\mathcal{X},\mathcal{S},\mathcal{D})}{f(M,\mathcal{X},\mathcal{S})} \quad (3)$$

where M is maintained to be positive semidefinite, and $f(\cdot)$ and $g(\cdot)$ are two proper objective functions defined over the given data and constraints.

Given the above definition, the theme to attack metric learning is to design appropriate objective functions f and g, and afterward find an efficient algorithm to solve the optimization problem. In the following subsections, we will discuss some principles for formulating reasonable optimization models. Importantly, we have to emphasize that it is critical to avoid overfitting when solving real-world metric learning problems.

(a) Distance Gap Maximization. It is very intuitive to formulate g to be maximized as the gap between the average squared distance among dissimilar data pairs in the set $\mathcal{D}$ and the average squared distance among similar data pairs in the set $\mathcal{S}$, that is

$$\begin{aligned} & g(M,\mathcal{X},\mathcal{S},\mathcal{D}) \\ &= \frac{1}{|\mathcal{D}|}\sum_{(i,j)\in\mathcal{D}} \|\mathbf{x}_i - \mathbf{x}_j\|_M^2 - \frac{\gamma}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}} \|\mathbf{x}_i - \mathbf{x}_j\|_M^2, \end{aligned} \quad (4)$$

where $\gamma \geq 1$ is the gap factor.

To learn a distance metric, one can assume there exits a corresponding linear mapping $U^\top : \mathbb{R}^m \rightarrow \mathbb{R}^r$, where $U = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \in \mathbb{R}^{m\times r}$ and $M = UU^\top$. We require that $\mathbf{u}_1, \ldots, \mathbf{u}_r$ be linearly independent so that r is the rank of the target metric matrix M. Then the distance under M between two inputs can be computed as:

$$\begin{aligned} \|\mathbf{x}_i - \mathbf{x}_j\|_M &= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top M(\mathbf{x}_i - \mathbf{x}_j)} \\ &= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top UU^\top(\mathbf{x}_i - \mathbf{x}_j)} \\ &= \left\|U^\top(\mathbf{x}_i - \mathbf{x}_j)\right\|. \end{aligned}$$

Actually, the target metric M is usually low-rank in high-dimensional data spaces (Davis and Dhillon 2008). Hence, we seek the subspace U instead of the full square matrix M.

In terms of U, the distance gap function defined in eq. (4) can be reformulated as

$$\begin{aligned} g(U) &= \frac{1}{|\mathcal{D}|}\sum_{(i,j)\in\mathcal{D}}(\mathbf{x}_i-\mathbf{x}_j)^\top UU^\top(\mathbf{x}_i-\mathbf{x}_j)- \\ &\quad \frac{\gamma}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}(\mathbf{x}_i-\mathbf{x}_j)^\top UU^\top(\mathbf{x}_i-\mathbf{x}_j) \\ &= \frac{1}{|\mathcal{D}|}\sum_{(i,j)\in\mathcal{D}}\mathrm{tr}\left(U^\top(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top U\right)- \\ &\quad \frac{\gamma}{|\mathcal{S}|}\sum_{(i,j)\in\mathcal{S}}\mathrm{tr}\left(U^\top(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top U\right) \\ &= \mathrm{tr}\left(U^\top C_g U\right), \end{aligned} \tag{5}$$

in which we define an $m \times m$ matrix by

$$\begin{aligned} C_g &= \frac{\sum_{(i,j)\in\mathcal{D}}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top}{|\mathcal{D}|} \\ &\quad -\gamma\frac{\sum_{(i,j)\in\mathcal{S}}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top}{|\mathcal{S}|}. \end{aligned} \tag{6}$$

$\mathrm{tr}(\cdot)$ stands for the *trace* operator.

One common principle for metric learning is minimizing the distances among the data points of similar constraints and meanwhile maximizing the distances among the data points of dissimilar constraints. We refer to it as a *min-max* principle. Some existing metric learning works such as (Xing et al. 2003) and (Weinberger, Blitzer, and Saul 2006) can be interpreted via this min-max principle. Obviously, our principle *distance gap maximization* coincides with the min-max principle. Unlike (Davis et al. 2007), we do not enforce any constraint on the values of the distances on the two sets $\mathcal{S}$ and $\mathcal{D}$, which is because of limited and possibly noisy log data. More naturally, we choose to maximize the gap between two average squared distances on two constraint sets $\mathcal{S}$ and $\mathcal{D}$.

