ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414630

# FAST LOCAL REPRESENTATION LEARNING WITH ADAPTIVE ANCHOR GRAPH

*Canyu Zhang*[1] *Feiping Nie*[1†] *Zheng Wang*[1] *Rong Wang*[1,2] *Xuelong Li*[1]

[1] School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an, 710072, Shaanxi, P. R. China.
[2] School of Cybersecurity, Northwestern Polytechnical University, Xi'an, 710072, Shaanxi, P. R. China

## ABSTRACT

Dimension reduction is an effective technology to embed data with high dimension to lower dimension space, where Linear Discriminant Analysis (LDA), one of representative methods, only works with Gaussian distribution data. However, in order to solve non-Gaussian issue that only one cluster cannot well fit the distribution of same class, many graph-based discriminant analysis methods are proposed which capture local structure through measuring each pairwise distance. This is expense of time complexity because of the full-connections. In order to solve this issue, we propose a fast local representation learning with adaptive anchor graph to learn local structure information through similarity matrix in anchor-based graph. Notably, anchor points and similarity matrix are updated in subspace which is more precisely to capture local discriminant information. Experimental results on several synthetic and well-known datasets demonstrate the advantages of our method over the state-of-the-art methods.

***Index Terms***— Fast Local Representation Learning, Adaptive Anchor Graph, Linear Discriminant Analysis

## 1. INTRODUCTION

Dimension reduction is an effective method to embed high dimension data into low-space in order to solve *"curse of dimensionality"* [1], which is widely applied in classification [2][3], face recognition [4][5], biology [6]. According to the label whether used, dimension reduction methods can divided into three major, unsupervised method [7][8], semi-supervised method [9][10] and supervised method [11][12].

LDA [13] is one of popular supervised linear transformation methods, in which the aim of LDA is to make same class points closer to the center point and separate different classes far away from each other in subspace. However, LDA is not suitable for non-Gaussian data [14] with the traditional strategy in what we mentioned above owing to the fact that center point cannot represent data with non-Gaussian distribution, which called non-Gaussian issue [15].

In order to solve the non-Gaussian issue, many graph-based methods which focus on local structure information through establishing fully-connected graph, such as Locality Sensitive Discriminant Analysis (LSDA) [16], Adaptive Discriminative Analysis (ADA) [17] and Adaptive Local Linear Discriminant Analysis (ALLDA) [18]. The local structure of the data can be better captured in the full-connections graph through similarity matrix. When the distance of pairwise points is smaller, the value of similarity matrix is larger. However, the subsequent cost is heavy computation complexity, because the similarity of each pairwise need to calculate. Another way to solve non-Gaussian issue is redividing each class into several sub-clsuters, such as Subclass Discriminant Analysis (SDA) [19], Separability-oriented Subclass Discriminant Analysis (SSDA) [20] and Mixture Subclassiscrim Dinant Analysis (MSDA) [21]. Nevertheless, the partition of each sub-cluster is carried out in original space which suffers the consequence of noises and redundant features.

To get around these issues, we proposed a novel method called Fast Local Representation Learning with Adaptive Anchor Graph (FLRL) which can effectively capture local structure information with reducing heavy computation complexity and free from being effected by noises. In a word, we summarize the contributions from following three aspects:

1. We construct an efficient adaptive anchor graph to learn local structure information of data by similarity matrix and reduce computation complexity. And we apply $\ell_0-$norm constraint on each point to ensure our model more precisely and sensitively to local structure.

2. Bridging with anchor graph, we propose a fast local representation learning method called FLRL to learn discriminant information to solve non-Gaussian issues.

3. We propose an effective algorithm to update similarity matrix and anchor points in subspace and learn subspace alternately. Our alternative optimization algorithm can eliminate the noises of original space and obtain more truly relationships between data points and anchor points.

[†] Corresponding author. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1403501, in part by the National Natural Science Foundation of China under Grant 61936014, Grant 61772427 and Grant 61751202, and in part by the Fundamental Research Funds for the Central Universities under Grant G2019KY0501.

