EMBEDDED CLUSTERING VIA ROBUST ORTHOGONAL LEAST SQUARE DISCRIMINANT ANALYSIS

Rui Zhang, Feiping Nie, and Xuelong Li*

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL)
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China.

ABSTRACT

In this paper, a novel embedded clustering (EC) method is derived from the perspective of extending the supervised orthogonal least square discriminant analysis (OLSDA) method to the unsupervised case, which proves to be closely related to k-means. To achieve more statistical and structural properties, the robust learning of unsupervised OLSDA is investigated to further derive the unsupervised robust OLSDA (ROLSDA) problem. For the convenience of solving the proposed ROLSDA problem, re-weighted counterpart of ROLSDA is utilized with self-adaptive weight, such that the smaller weight would be assigned to the term with larger outliers automatically. Consequently, aforementioned EC method is proposed with not only the robust outliers but also the optimal weighted cluster centroids. Comparative experiments are presented to show the effectiveness of the EC method under the proposed ROLSDA problem.

***Index Terms*—** Embedded clustering, least square discriminant analysis, robust learning, re-weighted problem.

1. INTRODUCTION

Orthogonal least square discriminant analysis (OLSDA) [1] serves as a pretty significant supervised technique for dimensionality reduction [2, 3, 4, 5, 6, 7, 8]. OLSDA method is derived from minimizing the discriminative information in data. In other words, OLSDA method virtually minimizes the within-class scatter matrix from the viewpoint of the least square regression. To achieve great discriminative power, the projected data in the same class are expected to be regressed to a single vector with the purpose of seeking an orthogonal projection, such that the error of square regression is minimized.

In this paper, former supervised OLSDA method is further extended to the unsupervised case. By virtue of revisiting both OLSDA and orthogonal centroid method (OCM) [9], a brand new least square form of unsupervised OLSDA can be derived. Besides, the proposed unsupervised OLSDA problem is proved to be closely related to k-means [10]. Moreover, robust OLSDA (ROLSDA) problem is further proposed and investigated via robust learning[11, 12, 13, 14]. To unravel the proposed ROLSDA problem, re-weighted counterpart of ROLSDA is utilized with self-adaptive weight, such that the smaller weight would be assigned to the term with larger outliers automatically. Accordingly, a novel embedded clustering (EC) method is derived from the proposed ROLSDA problem with both the robust outliers and the optimal weighted cluster centroids.

The rest of the paper is organized as follows. Section 2 revisits the OLSDA and the OCM methods. Section 3 extends supervised OLSDA to the unsupervised OLSDA with a novel least square form. Consequently, EC method is derived from the unsupervised robust OLSDA problem. In Section 4, comparative experiments are presented to show the effectiveness of the proposed EC method. Section 5 concludes the paper.

2. ORTHOGONAL LEAST SQUARE DISCRIMINANT ANALYSIS REVISITED

Given the training dataset $\mathscr{X} = \{(\mathbf{x_i}, \mathbf{c_i})|\mathbf{x_i} \in \mathbb{R}^{\mathbf{d}\times\mathbf{1}};\ \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{n}\}$ and related data matrix $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}] \in \mathbb{R}^{\mathbf{d}\times\mathbf{n}}$ which are distributed in $\mathbf{c}$ different classes with dimension $\mathbf{d}$ and data number $\mathbf{n}$, each data point $\mathbf{x_i}$ is associated with a class label $\mathbf{c_i} \in \{\mathbf{1}, \mathbf{2}, \ldots \mathbf{c}\}$. Denote $\mathscr{X}_\mathbf{i}$ is the dataset of $\mathbf{i}$-th class and $\mathbf{n_i}$ is the number of data points in $\mathbf{i}$-th class, then the within-class scatter matrix $\mathbf{S_w}$, the between-class scatter matrix $\mathbf{S_b}$ and the total-class scatter matrix $\mathbf{S_t}$ are defined as follows:

$$\begin{cases} \mathbf{S_w} = \sum\limits_{\mathbf{i=1}}^{\mathbf{c}} \sum\limits_{\mathbf{x}\in\mathscr{X}_\mathbf{i}} (\mathbf{x} - \bar{\mathbf{x}}_\mathbf{i})(\mathbf{x} - \bar{\mathbf{x}}_\mathbf{i})^\mathbf{T} \\ \mathbf{S_b} = \sum\limits_{\mathbf{i=1}}^{\mathbf{c}} \mathbf{n_i}(\bar{\mathbf{x}}_\mathbf{i} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\mathbf{i} - \bar{\mathbf{x}})^\mathbf{T} \\ \mathbf{S_t} = \sum\limits_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^\mathbf{T} \end{cases} \tag{1}$$

