

Cutting L1-Norm Distance Discriminant Analysis with Sample Reconstruction

Guowan SHAO ^{a,1}, Chunjiang PENG ^a, Wenchu OU ^a and Kai DUAN ^b

^a College of Mechanical and Electrical Engineering, Hunan University of Science and Technology, Xiangtan 411201, Hunan, PR China

^b Engineering Research Center for Advanced Mine Equipment of Ministry of Education, Hunan University of Science and Technology, Xiangtan 411201, Hunan, PR China

Abstract. Dimensionality reduction plays an important role in the fields of pattern recognition and computer vision. Recursive discriminative subspace learning with an L1-norm distance constraint (RDSL) is proposed to robustly extract features from contaminated data and L1-norm and slack variables are utilized for accomplishing the goal. However, its performance may decline when too many outliers are available. Moreover, the method ignores the global structure of the data. In this paper, we propose cutting L1-norm distance discriminant analysis with sample reconstruction (C-L1-DDA) to solve the two problems. We apply cutting L1-norm to measure within-class and between-class distances and thus outliers may be strongly suppressed. Moreover, we use cutting squared L2-norm to measure reconstruction errors. In this way, outliers may be constrained and the global structure of data may be approximately preserved. Finally, we give an alternating iterative algorithm to extract feature vectors. Experimental results on two publicly available real databases verify the feasibility and effectiveness of the proposed method.

Keywords. Discriminant analysis, dimensionality reduction, cutting L1-norm, sample reconstruction

1. Introduction

Dimensionality reduction is utilized for preprocessing high dimensional data, which plays an irreplaceable role in the fields of pattern recognition and computer vision. In the past decades, it has been extensively studied and a variety of methods have been presented, such as linear discriminant analysis (LDA) [1], locality preserving projections [2] and principal component analysis (PCA) [3]. Among them, LDA and PCA are two basic methods, which are widely used in practice. The former is a supervised method and based on the criterion that the distances between samples from the same class are minimized and the distances between samples from different classes are maximized simultaneously.

LDA is an effective method for dimensionality reduction. However, its performance may decline when the data are contaminated by outliers or noise. Many methods have

¹ Corresponding Author, Guowan SHAO, College of Mechanical and Electrical Engineering, Hunan University of Science and Technology, Xiangtan 411201, Hunan, PR China; E-mail: gwshao_ezhou@163.com, 821211624@qq.com.

been presented to deal with the problem and enhance the performance [4-6]. From the view of mathematics, L1-norm is more robust to outliers or noise than L2-norm and hence some methods replace the latter with the former and improve the robustness [7-11]. Rotational invariant LDA (RILDA) [7] constructs a unified rotational invariant framework with L2,1-norm and generalizes Laplacian eigenmaps for better robustness. Based on LDA, L1-norm LDA [8] substitutes L1-norm for L2-norm and outperforms RILDA. However, the gradient ascending iterative algorithm may not guarantee to obtain an optimal solution. L1-LDA [9] follows the theoretical framework of Bayes optimality, but the method actually cannot guarantee the Bayes optimality for it introduces the alternating optimization strategy to obtain a solution. Moreover, it is not more robust to outliers or noise for the same reason. In order to avoid the problems, non-greedy L1-LDA [10] uses a non-greedy iterative algorithm to optimize the trace ratio form of L1-LDA. However, the method is sensitive to initialization for the introduction of the learning rate, and it is true to L1-norm distance-based effective LDA [11].

Recently, a novel recursive discriminative subspace learning method with an L1-norm distance constraint (RDSL) has been proposed to obtain a robust and discriminative subspace [12]. The method formulates LDA with the maximum margin criterion and avoids the introduction of the learning rate. It introduces slack variables in order to obtain better robustness. However, its performance may decline when too many outliers are available in training samples. Moreover, the method ignores the global structure of the data and thus there is still room for performance improvement.