## 2. FAST LOCAL REPRESENTATION LEARNING WITH ADAPTIVE ANCHOR GRAPH

Given a data matrix $X \in R^{d\times n}$, each data point can be viewed as nodes in graph and the edges are represented as similarity matrix $A \in R^{n\times n}$ to describe their own relationships. Even though full-connections graph what we above mentioned can capture local structure, the computation complexity is much heavy because the number of pairwise connections is reached to $n^2$. Different from these full-connection strategy, we propose a novel method to leverage anchor points and construct anchor-based graph which only need to learn relationship between data points with anchor points. It is obvious that we only need to calculate $nm$ pairwise connections where $m$ is the number of anchor points.

**Adaptive Anchor Graph** The similarity matrix $A$ to capture relationship of pairwise in graph is depend on the distances of each pairwise. When the distance between data point $x_j^i$ and anchor point $z_h^i$ is smaller, the value of similarity $A_{jh}^i$ becomes large. In order to depict the relationships of data points and anchor points and reveal the local structure in quantitative, we can construct the graph as follows:

$$\min_{A_j^i\mathbf{1}=1}\sum_{j=1}^{n_i}\sum_{h=1}^{m_i}A_{jh}^i\left\|x_j^i - z_h^i\right\|_2^2, \tag{1}$$

where $A_{jh}^i \in R^{n_i\times m_i}$ is a similarity matrix of anchor graph and $\mathbf{1}$ is a vector that all elements are 1.

However, according to Eq. (1), the solution of $A_{jh}^i$ always meet to trivial solution [22]. In order to solve the issue and extend the model, we have made changes in two aspects. One is to add exponential decay factor $r$ on similarity matrix and one is to apply $\ell_0$-norm constraint on anchor points which ensures each point connected with $k$ anchor points adaptively. Based on above-mentioned, Eq. (1) can be rewritten as follows

$$\min_{A_j^i\mathbf{1}=1,\left\|A_j^i\right\|_0=k}\sum_{j=1}^{n_i}\sum_{h=1}^{m_i}(A_{jh}^i)^r\left\|x_j^i - z_h^i\right\|_2^2. \tag{2}$$

**Fast Local Representation Learning** Bridging with adaptive anchor graph, we apply this into classical LDA method, then the objective function can be transformed as:

$$\begin{aligned}\min_{W,A,Z}&\sum_{i=1}^{C}\sum_{j=1}^{n_i}\sum_{h=1}^{m_i}\left(A_{jh}^i\right)^r\left\|W^Tx_j^i - z_h^i\right\|_2^2\\ \textit{s.t.}\quad & A_j^i\mathbf{1}=1,\left\|A_j^i\right\|_0=k, W^TS_tW=I,\end{aligned} \tag{3}$$

where the total-class scatter matrix is defined as $S_t = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)(x_i - x_j)^T$ and the last constraint is to keep total-class scatter matrix $S_t$ as a certain level.

In addition, considering with the effect of noise and redundant feature, we set anchor point as variables to update in lower-dimension space. In this way, anchor points can become more precisely to reflect the distribution of data.

## 3. OPTIMIZATION

In this section, we propose an iterative algorithm to update transformation matrix $W$, anchor point $Z$ and similarity matrix $A$. And we summarize its details in Algorithm 1.

When $W$ and $A$ are fixed, considering anchor point of different class won't interfere with each other, Eq. (3) becomes

$$\min_{Z}\quad\sum_{h=1}^{m_i}\left(A_{jh}^i\right)^r\left\|W^Tx_j^i - z_h^i\right\|_2^2. \tag{4}$$

Based on this, we can get the solution of Eq. (4) as follows:

$$z_h^i = W^T\widehat{z}_h^i = \frac{\sum_{j=1}^{n_i}\left(A_{jh}^i\right)^r W^Tx_j^i}{\sum_{j=1}^{n_i}\left(A_{jh}^i\right)^r}, \tag{5}$$

When $W$ and $Z$ are fixed, we can transform the Eq. (3) to

$$\begin{aligned}\min_{A}&\sum_{h=1}^{m_i}\left(A_{jh}^i\right)^r\left\|W^Tx_j^i - z_h^i\right\|_2^2\\ \textit{s.t.}\quad & A_j^i\mathbf{1}=1,\left\|A_j^i\right\|_0=k.\end{aligned} \tag{6}$$

After calculated distances between anchor point and data point in subspace, the optimal solution of similarity matrix is

$$A_{jt}^i = \begin{cases}\frac{\left(\left\|W^Tx_j^i - z_t^i\right\|_2^2\right)^{\frac{1}{1-r}}}{\sum_{l=1}^{k}\left(\left\|W^Tx_j^i - z_{h_l^*}^i\right\|_2^2\right)^{\frac{1}{1-r}}}, & \text{if}\quad t=h_l^*\\ 0, & \text{otherwise,}\end{cases} \tag{7}$$

where $h_l^*$ is the index of neighbor anchor point for each point.