where $\bar{\mathbf{x}}_\mathbf{i} = \frac{\mathbf{1}}{\mathbf{n_i}} \sum_{\mathbf{x_j}\in\mathscr{X}_\mathbf{i}} \mathbf{x_j}$ is the class-specific mean of the $\mathbf{i}$-th class and $\bar{\mathbf{x}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i}$ is the global mean. The least

*Corresponding author. Email: feipingnie@gmail.com.

squared loss function could be illustrated as:

$$\varepsilon = \|\mathbf{T_1} - \mathbf{T_2}\|_{\mathbf{F}}^{\mathbf{2}} \tag{2}$$

where $\|\cdot\|_{\mathbf{F}}^{\mathbf{2}}$ represents Frobenius norm. Define $\mathbf{A}^{(\mathbf{t})} = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{1}\mathbf{1}^{\mathbf{T}}$ and

$$\mathbf{A}_{\mathbf{ij}}^{(\mathbf{w})} = \begin{cases} \frac{\mathbf{1}}{\mathbf{n_{c_i}}} & \mathbf{c_i} = \mathbf{c_j} \\ \mathbf{0} & otherwise \end{cases}$$

where $\mathbf{1} = (\mathbf{1}, \mathbf{1}, \ldots, \mathbf{1})^{\mathbf{T}} \in \mathbb{R}^{\mathbf{n}\times\mathbf{1}}$. By substituting $\mathbf{T_1} = \mathbf{W^T X}$ and $\mathbf{T_2} = \mathbf{W^T X A^{(w)}}$ in Eq. (2) with orthogonal constraint $\mathbf{W^T W} = \mathbf{I}, \mathbf{W} \in \mathbb{R}^{\mathbf{d}\times\mathbf{k}}$, we have

$$\begin{aligned} \|\mathbf{W^T X} - \mathbf{W^T X A^{(w)}}\|_{\mathbf{F}}^{\mathbf{2}} &= \mathbf{Tr}(\mathbf{W^T X}(\mathbf{I} - \mathbf{A^{(w)}})^{\mathbf{2}}\mathbf{X^T W}) \\ &= \mathbf{Tr}(\mathbf{W^T S_w W}) \end{aligned} \tag{3}$$

which is the objective function of OLSDA discussed in [1]. Similarly, we could further set $\mathbf{T_1} = \mathbf{W^T X}$ and $\mathbf{T_2} = \mathbf{W^T X A^{(t)}}$ in Eq. (2) as

$$\begin{aligned} \|\mathbf{W^T X} - \mathbf{W^T X A^{(t)}}\|_{\mathbf{F}}^{\mathbf{2}} &= \mathbf{Tr}(\mathbf{W^T X}(\mathbf{I} - \mathbf{A^{(t)}})^{\mathbf{2}}\mathbf{X^T W}) \\ &= \mathbf{Tr}(\mathbf{W^T S_t W}) \end{aligned} \tag{4}$$

with orthogonal constraint $\mathbf{W^T W} = \mathbf{I}$.

In terms of Eqs. (3) and (4), the objective function of OCM [9] could be rewritten as

$$\begin{aligned} \mathbf{Tr}(\mathbf{W^T S_b W}) &= \mathbf{Tr}(\mathbf{W^T}(\mathbf{S_t} - \mathbf{S_w})\mathbf{W}) \\ &= \mathbf{Tr}(\mathbf{W^T X}(\mathbf{I} - \mathbf{A^{(t)}} - \mathbf{I} + \mathbf{A^{(w)}})\mathbf{X^T W}) \\ &= \mathbf{Tr}(\mathbf{W^T X}(\mathbf{A^{(w)}} - \mathbf{A^{(t)}})\mathbf{X^T W}) \end{aligned} \tag{5}$$

due to the idempotent matrices $\mathbf{A^{(t)}}$ and $\mathbf{A^{(w)}}$, i.e., $(\mathbf{A^{(t)}})^{\mathbf{2}} = \mathbf{A^{(t)}}$ and $(\mathbf{A^{(w)}})^{\mathbf{2}} = \mathbf{A^{(w)}}$.

