Adaptive Feature-weighted Local-global Clustering

Mimi Jin, Yiyan Wang, Yong Peng, Feiping Nie, and Andrzej Cichocki, *Fellow, IEEE*

***Abstract*—Clustering has long been a fundamental problem in machine learning and data mining, with the aim of grouping data samples on the basis of their intrinsic similarity. However, the consensus that different features often exhibit varying levels of discriminative power in clustering model learning is under explored sufficiently in collaboration with the pseudo-label guided unsupervised discriminative analysis. To this end, we propose an Adaptive Feature-Weighted Local-global data Clustering (AFW-LGC) model which is featured by two improvements. First, AFW-LGC takes into account both global separability (between-cluster scatter) and local compactness (within-cluster scatter) whose impacts are mediated by a learnable parameter. Second, the different contributions of features are adaptively learned in AFW-LGC for further discriminative ability enhancement. Both improvements are seamlessly integrated for feature-weighted unsupervised discriminative subspace clustering nature of AFW-LGC. Extensive experiments on eight data sets demonstrate the superior clustering performance of AFW-LGC over some SOTA methods as well as the rationality of our proposed feature importance exploration strategy.**

***Index Terms*—Clustering, feature-weighting, local-global clustering, unsupervised discriminative analysis**

I. Introduction

In the fields of machine learning and data mining, clustering acts as an unsupervised learning technique to partition data samples into different groups based on their similarities without relying on labeled training data [1]. Clustering methods have appeared in diverse applications such as biomedical data analysis, image processing, and text mining [2]. However, the increasing data dimensionality not only obscures the essential data structure but also significantly increases the storage and computational burden; besides, more challenges are imposed on the clustering models. Over the past few decades, researchers have proposed various clustering methods including the graph-based [3], [4], hierarchical [5], density-based [6], and partitioning ones [7]. However, the clustering performance still needs to be improved in handling high-dimensional data.

To address the challenges posed by high-dimensional data, it is intuitively to project data into a lower-dimensional subspace for dimensionality reduction and discriminative ability enhancement. For example, Wang *et al.* proposed an unsupervised discriminative projection method by jointly subspace learning and informative features exploration in clustering [8]. Similarly, a one-step adaptive spectral clustering network was proposed by embedding the affinity matrix learning, spectral embedding and cluster indicator learning into a unified deep subspace framework [9]. By combining t-SNE based dimensionality reduction and adaptive density clustering, a subspace-based model was achieved in a 2D subspace without prior knowledge [10]. By adaptively learning pseudo-label to guide the structural subspace learning, a unified framework for joint unsupervised feature selection and probability graph construction was achieved in [11]. To enhance the nonlinear learning ability, an adaptive graph convolutional subspace clustering model was proposed to jointly learn a feature extraction mechanism and impose constraints on the coefficient matrix to improve clustering performance [12]. In the context of incomplete multi-view clustering, a robust subspace-based method was proposed by constructing reliable similarity graphs in low-dimensional subspaces and incorporating mixed-order information to enhance clustering performance [13].

Adaptive feature importance learning also enhances the model's discriminative ability by identifying and weighting discriminative features, thereby guiding the clustering process. Inspired by the attention mechanisms prevalent in neural networks, a feature weighting approach was integrated into the sparse fuzzy k-means clustering model, and significantly improved the clustering performance [14]. Within the semi-supervised linear square regression framework, Chen *et al.* theoretically built the underlying equivalence between a specific feature weighting vector and the $\ell_{2,1}$-norm induced feature sparsity [15]. In [16], a self-weighted orthogonal projection with sparsity-inducing regularization was introduced to effectively select discriminative features while minimizing redundancy. Based on the explicit Euler kernel mapping function, a self-weighted Euler k-means clustering model was proposed by assigning adaptive weights to different features of samples in RKHS-based representations [17]. Peng *et al.* incorporated adaptive feature importance learning into the domain of EEG-based brain-computer interfaces, aiming to identify the critical frequency bands and channels in emotion recognition [18], [19]. Sometimes, the two strategies of subspace learning and feature selection can be combined; for example, Nie *et al.* proposed using the $\ell_{2,0}$-norm to induce row sparsity of the projection matrix for subspace learning, leading to the subspace sparsity discriminative feature selection [20]. Similarly, Guo *et al.* proposed a unified subspace learning framework that incorporates adaptive view weighting to enhance discriminative representation for clustering [21].