When $A$ and $Z$ are fixed, substituting Eq. (5) into Eq. (3), Eq. (3) is transformed into the following:

$$\begin{aligned}\min_{W}&\sum_{i=1}^{C}\sum_{j=1}^{n_i}\sum_{h=1}^{m_i}\left(A_{jh}^i\right)^r\left\|W^Tx_j^i - W^T\widehat{z}_h^i\right\|_2^2\\ \textit{s.t.}\quad & W^TS_tW=I.\end{aligned} \tag{8}$$

Now we define a new matrix $\widehat{X} = [X \quad \widehat{Z}] \in R^{d\times(n+m)}$, $P_{jh}^i = \left(A_{jh}^i\right)^r$, so the matrix $S$ can be defined as $S = \begin{bmatrix} & P\\ P^T & \end{bmatrix}$. Then the degree matrix $D_S$ is a diagonal matrix and the $j$-th diagonal element is $\sum_{h=1}^{n+m}S_{jh}$.

Based on above definition, Eq. (8) reduces to solve

$$\begin{aligned}\min_{W}& Tr(W^T\widehat{X}L_S\widehat{X}^TW)\\ \textit{s.t.}\quad & W^TS_tW=I,\end{aligned} \tag{9}$$

where Laplacian matrix $L_S = D_S - S$. The optimal transformed matrix $W$ is composed by the $m$ eigenvectors of $S_t^{-1}\widehat{X}L_s\widehat{X}^T$ corresponding to the $m$ small eigenvalues.

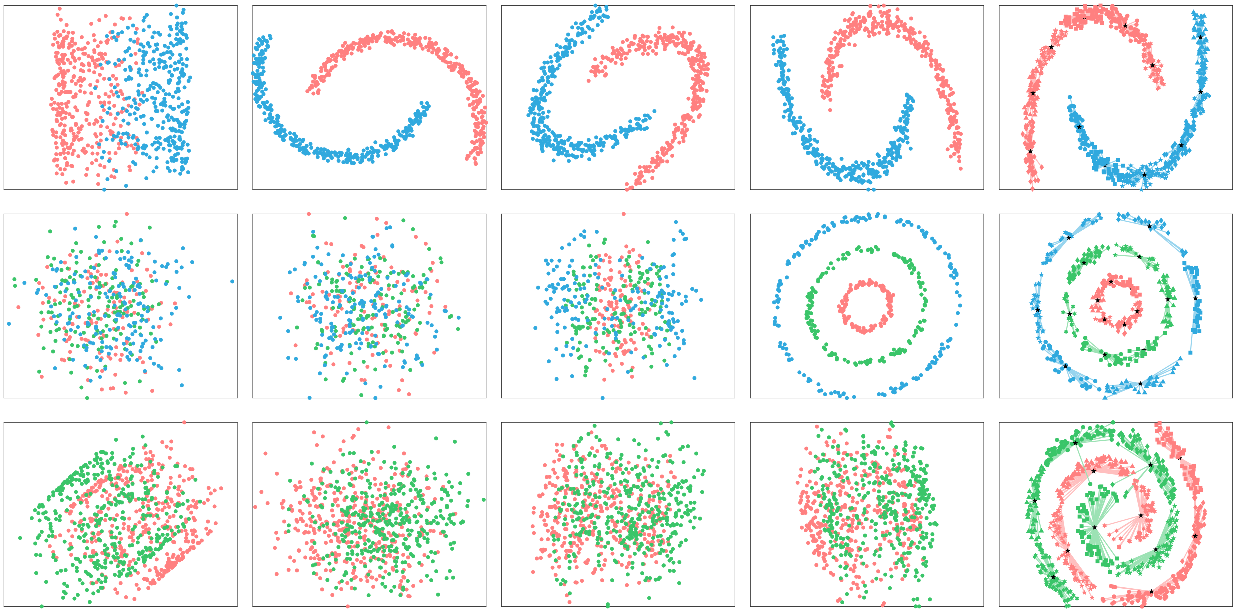

**Fig. 1**. Visualization the classification of synthetic toy datasets in LDA, SSDA, LADA, ALLDA and FLRL in turns from left to right. The first two dimensions of these toy datasets are distributed in two moons, three circles and two spirals shapes respectively. The black points mean anchor points and the samples connected with same anchor point is set to the same symbols in FLRL.