In this paper, we present cutting L1-norm distance discriminant analysis with sample reconstruction (C-L1-DDA) to solve the two problems existing in RDSL. Based on RDSL, we apply cutting L1-norm to measure within-class and between-class distances, thus strongly suppressing outliers. Furthermore, we construct a model of sample reconstruction as a regularization term, which uses cutting squared L2-norm to measure reconstruction errors. The model can also suppress outliers and approximately preserve the global structure of data. Finally, we use an alternating iterative technique to optimize the objective function. Experimental results on two publicly available real datasets demonstrate that our method obtains better classification performance than RDSL.

2. Recursive Discriminative Subspace Learning

As previously mentioned, RDSL reformulates LDA according to the maximum margin criterion and avoids the sensitiveness to initialization. Here we give a brief review of RDSL for our method is based on it.

Let there be a data set $X \in R^{m \times n}$, which includes n samples from c classes in an m dimensional space. Let \bar{x}_i be the mean of class i , $i = 1, 2, \dots, c$, and \bar{x} the mean of all the classes. Let x_{ij} be the j th sample from class i and n_i the number of samples from the same class. Then RDSL is formulated as

$$\begin{aligned} \min_{w_q, \xi_i} & \sum_{i=1}^c \sum_{j=1}^{n_i} \|w_q^T (x_{ij} - \bar{x}_i)\|_1 + \lambda \sum_{i=1}^c \xi_i \\ \text{s.t.} & \|w_q^T (\bar{x}_i - \bar{x})\|_1 + \xi_i \geq 1 \\ & \xi_i \geq 0, \quad i = 1, 2, \dots, c \end{aligned} \quad (1)$$

where w_q is the q th projection vector, $1 \leq q \leq d$, and d is the dimensionality of the subspace. Accordingly, the projection matrix $W = [w_1, w_2, \dots, w_d]$. Moreover, ξ_i is a slack variable, which is used to improve the generalization or the robustness of RDSL, and λ is the regularization parameter.

RDSL simultaneously considers within-class compactness and between-class separability and uses L1-norm and slack variables to improve the robustness to outliers or noise. However, it is difficult to directly solve Eq. (1) for the use of L1-norm. Fortunately, the problem can be transformed into a linear or quadratic programming form and then an iterative technique is used to obtain the solution [12].

3. Cutting L1-Norm Distance Discriminant Analysis

RDSL can obtain excellent performance when there are a few outliers in training samples. However, its performance may decline when too many outliers are available since, in this situation, L1-norm and slack variables may not eliminate the adverse effects of the outliers to a large extent.

Inspired by the success of the cutting L1-norm loss function in support vector machines [13], we first apply cutting L1-norm to RDSL in order to eliminate the effects of outliers. Our method is formulated as follows:

$$\begin{aligned} \min_{w_q, \xi_i} \sum_{i=1}^c \sum_{j=1}^{n_i} \min \left(s, \left\| w_q^T (x_{ij} - \bar{x}_i) \right\|_1 \right) + \lambda \sum_{i=1}^c \xi_i \\ \text{s.t. } \min \left(h, n_i \left\| w_q^T (\bar{x}_i - \bar{x}) \right\|_1 \right) + \xi_i \geq 1 \\ \xi_i \geq 0, \quad i = 1, 2, \dots, c \end{aligned} \quad (2)$$

where s and h are two given values, $s > 0$, $h > 0$.

By applying cutting L1-norm in the objective function and the constraint, our method can strongly suppress outliers. On the one hand, the distances between outliers and their class means are usually larger than those between normal samples and the means. The larger distances may dominate the objective function when too many outliers are available, which may lead to the drift of projection directions. It may be the reason that the performance of RDSL declines when training samples include too many outliers. We use cutting L1-norm to constrain the larger distances and thus may greatly improve the robustness to outliers. On the other hand, the distances between the class means and the mean of all the classes may also be severely influenced by outliers. In other words, the distances may be significantly increased. In this case, the drift of projection directions may occur for the constraint in Eq. (1). Therefore, we apply cutting L1-norm to constrain the terms with larger distances, which may alleviate the problem.

