Embedded Clustering via Robust Orthogonal Least Square Discriminant Analysis

Rui Zhang, Feiping Nie, and Xuelong Li

Northwestern Polytechnical University

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017

Outline

Conventional definitions of scatter matrices

The least squared loss function could be illustrated as:

$$\varepsilon = \| T_1 - T_2 \|_F^2 \quad (1)$$

Denote $\mathscr{X}_i$ is the dataset of i-th class and n_i is the number of data points in i-th class, then the within-class scatter matrix S_w, the between-class scatter matrix S_b and the total-class scatter matrix S_t are defined as follows:

- $S_w = \sum_{i=1}^{c} \sum_{x \in \mathscr{X}_i} (x - \bar{x}_i)(x - \bar{x}_i)^T$
- $S_b = \sum_{i=1}^{c} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$
- $S_t = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$

Reformulation of discriminant problems in least square forms

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

Orthogonal Least Square Discriminant Analysis (OLSDA)

By substituting $T_1 = W^T X$ and $T_2 = W^T X A^{(w)}$, we have

$$\begin{aligned} \|W^T X - W^T X A^{(w)}\|_F^2 &= Tr(W^T X (I - A^{(w)})^2 X^T W) \\ &= Tr(W^T S_w W) \end{aligned} \quad (2)$$

which is the objective function of OLSDA.

- Similarly, by replacing $T_1 = W^T X$ and $T_2 = W^T X A^{(t)}$, we have

$$\begin{aligned} \|W^T X - W^T X A^{(t)}\|_F^2 &= Tr(W^T X (I - A^{(t)})^2 X^T W) \\ &= Tr(W^T S_t W) \end{aligned} \quad (3)$$

Reformulation of discriminant problems in least square forms

New form of between-class scatter S_b

According to the results in (2) and (3), we have

$$\begin{aligned} S_b = S_t - S_w &= X(I - A^{(t)} - I + A^{(w)})X^T \\ &= X(A^{(w)} - A^{(t)})X^T \\ &= XHY(Y^TY)^{-1}Y^THX^T \end{aligned} \tag{4}$$

Moreover, $S_t = X(I - A^{(t)})X^T = XHX^T$ due to Eq. (3). In sum, we have

$$\begin{cases} S_t = XHX^T \\ S_b = XHY(Y^TY)^{-1}Y^THX^T \end{cases} \tag{5}$$

Interesting observations of OLSDA and k-means

OLSDA in a brand new form

$$\begin{aligned}\min_{W^TW=I} Tr(W^T S_w W) &= \min_{W^TW=I} Tr(W^T(S_t - S_b)W) \\ &= \min_{W^TW=I} Tr(W^T XH(I - Y(Y^TY)^{-1}Y^T)HX^TW) \\ &= \min_{W^TW=I} \|W^T XH(I - Y(Y^TY)^{-1}Y^T)\|_F^2\end{aligned} \quad (6)$$

- The k-means problem: $\min_{F,G\in ind} \|T - FG^T\|_F^2$

Supervised k-means

If the associated label is known, i.e., indicative matrix G is fixed as binary label Y, the k-means problem degenerates to

$$\min_F \|T - FY^T\|_F^2 = \|T - TY(Y^TY)^{-1}Y^T\|_F^2. \quad (7)$$

Equivalence between OLSDA and k-means

- By further replacing the data T with the centralized projected data $W^T XH \in \mathbb{R}^{k \times n}$ in Eq. (7), we notice that the problem (7) is same as the problem (6).

Theorem 1

OLSDA in (6) is equivalent to k-means problem when $T = W^T XH$ and $G = Y$.

Unsupervised OLSDA

Due to theorem 1, we could extend OLSDA to the unsupervised case.

Unsupervised OLSDA

Accordingly, OLSDA in (6) could be naturally extended to the unsupervised case as

$$\min_{W^TW=I,F,G\in ind} \|W^TXH - FG^T\|_F^2. \tag{8}$$

How to further modify unsupervised OLSDA in (8)

- Enhancing the robustness of OLSDA has the following superiorities.
  1. Insensitive to the outliers.
  2. Weighted cluster centroids.