**Algorithm 1** An efficient algorithm to solve problem (3).

```
1. Input: Data $X \in \mathbb{R}^{d\times n}$, the number of anchor points $m$,
reduced dimension $q$.
2.Initialisation: Anchor point $Z$ and similarity matrix $A$
While not converge do
3. Calculate $L_S = D_S - S$.
4. Update $W$ according to Eq. (9).
5. Update $A$ according to Eq. (7).
6. Update $Z$ according to Eq. (5).
end while
7. Output: Transformation matrix $W^* \in \mathbb{R}^{d\times q}$.
```

To summarize, the computational complexity of Algorithm 1 are concentrated on step 4 to step 6 in each iteration. The computation complexity in step 4 to calculate transformation matrix $W$ is $O(d^3 + nmd)$. The computation complexity in step 5 is $O(nmd)$. Calculated $Z$ takes computation complexity $O(nmd)$. Considering $n$ is much larger than $d$ and $m$, the total computational complexity turns to $O(nmd)$.

## 4. EXPERIMENTS

In this section, we will show the performance of the our method on three toy datasets and seven real-world datasets. And we also record the running times to validate computation complexity and show the effectiveness of our method.

**Table 1**. Descriptions of datasets

| Datasets | Instance | Dimension | Class | Domain |
|---|---|---|---|---|
| TOX | 171 | 5748 | 4 | Biology |
| ProstateMS | 322 | 15154 | 3 | Biology |
| German | 1000 | 20 | 2 | other |
| Coil20 | 1440 | 1024 | 20 | Object |
| Qsar | 8992 | 1024 | 2 | Chemistry |
| CMU-PIE | 11554 | 6996 | 68 | Face |
| Emnist | 402953 | 784 | 10 | Digits |

### 4.1. Experimental results on toy dataset

In this part, we generate three synthetic toy datasets respectively to experiment visually in order to verify the ability of the algorithms to capture local structure. The toy datasets are all 12 dimensions which we set two moons, three circles and two spirals in the first 2-dimensions we need to focus on and the rest noisy dimensions are generated by using uniform distribution. And we used transformation matrix $W$ to embed high dimension data into 2 dimensions space. These five columns are LDA, SSDA [20], LADA [23], ALLDA [18] and our method in turn from left to right. According to Fig. 1, LDA is failed to reconstruct the original distribution because of neglect local structure. Even through SSDA, LADA and ALLDA can mine local information to some extent, noisy or full-connections are still lead to failure to reconstruct original distribution in subspace precisely. Visual experiments

**Table 2**. Experimental results (mean accuracy±standard deviations%) with optimal dimensions on German, Coil20, Qsar, CMU-PIE, TOX and ProstateMS datasets by 1NN. The bold numbers are the highest in statistical view.

| Datasets | German | Coil20 | Qsar | CMU-PIE | TOX | ProstateMS |
|---|---|---|---|---|---|---|
| LDA | 64.73±1.86(1) | 99.09±0.43(16) | 88.76±0.41(1) | 94.99±0.24(30) | 74.95±6.82(3) | 90.65±2.05(2) |
| MMC | 63.60±1.72(19) | 98.31±0.52(25) | 91.55±0.27(121) | 81.29±0.50(45) | 69.80±5.85(43) | 83.08±2.13(37) |
| LSDA | 64.73±2.02(13) | 98.95±0.46(16) | 91.78±0.34(85) | 94.88±0.26(30) | 74.31±6.14(23) | 91.30±2.14(21) |
| LADA | 66.68±1.84(15) | 99.06±0.54(13) | 91.66±0.32(25) | 95.43±0.27(40) | 78.04±5.92(31) | 90.39±2.42(21) |
| SSDA | 63.70±1.92(9) | 99.53±0.31(19) | 91.87±0.32(109) | 95.49±0.22(30) | 81.52±5.33(23) | 89.02±2.01(21) |
| ADA | 64.18±1.96(11) | 98.72±0.57(22) | 92.50±0.29(37) | 87.13±0.47(40) | 73.19±5.79(35) | 85.65±1.83(21) |
| ALLDA | 67.00±0.93(13) | 99.48±0.39(10) | 92.07±0.26(109) | 95.30±0.29(35) | 81.47±4.37(27) | 90.62±2.11(13) |
| FLRL | **69.23±1.33**(11) | **99.74±0.30**(13) | **92.62±0.26**(61) | **95.63±0.21**(30) | **82.70±3.14**(19) | **93.24±2.03**(9) |