(b) Neighborhood Smoothness. In the prototype eq. (3), we also hope for minimization of some function f. The straightforward way is to minimize the sum of squared distances between all similar pairs in $\mathcal{S}$, but such an optimization will overfit the log data that are scarce in real-world applications. To remedy that, we aim at taking full advantage of unlabeled data that are demonstrated to be quite beneficial to the semi-supervised learning problem. Due to this consideration, we define f based on the notion of neighborhood preserving (He and Niyogi 2004).

Given the collection of n data points $\mathcal{X}$ including the log data and the unlabeled data, we can define a neighborhood indicator matrix $W \in \mathbb{R}^{n\times n}$ on $\mathcal{X}$:

$$W_{ij} = \begin{cases} 1, & \mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $\mathcal{N}(\mathbf{x}_i)$ denotes the list composed of k nearest neighbors of the data point $\mathbf{x}_i$ using the Euclidean metric. Actually, such a matrix W holds *weak* (probably correct) similarities between all data pairs. Note that W is asymmetric and $W_{ii} = 0$ for $i = 1, \cdots, n$.

Through absorbing all data points $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ and utilizing all weak similarities W, we formulate f as follows:

$$\begin{aligned} f(M,\mathcal{X},\mathcal{S}) &= \frac{1}{2}\sum_{i,j=1}^{n}\|\mathbf{x}_i-\mathbf{x}_j\|_M^2 W_{ij} \\ &= \frac{1}{2}\sum_{i,j=1}^{n}\left\|U^\top(\mathbf{x}_i-\mathbf{x}_j)\right\|^2 W_{ij}. \end{aligned} \tag{8}$$

Importantly, $f(\cdot)$ provides a smoothness measure of data neighborhoods under the chosen distance metric M.

Similarity Propagation

To enable metric learning techniques to work for practical applications such as classification and retrieval, we should shrink the distances between as many similar pairs as possible. Although the neighborhood smoothness function f has incorporated all unlabeled data, it does not emphasize the distances between "real" similar data pairs. Since the similar constraint set $\mathcal{S}$ is available, we desire to propagate the limited real similar constraints to all data pairs via the found neighborhoods. Specifically, we intend to learn a new similarity matrix $\tilde{W}$ such that $\tilde{W}_{ij}$ reflects the extent of real similarity between data pair $(\mathbf{x}_i, \mathbf{x}_j)$.

Let us begin with the *strong* (definitely correct) similarity matrix $S^0 \in \mathbb{R}^{n\times n}$ where we set $S^0_{ii} = 1$ for any i and $S^0_{ij} = 1$ for any similar constraint $(i,j) \in \mathcal{S}$. If we conceived 1-entries in S^0 as positive energies, our purpose would be to propagate energies in S^0 to its 0-entries. The propagation path follows the neighborhood structures residing on each data point, so we pose the similarity propagation criterion as the locally linear energy mixture, i.e.

$$S^{(t+1)}_{i.} = (1-\alpha)S^{(0)}_{i.} + \alpha\frac{\sum_{j=1}^{n} W_{ij}S^{(t)}_{j.}}{\sum_{j=1}^{n} W_{ij}}, \tag{9}$$

where $S^{(t)}_{i.}$ denotes the ith row of $S^{(t)}$ and $t = 0, 1, 2, \cdots$ is time stamp. We write the matrix form of eq. (9) as

$$S^{(t+1)} = (1-\alpha)S^{(0)} + \alpha P S^{(t)}, \tag{10}$$

where $0 < \alpha < 1$ is the trade-off parameter and $P = D^{-1}W$ (D is a diagonal matrix whose diagonal elements equal the sums of row entries of W, i.e., $D_{ii} = \sum_{j=1}^{n} W_{ij}$) is the transition probability matrix widely used in Markov random walk models.