4. Cutting L1-Norm Distance Discriminant Analysis with Sample Reconstruction

RDSL is designed to robustly learn a series of projection vectors from noisy data. It realizes within-class compactness by the objective function and between-class separability by the constraint. However, the method ignores the global structure of the

data. As a result, the projection vectors may not extract as much discriminative information as possible.

In order to solve the problem, we utilize sample reconstruction for preserving the global structure. More specifically, we minimize the reconstruction errors of training samples, and cutting L1-norm is applied to accomplish the goal. The modified form is formulated in the following:

$$\begin{aligned} \min_{w_q, u_q, \xi_i} \sum_{i=1}^c \sum_{j=1}^{n_i} \min \left(s, \|w_q^T (x_{ij} - \bar{x}_i)\|_1 \right) &+ \alpha \sum_{i=1}^c \sum_{j=1}^{n_i} \min \left(k, \|x_{ij} - u_q w_q^T x_{ij}\|_1 \right) + \lambda \sum_{i=1}^c \xi_i \\ \text{s.t. } \min \left(h, n_i \|w_q^T (\bar{x}_i - \bar{x})\|_1 \right) &+ \xi_i \geq 1 \\ \xi_i \geq 0, \quad i = 1, 2, \dots, c \end{aligned} \quad (3)$$

where u_q is the q th reconstruction vector, $1 \leq q \leq d$, α is a regularization parameter, $\alpha \geq 0$, and k is a given value, $k > 0$. The reconstruction regularization term takes the form of cutting L1-norm and thus can improve the robustness to a certain extent. However, the form may lead to a high computational cost in the optimization when the number of training samples is large.

Therefore, we replace L1-norm with squared L2-norm in order to reduce the computational cost and the modified version is described as follows:

$$\begin{aligned} \min_{w_q, u_q, \xi_i} \sum_{i=1}^c \sum_{j=1}^{n_i} \min \left(s, \|w_q^T (x_{ij} - \bar{x}_i)\|_1 \right) &+ \alpha \sum_{i=1}^c \sum_{j=1}^{n_i} \min \left(k, \|x_{ij} - u_q w_q^T x_{ij}\|_2^2 \right) + \lambda \sum_{i=1}^c \xi_i \\ \text{s.t. } \min \left(h, n_i \|w_q^T (\bar{x}_i - \bar{x})\|_1 \right) &+ \xi_i \geq 1 \\ \xi_i \geq 0, \quad i = 1, 2, \dots, c \end{aligned} \quad (4)$$

Although L2-norm is sensitive to noise, it is possible to evade the problem in the form of cutting squared L2-norm. That is to say, the samples with larger reconstruction residuals are regarded as outliers and constrained.

5. Optimization

Eq. (4) involves the variables: w_q , u_q and ξ_i and we adopt the alternating iterative method [14] to obtain the solutions.

5.1. Reformulation of Minimization Terms

Eq. (4) involves more complex minimization terms and it is difficult to directly optimize the problem. In order to solve the difficulty, we reformulate the terms in the following.

We first reformulate the constrain in Eq. (4). Let $p_i = n_i (\bar{x}_i - \bar{x})$, then

$$\begin{aligned} \min \left(h, n_i \|w_q^T (\bar{x}_i - \bar{x})\|_1 \right) &= \min \left(h, \|w_q^T p_i\|_1 \right) \\ &= \min \left(h, \max \left(0, w_q^T p_i \right) \right) + \min \left(h, \max \left(0, -w_q^T p_i \right) \right) \end{aligned}$$

and

$$\begin{aligned}\min(h, \max(0, w_q^T p_i)) &= \max(0, w_q^T p_i) - \max(0, w_q^T p_i - h) \\ \min(h, \max(0, -w_q^T p_i)) &= \max(0, -w_q^T p_i) - \max(0, -w_q^T p_i - h)\end{aligned}$$