prove the effectiveness of our algorithm while leverage anchor points to construct anchor-based graph through similarity matrix to capture local structure.

### 4.2. Experimental results on real-world datasets

In our experiments, we use six datasets to evaluate the performance for some dimension reduction methods on the classification task. These datasets include two UCI datasets[1], named German and Qsar, and other four datasets, such as Coil20[2], TOX[2], CMU-PIE[3] and ProstateMS[4]. And for each grey image, we simply resize them to $16 \times 16$ pixels.

Firstly, we pre-process all datasets with normalization and PCA to preserve major information. For each dataset, 40% of samples per class are randomly selected as the training data and the remaining data are used as the testing data. For SSDA [20], the number of max subclass is set as 10 or half of training number in each class. For ADA [17], the scale parameter $\delta$ is set as $[0.1 : 0.1 : 1]$. For ALLDA [18], the number of neighbor points $h$ is search in the range of $[1 : 10]$. For FLRL, the number of anchor number $m_i$ is in range of $[2 : 10]$, and the parameter $k$ is set as $[1 : m_i - 1]$. And for other methods, the parameters are as default. The reducing dimension of LDA is searched in $[1 : c - 1]$. In contrast, our method and others are searched in $[1 : d - 1]$ where $d$ is the dimension of data after precessing. Notably, we use 1-nearest neighbor classifier [24] to verify these methods. All experiments are performed 20 runs to obtain convinced and stable results.

The results including mean accuracy, standard deviations and optional dimension are provided in Table.2. Apparently, our method can achieve best performance on the six datasets than other methods in optimal dimension. That proves the effectiveness of our method which use anchor point to learn local structure. We also compared our method with three local methods which have to optimize iterative variables alternately on the emist dataset. We conduct 100 iterations to record the running times of each method in Fig. 2.Obviously, our method can achieve the smallest time in all method and still can work with large scale dataset while other methods are out of memory when the training number are 50k and 70k respectively.

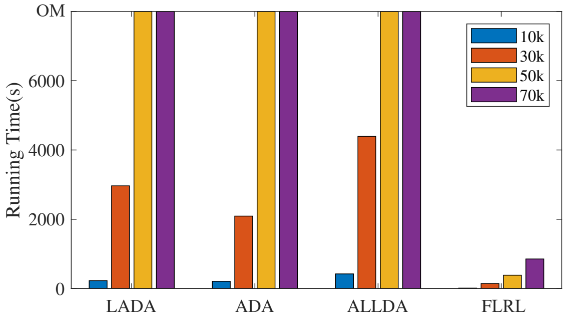

**Fig. 2**. Comparison of running time (sec) about LADA, ADA and ALLDA methods (full connection graph) with our method (anchor-based graph). OM means out of memory.

## 5. CONCLUSION

In this paper, we propose a fast and effective Local Representation Learning (FLRL) with Adaptive Anchor Graph method. We establish anchor-based graph to replace full-connections graph to explore the information of data's local structure. We also impose $\ell_0$-norm constraint on anchor point in our method. At last, we derive an iterative optimization algorithm to solve similarity matrix and transformation which can alleviate the affect of noises and redundant features in original space. The effectiveness of our method to capture local informations and reduce high computation complexity are proved in experiments. In future work, we will focus on how to set the number of anchor points for each class adaptively in order to give model more autonomous and no need to restrict obey the fixed number.