Because $0 < \alpha < 1$ and the eigenvalues of P are in $[-1, 1]$, the limit $S^* = \lim_{t\to\infty} S^{(t)}$ exists. It suffices to solve the limit as

$$S^* = (1-\alpha)(I-\alpha P)^{-1}S^{(0)}. \tag{11}$$

Then, we build the new similarity matrix by symmetrizing the converged similarity matrix S^* and removing small similarity values, that is

$$\tilde{W} = \left\lfloor \frac{S^* + S^{*\top}}{2} \right\rfloor_{\geq\theta}, \tag{12}$$

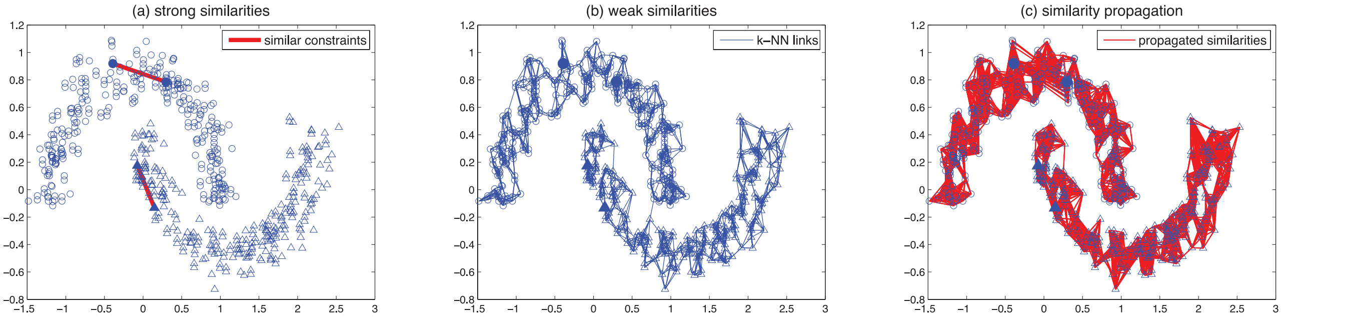

Figure 1: The two-moon problem of 600 2D points. (a) Strong similarities, i.e., two similar constraints among four points of which each two come from the same class; (b) weak similarities provided by k-NN ($k = 6$) search; (c) similarity propagation with $\alpha = 0.9$ and $\theta = 0.01$.

in which the operator $\lfloor S \rfloor_{\geq\theta}$ zeros out the entries of S smaller than $0 < \theta < 1$.

Subsequently, we define a new smoothness function f using the learned similarity matrix $\tilde{W}$ as

$$\begin{aligned} f(U) &= \frac{1}{2}\sum_{i,j=1}^{n} \|\mathbf{x}_i - \mathbf{x}_j\|_M^2 \tilde{W}_{ij} \\ &= \frac{1}{2}\sum_{i,j=1}^{n} \left\|U^\top(\mathbf{x}_i - \mathbf{x}_j)\right\|^2 \tilde{W}_{ij} \\ &= \sum_{d=1}^{r} \mathbf{u}_d^\top X(\tilde{D} - \tilde{W})X^\top \mathbf{u}_d = \sum_{d=1}^{r} \mathbf{u}_d^\top X\tilde{L}X^\top \mathbf{u}_d \\ &= \mathrm{tr}\left(U^\top X\tilde{L}X^\top U\right), \end{aligned} \quad (13)$$

where $\tilde{D}$ is a diagonal matrix whose diagonal elements equal the sums of row entries of $\tilde{W}$, i.e., $\tilde{D}_{ii} = \sum_{j=1}^{n} \tilde{W}_{ij}$, and $\tilde{L} = \tilde{D} - \tilde{W}$ is known as the *graph Laplacian*. At this time, the smoothness function $f(\cdot)$ unifies low-level neighborhood structures $\{\mathcal{N}(\mathbf{x}_i)\}$ and high-level similar constraints in $\mathcal{S}$. In summary, we are capable of learning a reliable similarity matrix provided with heuristic k-NN search and a genuine similar constraint set. The technical essence of similarity propagation subject to real similar constraints is enhancing the generalization and robustness of the smoothness function f.

A toy example to illustrate the effect of similarity propagation from strong similarities over the basis of weak similarities is show in Fig. 1. It is clear that we produce more reliable similarities (see Fig. 1(c)) than those weak ones that introduce a few wrong links (see Fig. 1(b)).

