On Mitigating Hard Clusters for Face Clustering

Yingjie Chen[1,2⋆], Huasong Zhong[2⋆], Chong Chen[2⋆⋆], Chen Shen[2], Jianqiang Huang[2], Tao Wang[1], Yun Liang[1], and Qianru Sun[3]

[1] Peking University, Beijing, China {chenyingjie,wangtao,ericlyun}@pku.edu.cn
[2] DAMO Academy, Alibaba Group, China
cheung.cc@alibaba-inc.com,{zjushenchen,zhonghsuestc,jianqiang.jqh}@gmail.com
[3] Singapore Management University, Singapore qianrusun@smu.edu.sg

1 Appendix

1.1 Proof of Theorem 1

Theorem 1. *Assume the dataset* $\boldsymbol{X}$ *can be split into* m *disjoint clusters:* i.e., $\boldsymbol{X} = \boldsymbol{X}_1 \bigcup ... \bigcup \boldsymbol{X}_m$. *Define* $\bar{\rho}_i = \frac{\sum_{j\in \boldsymbol{X}_i} \rho(j)}{|\boldsymbol{X}_i|}$ *is the average density of* $\boldsymbol{X}_i$*, then we have*

$$\bar{\rho}_1 = \cdots = \bar{\rho}_m = 1,$$

where $|\boldsymbol{X}_i|$ *is the number of samples in* $\boldsymbol{X}_i$*.*

Proof. Since $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_m\}$ are disjoint, we have $p_{ij} = 0$ if point i and point j belong to different clusters. By the definition of transition matrix $\boldsymbol{P}$, for each $i \in \boldsymbol{X}_l$, we have

$$\sum_j p_{ij} = 1$$

which implies that

$$\sum_{i\in \boldsymbol{X}_l} \sum_{j\in \boldsymbol{X}_l} p_{ij} = |\boldsymbol{X}_l|.$$

Therefore,

$$\bar{\rho}_l |\boldsymbol{X}_l| = \sum_{i\in \boldsymbol{X}_l} \sum_{j\in \boldsymbol{X}_l} p_{ij} = |\boldsymbol{X}_l|,$$

which implies that $\bar{\rho}_l = 1$ for any $l = 1, , \ldots, m$. □

1.2 Pseudo code for DPC

Here we provide a pseudo code for Density Peak Clustering (DPC) [5] in Algorithm 1.

⋆ Equal contribution.
⋆⋆ Corresponding author.

Algorithm 1: Pseudocode for DPC

Input: Point-wise density $\rho_{N\times 1}$, pair-wise distance $(d_{ij})_{N\times N}$, connecting threshold τ, number of nearest neighbors K.
Output: clusters $\mathcal{C}$.

for each sample i:
    Obtain its K-nearest neighbors $\mathrm{nbr}_i = \{i_k\}_{k=1}^{K}$;
    Find $\hat{j} = \mathrm{argmin}_{\{j \in \mathrm{nbr}_i \mid \rho_j > \rho_i\}} d_{ij}$,;
    if $\hat{j}$ exists and $d_{i\hat{j}} < \tau$:
        Connect i to $\hat{j}$;
Use BFS algorithm to obtain separated trees as clusters $\mathcal{C}$;

1.3 Metrics

In our experiments, Pairwise F-score [2] and BCubed F-score [1] are used as evaluation metrics. We denote the ground-truth clusters as $\{L_1, L_2, \dots, L_M\}$ and the predicted clusters as $\{C_1, C_2, \dots, C_N\}$. And for each point, we denote g_i and p_i as the ground-truth cluster label and the predicted cluster label of point i, respectively. For each pair of ground-truth cluster L_m and predicted cluster C_n, we define $TP(m,n) = |\{(i,j) | g_i = g_j = m \wedge p_i = p_j = n\}|$. Pairwise Precision and Pairwise Recall are defined in Eq. 1, and BCubed Precision and BCubed Recall are defined in Eq. 2. Both Pairwise F-score and Bcubed F-score are the harmonic mean between the corresponding precision and recall, which are computed as in Eq. 3.

$$\begin{aligned} \mathit{Pairwise\ Precision} &= \frac{\sum_{m,n} TP(m,n)^2 - \sum_n |C_n|}{\sum_n |C_n|^2 - \sum_n |C_n|}, \\ \mathit{Pairwise\ Recall} &= \frac{\sum_{m,n} TP(m,n)^2 - \sum_m |L_m|}{\sum_m |L_m|^2 - \sum_m |L_m|}. \end{aligned} \tag{1}$$

$$\begin{aligned} \mathit{Bcubed\ Precision} &= \mathbb{E}_L\Big(\sum_n \frac{TP(m,n)^2}{|L_m|^2}\Big), \\ \mathit{Bcubed\ Recall} &= \mathbb{E}_C\Big(\sum_n \frac{TP(m,n)^2}{|C_n|^2}\Big). \end{aligned} \tag{2}$$

$$\mathit{F\text{-}score} = \frac{2 \times \mathit{Precision} \times \mathit{Recall}}{\mathit{Precision} + \mathit{Recall}}. \tag{3}$$

1.4 Results on Emore Dataset

To further show the effectiveness of the proposed NDDe and TPDi, we also conduct experiments on Emore dataset, which contains 2,577 identities with 200,000 images, following the same experimental protocol mentioned in [7].

For a fair comparison with these traditional methods, we conduct experiments without the training process on Emore dataset. We perform the clustering procedure just using the original features to compute the cosine similarity for the construction of transition matrix P. As shown in Table 1, our method significantly outperforms these traditional methods. And the number of clusters achieved by our method is quite close to the ground-truth (note that K-means and HAC take the ground-truth number of clusters as an input). It is worth noticing that by using NDDe and TPDi, Bcubed Recall of our method outperforms others by a larger margin, resulting in the highest F_B, which shows the effectiveness of the proposed NDDe and TPDi in mitigating hard clusters.

Method	#Clusters	Precision	Recall	F_B
K-means [3]	2577	94.24	74.89	83.45
HAC [6]	2577	**97.74**	88.02	92.62
ARO [4]	85150	52.96	16.93	25.66
CDP [7]	-	89.35	88.98	89.16
Ours	2569	96.78	**97.51**	**97.14**

Table 1. Comparison on Emore dataset. BCubed Precision, BCubed Recall and F_B are reported. The best and the second highest results are highlighted with **bold** and underline, respectively.

1.5 Code

Our code is at https://github.com/echoanran/On-Mitigating-Hard-Clusters.

References

1. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval **12**(4), 461–486 (2009)
2. Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S., Mooney, R.J.: Model-based overlapping clustering. In: Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. pp. 532–537 (2005)
3. Lloyd, S.: Least squares quantization in pcm. TIP (1982)
4. Otto, C., Wang, D., Jain, A.K.: Clustering millions of faces by identity. TPAMI (2017)
5. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science **344**(6191), 1492–1496 (2014)
6. Sibson, R.: Slink: an optimally efficient algorithm for the single-link cluster method. The Computer Journal (1973)
7. Zhan, X., Liu, Z., Yan, J., Lin, D., Loy, C.C.: Consensus-driven propagation in massive unlabeled data for face recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 568–583 (2018)