Embedded clustering via robust OLSDA

Robust OLSDA

Based on the unsupervised OLSDA in (8), robust OLSDA (ROLSDA) could be proposed as

$$\min_{W^T W=I,F,G\in ind} \|W^T XH - FG^T\|_{2,1}. \tag{9}$$

How to solve the ROLSDA in (9)

Re-weighted counterpart of ROLSDA in (9) is utilized as

$$\begin{aligned} &\min_{W^T W=I,F,G\in ind} \|(W^T XH - FG^T)D^{\frac{1}{2}}\|_F^2 \\ =& \min_{W^T W=I,F,G\in ind} \sum_{i=1}^{n} D_{ii}\|W^T x_i^{(H)} - Fg_i\|_2^2. \end{aligned} \tag{10}$$

Associated Karush-Kuhn-Tucker (KKT) conditions

The Lagrangian function is represented as

$$\mathscr{L}(W,F) = \|(W^T XH - FG^T)D^{\frac{1}{2}}\|_F^2 - Tr(\Lambda(W^T W - I)). \quad (11)$$

Closed form solution

1) $\frac{\partial \mathscr{L}(W,F)}{\partial F} = 0 \Rightarrow F = W^T XHDG(G^T DG)^{-1}$, which is the weighted form of cluster centroids.

2) $\frac{\partial \mathscr{L}(W,F)}{\partial W} = 0 \Rightarrow W^T(S_t^{(D)} - S_b^{(D)})W = \Lambda$, which implies that W is the matrix of eigenvector corresponding to the first k smallest eigenvalues of $S_t^{(D)} - S_b^{(D)}$ with

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

Pseudo-code

How to determine the hard label G?

This question could be answered by individually solving

$$\min_{g_i \in ind} \|W^T x_i^{(H)} - F g_i\|_2^2 \;\; s.t. \;\; \mathbf{1_c}^T g_i = 1. \tag{12}$$

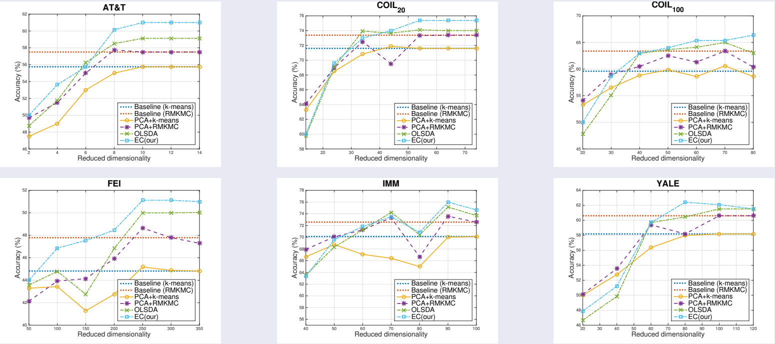
The algorithm could be summarized as

Embedded clustering (EC)

Initialize $D = I$ with random $G \in ind$ and orthogonal W

1. Update cluster centroids by $F \leftarrow W^T XHDG(G^T DG)^{-1}$.
2. Update hard label G by by individually solving Eq. (12).
3. Update $D_{ii} \leftarrow \frac{1}{2\|W^T x_i^{(H)} - F g_i\|_2}$.
4. Update $S_t^{(D)}, S_b^{(D)}$.
5. Compute W by solving $\min_{W^T W = I} W^T (S_t^{(D)} - S_b^{(D)}) W$.

Experiments



Interpretation

The clustering accuracy comparisons are performed for PCA+k-means method, PCA+RMKMC method, unsupervised OLSDA method and proposed EC method.

Conclusions and Future Works

- We discover an interesting theorem about the equivalence between OLSDA and k-means.
- Based on the robust OLSDA, we propose EC method to deal with unlabeled data sets efficiently.
- Some further progress on anchor generation strategy are needed. Recently, we propose a pretty efficient and effective method to replace k-means method.

Thanks!