In this paper, we propose an adaptive feature-weighted local-global data clustering (AFW-LGC) model. First, inspired by the pseudo-label guided unsupervised discriminative clustering, both global separability and local compactness in

This work was supported in part by the 'Pioneer' and 'Leading Goose' R&D Program of Zhejiang under Grant 2025C04001, and National Key Research and Development Program of China under Grant 2023YFE0114900 (*Corresponding author: Yong Peng*).

M. Jin, Y. Wang, Y. Peng are with School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (email: yongpeng@hdu.edu.cn)

F. Nie is with School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China

Andrzej Cichocki is with the Systems Research Institute of Polish Academy of Sciences, Newelska 6, Warsaw 01-447, Poland, and also with the Warsaw University of Technology, Warsaw 00-661, Poland.

data clustering are considered and a learnable parameter is introduced to balance their in-between impacts. Second, based on the consensus that different features have different discriminate abilities, we propose to adaptively explore the different contributions of features in data clustering. As a result, AFW-LGC demonstrates competitive clustering performance and rational feature importance exploration results. Besides, it is free from manually adjusted parameters.

II. The Proposed Model

A. Revisit to Subspace Clustering and Discriminative Analysis

Subspace clustering jointly completes subspace exploration and clustering, aiming to find the subspace projection matrix and cluster centroids collaboratively to achieve better clustering performance [22]. Suppose we are given a data set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where n and d are respectively the sample size and feature dimensionality. Our task is to estimate the cluster indicator matrix $\mathbf{G} = \left[\mathbf{g}^1; \mathbf{g}^2; \cdots; \mathbf{g}^n\right] \in \text{Ind}^{n\times c}$, where c represents the number of clusters. If $g_{ij} = 1$, it indicates that the i-th sample belongs to the j-th cluster. By defining the projection matrix as $\mathbf{W} \in \mathbb{R}^{d\times m}$ (i.e., m is the subspace dimensionality), the subspace cluster centroid matrix as $\mathbf{F} \in \mathbb{R}^{m\times c}$, subspace clustering aims to solve

$$\min_{\mathbf{G}\in\text{Ind},\mathbf{F},\mathbf{W}} \|\mathbf{W}^T\mathbf{X} - \mathbf{F}\mathbf{G}^T\|_F^2, s.t.\ \mathcal{C}(\mathbf{W}), \tag{1}$$

where the variables are optimized alternately.

Discriminative analysis (i.e., Fisher criterion) aims to project data into a lower-dimensional space by maximizing the separability, which is often achieved by maximizing the between-class scatter and simultaneously minimizing the within-class scatter. By using the trace ratio formula [23], discriminative projection can be achieved by maximizing the following objective function

$$\max_{\mathbf{W}} \frac{\text{Tr}(\mathbf{W}^T\mathbf{S}_b\mathbf{W})}{\text{Tr}(\mathbf{W}^T\mathbf{S}_w\mathbf{W})},\ s.t.\ \mathcal{C}(\mathbf{W}). \tag{2}$$

In both (1) and (2), $\mathcal{C}(\mathbf{W})$ represents the constraint(s) defined on $\mathbf{W}$, i.e., orthogonality ($\mathbf{W}^T\mathbf{W} = \mathbf{I}$).

B. Model Formulations

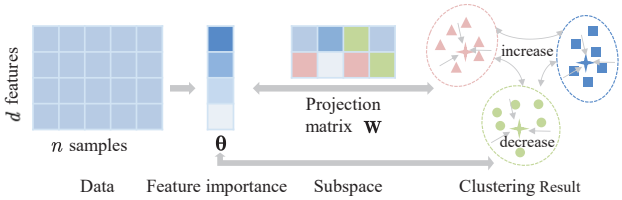

Fig. 1. The general framework of our proposed AFW-LGC model.

