Robust Distance Metric Learning via Simultaneous ℓ_1-Norm Minimization and Maximization

Hua Wang HUAWANGCS@GMAIL.COM
Colorado School of Mines, Department of Electrical Engineering and Computer Science, Golden, Colorado 80401

Feiping Nie FEIPINGNIE@GMAIL.COM
Heng Huang HENG@UTA.EDU
Computer Science and Engineering Department, University of Texas at Arlington, Arlington, TX, 76019

Abstract

Traditional distance metric learning with side information usually formulates the objectives using the covariance matrices of the data point pairs in the two constraint sets of must-links and cannot-links. Because the covariance matrix computes the sum of the squared ℓ_2-norm distances, it is prone to both outlier samples and outlier features. To develop a robust distance metric learning method, we propose a new objective for distance metric learning using the ℓ_1-norm distances. The resulted objective is challenging to solve, because it simultaneously minimizes and maximizes (minmax) a number of non-smooth ℓ_1-norm terms. As an important theoretical contribution of this paper, we systematically derive an efficient iterative algorithm to solve the general ℓ_1-norm minmax problem. We performed extensive empirical evaluations, where our new distance metric learning method outperforms related state-of-the-art methods in a variety of experimental settings.

1. Introduction

Distance metric plays a critical role in a number of machine learning algorithms. For example, K-means clustering and k-Nearest Neighbor (k-NN) classification need to be supplied with a suitable distance metric, through which neighboring data points can be identified. The commonly used Euclidean distance metric assumes that every feature of the input data is equally important and independent from others. This assumption, however, may not be always satisfied in real world applications, especially when we deal with high dimensional data where some features may not be tightly related to the topic of interest. In contrast, a distance metric with good quality should identify important features and discriminate relevant and irrelevant features. Thus, supplying such a distance metric is highly problem-specific and determines the success or failure of a learning algorithm (Xing et al., 2002; Hoi et al., 2006; Xiang et al., 2008; Weinberger et al., 2006; Davis et al., 2007).

In this paper, we address the issue of the robustness of a distance metric in the presence of both outlier samples and outlier features. The former is defined as the data points that deviates significantly from the majority of the data points, and the latter is defined as features that do not have a regular distribution over the data points. Given the side information of the must-links and cannon-links (Xing et al., 2002; Hoi et al., 2006; Xiang et al., 2008), traditional distance metric learning approaches often formulate the objectives using the covariance matrices of the data point pairs in the two constraint sets. However, because these estimates are defined as the sum of the squared ℓ_2-norm distances, they could be highly influenced by outlying observations and features. That is, these measurements become inappropriate on contaminated data sets, because large errors squared dominate the sum.

Many previous works have been done to improve the robustness of machine learning models through using the ℓ_1-norm formulations (Ding et al., 2006; Cayton & Dasgupta, 2006; Gao, 2008; Ke & Kanade, 2004; Kwak, 2008; Wright et al., 2009; Nie et al., 2011; Wang et al., 2013b;a). However, most, if not all, these methods tried to improve the robustness of the models like Principal Component Analysis (PCA). Thus far, to our best knowledge, there exists no work to utilize ℓ_1-norm based objective for distance metric learning. Although it is easy and straightforward to derive a robust formulation for metric learning using the ℓ_1-norm distances, it is non-trivial and difficult to solve the resulted optimization problems involving the ℓ_1-

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

norm distances. Because most metric learning objectives (either ratio or substraction) have to simultaneously minimize the distances of the point pairs in the must-links and maximize those in the cannot-links, all current optimization methods in sparse learning, such as gradient projection, homotopy, iterative shrinkage-thresholding, proximal gradient, and augmented lagrange multiplier methods, cannot efficiently solve the ℓ_1-norm based distance metric learning objectives. Zha et al. (2009) made a successful attempt to learn a robust distance metric, which, however, achieved its goal by utilizing additional knowledge from auxiliary data yet did not replace the traditional distance metric learning objective by any new robust formulation.

In this paper, we propose a new robust distance metric learning objective that utilizes the ℓ_1-norm distances and directly provides the robustness against outlier samples and outlier features. However, because of using the ℓ_1-norm distances, the resulted objective ends up to be a simultaneous ℓ_1-norm minimization and maximization (minmax) problem. Although there exist a large number of optimization algorithms to solve the objectives involving ℓ_1-norm or $\ell_{2,1}$-norm minimizations, how to efficiently solve the ℓ_1-norm minmax problem is scarcely studied in literature. As an important theoretical contribution of this paper, we systematically derive an efficient iterative algorithm to solve the general ℓ_1-norm minmax problem, whose convergence is guaranteed by the rigorous theoretical analysis. We have performed extensive experiments to apply our new distance metric learning method on five benchmark data sets to evaluate its robustness against both outlier samples and outlier features, in which we achieved very promising results that are consistent with our theoretical analysis.

2. Learning A Robust Distance Metric

In this section, we will first develop a robust objective for distance metric learning using the ℓ_1-norm distances, which, though, is highly non-smooth and difficult to solve. In the next section, we will present an efficient iterative solution algorithm with a rigorous proof on its convergence.

Notations. Throughout this paper, we write matrices as bold uppercase characters and vectors as bold lowercase characters. Given a matrix $\mathbf{M} = [m_{ij}]$, we denote its i-th row and its j-th column as $\mathbf{m}^i$ and $\mathbf{m}_j$, respectively. The ℓ_1-Norm of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_1 = \sum_{i,j} |m_{ij}|$.

Problem formalization. Assume that we have a set of n data points $\mathcal{X} = \left\{\mathbf{x}_i \in \mathbb{R}^d\right\}_{i=1}^n$. For convenience, we write $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$. Besides, we are also supplied with two sets of pairwise constraints among the data points, which are manually labeled by users under certain application context:

$$\begin{cases} \mathcal{S} = \{(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are in the same class}\} \quad , \\ \mathcal{D} = \{(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are not in the same class}\} \quad , \end{cases} \tag{1}$$

where we call $\mathcal{S}$ as must-links and $\mathcal{D}$ as cannot-links (Xing et al., 2002). Note that it is not necessary for all data points in $\mathcal{X}$ to be involved in $\mathcal{S}$ or $\mathcal{D}$.