3. ROBUST ORTHOGONAL LEAST SQUARE DISCRIMINANT ANALYSIS

Based on Eqs. (4) and (5), the total-class scatter $\mathbf{S_t}$ and the between-class scatter $\mathbf{S_b}$ in (1) could be further reformulated into

$$\begin{cases} \mathbf{S_t} = \mathbf{X H X^T} \\ \mathbf{S_b} = \mathbf{X H Y}(\mathbf{Y^T Y})^{-\mathbf{1}}\mathbf{Y^T H X^T} \end{cases} \tag{6}$$

where $\mathbf{Y} \in \mathbb{R}^{\mathbf{n}\times\mathbf{c}}$ is the binary label matrix and $\mathbf{H} = \mathbf{I} - \mathbf{A^{(t)}}$ is the centering matrix. Based on Eq. (6), OLSDA in (3) could be rewritten as

$$\begin{aligned} &\min_{\mathbf{W^T W}=\mathbf{I}} \mathbf{Tr}(\mathbf{W^T S_w W}) = \min_{\mathbf{W^T W}=\mathbf{I}} \mathbf{Tr}(\mathbf{W^T}(\mathbf{S_t} - \mathbf{S_b})\mathbf{W}) \\ &= \min_{\mathbf{W^T W}=\mathbf{I}} \mathbf{Tr}(\mathbf{W^T X H}(\mathbf{I} - \mathbf{Y}(\mathbf{Y^T Y})^{-\mathbf{1}}\mathbf{Y^T})\mathbf{H X^T W}) \\ &= \min_{\mathbf{W^T W}=\mathbf{I}} \|\mathbf{W^T X H}(\mathbf{I} - \mathbf{Y}(\mathbf{Y^T Y})^{-\mathbf{1}}\mathbf{Y^T})\|_{\mathbf{F}}^{\mathbf{2}} \end{aligned} \tag{7}$$

which serves as a special case of problem (2) by letting $\mathbf{T_1} = \mathbf{W^T X H}$ and $\mathbf{T_2} = \mathbf{W^T X H Y}(\mathbf{Y^T Y})^{-\mathbf{1}}\mathbf{Y^T}$, respectively.

• Relationship between OLSDA in (7) and k-means.

The k-means problem [10] can be recapitulated as

$$\min_{\mathbf{F},\mathbf{G}\in\mathbf{ind}} \|\mathbf{T} - \mathbf{F G^T}\|_{\mathbf{F}}^{\mathbf{2}} \tag{8}$$

where each column of matrix $\mathbf{F} \in \mathbb{R}^{\mathbf{k}\times\mathbf{c}}$ represents the cluster centroid and each row of the indicative matrix $\mathbf{G} \in \mathbb{R}^{\mathbf{n}\times\mathbf{c}}$ demonstrates the binary label. If the associated label of data $\mathbf{T} \in \mathbb{R}^{\mathbf{k}\times\mathbf{n}}$ is known, i.e., indicative matrix $\mathbf{G}$ is fixed as binary label $\mathbf{Y}$ in (8), the k-means problem degenerates to

$$\min_{\mathbf{F}} \|\mathbf{T} - \mathbf{F Y^T}\|_{\mathbf{F}}^{\mathbf{2}} = \|\mathbf{T} - \mathbf{T Y}(\mathbf{Y^T Y})^{-\mathbf{1}}\mathbf{Y^T}\|_{\mathbf{F}}^{\mathbf{2}}. \tag{9}$$

By further replacing the data $\mathbf{T}$ with the centralized projected data $\mathbf{W^T X H} \in \mathbb{R}^{\mathbf{k}\times\mathbf{n}}$ in Eq. (9), we notice that the problem (9) is same as the problem (7). In other words, OLSDA in (7) is equivalent to k-means in (8) when $\mathbf{T} = \mathbf{W^T X H}$ and $\mathbf{G} = \mathbf{Y}$. Accordingly, OLSDA in (7) could be naturally extended to the unsupervised case as

$$\min_{\mathbf{W^T W}=\mathbf{I},\mathbf{F},\mathbf{G}\in\mathbf{ind}} \|\mathbf{W^T X H} - \mathbf{F G^T}\|_{\mathbf{F}}^{\mathbf{2}}. \tag{10}$$

• Embedded clustering via Robust OLSDA.