Fig. 1 provides the general framework of AFW-LGC, which is featured as two aspects, i.e., 1) adaptive feature importance learning to enhance the model discriminative ability, and 2) pseudo-label guided the adaptive between-cluster (global) and within-cluster (local) compatibility. The first improvement is the introduction of $\boldsymbol{\theta} = [\theta_i]_{i=1}^d$ (i.e., $\mathcal{C}(\boldsymbol{\theta}) \triangleq \boldsymbol{\theta} \geq \mathbf{0}, \boldsymbol{\theta}^T\mathbf{1} = 1$) to characterize the different contributions of different features in clustering. The second improvement is the joint minimization of the subspace clustering objective and maximization the total scatter of data in the projected subspace. Certainly, the subspace learning is mediated by the feature importance learning. Accordingly, we formulate the objective function of AFW-LGC as

$$\begin{aligned}\min_{\mathbf{W},\mathbf{G},\boldsymbol{\Theta},\mathbf{F},\lambda} &\lambda^2\|\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X} - \mathbf{F}\mathbf{G}^T\|_F^2 - \lambda\text{Tr}(\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W}) \\ &s.t.\ \mathbf{W}^T\mathbf{W} = \mathbf{I}, \mathbf{G} \in \text{Ind}^{n\times c}, \boldsymbol{\Theta} = \text{diag}(\boldsymbol{\theta}), \mathcal{C}(\boldsymbol{\theta}),\end{aligned} \tag{3}$$

where the first term takes the form of decomposing subspace-represented data into the product of centroid and cluster indicator matrices to achieve subspace clustering with explicit feature importance consideration, thereby promoting within-cluster compactness. The second term encourages the preserving of the total scatter $\tilde{\mathbf{S}}_t$ of data in the subspace, which helps retain the global data structure. The parameter λ is treated as a variable that can be adaptively optimized without manual adjustment to balance the contributions of both terms.

Below we show how (3) achieves discriminative subspace exploration, under the mediation of feature importance. By taking the derivative w.r.t. $\mathbf{F}$ and setting it to zero, we get

$$\mathbf{F} = \mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}. \tag{4}$$

Next, by substituting (4) into the first term of (3) and defining $\tilde{\mathbf{G}} = \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T$, we have

$$\begin{aligned}&\|\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X} - \mathbf{F}\mathbf{G}^T\|_F^2 \\ &= \text{Tr}(\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W}\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}) - 2\text{Tr}(\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W}\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}\tilde{\mathbf{G}}) \\ &+ \text{Tr}(\mathbf{G}^T\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W}\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}\mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1}) \\ &= \text{Tr}(\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W}) - \text{Tr}(\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}\tilde{\mathbf{G}}\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W}) \\ &= \text{Tr}(\mathbf{W}^T(\boldsymbol{\Theta}\mathbf{X}\mathbf{X}^T\boldsymbol{\Theta} - \boldsymbol{\Theta}\mathbf{X}\tilde{\mathbf{G}}\mathbf{X}^T\boldsymbol{\Theta})\mathbf{W}) \\ &= \text{Tr}(\mathbf{W}^T\tilde{\mathbf{S}}_w\mathbf{W}).\end{aligned} \tag{5}$$

Here, $\tilde{\mathbf{S}}_b = \boldsymbol{\Theta}\mathbf{X}\tilde{\mathbf{G}}\mathbf{X}^T\boldsymbol{\Theta}$ and $\tilde{\mathbf{S}}_t = \boldsymbol{\Theta}\mathbf{X}\mathbf{X}^T\boldsymbol{\Theta}$. Since $\tilde{\mathbf{S}}_t = \tilde{\mathbf{S}}_w + \tilde{\mathbf{S}}_b$, minimizing the first term is equivalent to minimizing $\tilde{\mathbf{S}}_w$. Then, the second term in (3) is equivalent to

$$\begin{aligned}&\min\ -\text{Tr}(\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}\mathbf{X}^T\boldsymbol{\Theta}^T\mathbf{W}) \\ \Leftrightarrow &\max\ \text{Tr}(\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}\mathbf{X}^T\boldsymbol{\Theta}^T\mathbf{W}) = \max\ \text{Tr}(\mathbf{W}^T\tilde{\mathbf{S}}_t\mathbf{W}).\end{aligned} \tag{6}$$

Finally, by substituting (5) and (6) into (3), we obtain

$$\begin{aligned}&\min\ \lambda^2\|\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X} - \mathbf{F}\mathbf{G}^T\|_F^2 - \lambda\text{Tr}(\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W}) \\ &= \min\ \lambda^2\text{Tr}(\mathbf{W}^T\tilde{\mathbf{S}}_w\mathbf{W}) - \lambda\text{Tr}(\mathbf{W}^T(\tilde{\mathbf{S}}_w + \tilde{\mathbf{S}}_b)\mathbf{W}) \\ &= \min\ \text{Tr}(\mathbf{W}^T((\lambda^2 - \lambda)\tilde{\mathbf{S}}_w - \lambda\tilde{\mathbf{S}}_b)\mathbf{W}).\end{aligned} \tag{7}$$

As shown in (7), AFW-LGC model essentially performs pseudo-label guided discriminative subspace learning; however, the merits of AFW-LGC lie in two aspects; one is that the discriminative analysis is achieved in the adaptive feature-weighted space, and the other is that the relative importance between within-class and between-class data scatter is adaptively mediated (i.e., by λ).