Given any two data points $\mathbf{x}_i$ and $\mathbf{x}_j$, a Mahalanobis distance between them can be computed as following:

$$\|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{M}} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)} \ , \tag{2}$$

where $\mathbf{M} \in \mathbb{R}^{d\times d}$ is the Mahalanobis distance metric, a symmetric matrix of size $d \times d$. In general, $\mathbf{M}$ is a valid metric if and only if $\mathbf{M}$ is a positive semi-definite matrix by satisfying the non-negativity and the triangle inequality conditions, *i.e.*, $\mathbf{M} \succeq 0$. When setting $\mathbf{M}$ to be the identity matrix $\mathbf{I}_{d\times d}$, the distance computed by Eq. (2) becomes the Euclidean distance. Our goal in robust distance metric learning is to learn an optimal square matrix $\mathbf{M}$ from a collection of data points $\mathcal{X}$ in the presence of outliers, such that the distances between the data point pairs in $\mathcal{S}$ are as small as possible, whilst those in $\mathcal{D}$ are as large as possible.

Learning a robust distance metric using the ℓ_1-norm minmax formulation. Xing *et al.* (Xing et al., 2002) first studied the problem of learning a distance metric from must-links and cannot-links, in which the sum of the Mahalanobis distances between the point pairs in the must-links is minimized under the constraints developed from the point pairs in the cannot-links. Despite its effectiveness, it is computationally inefficient when dealing with high-dimensional data (Xiang et al., 2008). Relevance Component Analysis (RCA) (Bar-Hillel et al., 2003) was then proposed to solve the inverse matrix of the covariance matrix of the data point pairs in the chunklets (must-links), which, though, may not exist in high dimensional data. Disciminative Component Analysis (DCA) and Kernel DCA (Hoi et al., 2006) improved RCA by exploiting negative constraints and aimed to capture nonlinear relationships using contextual information. Both RCA and DCA, however, faced the singular problem when computing the covariance matrix for the data point pairs in the must-links in the case of high dimensionality (Xiang et al., 2008). To tackle this, Xiang et al. (2008) proposed to formulate the distance metric learning problem as a trace ratio minimization problem as following.

Because $\mathbf{M}$ is a positive semi-definite matrix, we can reasonably write $\mathbf{M} = \mathbf{W}\mathbf{W}^T$ by eigen-decomposition, where $\mathbf{W} \in \mathbb{R}^{d\times r}$ with $r \leq d$. Thus the Mahalanobis distance under the metric $\mathbf{M}$ can be computed as $\|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{M}} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}\mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j)} =$

$\left\|\mathbf{W}^T\left(\mathbf{x}_i - \mathbf{x}_j\right)\right\|_2$, which indeed defines a transformation of $\mathbf{y} = \mathbf{W}^T\mathbf{x}$ under the projection matrix $\mathbf{W}$. Then denote the covariance matrix of the data point pairs in the must-links as $\mathbf{S}_w = \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}}\left(\mathbf{x}_i - \mathbf{x}_j\right)\left(\mathbf{x}_i - \mathbf{x}_j\right)^T$ and the covariance matrix of the data point pairs in the cannot-links as $\mathbf{S}_b = \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}}\left(\mathbf{x}_i - \mathbf{x}_j\right)\left(\mathbf{x}_i - \mathbf{x}_j\right)^T$, Xiang et al. (2008) proposed to learn the transformation matrix $\mathbf{W}$ by solving the following objective:

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\mathbf{tr}\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)}{\mathbf{tr}\left(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right)} = \frac{\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}}\left\|\left(\mathbf{x}_i - \mathbf{x}_j\right)^T\mathbf{W}\right\|_2^2}{\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}}\left\|\left(\mathbf{x}_i - \mathbf{x}_j\right)^T\mathbf{W}\right\|_2^2} . \tag{3}$$

The objective in Eq. (3) measures the ratio of the sums of the two sets of the squared ℓ_2-norm distances, one set for the data point pairs in the must-links and the other for those in the cannot-links. As a result, same as other least square minimization based models in machine learning and statistics, Eq. (3) is sensitive to the presence of outliers. Recent progress (Gao, 2008; Ke & Kanade, 2004; Ding et al., 2006; Kwak, 2008; Wright et al., 2009) has shown that the ℓ_1-norm distance can introduce robustness against both outlier samples as well as outlier features, which have been widely applied to replace the squared ℓ_2-norm distance in many machine learning methods, such as PCA (Wright et al., 2009). Following the same motivations as these prior studies, we propose to replace the squared ℓ_2-norm distances in the distance metric learning objective in Eq. (3) by the ℓ_1-norm distances to promote the robustness, which leads to the following optimization problem:

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}}\left\|\left(\mathbf{x}_i - \mathbf{x}_j\right)^T\mathbf{W}\right\|_1}{\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{D}}\left\|\left(\mathbf{x}_i - \mathbf{x}_j\right)^T\mathbf{W}\right\|_1} = \frac{\|\mathbf{A}\mathbf{W}\|_1}{\|\mathbf{B}\mathbf{W}\|_1} , \tag{4}$$

where each row of $\mathbf{A}$ is one $\left(\mathbf{x}_i - \mathbf{x}_j\right)^T$ that satisfies $\left(\mathbf{x}_i, \mathbf{x}_j\right) \in \mathcal{S}$, and similarly each row of $\mathbf{B}$ is one $\left(\mathbf{x}_i - \mathbf{x}_j\right)^T$ that satisfies $\left(\mathbf{x}_i, \mathbf{x}_j\right) \in \mathcal{D}$.

Note that Eq. (4) can be rewritten as following:

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\|\mathbf{A}\mathbf{W}\|_1}{\|\mathbf{B}\mathbf{W}\|_1} = \frac{\sum_{i=1}^r\|\mathbf{A}\mathbf{w}_i\|_1}{\sum_{i=1}^r\|\mathbf{B}\mathbf{w}_i\|_1} , \tag{5}$$

which minimizes the ratio between the overall ℓ_1-norm distances of the data point pairs in the must-links along all the projecting directions and those in the cannot-links. A potential problem is that the ratio between the ℓ_1-norm distances of the data point pairs in the must-links and those in the cannot-links may not be minimized along each individual projecting direction. Thus, we turn to minimize the following better objective which emphasizes the projection performance along every projecting direction:

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \sum_{i=1}^r \frac{\|\mathbf{A}\mathbf{w}_i\|_1}{\|\mathbf{B}\mathbf{w}_i\|_1} , \tag{6}$$

Because Eq. (6) does not square the distances from either data point or feature perspective, the outlying samples and features will have less influence. As a result, same as other ℓ_1-norm distance based learning models, the objective in Eq. (6) is more robust against outliers than its counterpart defined in Eq. (3). Thus, we call Eq. (6) as the proposed *robust distance metric learning* model.

Upon solution, the distance metric can be obtained by computing $\mathbf{M} = \mathbf{W}\mathbf{W}^T$.