Based on the unsupervised OLSDA in (10), robust OLSDA (ROLSDA) could be proposed as

$$\min_{\mathbf{W^T W}=\mathbf{I},\mathbf{F},\mathbf{G}\in\mathbf{ind}} \|\mathbf{W^T X H} - \mathbf{F G^T}\|_{\mathbf{2,1}}. \tag{11}$$

Motivated by [11, 12], we could utilize the re-weighted counterpart of ROLSDA in (11) for the convenience as

$$\begin{aligned} &\min_{\mathbf{W^T W}=\mathbf{I},\mathbf{F},\mathbf{G}\in\mathbf{ind}} \|(\mathbf{W^T X H} - \mathbf{F G^T})\mathbf{D}^{\frac{\mathbf{1}}{\mathbf{2}}}\|_{\mathbf{F}}^{\mathbf{2}} \\ &= \min_{\mathbf{W^T W}=\mathbf{I},\mathbf{F},\mathbf{G}\in\mathbf{ind}} \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} \mathbf{D_{ii}} \|\mathbf{W^T x_i^{(H)}} - \mathbf{F g_i}\|_{\mathbf{2}}^{\mathbf{2}} \end{aligned} \tag{12}$$

where diagonal weight matrix $\mathbf{D}$ is to be updated iteratively in the algorithm with $\mathbf{D_{ii}} \leftarrow \frac{\mathbf{1}}{\mathbf{2}\|\mathbf{W^T x_i^{(H)}} - \mathbf{F g_i}\|_{\mathbf{2}}}$. Besides, $\mathbf{x_i^{(H)}}$ and $\mathbf{g_i}$ are $\mathbf{i}$-th columns of $\mathbf{XH}$ and $\mathbf{G^T}$, respectively. Actually, the proposed re-weighted ROLSDA in (12) is connected with self-adaptive weight $\mathbf{D}$ such that re-weighted ROLSDA not only has robust outliers but achieves the optimal weighted cluster centroids as well.

Since the centroid matrix $\mathbf{F}$ is free from any constraint in (12), the extreme value condition w.r.t. $\mathbf{F}$ could be inferred that

$$\begin{aligned} &\frac{\partial \|(\mathbf{W^T X H} - \mathbf{F G^T})\mathbf{D}^{\frac{\mathbf{1}}{\mathbf{2}}}\|_{\mathbf{F}}^{\mathbf{2}}}{\partial \mathbf{F}} = \mathbf{0} \\ &\Rightarrow \frac{\partial \mathbf{Tr}(\mathbf{F G^T D G F^T} - \mathbf{2 W^T X H D G F^T})}{\partial \mathbf{F}} = \mathbf{0} \\ &\Rightarrow \mathbf{F} = \mathbf{W^T X H D G}(\mathbf{G^T D G})^{-\mathbf{1}}. \end{aligned} \tag{13}$$

Table 1. Comparisons of error rate and NMI for 4 methods as k-means[10], RMKMC[16] and unsupervised OLSDA[1] over 3 different datasets are performed .

Error rate			
Dataset	AR	GT	USPS
k-means[10]	0.3975	0.4117	0.4530
RMKMC[16]	0.3771	0.3912	0.4228
OLSDA[1]	0.3826	0.4006	0.4219
EC(our)	**0.3611**	**0.3773**	**0.4200**
NMI			
Dataset	AR	GT	USPS
k-means[10]	0.6556	0.6942	0.6258
RMKMC[16]	0.6723	0.7250	0.6547
OLSDA[1]	0.6730	0.7334	0.6124
EC(our)	**0.6882**	**0.7483**	**0.6609**

Consequently, re-weighted ROLSDA in (12) degenerates to

$$\min_{\mathbf{W^T W=I}, \mathbf{G}\in\mathbf{ind}} \|(\mathbf{W^T XH} - \mathbf{FG^T})\mathbf{D}^{\frac{1}{2}}\|_{\mathbf{F}}^{\mathbf{2}} \tag{14}$$

where $\mathbf{F} = \mathbf{W^T XHDG}(\mathbf{G^T DG})^{-1}$. Via the coordinate blocking method, i.e., fixing each variable alternatively, re-weighted ROLSDA in (14) could be solved correspondingly.