C. Model Optimization

Below we solve the four variables in (3) in an alternate way.

■ Update $\mathbf{W}$. Considering the constraint, we have

$$\begin{aligned}\min_{\mathbf{W}}\ &\lambda^2\|\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X} - \mathbf{F}\mathbf{G}^T\|_F^2 - \lambda\text{Tr}(\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W}) \\ &- \text{Tr}(\boldsymbol{\Lambda}(\mathbf{W}^T\mathbf{W} - \mathbf{I})).\end{aligned} \tag{8}$$

After taking the derivative of (8) with respect to $\mathbf{W}$ and setting it to zero, we obtain

$$\mathbf{W} = \min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \mathbf{W}^T(\mathbf{\Theta X}(\mathbf{A}-\mathbf{B})\mathbf{X}^T\mathbf{\Theta})\mathbf{W}, \tag{9}$$

from which we know that the columns of $\mathbf{W}$ are corresponding to the eigenvectors in terms of the first m smallest eigenvalues of $\mathbf{\Theta X}(\mathbf{A}-\mathbf{B})\mathbf{X}^T\mathbf{\Theta}$. Here $\mathbf{A} = (\lambda^2 - \lambda)\mathbf{I}$, and $\mathbf{B} = \lambda^2\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T$.

■ Update $\mathbf{G}$. Considering that each row of matrix $\mathbf{G}$ is independent, and there is only one element equal to 1 in each row with all the other elements being 0, we update $\mathbf{G}$ for each row independently as

$$g_{ij} = \begin{cases} 1, & j = \arg\min_k \|(\mathbf{W}^T\mathbf{\Theta X})_i - \mathbf{F}_k\|_2^2 \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

■ Update Θ. We transform the objective function into

$$\min_{\boldsymbol{\theta}} \ \boldsymbol{\theta}^T\mathbf{D}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{c}, s.t.\ \boldsymbol{\theta} \geq \mathbf{0}, \mathbf{1}^T\boldsymbol{\theta} = 1, \tag{11}$$

where $\mathbf{D} = ((1-\lambda)\mathbf{X}\mathbf{X}^T) \circ (\mathbf{W}\mathbf{W}^T)$ and $\mathbf{c} = diag(\mathbf{W}\mathbf{F}\mathbf{G}^T\mathbf{X}^T)$. Subsequently, we adopt the quadratic programming with simplex constraint approach to solve (11) [18].

■ Update $\mathbf{F}$. By taking the derivative of (3) w.r.t. $\mathbf{F}$ and setting it to zero, we have the same solution as in (4).

■ Update λ. Variable λ in (3) serves as a mediation parameter between the within-cluster and between-cluster scatters. Taking the derivative of (3) w.r.t λ and set it to zero, we have

$$\lambda = \frac{\mathrm{Tr}(\mathbf{W}^T\mathbf{\Theta}\mathbf{X}\mathbf{X}^T\mathbf{\Theta}\mathbf{W})}{2\|\mathbf{W}^T\mathbf{\Theta X} - \mathbf{F}\mathbf{G}^T\|_F^2}. \tag{12}$$

This strategy has also been employed in the research [24].

The above updating rules are summarized in Algorithm 1, whose total time complexity is $\mathcal{O}(tdn^2)$ by considering the general case that $n > d > m \gg c$.

Algorithm 1 Optimization method to AFW-LGC objective (3)

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, number of clusters c, parameter λ, and the subspace dimensionality m;

Output: Cluster indicator matrix $\mathbf{G} \in \mathrm{Ind}^{n\times c}$, and the feature importance descriptor $\boldsymbol{\theta}$.