3. An Efficient Algorithm to Solve ℓ_1-Norm Minmax Problem

Despite the straightforward replacement from the squared ℓ_2-norm distances in Eq. (3) to the ℓ_1-norm distances in Eq. (6) to promote robustness against outliers, solving Eq. (6) is very challenging, because it simultaneously minimizes and maximizes a number of non-smooth ℓ_1-norm terms. Although there exist in literature a plethora of algorithms (Argyriou et al., 2008; Nie et al., 2010; Wang et al., 2011; 2012) that can minimize the objectives involving ℓ_1-norm or $\ell_{2,1}$-norm terms, how to efficiently solve the objectives that simultaneously minimize and maximize ℓ_1-norm terms are rarely studied. As an important theoretical contribution of this paper, we will derive an efficient algorithm to solve the general ℓ_1-norm minmax problem in Eq. (6) in the rest of this section.

Before deriving our new solution algorithm, we first introduce the following useful proposition.

Proposition 1 *For a linear learning model, suppose the output projection vector* $\mathbf{u}$ *can be linearly represented by the input data* $\mathbf{X}$*, and suppose an orthonormal projection matrix* $\mathbf{V}$ *has been learned by the same model,* i.e., $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. *If we learn another projection vector* $\mathbf{u}$ *from*

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{V}\mathbf{V}^T\mathbf{X} \tag{7}$$

by the same learning model, then $\mathbf{u}$ *is orthogonal to* $\mathbf{V}$.

Proof. Since $\mathbf{u}$ can be linearly represented by the input data $\widetilde{\mathbf{X}}$, we have $\mathbf{u} = \widetilde{\mathbf{X}}\boldsymbol{\alpha}$ for certain $\boldsymbol{\alpha}$. Thus $\mathbf{V}^T\mathbf{u} = \mathbf{V}^T\left(\mathbf{X} - \mathbf{V}\mathbf{V}^T\mathbf{X}\right)\boldsymbol{\alpha} = \left(\mathbf{V}^T\mathbf{X} - \mathbf{V}^T\mathbf{V}\mathbf{V}^T\mathbf{X}\right)\boldsymbol{\alpha} = \left(\mathbf{V}^T\mathbf{X} - \mathbf{V}^T\mathbf{X}\right)\boldsymbol{\alpha} = 0$, which completes the proof of Proposition 1. ■

The main strength of Proposition 1 lies in that it converts a difficult orthogonally constrained optimization problem to a series of unconstrained optimization problems by using the reconstruction residues from the learned projections in Eq. (7), which, fortunately, is usually much easier to solve.

Equipped with Proposition 1, we derive the solution algorithm to the problem in Eq. (6) in Algorithm 1.

Algorithm 1 An efficient iterative algorithm to solve the general ℓ_1-norm minmax problem with orthogonal constraint in Eq. (6).

```
Input: Input data $\mathcal{X}$, and the must-links $\mathcal{S}$ and cannot-links $\mathcal{D}$.
1. Let $\mathbf{X}_1 = \mathbf{X}$. Construct $\mathbf{A}_1$ and $\mathbf{B}_1$ from $\mathbf{X}_1$, $\mathcal{S}$ and $\mathcal{D}$, and compute $\mathbf{w}_1$
by solving the problem $\min_{\mathbf{w}_1} \frac{\|\mathbf{A}\mathbf{w}_1\|_1}{\|\mathbf{B}\mathbf{w}_1\|_1}$.
2. $j = 2$.
while $j \leq r$ do
    3. Compute $\mathbf{X}_j = \mathbf{X}_{j-1} - \mathbf{w}_{j-1}\mathbf{w}_{j-1}^T\mathbf{X}_{j-1}$.
    4. Construct $\mathbf{A}_j$ and $\mathbf{B}_j$ from $\mathbf{X}_j$, $\mathcal{S}$ and $\mathcal{D}$.
    5. Compute $\mathbf{w}_j$ by solving the problem $\min_{\mathbf{w}_j} \frac{\|\mathbf{A}_j\mathbf{w}_j\|_1}{\|\mathbf{B}_j\mathbf{w}_j\|_1}$.
    6. $j = j + 1$.
end while
Output: The learned projection matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_r] \in \mathbb{R}^{d \times r}$.
```

In Algorithm 1, we learn one projection vector at a time. Because we learn every projection vector on the reconstruction residue from the previously learned projections, according to Proposition 1, the learned projection vectors $\mathbf{w}_j$ $(1 \leq j \leq r)$ are orthogonal to each other.

Obviously, Step 1 and Step 5 are the key steps of Algorithm 1, which solve the following problem:

$$\min_{\mathbf{w}} \frac{\|\mathbf{A}\mathbf{w}\|_1}{\|\mathbf{B}\mathbf{w}\|_1} \ . \tag{8}$$

In the following, we will derive the solution to Eq. (8). The detailed procedures are summarized in Algorithm 3.

3.1. Useful Theorems

Theorem 1 *The global solution of the following general optimization problem:*

$$\min_{\mathbf{v}\in\mathcal{C}} \frac{f(\mathbf{v})}{g(\mathbf{v})} \ , \quad \textit{where} \ \ g(\mathbf{v}) > 0 \ \ (\forall \ \mathbf{v} \in \mathcal{C}) \ , \tag{9}$$

is computed by the root of the following function:

$$h(\lambda) = \min_{\mathbf{v}\in\mathcal{C}} \ f(\mathbf{v}) - \lambda g(\mathbf{v}) \ , \tag{10}$$

given that $f(\mathbf{v}) - \lambda g(\mathbf{v})$ is lower bounded.

Proof. Suppose $\mathbf{v}^*$ is the global solution of the problem in Eq. (9), and λ^* is the corresponding global minimal objective value, the following holds: $\frac{f(\mathbf{v}^*)}{g(\mathbf{v}^*)} = \lambda^*$. Thus $\forall \ \mathbf{v} \in \mathcal{C}$, we have $\frac{f(\mathbf{v})}{g(\mathbf{v})} \geq \lambda^*$. Because we know that $g(\mathbf{v}) > 0$, we can derive $f(\mathbf{v}) - \lambda^* g(\mathbf{v}) \geq 0$, which means:

$$\min_{\mathbf{v} \ \in\mathcal{C}} \ f(\mathbf{v}) - \lambda^* g(\mathbf{v}) = 0 \quad \Longleftrightarrow \quad h(\lambda^*) = 0 \ . \tag{11}$$

That is, the global minimal objective value λ^* of the problem in Eq. (9) is the root of the function $h(\lambda)$, which completes the proof of Theorem 1. ■

To solve the problem in Eq. (9), we propose an iterative algorithm as summarized in Algorithm 2, whose convergence is guaranteed by Theorem 2 and whose computational efficiency is guaranteed by Theorem 3.