Generalized Eigenvalue Problem

After designing two functions $g(U)$ and $f(U)$, we propose a novel distance metric learning technique, Constrained Metric Learning (CML), by implementing the following simple optimization framework

$$\max_{U\in\mathbb{R}^{m\times r}} \frac{g(U)}{f(U)} = \max_{U\in\mathbb{R}^{m\times r}} \frac{\mathrm{tr}\left(U^\top C_g U\right)}{\mathrm{tr}\left(U^\top X\tilde{L}X^\top U\right)}. \quad (14)$$

The optimal subspace $U = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$ that maximizes eq. (14) is offered by the maximal eigenvalue solution to the generalized eigenvalue problem:

$$C_g\mathbf{u}_d = \lambda_d X\tilde{L}X^\top\mathbf{u}_d, \quad d = 1, 2, \cdots, r. \quad (15)$$

Note that we retain r eigenvectors $\mathbf{u}_d$'s corresponding to r largest positive eigenvalues λ_d's in order to make the distance gap under the learned metric $M = UU^\top$ be

$$g(U) = \mathrm{tr}\left(U^\top C_g U\right) = \sum_{d=1}^{r} \mathbf{u}_d^\top C_g \mathbf{u}_d = \sum_{d=1}^{r} \lambda_d > 0. \quad (16)$$

Accordingly, the proposed CML approach always keeps a positive gap between the average squared distance over the dissimilar constraints $\mathcal{D}$ and the average squared distance over the similar constraints $\mathcal{S}$.

It is delightful to see that CML does not invoke expensive optimization such as semidefinite programming (SDP) (Boyd and Vandenberge 2003) which was usually applied to solve metric learning problems (Weinberger, Blitzer, and Saul 2006). While solving metric learning by an SDP solver is feasible for small-scale problems, it often becomes impractical for real applications, even for moderate-scale datasets. This is because the time complexity of a general SDP solver may be quite expensive, which is clearly inefficient and not scalable for real applications. In contrast, the proposed CML method is simple and significantly fast, avoiding to invoke costly SDP. We expect that CML would have widespread applicability because CML only involves simple matrix inversion and eigen-decomposition.

One noticeable issue is the singularity of the matrix $X\tilde{L}X^\top \in \mathbb{R}^{m\times m}$. $\tilde{L}$ is a singular matrix with $\mathrm{Rank}(\tilde{L}) \leq n - 1$, so $\mathrm{Rank}(X\tilde{L}X^\top) \leq \min(m, n-1)$. If the data dimension m is larger than the data size n, $X\tilde{L}X^\top$ must be singular. In this case, the generalized eigenvalue decomposition eq. (15) leads to many zero eigenvalues as well as the corresponding useless eigenvectors. We may apply PCA to reduce the data dimension to $m_1 \leq n - 1$ to guarantee that eq. (15) provides meaningful eigenvectors. As a broadly adopted strategy for data dimensionality reduction, this PCA preprocessing step has been engaged into high-dimensional metric learning (Weinberger and Saul 2008)(Davis and Dhillon 2008).

Table 1: Dataset Information: the numbers of dimensions, samples and classes.

Dataset	# Dimensions	# Samples	# Classes
WINE	13	178	3
BREAST CANCER	30	569	2
USPS	256	2007	10

Experiments

In this section, we investigate the power of the proposed metric learning approach CML on two UCI datasets[1] and one image dataset USPS[2]. Table 1 describes the fundamental information about these benchmark datasets. In detail, we compare CML of two versions with the baseline Euclidean and Mahalanobis metrics and two recently published metric learning methods LMNN (Weinberger, Blitzer, and Saul 2006) and ITML (Davis et al. 2007). We know that the success of semi-supervised learning stems from making use of unlabeled data for performance gain. In many application domains, unlabeled data are plentiful, such as images, documents, etc. On the other hand, in many cases it is often convenient to collect unlabeled data. Therefore, the proposed CML method explores the potential of unlabeled data by incorporating them into conventional supervised metric learning problems, exhibiting superior performance in classification and retrieval. As the unified computation platform, we let kNN cooperate with various distance metrics.

Let us fix three simple parameters as $k = 6$ (for heuristic k-NN search), $\alpha = 0.9$, and $\theta = 0.005$. What remains to do is to tune the distance gap factor $\gamma \geq 1$ to the best values on each dataset. In the following experiments, we evaluate the effectiveness of the proposed CML method applied to semi-supervised classification and content-based image retrieval (CBIR).

Compared Methods

We compare CML extensively with three classes of representative distance metric techniques: two unsupervised approaches, one supervised approach, and one weakly supervised approach. Although it may be unfair to compare the unsupervised approaches with supervised ones, we still report the unsupervised results which can help us comprehend how effective is the proposed method compared to traditional. The compared approaches include:

Euclidean: the well-known Euclidean metric denoted as "EU" in short.

Mahalanobis: a standard Mahalanobis metric denoted as "Mah" in short. Specifically, the metric matrix $A = \mathrm{Cov}^{-1}$ where Cov is the sample covariance matrix.