1: Initialize λ=2, randomly initialize $\mathbf{G}$ to satisfy the requirements to be a cluster indicator matrix;
2: **while** not converged **do**
3: Update $\mathbf{W}$ by equation (9);
4: Update $\mathbf{G}$ by equation (10);
5: Update $\mathbf{\Theta}$ by solving equation (11);
6: Update $\mathbf{F}$ by solving equation (4);
7: Update parameter λ by equation (12);
8: **end while**

III. Experiments

A. Experiments on Benchmark Data Sets

1) Data Sets and Experimental Setup: Eight popular data sets were used in the experiments, including the dig, wine, dermatology, CBCL, USPS, Yale, Palm and UMIST. Table I summarizes their main properties. To evaluate the clustering performance of our AFW-LGC method, we selected five clustering models including k-means, AWFKM [14], PCIP [25], SWULDA [24] and SWCAN [26] in the comparative studies. For fair comparison, model parameters were set as per the original papers.

2) Results and Analysis: Table II shows the clustering results of different models on the eight data sets in terms of the seven metrics that are commonly used for clustering performance evaluation, *i.e.*, ACC, NMI, Purity, ARI, Precision, Recall and F-score. The best results for each model on each metric are highlighted in bold, and most of these bolded results were achieved by our proposed AFW-LGC model. The following conclusions are drawn from these results.

TABLE I
BASIC PROPERTIES OF THE DATA SETS IN THE EXPERIMENTS.

Dataset	#sample	#feat.	#clust.	Dataset	#sample	#feat.	#clust.
dig	1797	64	10	USPS	2007	256	10
wine	178	13	3	Yale	165	1024	15
dermatology	366	34	6	Palm	2000	256	100
CBCL	2000	361	2	UMIST	575	1024	20

- AFW-LGC consistently outperforms k-means, AWFKM, PCIP, and SWULDA across all the data sets. AWFKM considers the feature importance in fuzzy clustering, which neglects fully considering discriminative information. In PCIP, outliers are filtered by a pre-defined abnormal detection algorithm, whose connection with the subsequent clustering process is separated. Besides, feature importance is not explored in PCIP. Compared with SWULDA, the superiority of AFW-LGC demonstrates the necessity of learning feature importance in further improving the model discriminative ability.
- Although AFW-LGC performs slightly worse than SWCAN in terms of very few metrics (i.e., NMI and Purity) on USPS, the difference is minor. Furthermore, our method does not require calculating a similarity matrix, thus avoiding the potential negative impact that this process might have on performance. Generally, graph-based methods tend to outperform non-graph-based methods because they utilize more prior information [24]. Although AFW-LGC is a non-graph-based approach, it still achieved competitive clustering performance.

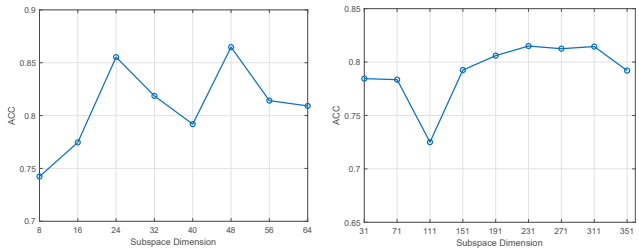

Fig. 2. Performance vs. subspace dimensions on dig (left) and CBCL (right).

3) Parameters Analysis: There is no explicit parameters that need to be manually adjusted in the AFW-LGC objective function (3). For the parameter λ, we only need to initialize it (i.e., λ=2), which is then iteratively updated during the optimization process. On sensitivity to the subspace dimensionality, in Fig. 2, we present the variation of ACC on dig