Algorithm 2 The algorithm to solve Eq. (9).

```
1. $t = 1$. Initialize $\mathbf{v}_t \in \mathcal{C}$.
while not converge do
    2. Calculate $\lambda_t = \frac{f(\mathbf{v}_t)}{g(\mathbf{v}_t)}$.
    3. Calculate $\mathbf{v}_{t+1} = \arg\min_{\mathbf{v}\in\mathcal{C}} f(\mathbf{v}) - \lambda_t g(\mathbf{v})$.
    4. $t = t + 1$.
end while
```

Theorem 2 *Algorithm 2 decreases the objective value of the problem in Eq. (9) in each iteration till converges.*

Proof. In Algorithm 2, from step 2 we know that $f(\mathbf{v}_t) - \lambda_t g(\mathbf{v}_t) = 0$. According to step 3, we know that $f(\mathbf{v}_{t+1}) - \lambda_t g(\mathbf{v}_{t+1}) \leq f(\mathbf{v}_t) - \lambda_t g(\mathbf{v}_t)$. Combining the above two inequalities, we have $f(\mathbf{v}_{t+1}) - \lambda_t g(\mathbf{v}_{t+1}) \leq 0$, which indicates $\frac{f(\mathbf{v}_{t+1})}{g(\mathbf{v}_{t+1})} \leq \lambda_t = \frac{f(\mathbf{v}_t)}{g(\mathbf{v}_t)}$. That is, Algorithm 2 decreases the objective value of Eq. (9) in each iteration, which completes the proof of Theorem 2. ■

Theorem 3 *Algorithm 2 is a Newton's method that finds the root of the function $h(\lambda)$ in Eq. (10).*

Proof. From step 3 in Algorithm 2 we know that

$$h(\lambda_t) = f(\mathbf{v}_{t+1}) - \lambda_t g(\mathbf{v}_{t+1}) \ . \tag{12}$$

Thus $h'(\lambda_t) = -g(\mathbf{v}_{t+1})$.

In Newton's method, the updated solution should be

$$\begin{aligned} \lambda_{t+1} &= \lambda_t - \frac{h(\lambda_t)}{h'(\lambda_t)} \\ &= \lambda_t - \frac{f(\mathbf{v}_{t+1}) - \lambda_t g(\mathbf{v}_{t+1})}{-g(\mathbf{v}_{t+1})} = \frac{f(\mathbf{v}_{t+1})}{g(\mathbf{v}_{t+1})} \ , \end{aligned} \tag{13}$$

which is exactly the step 2 in Algorithm 2. That is, Algorithm 2 is a Newton's method to find the root of the function $h(\lambda)$. ■

Theorems (1–3) present a complete framework to solve the general optimization problem in Eq. (9), where an efficient iterative algorithm is provided in Algorithm 2 with rigorously proved convergence and satisfactory computational efficiency. It is worth to noting that, besides applying it to solve the ℓ_1-norm minmax problem as in our objective in Eq. (6), we can also employ this framework to efficiently solve many other optimization problems that widely appear in machine learning and statistics, *e.g.*, the trace-ratio minimization problem in Eq. (3) (Jia et al., 2009; Wang et al., 2014). Therefore, Theorems (1–3) are considered as one of the most important theoretical contribution of this paper.

3.2. Derivation of Algorithm 3

Now we derive the solution algorithm to Eq. (8) using the optimization framework described in Theorems (1–3).

Because the problem in Eq. (8) is a special case of the general optimization problem in Eq. (9), we can derive its solution algorithm using Algorithm 2, in which the key step is to solve the problem in Step 2. Given the λ computed from Step 1 of Algorithm 2, according to Step 2 of Algorithm 2, instead of solving the original problem in Eq. (8), we turn to solve the following optimization problem:

$$\min_{\mathbf{w}} \|\mathbf{A}\mathbf{w}\|_1 - \lambda \|\mathbf{B}\mathbf{w}\|_1 \quad . \tag{14}$$

Solving the problem in Eq. (14), again, is challenging, because it is non-smooth due to involving the non-smooth ℓ_1-norm terms. We derive its solution as following.

By taking the derivative of Eq. (14) with respect $\mathbf{w}$ and setting it as 0, we obtain[1]:

$$\sum_i \frac{2\left(\mathbf{a}^i\right)^T \mathbf{a}^i \mathbf{w}}{2|\mathbf{a}^i \mathbf{w}|} - \lambda \sum_i \frac{\left(\mathbf{b}^i\right)^T \mathbf{b}^i \mathbf{w}}{|\mathbf{b}^i \mathbf{w}|} = 0 \quad . \tag{15}$$

Let $d_{ii} = \frac{1}{2|\mathbf{a}^i \mathbf{w}|}$ and construct the diagonal matrix $\mathbf{D}$ with its diagonal entries as d_{ii}, and let $s_i = \frac{\mathbf{b}^i \mathbf{w}}{|\mathbf{b}^i \mathbf{w}|}$ and construct the vector $\mathbf{s}$ with its i-th element as s_i, we can rewrite Eq. (15) as following:

$$2\mathbf{A}^T \mathbf{D}\mathbf{A}\mathbf{w} - \lambda \mathbf{B}^T \mathbf{s} = \mathbf{0} \quad . \tag{16}$$

Thus, we can compute $\mathbf{w}$ by solving a system of linear equations in Eq. (16)[2].

Note that, in Eq. (16) both $\mathbf{D}$ and $\mathbf{s}$ are dependent on $\mathbf{w}$. Therefore they are unknown variables and can be seen as the latent variables of the objective in Eq. (14), which can be solved under the same iterative framework by alternatively optimizing them (Nie et al., 2010). Specifically, we compute $\mathbf{D}$ and $\mathbf{s}$ upon the current $\mathbf{w}$ obtained in the last iteration. Finally, we summarize the whole computation procedures in Algorithm 3, which, to our best knowledge, solves the ℓ_1-norm minmax problem for the first time[3].