Therefore, we have

$$\min(h, \|w_q^T p_i\|_1) = \|w_q^T p_i\|_1 - \max(0, w_q^T p_i - h) - \max(0, -w_q^T p_i - h)$$

Namely, the constraint term $\min(h, n_i \|w_q^T (\bar{x}_i - \bar{x})\|_1) + \xi_i \geq 1$ can be transformed into as follows:

$$\|w_q^T p_i\|_1 - \max(0, w_q^T p_i - h) - \max(0, -w_q^T p_i - h) + \xi_i \geq 1$$

Let $\pi_{i1} = \max(0, w_q^T p_i - h)$ and $\pi_{i2} = \max(0, -w_q^T p_i - h)$, then the constraint term can be written as follows:

$$\begin{aligned}\|w_q^T p_i\|_1 - \pi_{i1} - \pi_{i2} + \xi_i &\geq 1 \\ h + \pi_{i1} &\geq w_q^T p_i \\ h + \pi_{i2} &\geq -w_q^T p_i \\ \pi_{i1} &\geq 0, \pi_{i2} \geq 0\end{aligned}$$

Clearly, $\|w_q^T p_i\|_1 = \text{sign}(w_q^T p_i) w_q^T p_i = \text{sign}(w_q^T p_i) p_i^T w_q$, where $\text{sign}(\cdot)$ is the sign function. Let $a_i = \text{sign}(w_q^T p_i) p_i$, $A_q = [a_1, a_2, \dots, a_c]^T$ and $B = [p_1, p_2, \dots, p_c]^T$. Let e_c denote a column vector and its elements consist of c ones. In addition, let $\gamma_1 = [\pi_{11}, \pi_{21}, \dots, \pi_{c1}]^T$, $\gamma_2 = [\pi_{12}, \pi_{22}, \dots, \pi_{c2}]^T$ and $\gamma_3 = [\xi_1, \xi_2, \dots, \xi_c]^T$. Accordingly, the constraint in Eq. (4) can be reformulated as the following form:

$$\begin{aligned}A_q w_q - \gamma_1 - \gamma_2 + \gamma_3 &\geq e_c \\ h e_c + \gamma_1 &\geq B w_q \\ h e_c + \gamma_2 &\geq -B w_q \\ \gamma_1 &\geq 0, \gamma_2 \geq 0, \gamma_3 \geq 0\end{aligned}$$

We then reformulate the first term of the objective function in Eq. (4). Let $f_{ij} = x_{ij} - \bar{x}_i$, then

$$\|w_q^T (x_{ij} - \bar{x}_i)\|_1 = \|w_q^T f_{ij}\|_1 = \frac{w_q^T f_{ij} f_{ij}^T w_q}{\|w_q^T f_{ij}\|_1}$$

Accordingly,

$$\min(s, \|w_q^T (x_{ij} - \bar{x}_i)\|_1) = \min(s, \|w_q^T f_{ij}\|_1) = \rho_{ij} w_q^T f_{ij} f_{ij}^T w_q + \theta_{ij} s$$

where $\rho_{ij} = 1/\|w_q^T f_{ij}\|_1$ and $\theta_{ij} = 0$ if $\|w_q^T f_{ij}\|_1 < s$, otherwise $\rho_{ij} = 0$, $\theta_{ij} = 1$. Therefore, we have

$$\sum_{i=1}^c \sum_{j=1}^{n_i} \min(s, \|w_q^T (x_{ij} - \bar{x}_i)\|_1) = w_q^T F_q w_q + z_q$$

where $F_q = \sum_{i=1}^c \sum_{j=1}^{n_i} \rho_{ij} f_{ij} f_{ij}^T$, and $z_q = \sum_{i=1}^c \sum_{j=1}^{n_i} \theta_{ij} s$.