TABLE II
CLUSTERING PERFORMANCE OF DIFFERENT METHODS ON THE EIGHT BENCHMARK DATA SETS

Metrics	ACC	NMI	Purity	ARI	Precision	Recall	Fscore	ACC	NMI	Purity	ARI	Precision	Recall	Fscore
	dig							wine						
k-means	0.6989	0.7060	0.7284	0.3039	0.4796	0.3590	0.6897	0.9494	0.8447	0.9494	0.7987	0.9342	0.9385	0.9359
AWFKM	0.5242	0.4624	0.5292	0.2990	0.4771	0.4041	0.3347	0.9213	0.7559	0.9213	0.7597	0.9170	0.9296	0.9163
PCIP	0.8141	0.7218	0.8141	0.6615	0.7044	0.6935	0.6904	0.8146	0.5320	0.8146	0.4520	0.8360	0.7965	0.7836
SWULDA	0.7935	0.7390	0.7935	0.6254	0.6937	0.6623	0.6653	0.9663	0.8746	0.9663	0.8498	0.9474	0.9577	0.9500
SWCAN	0.8431	0.7612	0.8431	0.7021	0.6928	0.7374	0.6897	0.9438	0.8176	0.9438	0.8819	0.9582	0.9671	0.9611
ours	**0.8620**	**0.7752**	**0.8620**	**0.7298**	**0.7742**	**0.7640**	**0.7649**	**0.9719**	**0.8896**	**0.9719**	**0.9134**	**0.9705**	**0.9765**	**0.9728**
	dermatology							CBCL						
k-means	0.8661	0.8415	0.8661	0.8182	0.8838	0.8703	0.8536	0.5900	0.0256	0.5900	0.0319	0.5942	0.5900	0.5854
AWFKM	0.8852	0.8412	0.8852	0.8176	0.8308	0.8009	0.7318	0.5800	0.0194	0.5914	0.0428	0.6042	0.6040	0.6038
PCIP	0.9617	0.9228	0.9617	0.9350	0.9617	0.9648	0.9620	0.6595	0.1212	0.6595	0.2577	0.7570	0.7540	0.7533
SWULDA	0.9672	0.9374	0.9682	0.8473	0.8958	0.8771	0.8606	0.7985	0.2794	0.7985	0.3561	0.8017	0.7985	0.7980
SWCAN	0.9563	0.9175	0.9563	0.9309	0.9583	0.9614	0.9589	0.5955	0.0269	0.5955	0.0352	0.5964	0.5945	0.5925
ours	**0.9754**	**0.9401**	**0.9754**	**0.9523**	**0.9727**	**0.9751**	**0.9738**	**0.8205**	**0.3265**	**0.8205**	**0.4106**	**0.8239**	**0.8205**	**0.8200**
	USPS							Yale						
k-means	0.4614	0.4660	0.5595	0.4503	0.5560	0.5863	0.5414	0.4303	0.4913	0.4303	0.1952	0.3681	0.3636	0.3459
AWFKM	0.4978	0.3382	0.5072	0.2039	0.4383	0.3467	0.3141	0.5122	0.5376	0.5122	0.2381	0.4309	0.4000	0.3562
PCIP	0.5745	0.4674	0.5774	0.4225	0.6130	0.5659	0.5718	0.3273	0.3731	0.3273	0.0922	0.3256	0.3152	0.3038
SWULDA	0.5750	0.5529	0.6557	0.4395	0.5771	0.5738	0.5528	0.5152	0.5529	0.5273	0.1871	0.4154	0.4000	0.3938
SWCAN	0.6079	**0.6360**	**0.7265**	0.5239	0.6281	0.6209	0.5980	0.5152	0.5642	0.5333	0.3280	0.5550	0.4970	0.4906
ours	**0.6507**	0.5636	0.7090	**0.5725**	**0.6491**	**0.6394**	**0.6320**	**0.5455**	**0.5671**	**0.5455**	**0.3311**	**0.5583**	**0.5455**	**0.5370**
	Palm							Umist						
k-means	0.7135	0.8989	0.7370	0.6432	0.7206	0.6985	0.6841	0.4313	0.6285	0.4957	0.3046	0.4168	0.3771	0.3670
AWFKM	0.7025	0.9079	0.7770	0.6544	0.6936	0.6850	0.6580	0.4209	0.6305	0.5078	0.2958	0.3213	0.3445	0.3068
PCIP	0.6570	0.8788	0.6570	0.5884	0.6814	0.6475	0.6279	0.4122	0.5428	0.4574	0.2457	0.4994	0.4069	0.4112
SWULDA	0.7255	0.9013	0.7530	0.5321	0.6377	0.6690	0.6240	0.4957	0.6661	0.5461	0.2814	0.3999	0.3603	0.3598
SWCAN	0.8105	0.9458	0.8550	0.7606	0.8324	0.7955	0.7860	0.5217	0.6837	0.5461	0.4007	0.5137	0.4735	0.4660
ours	**0.8240**	**0.9257**	**0.8455**	**0.7717**	**0.8357**	**0.8240**	**0.8146**	**0.5322**	**0.7558**	**0.6139**	**0.5053**	**0.5364**	**0.5112**	**0.4854**

and CBCL under different subspace dimensions. The subspace dimension m is selected by manually varying it from 10% to 100% of the original dimension d, with a fixed step size. When recording the results, we selected the values corresponding to the optimal subspace dimension for each data set.