Algorithm 3 An efficient iterative algorithm to solve the general ℓ_1-norm minmax problem in Eq. (8).

```
Input: Matrices $\mathbf{A}$ and $\mathbf{B}$.
1. Initialize $\mathbf{w}$ by a random guess.
repeat
    2. Compute $\lambda = \frac{\|\mathbf{A}\mathbf{w}\|_1}{\|\mathbf{B}\mathbf{w}\|_1}$.
    3. Compute the diagonal matrix $\mathbf{D}$ with its $i$-th diagonal entry as $d_{ii} = \frac{1}{2|\mathbf{a}^i\mathbf{w}|}$.
    4. Compute the vector $\mathbf{s}$ with its $i$-th entry as $s_i = \frac{\mathbf{b}^i\mathbf{w}}{|\mathbf{b}^i\mathbf{w}|}$.
    5. Compute $\mathbf{w}$ by solving the linear equations $2\mathbf{A}^T\mathbf{D}\mathbf{A}\mathbf{w} = \lambda\mathbf{B}^T\mathbf{s}$.
until Converges
Output: The learned projection vector $\mathbf{w}$.
```

3.3. Analysis of Algorithm 3

Convergence analysis of the algorithm. The following theorem guarantees the convergence of Algorithm 3.

Theorem 4 *Algorithm 3 decreases the objective value of the problem in Eq. (8) in each iteration till converges.*

Proof. First, it is obvious that Step 5 of Algorithm 3 computes the optimal solution of the following problem:

$$\min_{\mathbf{w}} \; \mathbf{w}^T \mathbf{A}^T \mathbf{D}\mathbf{A}\mathbf{w} - \lambda \mathbf{w}^T \mathbf{B}^T \mathbf{s} \quad . \tag{17}$$

For each iteration, we denote by $\widetilde{\mathbf{w}}$ the updated $\mathbf{w}$ by Algorithm 3. According to Step 5 of Algorithm 3 and Eq. (17), we have the following inequality:

$$\begin{aligned} &\widetilde{\mathbf{w}}^T \mathbf{A}^T \mathbf{D}\mathbf{A}\widetilde{\mathbf{w}} - \lambda \widetilde{\mathbf{w}}^T \mathbf{B}^T \mathbf{s} \leq \\ &\mathbf{w}^T \mathbf{A}^T \mathbf{D}\mathbf{A}\mathbf{w} - \lambda \mathbf{w}^T \mathbf{B}^T \mathbf{s} \quad . \end{aligned} \tag{18}$$

Applying the definitions of $\mathbf{A}$ and $\mathbf{s}$ in Steps 3–4, we can equivalently rewrite Eq. (18) by decoupling the computation for each row of $\mathbf{A}$ and $\mathbf{B}$ as following:

$$\begin{aligned} &\sum_i \frac{\left(\mathbf{a}^i \widetilde{\mathbf{w}}\right)^2}{2|\mathbf{a}^i \mathbf{w}|} - \lambda \sum_i \frac{\widetilde{\mathbf{w}}^T \left(\mathbf{b}^i\right)^T \mathbf{b}^i \mathbf{w}}{|\mathbf{b}^i \mathbf{w}|} \leq \\ &\sum_i \frac{\left(\mathbf{a}^i \mathbf{w}\right)^2}{2|\mathbf{a}^i \mathbf{w}|} - \lambda \sum_i \frac{\mathbf{w}^T \left(\mathbf{b}^i\right)^T \mathbf{b}^i \mathbf{w}}{|\mathbf{b}^i \mathbf{w}|} \quad , \end{aligned} \tag{19}$$

where we note that $\mathbf{a}^i \widetilde{\mathbf{w}}$ is a scalar and $\left(\mathbf{a}^i \mathbf{w}\right)^T \left(\mathbf{a}^i \mathbf{w}\right) = |\mathbf{a}^i \widetilde{\mathbf{w}}|^2 = (\mathbf{a}^i \widetilde{\mathbf{w}})^2$.

Because it can verified that for function $z(x) = x - \frac{x^2}{2\alpha}$, given any $x \neq \alpha \in \mathbb{R}^n$, $z(x) \leq z(\alpha)$ holds, we have:

$$|\mathbf{a}^i \widetilde{\mathbf{w}}| - \frac{\left(\mathbf{a}^i \widetilde{\mathbf{w}}\right)^2}{2|\mathbf{a}^i \mathbf{w}|} \leq |\mathbf{a}^i \mathbf{w}| - \frac{\left(\mathbf{a}^i \mathbf{w}\right)^2}{2|\mathbf{a}^i \mathbf{w}|} \quad . \tag{20}$$

[1] Because the ℓ_1-norm is non-smooth, following (Argyriou et al., 2008), we address this by introducing a small perturbation to replace $\|\mathbf{v}\|_1$ by $\sum_i \sqrt{v_i^2 + \zeta}$, where $\mathbf{v}$ is a general vector and ζ is a small positive constant. Apparently, $\sum_i \sqrt{v_i^2 + \zeta}$ reduces to $\|\mathbf{v}\|_1$ when $\zeta \rightarrow 0$. In the rest of this paper, this replacement is always implicitly applied in the definition of $\|\cdot\|_1$, unless otherwise stated.

Similarly, given a scalar variable v, we always replace $|v|$ by $\sqrt{v^2 + \zeta}$ to address the non-smoothness of $|\cdot|$.

[2] Note that, in real problems $\mathbf{A}^T\mathbf{D}\mathbf{A}$ could be rank deficient, because the number of constraints in the must-links could be smaller than the dimensionality of the feature space. Therefore, in practice we compute $\mathbf{w}$ by solving $\left(2\mathbf{A}^T\mathbf{D}\mathbf{A} + \zeta\mathbf{I}\right)\mathbf{w} = \lambda\mathbf{B}^T\mathbf{s}$, where $\mathbf{I}$ is the identity matrix and ζ is a small pertubation.

[3] The objective of the ℓ_1-norm minmax problem in Eq. (14) is non-convex, thus the algorithm is guaranteed to converge to a local optima of the objective. A proper initialization could effectively avoid being trapped by local optima. In practice, we replace the ℓ_1-norm in Eq. (14) by the squared ℓ_2-norm and solve it as initialization, which empirically works very well in our experiments.