In addition, we reformulate the second term of the objective function. Clearly,

$$\|x_{ij} - u_q w_q^T x_{ij}\|_2^2 = x_{ij}^T x_{ij} - 2x_{ij}^T u_q x_{ij}^T w_q + w_q^T x_{ij} u_q^T u_q x_{ij}^T w_q$$

and

$$\min\left(k, \|x_{ij} - u_q w_q^T x_{ij}\|_2^2\right) = (1 - \mu_{ij})k + \mu_{ij} \|x_{ij} - u_q w_q^T x_{ij}\|_2^2$$

where $\mu_{ij} = 1$ if $\|x_{ij} - u_q w_q^T x_{ij}\|_2^2 < k$, otherwise $\mu_{ij} = 0$. Accordingly,

$$\sum_{i=1}^c \sum_{j=1}^{n_i} \min\left(k, \|x_{ij} - u_q w_q^T x_{ij}\|_2^2\right) = b_{q1} + a_{q1}^T w_q + w_q^T G_q w_q$$

where

$$b_{q1} = \sum_{i=1}^c \sum_{j=1}^{n_i} \left((1 - \mu_{ij})k + \mu_{ij} x_{ij}^T x_{ij} \right)$$

$$a_{q1} = -2 \sum_{i=1}^c \sum_{j=1}^{n_i} \mu_{ij} (x_{ij}^T u_q) x_{ij}$$

$$G_q = \sum_{i=1}^c \sum_{j=1}^{n_i} \mu_{ij} x_{ij} u_q^T u_q x_{ij}^T$$

5.2. Algorithm

After reformulating the constraint and the two terms of the objective function, we adopt an alternating iterative algorithm [14] to obtain the projection vector. Let w_q^t and u_q^t be the values of the projection vector and the reconstruction one at the t th iteration, and $w_q^{(t+1)}$ and $u_q^{(t+1)}$ the values at the $(t+1)$ th iteration.

The iterative method alternately performs the following two main steps:

Step 1. Fix u_q and compute the value of w_q in Eq. (4). More specifically, we can obtain $w_q^{(t+1)}$ by solving the following problem:

$$\begin{aligned} \min_{w_q^{(t+1)}, \gamma_1, \gamma_2, \gamma_3} & \left(w_q^{(t+1)} \right)^T \left(F_q + \alpha G_q \right) w_q^{(t+1)} + \alpha a_{q1}^T w_q^{(t+1)} + \lambda e_c^T \gamma_3 \\ \text{s.t. } & A_q w_q^{(t+1)} - \gamma_1 - \gamma_2 + \gamma_3 \geq e_c \\ & h e_c + \gamma_1 \geq B w_q^{(t+1)} \\ & h e_c + \gamma_2 \geq -B w_q^{(t+1)} \\ & \gamma_1 \geq 0, \gamma_2 \geq 0, \gamma_3 \geq 0 \end{aligned} \quad (5)$$

Eq. (5) is a quadratic programming problem and thus $w_q^{(t+1)}$ can be easily obtained.

Step 2. Fix w_q and ξ_i and compute the value of u_q in Eq. (4). In this situation, Eq. (4) reduces to the following form:

$$\min_{u_q} \sum_{i=1}^c \sum_{j=1}^{n_i} \min \left(k, \|x_{ij} - u_q w_q^T x_{ij}\|_2^2 \right) \quad (6)$$

Accordingly, $u_q^{(t+1)}$ can be easily obtained by solving the following quadratic programming problem:

$$\min_{u_q^{(t+1)}} g_q \left(u_q^{(t+1)} \right)^T u_q^{(t+1)} + a_{q2}^T u_q^{(t+1)} + b_{q2} \quad (7)$$

where

$$\begin{aligned} b_{q2} &= \sum_{i=1}^c \sum_{j=1}^{n_i} \left((1 - \mu_{ij}) k + \mu_{ij} x_{ij}^T x_{ij} \right) \\ a_{q2} &= -2 \sum_{i=1}^c \sum_{j=1}^{n_i} \mu_{ij} \left(x_{ij}^T w_q^{(t+1)} \right) x_{ij} \\ g_q &= \sum_{i=1}^c \sum_{j=1}^{n_i} \mu_{ij} \left(x_{ij}^T w_q^{(t+1)} \right)^2 \end{aligned}$$

After w_q is obtained by iteration, the training samples are updated by the factor $(I_m - WW^T)$, where I_m is the $m \times m$ dimensional identity matrix [12]. The detailed process of the algorithm is shown in Table 1.