4) Convergence: In Fig. 3, we present the convergence curves on two data sets, and the changes of the parameter λ during the iterations. It is observed that our method converges rapidly within dozens of iterations, and the value of λ also stabilizes gradually as the iteration increases. This confirms the desirable convergence property of AFW-LGC.

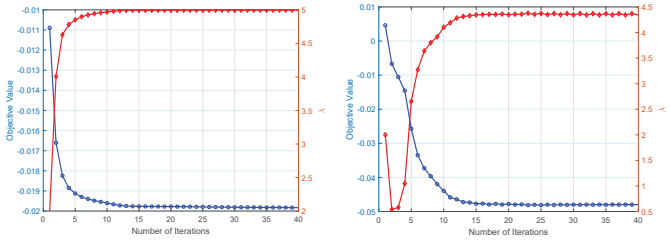

Fig. 3. Objective function value vs. λ on dig (left) and dermatology (right).

B. Rationality Evaluation of Feature Importance

We selected two face data sets (i.e., MSRA and CMU PIE) to demonstrate the effectiveness of AFW-LGC in feature importance exploration. MSRA consists of 768 samples from 12 clusters with 256 features, while CMUPIE comprises 2856 samples with 1024 features and is partitioned into 68 clusters. We first reshape the original facial features into an image representation, and then sort and visualize the learned feature importance. For MSRA, we present images with the top 10, 30, and 50 selected features, while for CMUPIE, we present images with the top 30, 50, and 100 selected features. Fig. 4 shows that the highlighted pixels are mainly concentrated around the eyes, nose, and mouth, indicating that these areas have higher feature weights and are crucial for describing facial contours. Therefore, we conclude that our method is capable of assigning higher weights to more discriminative features, indicating its rationality.

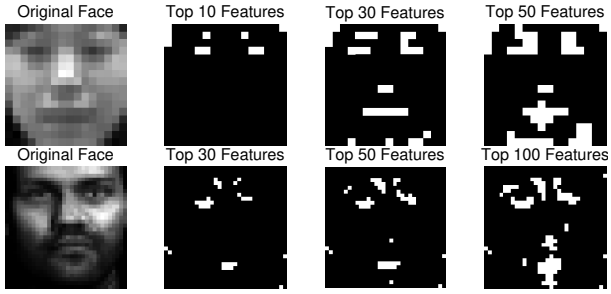

Fig. 4. Visualization of pixel value in MSRA (top) and CMUPIE (bottom).

IV. CONCLUSION

In this paper, we proposed an adaptive feature-weighted local-global clustering model termed AFW-LGC, in which on the one hand the discriminative subspace was identified by taking both maximizing between-cluster (global) and minimizing within-cluster (local) data scatters into consideration with learnable compatibility, and on the other hand, feature importance was adaptively explored to characterize the different contributions of features in clustering. Extensive experimental results demonstrated the effectiveness of AFW-LGC compared to other clustering models. Besides, the rationality of the explored feature importance was visually evaluated. As our future work, we will investigate the adaptive selection of λ on data sets with specific properties such as imbalance and inter-cluster overlap.

REFERENCES

[1] G. Chao, S. Sun, and J. Bi, "A survey on multiview clustering," *IEEE Trans Artif. Intell.*, vol. 2, no. 2, pp. 146–168, 2021.

[2] H. Yin, A. Aryani, S. Petrie, A. Nambissan, A. Astudillo, and S. Cao, "A rapid review of clustering algorithms," *arXiv preprint arXiv:2401.07389*, 2024.

[3] A. Tsitsulin, J. Palowitch, B. Perozzi, and E. Müller, "Graph clustering with graph neural networks," *J. Mach. Learn. Res.*, vol. 24, no. 127, pp. 1–21, 2023.

[4] Y. Peng, W. Huang, W. Kong, F. Nie, and B.-L. Lu, "JGSED: An end-to-end spectral clustering model for joint graph construction, spectral embedding and discretization," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 7, no. 6, pp. 1687–1701, 2023.

[5] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8219–8264, 2023.