[1] https://archive.ics.uci.edu/ml/datasets.php

[2] http://featureselection.asu.edu/datasets.php

[3] http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html

[4] https://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

## 6. REFERENCES

[1] Michel Verleysen and Damien François, "The curse of dimensionality in data mining and time series prediction," in *International work-conference on artificial neural networks*. Springer, 2005, pp. 758–770.

[2] Le Yang, Shiji Song, Yanshang Gong, Huang Gao, and Cheng Wu, "Nonparametric dimension reduction via maximizing pairwise separation probability," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3205–3210, 2019.

[3] Le Yang, Shiji Song, Shuang Li, Yiming Chen, and CL Philip Chen, "Discriminative dimension reduction via maximin separation probability analysis," *IEEE Transactions on Cybernetics*, 2019.

[4] Dapeng Tao, Yanan Guo, Yaotang Li, and Xinbo Gao, "Tensor rank preserving discriminant analysis for facial recognition," *IEEE transactions on image processing*, vol. 27, no. 1, pp. 325–334, 2017.

[5] Li Zhang, Tao Xiang, and Shaogang Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] Karl Øyvind Mikalsen, Cristina Soguero-Ruiz, Filippo Maria Bianchi, and Robert Jenssen, "Noisy multi-label semi-supervised dimensionality reduction," *Pattern Recognition*, vol. 90, pp. 257–270, 2019.

[7] Xiaofei He and Partha Niyogi, "Locality preserving projections," in *Advances in neural information processing systems*, 2004, pp. 153–160.

[8] Meihong Wang, Fei Sha, and Michael I Jordan, "Unsupervised kernel dimension reduction," in *Advances in Neural Information Processing Systems*, 2010, pp. 2379–2387.

[9] Sheng Wang, Jianfeng Lu, Xingjian Gu, Haishun Du, and Jingyu Yang, "Semi-supervised linear discriminant analysis for dimension reduction and classification," *Pattern Recognition*, vol. 57, pp. 179–189, 2016.

[10] Feiping Nie, Dong Xu, Ivor Wai-Hung Tsang, and Changshui 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.

[11] Masashi Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 905–912.

[12] Feiping Nie, Shiming Xiang, and Changshui Zhang, "Neighborhood minmax projections.," in *IJCAI*, 2007, pp. 993–998.

[13] Christopher M Bishop, *Pattern recognition and machine learning*, springer, 2006.

[14] Feiping Nie, Zheng Wang, Rong Wang, and Xuelong Li, "Submanifold-preserving discriminant analysis with an auto-optimized graph," *IEEE transactions on cybernetics*, 2019.

[15] Xipeng Qiu and Lide Wu, "Stepwise nearest neighbor discriminant analysis.," in *IJCAI*. Citeseer, 2005, pp. 829–834.

[16] Deng Cai, Xiaofei He, Kun Zhou, Jiawei Han, and Hujun Bao, "Locality sensitive discriminant analysis.," in *IJCAI*, 2007, vol. 2007, pp. 1713–1726.

[17] Tingjin Luo, Chenping Hou, Feiping Nie, and Dongyun Yi, "Dimension reduction for non-gaussian data by adaptive discriminative analysis," *IEEE transactions on cybernetics*, vol. 49, no. 3, pp. 933–946, 2018.

[18] Feiping Nie, Zheng Wang, Rong Wang, Zhen Wang, and Xuelong Li, "Adaptive local linear discriminant analysis," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 14, no. 1, pp. 1–19, 2020.

[19] Manli Zhu and Aleix M Martinez, "Subclass discriminant analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 8, pp. 1274–1286, 2006.

[20] Huan Wan, Hui Wang, Gongde Guo, and Xin Wei, "Separability-oriented subclass discriminant analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 2, pp. 409–422, 2017.

[21] Nikolaos Gkalelis, Vasileios Mezaris, and Ioannis Kompatsiaris, "Mixture subclass discriminant analysis," *IEEE Signal Processing Letters*, vol. 18, no. 5, pp. 319–322, 2011.

[22] Feiping Nie, Cheng-Long Wang, and Xuelong Li, "K-multiple-means: A multiple-means clustering method with specified k clusters," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 959–967.

[23] Xuelong Li, Mulin Chen, Feiping Nie, and Qi Wang, "Locality adaptive discriminant analysis," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI*, 2017, pp. 2201–2207.

[24] Thomas Cover and Peter Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.