In addition, according to the Cauchy-Schwarz inequality we have:

$$\left|\mathbf{b}^i\widetilde{\mathbf{w}}\right| \left|\mathbf{b}^i\mathbf{w}\right| \geq \left(\mathbf{b}^i\widetilde{\mathbf{w}}\right)^T \mathbf{b}^i\mathbf{w} \ , \tag{21}$$

which implies that

$$\begin{aligned} &\frac{\widetilde{\mathbf{w}}^T \left(\mathbf{b}^i\right)^T \mathbf{b}^i\mathbf{w}}{\left|\mathbf{b}^i\mathbf{w}\right|} - \left|\mathbf{b}^i\widetilde{\mathbf{w}}\right| \leq 0 = \left|\mathbf{b}^i\mathbf{w}\right| - \left|\mathbf{b}^i\mathbf{w}\right| = \\ &\frac{\mathbf{w}^T \left(\mathbf{b}^i\right)^T \mathbf{b}^i\mathbf{w}}{\left|\mathbf{b}^i\mathbf{w}\right|} - \left|\mathbf{b}^i\mathbf{w}\right| \ . \end{aligned} \tag{22}$$

Then by adding the three inequalities in Eq. (19), Eq. (20) and Eq. (22) in the both sides, we obtain:

$$\begin{aligned} &\sum_i \left|\mathbf{a}^i\widetilde{\mathbf{w}}\right| - \lambda \sum_i \left|\mathbf{b}^i\widetilde{\mathbf{w}}\right| \leq \sum_i \left|\mathbf{a}^i\mathbf{w}\right| - \lambda \sum_i \left|\mathbf{b}^i\mathbf{w}\right| = 0 \\ \Longrightarrow \ &\frac{\|\mathbf{A}\widetilde{\mathbf{w}}\|_1}{\|\mathbf{B}\widetilde{\mathbf{w}}\|_1} = \frac{\sum_i \left|\mathbf{a}^i\widetilde{\mathbf{w}}\right|}{\sum_i \left|\mathbf{b}^i\widetilde{\mathbf{w}}\right|} \leq \lambda = \frac{\sum_i \left|\mathbf{a}^i\mathbf{w}\right|}{\sum_i \left|\mathbf{b}^i\mathbf{w}\right|} = \frac{\|\mathbf{A}\mathbf{w}\|_1}{\|\mathbf{B}\mathbf{w}\|_1} \ . \end{aligned} \tag{23}$$

Note that, the equalities in Eqs. (19–23) hold if and only if the objective value converges. Therefore, the objective value of the problem in Eq. (8) is decreased in each iteration till converges, which completes the proof of Theorem 4. ■

Computational analysis of the algorithm. In Algorithm 3, Steps 2–4 are computationally trivial. Step 5 solves a system of linear equations, which is very well studied and efficient solution algorithms exist. Moreover, according to Theorem 3, Algorithm 3 implements a Newton's method to find the root of the function defined over the objective value of the problem in Eq. (8). Thus, Algorithm 3 converges very fast with the quadratic convergence rate, *i.e.*, the difference between the current objective value and the optimal objective value is smaller than $\frac{1}{c^{c^t}}$ at the t-th iteration, where $c > 1$ is a certain constant. In summary, Algorithm 3 scales very well to large-scale data sets, which adds to the practical value of the proposed method.

4. Experiments

In this section, we evaluate the proposed method in the tasks of data clustering, where our goal is to examine the robustness of our new method under the conditions when data outliers or feature outliers are present.

4.1. Data Preparation

We experiment with four benchmark data sets downloaded from the UCI machine learning data repository, including the **Breast**, **Diabetes**, **Iris** and **Protein** data sets, and one image data set downloaded from the **ORL** database, whose details are summarized in Table 1.

Following previous research, we generate the must-links and cannot-links for each data set as follows. For each constraint, we randomly pick up one pair of data points from the input data sets (the labels of which are available for evaluation purpose but unavailable for clustering). If the labels of this pair of data points are the same, we generate a must link. If the labels are different, a cannot link is generated. We pick up 100 constraints for each data set.

Table 1. Descriptions of the experimental data sets.

Data set	Breast	Diabetes	Iris	Protein	ORL
n	683	768	150	116	400
d	10	8	4	20	10304
c	5	4	2	8	40
r	10	8	4	16	80

n: number of samples. d: input dimensionality.
c: number of clusters. r: reduced dimensionality.

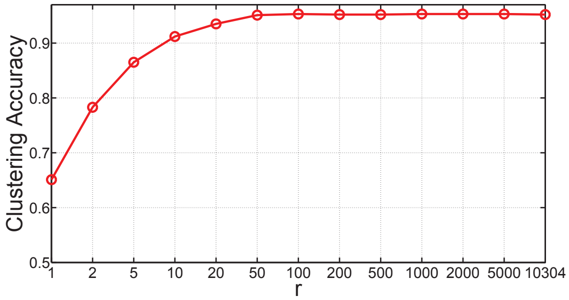


Figure 1. Clustering performance on the ORL data with respect to r (the dimensionality of the projected (sub)space by $\mathbf{W}$).

4.2. Parameter Selection of Our Method

The proposed method has only one parameter, *i.e.*, r of the reduced dimensionality of the projected (sub)space by the transformation of $\mathbf{W}$. In this subsection, we study its impact to the learned distance metrics by performing clustering on the ORL data, where we vary the value of r from its minimum possible value of 1 to its maximum possible value of 10304. For each experimental trial, first we learn a distance metric from the input data with a given value of the parameter r, then we perform K-means clustering using the learned distance metric. For each different value of r, we repeat the experiment for 100 times to eliminate the difference caused by the constraint pickup and the initialization of K-means clustering. The clustering performance measured by clustering accuracy averaged over 100 trials are reported in Figure 1.

A first glance at the results in Figure 1 shows that the clustering performance is very stable with respect to the parameter r in a considerably large range of $[50, 10304]$, which makes tuning parameter of our new method not a difficult task. In addition, we also notice that, when r is small, *e.g.*, when $r < 50$, the clustering performance is not satisfactory, which can be seen as follows. As discussed earlier, metric learning can be equivalently performed as subspace

Table 2. Clustering performances of the compared methods measured by clustering accuracy (mean ± std).