[6] Y. Wang, J. Qian, M. Hassan, X. Zhang, T. Zhang, C. Yang, X. Zhou, and F. Jia, "Density peak clustering algorithms: A review on the decade 2014–2023," *Expert Syst. Appl.*, vol. 238, p. 121860, 2024.

[7] J. Wang, Z. Li, C. Tang, S. Liu, X. Wan, and X. Liu, "Multiple kernel clustering with adaptive multi-scale partition selection," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 16, pp. 6641–6652, 2024.

[8] R. Wang, J. Bian, F. Nie, and X. Li, "Unsupervised discriminative projection for feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 2, pp. 942–953, 2022.

[9] F. Zhang, J. Zhao, X. Ye, and H. Chen, "One-step adaptive spectral clustering networks," *IEEE Signal Process. Lett.*, vol. 29, pp. 2263–2267, 2022.

[10] P. Lang, X. Fu, Z. Cui, C. Feng, and J. Chang, "Subspace decomposition based adaptive density peak clustering for radar signals sorting," *IEEE Signal Process. Lett.*, vol. 29, pp. 424–428, 2021.

[11] Z. Wang, Y. Yuan, R. Wang, F. Nie, Q. Huang, and X. Li, "Pseudo-label guided structural discriminative subspace learning for unsupervised feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 12, pp. 18 605–18 619, 2024.

[12] L. Wei, Z. Chen, J. Yin, C. Zhu, R. Zhou, and J. Liu, "Adaptive graph convolutional subspace clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6262–6271.

[13] W. Guo, H. Che, M.-F. Leung, L. Jin, and S. Wen, "Robust mixed-order graph learning for incomplete multi-view clustering," *Inf. Fusion*, vol. 115, p. 102776, 2024.

[14] F. Nie, W. Chang, X. Li, J. Xu, and G. Li, "Adaptive feature weight learning for robust clustering problem with sparse constraint," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 3125–3129.

[15] X. Chen, G. Yuan, F. Nie, and J. Z. Huang, "Semi-supervised feature selection via rescaled linear regression." in *Proc. Int. J. Conf. Artif. Intell.*, 2017, pp. 1525–1531.

[16] R. Zhang, F. Nie, and X. Li, "Self-weighted supervised discriminative feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3913–3918, 2018.

[17] H. Xin, Y. Lu, H. Tang, R. Wang, and F. Nie, "Self-weighted euler k-means clustering," *IEEE Signal Process. Lett.*, vol. 30, pp. 1127–1131, 2023.

[18] Y. Peng, F. Qin, W. Kong, Y. Ge, F. Nie, and A. Cichocki, "GFIL: A unified framework for the importance analysis of features, frequency bands, and channels in EEG-based emotion recognition," *IEEE Trans. Cogn. Develop. Syst.*, vol. 14, no. 3, pp. 935–947, 2022.

[19] X. Li, F. Shen, Y. Peng, W. Kong, and B.-L. Lu, "Efficient sample and feature importance mining in semi-supervised EEG emotion recognition," *IEEE Trans. Circuits Syst. II-Express Briefs*, vol. 69, no. 7, pp. 3349–3353, 2022.

[20] F. Nie, Z. Wang, L. Tian, R. Wang, and X. Li, "Subspace sparse discriminative feature selection," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 4221–4233, 2022.

[21] W. Guo, H. Che, M.-F. Leung, and Z. Yan, "Adaptive multi-view subspace learning based on distributed optimization," *Internet Things*, vol. 26, p. 101203, 2024.

[22] M. E. Timmerman, E. Ceulemans, K. De Roover, and K. Van Leeuwen, "Subspace k-means clustering," *Behav. Res. Methods*, vol. 45, pp. 1011–1023, 2013.

[23] Y. Jia, F. Nie, and C. Zhang, "Trace ratio problem revisited," *IEEE Trans. Neural Netw.*, vol. 20, no. 4, pp. 729–735, 2009.

[24] X. Li, Y. Zhang, and R. Zhang, "Self-weighted unsupervised LDA," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 3, pp. 1627–1632, 2023.

[25] J. Wang, Y. Wu, X. Huang, C. Zhang, and F. Nie, "Projected fuzzy c-means clustering algorithm with instance penalty," *Expert Syst. Appl.*, p. 124563, 2024.

[26] F. Nie, D. Wu, R. Wang, and X. Li, "Self-weighted clustering with adaptive neighbors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3428–3441, 2020.