Table 1. Algorithm 1. The alternating iterative algorithm

<p>Input: X --- labeled data set, $X \in R^{m \times n}$; α, λ --- regularization parameters; d --- number of projection vectors;</p> <p>Output: W --- projection matrix, $W \in R^{m \times d}$.</p>
<p>Procedure: Initiate the projection matrix $W = \phi$, where ϕ is the empty set; For $q = 1$ to d do Initiate w_q ; Compute u_q according to Eq. (7); While not converge do Update w_q according to Eq. (5); Update u_q according to Eq. (7); End while Let $W = [W, w_q]$, and normalize and orthogonalize it; Let $X = (I_m - WW^T)X$; End for</p>

6. Experiments

We conduct experiments on USPS [15] and COIL-20 [16] datasets to evaluate LDA, coined complete large margin linear discriminant analysis (CLMLDA) [17], RILDA, RDSL and C-L1-DDA. The grayscale of images from the datasets is normalized to $[0,1]$. After training samples are artificially contaminated and represented with the corresponding vectors, PCA is applied to reduce the dimensionality to $n-c$. In the experiments, the regularization parameters of CLMLDA, RDSL and C-L1-DDA are chosen from the set $[10^{-3}, 10^{-2}, \dots, 10^3]$. Our method has three cutting parameters, i.e., h , s and k , and it is difficult to specify them. As an alternative, the three parameters are empirically set in this way that five percent of training samples with the largest within-class distances are eliminated, which are regarded as outliers. The initial value of w_q is set to the solution of LDA. Basically, the convergence condition is set according to [12] and the difference lies in that the maximum number of iterations is set to fifty. In the experiments, the nearest neighbor classifier is employed in measuring classification performance, and it is directly used for preprocessed data by PCA as the baseline method. All the experiments are implemented ten times and the average lowest classification error rates are reported.

6.1. Experiments on the USPS Database

In this section, experiments are conducted with the USPS database. Ten handwritten digits from 0 to 9 of this database have 11,000 grayscale images whose size is 16×16 pixels. A subset is first constructed for the experiments and each digit includes 100 images which are randomly selected from the original database. For each digit, K samples are selected for training and the remaining ones are used for testing. In the experiments, K is specified for $\{10, 12, 14, 16\}$. Forty percent of training samples are contaminated by inserting occlusion with black or white rectangular noise at a random location and the size of the rectangle is at least 4×4 and up to 10×10 pixels.

Average lowest classification error rates with different dimensions are shown in Figure 1. From the experimental results, we can conclude that the proposed method is significantly effective for classification.

6.2. Experiments on the Coil-20 Database

The section deals with the experiments carried out on the Coil-20 database to test the performance of C-L1-DDA. The database consists of 1,440 grayscale images of twenty objects and each object has 72 images at an interval of five degrees of rotation. Each image is downsampled and resized to 32×32 pixels. Let $K = \{8, 10, 12, 14\}$, and for each object, K samples are used for training and the rest for testing. Similar to the experiments on the USPS database, forty percent of training samples are contaminated and the difference is that the size of the noise rectangle is at least 20×20 and up to 30×30 pixels.