	Original data	Noisy data	Perf. diff.	Original data	Noisy data	Perf. diff.
		Breast data			Diabetes data	
Eu	94.23 ± 1.23	90.53 ± 1.03	3.92%	55.83 ± 2.11	50.12 ± 2.23	10.40%
Mah	95.01 ± 0.21	91.42 ± 0.33	3.87%	58.12 ± 1.74	52.41 ± 1.81	9.82%
Xing's	94.24 ± 1.12	90.43 ± 0.96	4.04%	56.61 ± 1.88	51.11 ± 1.64	9.72%
RCA	93.36 ± 0.54	89.45 ± 0.62	4.19%	58.33 ± 1.21	53.01 ± 1.32	9.12%
DCA	92.07 ± 0.96	89.12 ± 0.88	3.59%	57.53 ± 1.63	50.23 ± 1.44	12.69%
Xiang's	94.48 ± 0.41	90.34 ± 0.44	4.38%	60.91 ± 0.64	54.15 ± 0.77	11.10%
LMNN	95.12 ± 0.18	91.56 ± 0.13	3.74%	60.44 ± 1.07	53.22 ± 1.11	11.94%
ITML	93.72 ± 1.01	90.24 ± 0.99	3.71%	60.21 ± 0.81	53.67 ± 1.03	10.86%
Our method	**95.94 ± 0.44**	**94.54 ± 0.51**	**1.46%**	**62.14 ± 0.37**	**60.67 ± 0.42**	**2.37%**
		Iris data			Protein data	
Eu	85.52 ± 2.45	80.41 ± 2.51	5.97%	66.22 ± 2.11	61.54 ± 2.65	7.11%
Mah	94.42 ± 1.15	89.44 ± 1.07	5.27%	68.02 ± 1.35	62.37 ± 1.41	8.31%
Xing's	92.36 ± 1.66	88.41 ± 1.95	4.27%	68.13 ± 1.62	63.41 ± 1.50	6.93%
RCA	95.91 ± 1.72	89.40 ± 1.85	6.79%	68.09 ± 1.11	62.17 ± 1.23	8.69%
DCA	96.54 ± 0.34	90.15 ± 0.74	6.62%	62.43 ± 2.11	58.41 ± 1.95	6.44%
Xiang's	96.60 ± 0.31	91.24 ± 0.96	5.55%	73.47 ± 0.41	65.42 ± 0.77	10.96%
LMNN	96.41 ± 0.39	90.60 ± 0.86	6.03%	72.15 ± 0.56	66.10 ± 0.86	8.39%
ITML	93.27 ± 0.74	88.95 ± 1.21	4.63%	73.94 ± 0.11	66.48 ± 0.69	10.09%
Our method	**97.03 ± 0.15**	**95.17 ± 0.31**	**1.91%**	**74.14 ± 0.19**	**72.15 ± 0.37**	**2.68%**

learning, because $\mathbf{W}$ is positive semi-definite. Therefore, when r is too small, the input data are projected into a very low-dimensional subspace, which might not have sufficient representative capability to properly express the clustering structures of the input data and lead to inferior clustering performances.

Based upon the above observations, we tentatively draw the following conclusion. As long as r is not too small, we can generally achieve decent clustering performances using the learned distance metrics. Empirically, we select $r = \min(d, 2c)$ in all our subsequent experiments, where d is the dimensionality of the original data space and c is the cluster number of the input data. The values of r for the five experimental data sets are listed in the last row of Table 1.

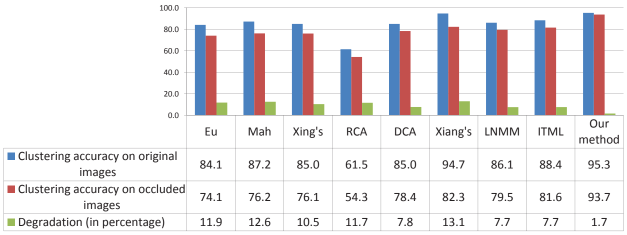
4.3. Clustering on Data with Outlier Samples

We evaluate the proposed method on noisy data with outlier samples using the four UCI data sets. We compare our method against its two closest counterparts, including (1) the Euclidean distance (**EU**) that sets $\mathbf{M} = \mathbf{I}$ and (2) the standard Mahalanobis distance (**Mah**) that sets the distance metric as the inverse of the sample covariance matrix, *i.e.* $\mathbf{M} = (\mathrm{Cov}(\mathbf{X}))^{-1}$. We also compare our method against several related and most recent metric learning methods, including (3) **Xing's** method (Xing et al., 2002), (4) **RCA** method (Bar-Hillel et al., 2003), (5) **DCA** method (Hoi et al., 2006), (6) **Xiang's** method (Xiang et al., 2008), (7) Large Margin Nearest Neighbor (**LMNN**) method (Weinberger et al., 2006) and (8) Information-Theoretic Metric Learning (**ITML**) method (Davis et al., 2007). We implement these compared methods following their original papers, and fine tune their parameters to achieve the best clustering accuracy in independent preliminary experiments. Again, once the distance metric is learned by a method on a data set, K-means clustering is performed on the same data set using the learned distance metric.

We conduct experiments in following two conditions: (1) original data and (2) noisy data with outlier samples. To emulate the outlier data samples, given the input data set $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, we corrupt it by a noise matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{d \times n}$ whose element are i.i.d. standard Gaussian variables. Then we carry out the same learning and clustering procedures on $\mathbf{X} + \sigma\tilde{\mathbf{X}}$ as those on the original data, where $\sigma = nf \frac{\|\mathbf{X}\|_F}{\|\tilde{\mathbf{X}}\|_F}$ and nf is a given noise factor. In all our experiments, we set $nf = 0.1$.

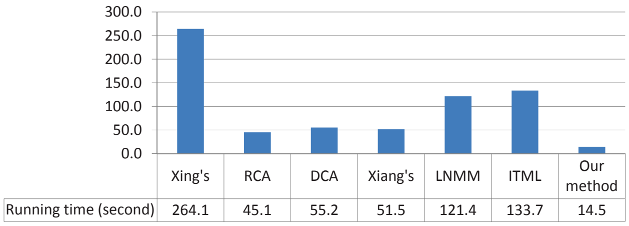
For every experimental case, the clustering performance measured by clustering accuracy are averaged over 100 trials to eliminate the difference caused picking up the constraints and initializing the K-means clustering procedures, which are reported in Table 2.

From Table 2 we have the following interesting observations. First, our method is consistently better than all other compared methods on all four experimental data sets, which demonstrate that our method is able to learn an effective distance metric that can improve the clustering performance. Second, although the improvements by our method over the competing methods on the original data are mediocre, the improvements by our method on the contaminated data with outlier data samples are considerably



	Eu	Mah	Xing's	RCA	DCA	Xiang's	LNMM	ITML	Our method
🟦 Clustering accuracy on original images	84.1	87.2	85.0	61.5	85.0	94.7	86.1	88.4	95.3
🟥 Clustering accuracy on occluded images	74.1	76.2	76.1	54.3	78.4	82.3	79.5	81.6	93.7
🟩 Degradation (in percentage)	11.9	12.6	10.5	11.7	7.8	13.1	7.7	7.7	1.7

Figure 2. Clustering performance of the compared methods on ORL image data set.