Case 1 (fixing G): The Lagrangian function of the problem (14) could be written as

$$\begin{aligned}\mathscr{L}(\mathbf{W}, \mathbf{\Lambda}) =& \|(\mathbf{W^T XH} - \mathbf{FG^T})\mathbf{D}^{\frac{1}{2}}\|_{\mathbf{F}}^{\mathbf{2}} - \\ & \mathbf{Tr}(\mathbf{\Lambda}(\mathbf{W^T W} - \mathbf{I})).\end{aligned} \tag{15}$$

Thus, the KKT condition can be derived as

$$\begin{aligned}&\frac{\partial \mathscr{L}(\mathbf{W}, \mathbf{\Lambda})}{\partial \mathbf{W}} = \mathbf{0} \\ &\Rightarrow \mathbf{2XHDHX^T W} - \mathbf{2XHDGF^T} = \mathbf{2W\Lambda} \\ &\Rightarrow \mathbf{W^T}(\mathbf{S_t^{(D)}} - \mathbf{S_b^{(D)}})\mathbf{W} = \mathbf{\Lambda}\end{aligned} \tag{16}$$

where

$$\begin{cases} \mathbf{S_t^{(D)}} = \mathbf{XHDHX^T} \\ \mathbf{S_b^{(D)}} = \mathbf{XHDG}(\mathbf{G^T DG})^{-1}\mathbf{G^T DHX^T} \end{cases}. \tag{17}$$

Eqs. (14) and (16) in whole imply that $\mathbf{W}$ is the matrix of eigenvector corresponding to the first $\mathbf{k}$ smallest eigenvalues of $\mathbf{S_t^{(D)}} - \mathbf{S_b^{(D)}}$ defined in (17).

Case 2 (fixing W): To obtain the indicative matrix $\mathbf{G}$, re-weighted ROLSDA in (14) could be further derived as

$$\begin{aligned}&\min_{\mathbf{G}\in\mathbf{ind}} \|(\mathbf{W^T XH} - \mathbf{FG^T})\mathbf{D}^{\frac{1}{2}}\|_{\mathbf{F}}^{\mathbf{2}} \\ =& \min_{\mathbf{g_i}\in\mathbf{ind}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{D_{ii}} \|\mathbf{W^T x_i^{(H)}} - \mathbf{Fg_i}\|_{\mathbf{2}}^{\mathbf{2}} \\ \geq& \sum_{\mathbf{i=1}}^{\mathbf{n}} \min_{\mathbf{g_i}\in\mathbf{ind}} \mathbf{D_{ii}} \|\mathbf{W^T x_i^{(H)}} - \mathbf{Fg_i}\|_{\mathbf{2}}^{\mathbf{2}}\end{aligned} \tag{18}$$

Input: data matrix $\mathbf{X} \in \mathbb{R}^{\mathbf{d}\times\mathbf{n}}$ and cluster number $\mathbf{c}$.
Output: indicative matrix $\mathbf{G} \in \mathbb{R}^{\mathbf{n}\times\mathbf{c}}$.

Initialize weight matrix $\mathbf{D} = \mathbf{I}$ with random initial guess $\mathbf{G} \in \mathbf{ind}$ and random orthogonal subspace $\mathbf{W} \in \mathbb{R}^{\mathbf{d}\times\mathbf{k}}$, i.e., $\mathbf{W^T W} = \mathbf{I}$;
while *not converge* **do**
  Update $\mathbf{F} \leftarrow \mathbf{W^T XHDG}(\mathbf{G^T DG})^{-1}$;
  for $\mathbf{i} = \mathbf{1} : \mathbf{n}$ **do**
    Update $\mathbf{g_i} \leftarrow \underset{\mathbf{g_i}\in\mathbf{ind}, \mathbf{1_c^T g_i}=\mathbf{1}}{\arg\min} \|\mathbf{W^T x_i^{(H)}} - \mathbf{Fg_i}\|_{\mathbf{2}}^{\mathbf{2}}$;
  end
  for $\mathbf{i} = \mathbf{1} : \mathbf{n}$ **do**
    Update $\mathbf{D_{ii}} \leftarrow \frac{\mathbf{1}}{\mathbf{2}\|\mathbf{W^T x_i^{(H)}} - \mathbf{Fg_i}\|_{\mathbf{2}}}$;
  end
  Update $\mathbf{S_t^{(D)}} \leftarrow \mathbf{XHDHX^T}$;
  Update $\mathbf{S_b^{(D)}} \leftarrow \mathbf{XHDG}(\mathbf{G^T DG})^{-1}\mathbf{G^T DHX^T}$;
  Update $\mathbf{W} \leftarrow \underset{\mathbf{W^T W}=\mathbf{I}}{\arg\min} \mathbf{Tr}(\mathbf{W^T}(\mathbf{S_t^{(D)}} - \mathbf{S_b^{(D)}})\mathbf{W})$;
end
return $\mathbf{G}$;