LMNN (Weinberger, Blitzer, and Saul 2006): Large Margin Nearest Neighbor which works under supervised settings where each sample has an exact class label.

ITML-1 (Davis et al. 2007): Information-Theoretic Metric Learning which works under pairwise relevance constraints but does not explicitly engage the unlabeled data. The initialized metric matrix is the identity matrix.

Table 2: Comparisons of classification error rates (%) on WINE. The last row shows the relative improvement of the proposed CML over the baseline metric EU.

Compared Methods	10 Labeled Samples	40 Labeled Samples
EU	34.61±6.34	31.33±3.07
Mah	33.62±6.97	19.03±4.42
LMNN	24.01±8.82	15.03±3.54
ITML-1	30.15±7.28	15.28±4.51
ITML-2	33.06±7.59	17.09±5.63
CML-1	19.79±6.53	8.75±6.11
CML-2	**19.40±6.51**	**5.57±2.62**
CML-2 Improve	**43.95%**	**82.22%**

ITML-2 (Davis et al. 2007): the same as ITML-1 except the initialized metric matrix is the inverse covariance matrix.

CML-1: the proposed weakly supervised metric learning method using pairwise relevance constraints and the unlabeled data, where the neighborhood smoothness function $f(\cdot)$ adopts the weak similarities described in eq. (7).

CML-2: the same as CML-1 except $f(\cdot)$ uses the learned similarities described in eq. (12).

Semi-Supervised Classification

We apply the above stated seven metric methods EU, Mah, LMNN, ITML-1, ITML-2, CML-1, and CML-2 on two UCI datasets: WINE and BREAST CANCER. To perform semi-supervised classification, we randomly choose at least one sample from each class as labeled data and treat the rest samples as unlabeled data. We evaluate 1NN classification performance in terms of error rates on unlabeled data. For each dataset, we repeat the evaluation process with the 7 methods 50 times, and take the average error rates for comparison. Table 2 and 3 list the comparative results. In contrast to unsupervised EU and Mah, supervised LMNN, and weakly supervised ITML-1, ITML-2, and CML-1, CML-2 achieves the lowest average error rates on both datasets. From Table 2 and 3, we find that CML-2 significantly outperforms the other methods and improves the baseline EU up to 82.22%.

Image Retrieval

In the USPS (test) handwritten digits dataset, each sample is a 16×16 image and there are ten types of digits 0, 1, 2, ..., 9 that are used as ten classes. There are 160 samples for each class at least, summing up to a total of 2007. For the setup of image retrieval on USPS, we follow a standard procedure for CBIR experiments. Specifically, a query image is picked from the dataset and then queried with the evaluated distance metrics. The retrieval performance is then evaluated based on the top ranked images ranging from top 10 to top 100 images. The query-averaged precision and recall, which are widely used in CBIR-related experiments, are utilized for the performance measures.

[1] http://archive.ics.uci.edu/ml/

[2] http://www-i6.informatik.rwth-aachen.de/ keysers/usps.html

Table 4: Query-averaged precision/recall (%) of 100 top ranked images over 1,607∼1,807 queries from ten classes. There are three groups of log data. The last column shows the relative improvement of the proposed CML over the baseline metric EU.

Settings	**EU**	**Mah**	**ITML-1**	**ITML-2**	**CML-1**	**CML-2**	**CML-2 Improve**
Precision (%) with 200 Log Images	62.08	61.79	66.57	64.45	70.59	**73.31**	**18.09%**
Recall (%) with 200 Log Images	28.65	28.47	31.15	30.38	33.45	**34.90**	**21.82%**
Precision (%) with 300 Log Images	62.53	62.25	68.91	65.23	70.87	**73.40**	**17.38%**
Recall (%) with 300 Log Images	28.74	28.56	32.32	30.50	33.52	**34.85**	**21.26%**
Precision (%) with 400 Log Images	62.83	62.54	72.07	71.26	72.08	**75.28**	**19.82%**
Recall (%) with 400 Log Images	28.68	28.49	33.75	32.90	34.01	**35.65**	**24.30%**

Table 3: Comparisons of classification error rates (%) on BREAST CANCER. The last row shows the relative improvement of the proposed CML over the baseline metric EU.