	Xing's	RCA	DCA	Xiang's	LNMM	ITML	Our method
Running time (second)	264.1	45.1	55.2	51.5	121.4	133.7	14.5

Figure 3. Running time of the compared methods to learn the distance metric matrix on the ORL image data set.

large. For example, on the noisy Diabetes data set, our new distance metric learning method improves the clustering performance over the simplest Euclidean distance method by $21.05\% = (60.67 - 50.12)/50.12$ and outperforms Xiang's method (with the best performance on the data set) by $12.04\% = (60.67 - 54.15)/54.15$. Finally, by a more careful examination on the experimental results, we also notice that the clustering performances for all the methods are degraded due to introducing outlier data samples, however, as shown in the columns of "Perf. diff." in Table 2, the degradations of the proposed method are much less than those of the other compared methods. The degradations of our method on all the four data sets are less than 3%. This important observation clearly demonstrates the robustness of our method against outlier data samples and empirically justifies our motivation to use the ℓ_1-norm distance to improve the distance metric learning.

4.4. Clustering on Data with Outlier Features

Because we replace the traditional squared ℓ_2-norm distance by the ℓ_1-norm distance in our learning objective, the learned distance metric is robust against not only outlier samples but also outliers features. Thus in this subsection, we evaluate the robustness against feature outliers of the proposed method on the ORL image data. The ORL data set includes 40 distinct individuals and each individual has 10 gray images with different expressions and facial details. The size of each image in this data set is 112×92. Besides performing clustering on the original ORL data following the same procedures as in the previous subsection, we also emulate corrupted features by occluding the images. For each image, we first randomly pick up a location and place a black square of size 25×25 onto the image. Then we perform clustering on the corrupted images. We still compare our method against the 8 competing methods using the same experimental settings as described in the previous subsection. We repeat each test case for 100 times and report the average clustering accuracy in Figure 2.

Figure 2 shows that our method is superior to all other competing methods on both the original images and the images with occlusions. Moreover, on the occluded images where the features are contaminated, compared to other methods, the performance degradations of our new method is very small, which provide one more concrete evidence to support the usefulness of the ℓ_1-norm distance in metric learning and confirm the correctness of the proposed method.

4.5. Computational Efficiency of Our Method

Finally, we evaluate the computational efficiency of the proposed method, because, as analyzed earlier, it is one of the most important advantage of the proposed method. We report in Figure 3 the running time of the compared methods on the ORL image data set, where we experiment on a Dell PowerEdge 2900 server, which has two quad-core Intel Xeon 5300 CPUs at 3.0 GHz and 48G bytes memory. Because the Euclidean distance (EU) method and the Mahalanobis distance (Mah) do not learn the distance metric matrices, they are not involved in this experiment. From Figure 3 we can see that our new metric learning method requires significantly less time to learn the distance metric matrix, which firmly demonstrates the computational advantage of the proposed method.

5. Conclusions

We proposed a robust distance metric learning method using the ℓ_1-norm distances, which formulated a simultaneous ℓ_1-norm minimization and maximization (minmax) problem. The new objective uses the ℓ_1-norm between both data points and features, thus our method is more robust to outliers. However, the new objective is much more challenging to optimize, to solve which we derived an efficient algorithm and rigorously proved its convergence. We have performed extensive experiments on both noiseless and noisy data, which have shown that the proposed methods are more effective and more robust against outlier samples and outlier features than traditional methods.

Acknowledgments

Corresponding Author: Heng Huang (heng@uta.edu). This work was partially supported by US NSF IIS-1117965, IIS-1302675, IIS-1344152.

References

Argyriou, A., Evgeniou, T., and Pontil, M. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

Bar-Hillel, A., Hertz, T., Shental, N., and Weinshall, D. Learning distance functions using equivalence relations. In *ICML*, 2003.

Cayton, L. and Dasgupta, S. Robust euclidean embedding. In *ICML*, pp. 169–176, 2006.

Davis, J.V., Kulis, B., Jain, P., Sra, S., and Dhillon, I.S. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pp. 209–216. ACM, 2007.

Ding, C., Zhou, D., He, X., and Zha, H. R1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *ICML*, 2006.

Gao, J. Robust l1 principal component analysis and its bayesian variational inference. *Neural Computation*, 20: 555–572, 2008.

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

Jia, Y., Nie, F., and Zhang, C. Trace ratio problem revisited. *IEEE Transactions on Neural Networks (TNN)*, 20(4): 729–735, 2009.

Ke, Q. and Kanade, T. Robust l1 norm factorization in the presence of outliers and missing data by alternative convex programming. *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 592–599, 2004.

Kwak, N. Principal component analysis based on l1-norm maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1672–1680, 2008.

Nie, F., Huang, H., Cai, X., and Ding, C. Efficient and Robust Feature Selection via Joint l2,1-Norms Minimization. In *NIPS*, 2010.

Nie, Feiping, Huang, Heng, Ding, Chris, Luo, Dijun, and Wang, Hua. Robust principal component analysis with non-greedy ℓ_1-norm maximization. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Two*, pp. 1433–1438. AAAI Press, 2011.

Wang, H., Nie, F., and Huang, H. Globally and Locally Consistent Unsupervised Projection. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*, 2014.

Wang, Hua, Nie, Feiping, Huang, Heng, Risacher, Shannon, Saykin, Andrew J, and Shen, Li. Identifying ad-sensitive and cognition-relevant imaging biomarkers via joint classification and regression. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2011)*, pp. 115–123. Springer, 2011.

Wang, Hua, Nie, Feiping, Huang, Heng, Yan, Jingwen, Kim, Sungeun, Risacher, Shannon L, Saykin, Andrew J, and Shen, Li. High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction. In *NIPS*, pp. 1286–1294, 2012.

Wang, Hua, Nie, Feiping, and Huang, Heng. Semi-Supervised Robust Dictionary Learning via Efficient $\ell_{2,0+}$-Norms Minimization . In *Proceedings of the 14th IEEE International Conference on Computer Vision (ICCV 2013)*, pp. 1145–1152, 2013a.

Wang, Hua, Nie, Feiping, and Huang, Heng. Robust and discriminative self-taught learning. In *Proceedings of The 30th International Conference on Machine Learning (ICML 2013)*, pp. 298–306, 2013b.

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

Wright, John, Ganesh, Arvind, Rao, Shankar, Peng, Yigang, and Ma, Yi. Robust principal component analysis: Exact recovery of corrupted. *Advances in Neural Information Processing Systems*, pp. 116, 2009.

Xiang, S., Nie, F., and Zhang, C. Learning a mahalanobis distance metric for data clustering and classification. *Pattern Recognition*, 41(12):3600–3612, 2008.

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

Zha, Z.J., Mei, T., Wang, M., Wang, Z., and Hua, X.S. Robust distance metric learning with auxiliary knowledge. In *IJCAI*, pp. 1327–1332, 2009.