Average lowest classification error rates with different dimensions are given in Figure 2. From the figure we conclude that the performance of C-L1-DDA is better than that of RDSL

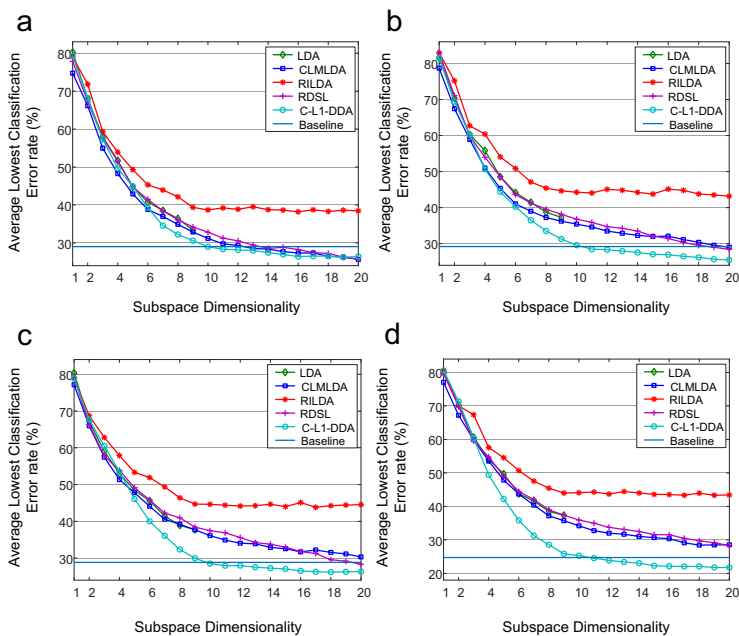


Figure 1. Average lowest classification error rates on the USPS database. (a) 10 samples, (b) 12 samples, (c) 14 samples, and (d) 16 samples.

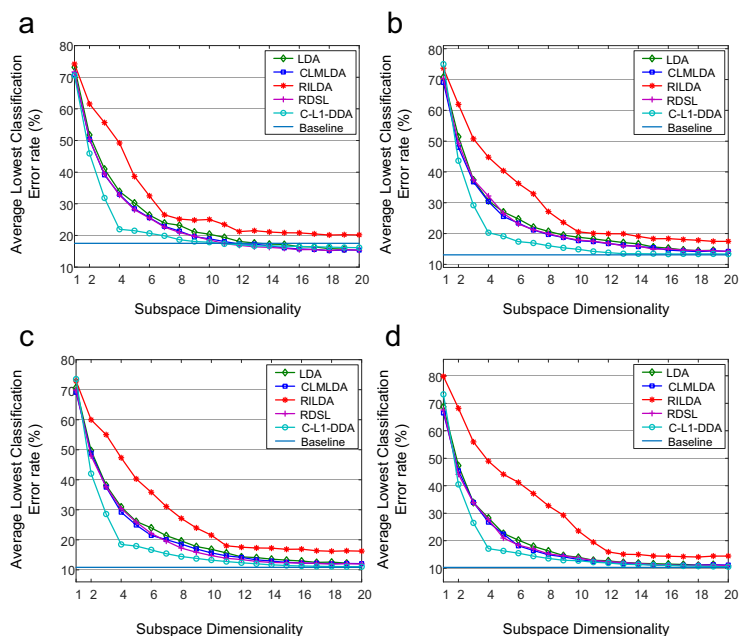


Figure 2. Average lowest classification error rates on the Coil-20 database. (a) 8 samples, (b) 10 samples, (c) 12 samples, and (d) 14 samples.

7. Conclusions

In this paper, we present a novel discriminant analysis method, cutting L1-norm distance discriminant analysis with sample reconstruction (C-L1-DDA), which utilizes cutting L1-norm for measuring within-class distances and between-class distances. It may strongly suppress outliers and thus the robustness may be improved. Moreover, our method uses cutting squared L2-norm to measure reconstruction errors. As a result, it may suppress the outliers and approximately preserve the global structure of data. Finally, we use an alternating iterative algorithm to obtain projection vectors. Experimental results on two real databases demonstrate the effectiveness of C-L1-DDA. In the future, we will test the performance of our method on more databases.