Compared Methods	20 Labeled Samples	80 Labeled Samples
EU	11.09±3.08	9.75±1.79
Mah	28.81±3.67	23.56±2.69
LMNN	10.12±2.74	7.15±1.54
ITML-1	10.45±2.75	7.53±1.34
ITML-2	27.61±3.58	18.56±7.54
CML-1	10.54±2.65	11.98±2.38
CML-2	**8.40±2.79**	**5.87±1.55**
CML-2 Improve	**24.26%**	**39.79%**

Note that we must provide the log data to run weakly supervised metric learning methods ITML-1, ITML-2, CML-1 and CML-2. Here we select 20, 30, and 40 samples uniformly from each class, and then we gather three groups of log data which contain 200, 300, and 400 log images respectively. The similar constraints are imposed on the same labeled log images, while the dissimilar constraints are imposed on the differently labeled log images. The query images are those samples outside the log data subset. Table 4 as well as Fig. 2 and 3 shows the experimental results with different groups of log data. From these results, we find that CML-2 consistently outperforms the other compared methods and its improvement over the baseline EU doubles that of ITML when only using 200 log images. Hence, it suffices to conclude that the proposed CML method is more effective to learn robust distance metrics by utilizing the unlabeled data, even with limited log data. The relative improvements of CML-2 over ITML in the case of 400 log images are less significant than those in the case of 200 log images, but CML-2 still achieves the best improvement among all compared methods.

Conclusion and Discussion

This paper studies the weakly supervised distance metric learning problem which works under pairwise similar and dissimilar constraints. In the context of image retrieval, real log data provide such constraints via user's relevance feedback. To robustly exploit the log data and smoothly incorporate the unlabeled data, we propose the constrained metric learning (CML) approach. CML offers a robust metric according to pairwise constraints exposed in the log data through maximizing the distance gap and learning the robust similarity matrix both of which can be carried out very efficiently. Extensive experiments have been conducted to evaluate semi-supervised classification and content-based image retrieval performances. The promising results show that the proposed CML approach is more effective than state-of-the-arts in learning reliable metrics with unlabeled data. As an advantage, CML is not limited to particular backgrounds and can accommodate itself to broad applications. For example, CML can be applied to constrained K-means clustering (Bennett, Bradley, and Demiriz 2000)(Wagstaff et al. 2001) since it supplies a good distance metric.

In future work, we plan to study techniques for learning nonlinear distance metrics. A simple kernelization trick can be used immediately to make CML produce a nonlinear metric. Another feasible research direction is to try metric learning under noisy pairwise constraints. Such a learning problem is confronted frequently in practice because some of user's relevance feedbacks are inaccurate. To this end, we need to design more appropriate learning principles in order to address the noisy constraints.

Acknowledgments

This work was supported by the Nanyang Technological University Nanyang SUG Grant (under project number M58020010) and the K. C. Wong Education Foundation Award. This work was also supported by grants from Natural Science Foundation of China (No. 60975029) and Shenzhen Bureau of Science Technology & Information, China (No. JC200903180635A).

References

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

Bennett, K.; Bradley, P.; and Demiriz, A. 2000. Constrained k-means clustering. *Technical Report 2000-65* Microsoft Research.

Boyd, S., and Vandenberge, L. 2003. *Convex Optimization*. Cambridge University Press, Cambridge, UK.