Algorithm 1: Embedded clustering (EC) method under the proposed ROLSDA problem in (11).

which indicates that indicative vector $\mathbf{g_i} \in \mathbb{R}^{\mathbf{c}\times\mathbf{1}}$ of $\mathbf{i}$-th data point could be determined by individually solving

$$\min_{\mathbf{g_i}\in\mathbf{ind}} \|\mathbf{W^T x_i^{(H)}} - \mathbf{Fg_i}\|_{\mathbf{2}}^{\mathbf{2}} \ \mathbf{s.t.}\ \mathbf{1_c^T g_i} = \mathbf{1} \tag{19}$$

where $\mathbf{1_c} = (\mathbf{1}, \ldots, \mathbf{1})^{\mathbf{T}} \in \mathbb{R}^{\mathbf{c}\times\mathbf{1}}$. Based on Eq. (14), Case 1 and 2, embedded clustering (EC) method could be summarized in Algorithm 1 to solve the proposed ROLSDA in (11).

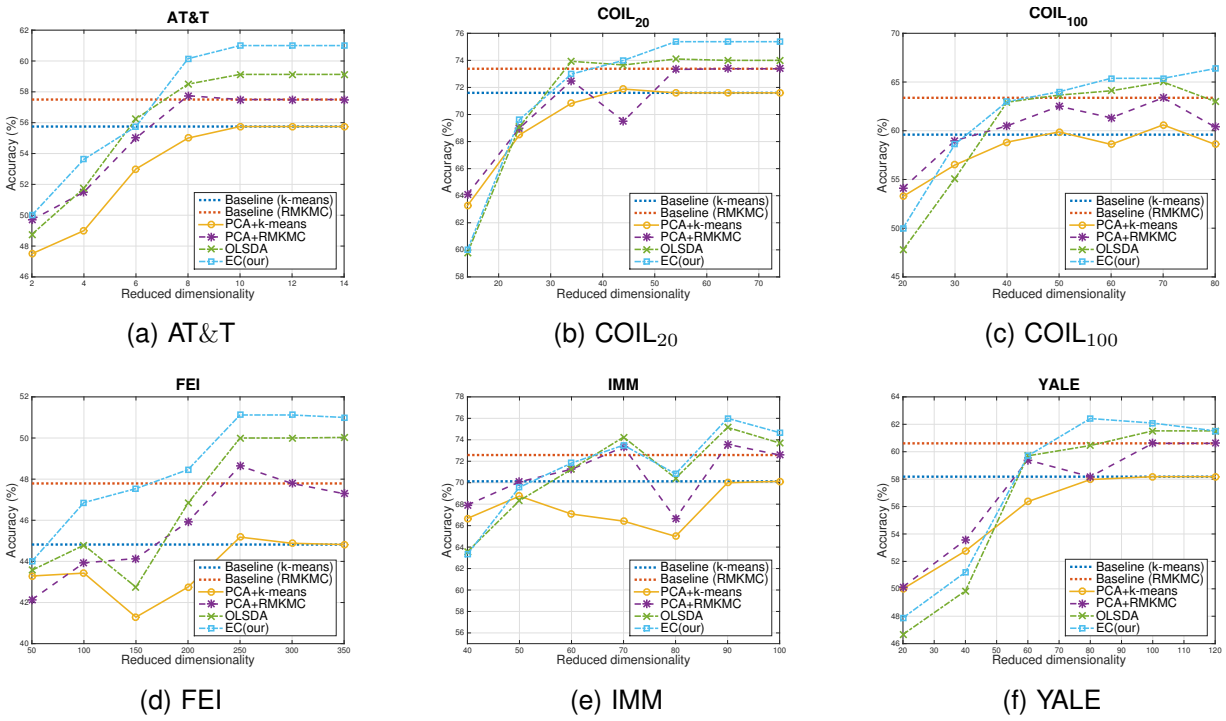
4. EXPERIMENT

We utilize 9 benchmark datasets as AT&T, COIL_{20}, COIL_{100}, FEI, IMM, YALE, AR, GT and USPS in the experimental part to further compare the clustering accuracy and NMI under different reduced dimensionality. Besides, clustering accuracy and NMI are defined as following.