Acknowledgement

The work is supported by Natural Science Foundation of Hunan Province (Grant No.2018JJ2133).

References

- [1] Belhumeur PN, Hespanha J, Kriegman J. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans Pattern Anal Mach Intell.* 1997 Jul; 19(7):711-20.
- [2] He X, Niyogi P. Locality preserving projections. In: Thrun S, Saul LK, Schölkopf B, editors. *Advances in neural information processing systems*; 2003 Dec 8-13; Cambridge, MA: MIT Press; c2004. p. 153-60.
- [3] Turk M, Pentland A. Eigenfaces for recognition. *Cognitive Neurosci.* 1991 Mar; 3(7):71-86.
- [4] Randles RH, Broffitt JD, Ramberg JS, Hogg RV. Generalized linear and quadratic discriminant functions using robust estimates. *J Amer Stat Assoc.* 1978 Sept;73(363):564-8.
- [5] Lanckriet GRG, El Ghaoui L, Bhattacharyya C, Jordan MI. A robust minimax approach to classification. *J Mach Learn Res.* 2002 Dec;3: 555-82.
- [6] Kim SJ, Magnani A, Boyd S. Robust Fisher discriminant analysis. In: Weiss Y, Schölkopf B, Platt J, editors. *Advances in Neural Information Processing Systems*; 2005 Dec 5-10; Cambridge, MA: MIT Press; c2006. p. 659-66.
- [7] Lai Z, Xu Y, Yang J, Shen L, Zhang D. Rotational invariant dimensionality reduction algorithms. *IEEE Trans Cybern.* 2017 Nov;47(11):3733-46.
- [8] Zhong F, Zhang J. Linear discriminant analysis based on L1-norm maximization. *IEEE Trans Image Process.* 2013 Aug;22(8):3018-27.
- [9] Zheng W, Lin Z, Wang H. L1-norm kernel discriminant analysis via Bayes error bound optimization for robust feature extraction. *IEEE Trans Neural Netw Learn Syst.* 2014 Apr; 25(4):793-805.
- [10] Liu Y, Gao Q, Miao S, Gao X, Nie F, Li Y. A non-greedy algorithm for L1-norm LDA. *IEEE Trans Image Process.* 2017 Feb;26(2):684-95.
- [11] Ye Q, Yang J, Liu F, Zhao C, Ye N, Yin T. L1-norm distance linear discriminant analysis based on an effective iterative algorithm. *IEEE Trans Circuits Syst Video Technol.* 2018 Jan;28(1):114-29.
- [12] Zhang D, Sun Y, Ye Q, Tang J. Recursive discriminative subspace learning with L1-norm distance constraint. *IEEE Trans Cybern.* 2020 May;50(5):2138-51.
- [13] Cevikalp H. Best fitting hyperplanes for classification. *IEEE Trans Pattern Anal Mach Intell.* 2017 Jun;39(6):1076-88.
- [14] Li X, Zhang H, Zhang R, Liu Y, Nie F. Generalized uncorrelated regression with adaptive graph for unsupervised feature selection. *IEEE Trans Neural Netw Learn Syst.* 2019 May;30(5):1587-95.
- [15] Simard P, LeCun Y, Denker JS. Efficient pattern recognition using a new transformation distance. In: Hanson S, Cowan J, Giles C, editors. *Advances in Neural Information Processing Systems*; 1992 Nov 30- Dec 3; San Francisco: Morgan-Kaufmann Press; c1993. p. 50-8.
- [16] Nene SA, Nayar SK, Murase H. 1996 Columbia object image library (COIL-20). Columbia Univ., New York, Tech. Rep. CUCS-005-96.
- [17] Chen X, Yang J, Zhang D, Liang J. Complete large margin linear discriminant analysis using mathematical programming approach. *Pattern Recogn.* 2013 Jun;46(6):1579-94.