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

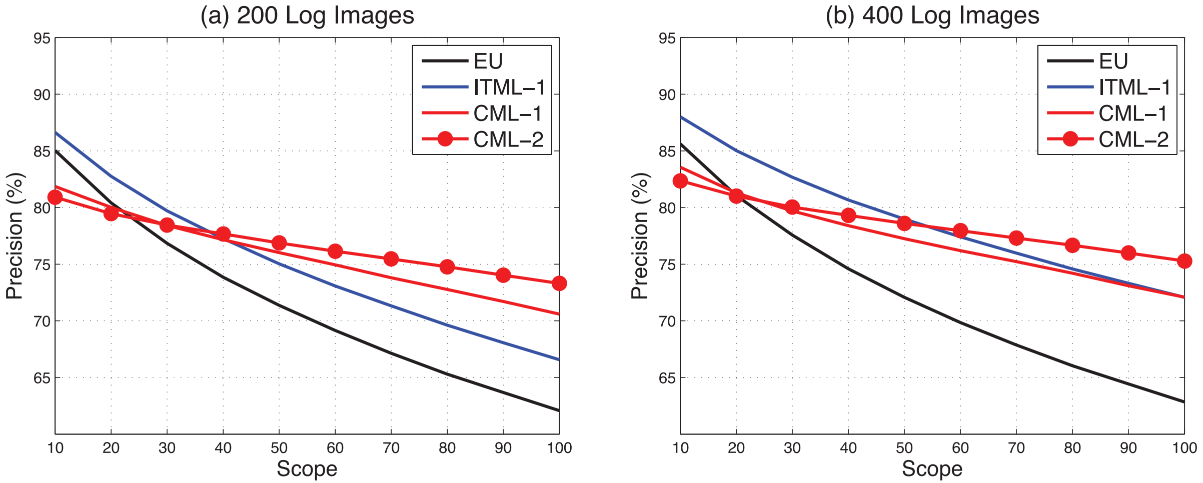

Figure 2: Query-averaged precision (%) of top ranked images with 200 and 400 log images respectively.

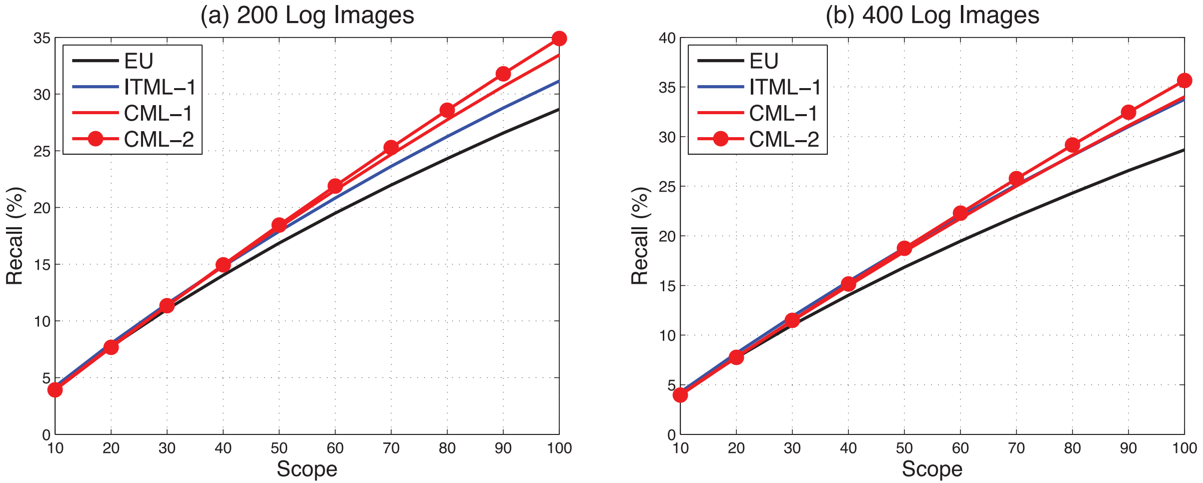

Figure 3: Query-averaged recall (%) of top ranked images with 200 and 400 log images respectively.

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

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*.

Globerson, A., and Roweis, S. 2006. Metric learning by collapsing classes. In *NIPS 18*.

Goldberger, G. H. J.; Roweis, S.; and Salakhutdinov, R. 2005. 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.

He, X., and Niyogi, P. 2004. Locality preserving projections. In *NIPS 16*.

Hoi, S. C.; Liu, W.; Lyu, M. R.; and Ma, W.-Y. 2006. Learning distance metrics with contextual constraints for image retrieval. In *Proc. CVPR*.

Liu, W.; Hoi, S. C.; and Liu, J. 2008. Output regularized metric learning with side information. In *Proc. ECCV*.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.

Smeulders, A. W. M.; Worring, M.; Santini, S.; Gupta, A.; and Jain, R. 2000. Content-based image retrieval at the end of the early years. *IEEE Trans. PAMI* 22(12):1349–1380.

Tenenbaum, J. B.; de Silva, V.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323.

Wagstaff, K.; Cardie, C.; Rogers, S.; and Schroedl, S. 2001. Constrained k-means clustering with background knowledge. In *Proc. ICML*.

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

Weinberger, K.; Blitzer, J.; and Saul, L. 2006. Distance metric learning for large margin nearest neighbor classification. In *NIPS 18*.

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; and Russell, S. 2003. Distance metric learning with application to clustering with side-information. In *NIPS 15*.

Yang, L.; Jin, R.; Sukthankar, R.; and Liu, Y. 2006. An efficient algorithm for local distance metric learning. In *Proc. AAAI*.