• **Clustering accuracy.** The clustering accuracy in the experiment can be computed as

$$\mathbf{Accuracy} = \frac{\mathbf{1}}{\mathbf{n}} \max\Big(\sum_{\mathbf{R_k}, \mathbf{O_m}} \mathbf{M}(\mathbf{R_k}, \mathbf{O_m})\Big)$$

where $\mathbf{R_k}$ stands for the $\mathbf{k}$-th cluster in the final result and $\mathbf{O_m}$ stands for the true $\mathbf{m}$-th class. $\mathbf{M}(\mathbf{R_k}, \mathbf{O_m})$ stands for the number of entities, which originally serve as the input data



(a) AT&T (b) COIL$_{20}$ (c) COIL$_{100}$

(d) FEI (e) IMM (f) YALE

Fig. 1. The clustering accuracy comparisons are performed for the PCA[15]+k-means[10] method, the PCA[15]+RMKMC[16] method, the unsupervised OLSDA[1] method and the proposed EC method with two baselines as the k-means[10] method and the RMKMC[16] method under 6 benchmark datasets. (a) AT&T. (b) COIL$_{20}$. (c) COIL$_{100}$. (d) FEI. (e) IMM. (f) YALE.

of the $\mathbf{m}$-th class and are assigned to the $\mathbf{k}$-th cluster in the final result.

• **Normalized mutual information.** The normalized mutual Information (NMI) serves as an index to determine the quality of the clusters. The $\mathbf{NMI}$ index is defined as

$$\mathbf{NMI} = \frac{\sum_{\mathbf{k_1}=1}^{\mathbf{c}} \sum_{\mathbf{k_2}=1}^{\mathbf{c}} \mathbf{n}_{\mathbf{C_{k_1}} \cap \mathbf{C_{k_2}}} \log\left(\frac{\mathbf{n}\mathbf{n}_{\mathbf{C_{k_1}} \cap \mathbf{C_{k_2}}}}{\mathbf{n}_{\mathbf{k_1}} \mathbf{n}_{\mathbf{k_2}}}\right)}{\sqrt{\sum_{\mathbf{k_1}=1}^{\mathbf{c}} \mathbf{n}_{\mathbf{k_1}} \log\left(\frac{\mathbf{n}_{\mathbf{k_1}}}{\mathbf{n}}\right)} \sqrt{\sum_{\mathbf{k_2}=1}^{\mathbf{c}} \mathbf{n}_{\mathbf{k_2}} \log\left(\frac{\mathbf{n}_{\mathbf{k_2}}}{\mathbf{n}}\right)}}$$

where $\mathbf{n}_{\mathbf{k}_1}, (\mathbf{1} \leq \mathbf{k_1} \leq \mathbf{c})$ denotes the number of data in cluster $\mathbf{C}_{\mathbf{k}_1}$ and $\mathbf{n}_{\mathbf{k}_2}, (\mathbf{1} \leq \mathbf{k_2} \leq \mathbf{c})$ denotes the number of data in cluster $\mathbf{C}_{\mathbf{k}_2}$ with $\mathbf{n}_{\mathbf{C}_{\mathbf{k}_1} \cap \mathbf{C}_{\mathbf{k}_2}}$ being the number of data in the intersection set $\mathbf{C}_{\mathbf{k}_1} \cap \mathbf{C}_{\mathbf{k}_2}$.

In Fig. 1, first six datasets of the selected datasets mentioned above are utilized to compare the clustering accuracy of four approaches as PCA[15] + k-means[10], PCA[15] + RMKMC[16], unsupervised OLSDA[1] and the proposed EC under different reduced dimensionality with k-means[10] and RMKMC[16] being the baselines. In particular, PCA[15] + k-means[10] and PCA[15] + RMKMC[16] denote that low-dimensional data $\mathbf{W}^{\mathbf{T}}\mathbf{X}$ is achieved in advance via PCA[15] for further clustering via the k-means[10] and RMKMC[16] methods, respectively.

In Tab. 1, average least error rates are recorded for unsupervised OLSDA and the proposed EC method to compare with the error rates obtained by the k-means and RMKMC methods under the rest of the datasets. Besides, the associated $\mathbf{NMI}$ index is also recorded in Tab. 1.

According to the results in Fig. 1 and Tab. 1, we could conclude that:

1. From Fig. 1, the clustering accuracy of the proposed EC method is better than those of other approaches including PCA + k-means, PCA + RMKMC, unsupervised OLSDA and two baselines under most reduced dimensionality.

2. From Tab. 1, the proposed EC method is consistently better than k-means, RMKMC and unsupervised OLSDA on both error rate and NMI index.

5. CONCLUSION

In this paper, a novel embedded clustering method is proposed via extending the supervised orthogonal least square discriminant analysis method to the robust unsupervised case, which is closely related to k-means. Besides, re-weighted counterpart of unsupervised robust orthogonal least square discriminant analysis problem is utilized for the convenient optimization with self-adaptive weight. Consequently, the embedded clustering method is proposed with robust outliers and optimal weighted cluster centroids. The effectiveness and the superiority of the proposed embedded clustering method are further verified both empirically and analytically.

6. REFERENCES

[1] F. Nie, S. Xiang, Y. Liu, C. Hou, and C. Zhang, "Orthogonal vs. uncorrelated least squares discriminant analysis for feature extraction," *Pattern Recognition Letters*, vol. 33, no. 5, pp. 485–491, 2012.

[2] F. Nie, H. Huang, C. Ding, D. Luo, and H. Wang, "Robust principal component analysis with non-greedy l1-norm maximization," in *Proceedings*. International Joint Conference on Artificial Intelligence, 2011, pp. 1433–1438.

[3] F. Nie, D. Xu, W. Tsang, and C. Zhang, "Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction," *IEEE Transactions on Image Processing*, vol. 19, no. 7, pp. 1921–1932, 2010.

[4] T. Zhou and D. Tao, "Double shrinking sparse dimension reduction," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 244–257, 2013.

[5] X. Li, Y. Pang, and Y. Yuan, "l1-norm based 2dpca," *IEEE Trans. on System, Man and Cybernetics, Part B*, vol. 40, no. 4, pp. 1170–1175, 2010.

[6] Y. Pang, D. Tao, Y. Yuan, and X. Li, "Binary two-dimensional pca," *IEEE Trans. on System, Man and Cybernetics, Part B*, vol. 38, no. 4, pp. 1176–1180, 2008.

[7] C. Zhang, F. Nie, and S. Xiang, "A general kernelization framework for learning algorithms based on kernel pca," *Neurocomputing*, vol. 73, no. 4-6, pp. 959–967, 2010.

[8] T. Luo, C. Hou, D. Yi, and J. Zhang, "Discriminative orthogonal elastic preserving projections for classification," *Neurocomputing*, vol. 179, pp. 54–58, 2016.

[9] H. Park, M. Jeon, and J. B. Rosen, "Lower dimensional representation of text data based on centroids and least squares," *BIT Numer. Math.*, vol. 43, no. 2, pp. 427–448, 2003.

[10] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings*. Berkeley Symposium on Mathematical Statistics and Probability, 1967, vol. 1, pp. 281–297.

[11] F. Nie, J. Yuan, and H. Huang, "Optimal mean robust principal component analysis," in *Proceedings*. International Conference on Machine Learning, 2014, pp. 2755–2763.

[12] W. Jiang, F. Nie, and H. Huang, "Robust dictionary learning with capped l1 -norm," in *Proceedings*. International Joint Conference on Artificial Intelligence, 2015, pp. 3590–3596.

[13] G. Lan, C. Hou, and D. Yi, "Robust feature selection via simultaneous capped l2-norm and l2;1-norm minimization," in *IEEE*. International Conference on Big Data Analysis, 2016.

[14] M. Qian and C. Zhai, "Robust unsupervised feature selection," in *Proceedings*. International Joint Conference on Artificial Intelligence, 2013, pp. 1621–1627.

[15] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, NY, USA, 2nd edition, 2002.

[16] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in *Proceedings*. International Joint Conference on Artificial Intelligence, 2013, pp. 2598–2604.