

Sharper Error Bounds in Late Fusion Multi-view Clustering with Eigenvalue Proportion Optimization

Liang Du¹, Henghui Jiang¹, Xiaodong Li¹, Yiqing Guo¹, Yan Chen², Feijiang Li¹, Peng Zhou³, Yuhua Qian^{4*}

¹School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, China

²College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, 030600, China

³School of Computer Science and Technology, Anhui University, Hefei, 230601, China

⁴Institute of Big Data Science and Industry, Shanxi University, Taiyuan, 030006, China
{duliang,fjli,jinchengqyh}@sxu.edu.cn,yanchen01@tyut.edu.cn,zhoupeng@ahu.edu.cn

Abstract

Multi-view clustering (MVC) aims to integrate complementary information from multiple views to enhance clustering performance. Late Fusion Multi-View Clustering (LFMVC) has shown promise by synthesizing diverse clustering results into a unified consensus. However, current LFMVC methods struggle with noisy and redundant partitions and often fail to capture high-order correlations across views. To address these limitations, we present a novel theoretical framework for analyzing the generalization error bounds of multiple kernel k -means, leveraging local Rademacher complexity and principal eigenvalue proportions. Our analysis establishes a convergence rate of $\mathcal{O}(1/n)$, significantly improving upon the existing rate in the order of $\mathcal{O}(\sqrt{k/n})$. Building on this insight, we propose a low-pass graph filtering strategy within a multiple linear K-means framework to mitigate noise and redundancy, further refining the principal eigenvalue proportion and enhancing clustering accuracy. Experimental results on benchmark datasets confirm that our approach outperforms state-of-the-art methods in clustering performance and robustness.

Introduction

Multi-view data is pervasive across numerous real-world scenarios, including image processing, bioinformatics, and social network analysis (Zhou, Du, and Li 2023a; Tang et al. 2022; Wang et al. 2021; Liu et al. 2022). By integrating complementary information from multiple views, state-of-the-art multi-view clustering (MVC) techniques have significantly advanced clustering performance (Zhou and Du 2023; Tang et al. 2020; Zhou et al. 2024; Li et al. 2021). Based on their fusion strategies, MVC methods can be broadly categorized into three paradigms: early fusion-based methods (Liu et al. 2023; Nie, Tian, and Li 2018), kernel fusion-based approaches (Liu 2023; Liu et al. 2020), and late fusion-based multi-view clustering (LFMVC) methods (Liu et al. 2021; Wang et al. 2023; Zhou et al. 2023).

LFMVC follows a two-stage process: first, it independently generates base clustering partitions for each view using methods such as kernel K-means; second, it synthesizes

these base partitions into a unified consensus result (Wang et al. 2019). This paradigm has gained considerable attention for its adaptability and effectiveness in handling diverse multi-view datasets. Over the years, several notable LFMVC methods have advanced the state of the art. For instance, OPLFMVC (Liu et al. 2021) integrates consensus partition learning and label generation into a unified optimization framework. ALMVC (Zhang et al. 2022) introduces orthogonally learned view-specific anchors to enhance clustering performance. MMLMVC (Li et al. 2024a) employs the min-max optimization framework from SimpleMKKM (Liu 2023) to refine late fusion strategies. sLGM (Wu et al. 2024) leverages a Grassmann manifold for partition fusion, preserving the topological structure in high-dimensional spaces. Lastly, RIWLF (Zhang et al. 2024) incorporates sample weights as prior knowledge to regularize the late fusion process, improving robustness and adaptability.

Despite their success, existing LFMVC approaches encounter persistent challenges. One major issue is the propagation of noise and redundancy in base clustering partitions, which compounds errors and affects the quality of the synthesized consensus. Another limitation is their inability to capture high-order correlations among samples and views, which is critical for fully leveraging the complementary information in multi-view data. These shortcomings restrict the effectiveness of LFMVC methods, particularly in complex scenarios requiring robust performance.

To overcome these limitations, we propose a novel theoretical and algorithmic framework to enhance late fusion-based multi-view clustering. Our contributions are twofold. First, we provide a novel generalization error bound analysis for multiple kernel K-means based on local Rademacher complexity and the principal eigenvalue proportion of the underlying data structure. Unlike previous works that achieve a convergence rate of $\mathcal{O}(\sqrt{k/n})$, our method reaches $\mathcal{O}(1/n)$, representing a significant improvement. Second, inspired by the theoretical analysis and recent advances in graph filtering, we develop a low-pass filtering strategy to refine clustering, aiming to improve the principal eigenvalue proportion, under the multiple linear K-means for LFMVC. We validate the effectiveness of our approach on benchmark datasets, demonstrating improvements over

*Corresponding author

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

state-of-the-art LFMVC methods in terms of clustering results. The main contributions of this paper are summarized as follows:

- We establish a novel generalization error bound for multiple kernel K-means based on local Rademacher complexity and principal eigenvalue proportion, achieving an unprecedented convergence rate of $\mathcal{O}(1/n)$ compared to the conventional $\mathcal{O}(\sqrt{k/n})$.
- Leveraging insights from our theoretical analysis, we propose a low-pass graph filtering-enhanced strategy within the multiple linear K-means framework, aiming to sharpen the clustering error bound by improving the principal eigenvalue proportion.

Related Work

In this section, we discuss two closely related works in LFMVC that provides generalization error analysis.

One Pass Late Fusion

Recognizing that the learning of the consensus partition matrix and the assignment of cluster labels are typically conducted as separate processes, which do not adequately inform each other, thus potentially undermining clustering performance, (Liu et al. 2021) addresses this issue by proposing an approach that seamlessly integrates both tasks into a unified optimization framework. The resulting optimization problem can be formulated as follows:

$$\begin{aligned} \max \quad & \text{tr} \left(\mathbf{C}^T \mathbf{Y}^T \sum_{p=1}^m \gamma_p \mathbf{H}_p \mathbf{W}_p \right) \\ \text{s.t.} \quad & \mathbf{C}^T \mathbf{C} = \mathbf{I}_c, \mathbf{W}_p^T \mathbf{W}_p = \mathbf{I}_c, \forall p, \\ & \mathbf{Y} \in \{0, 1\}^{n \times k}, \sum_{p=1}^m \gamma_p^2 = 1, \gamma_p \geq 0, \end{aligned} \quad (1)$$

where the objective denotes the alignment between the consensus partition matrix \mathbf{Y} and a group of pre-calculated base partition matrices $\{\mathbf{H}_p\}_{p=1}^m$, and \mathbf{W}_p is the p -th transformation matrix. It has been pointed out that the clustering generalization error bound based on the Rademacher complexity is of order $\mathcal{O}(k/\sqrt{n})$ on unseen samples.

Max-Min-Max Late Fusion

Motivated by the superior performance of kernel fusion-based methods, such as SimpleMKKM (Liu 2023), and the low computational complexity of late fusion-based approaches, (Li et al. 2024a) propose integrating the kernel fusion-based min-max learning paradigm into late fusion MVC, which can be achieved through a max-min-max optimization framework as follows,

$$\begin{aligned} \max_{\{\mathbf{W}_p\}_{p=1}^m} \min_{\gamma} \max_{\mathbf{H}} \quad & \text{tr} \left(\mathbf{H}^T \left(\sum_{p=1}^m \gamma_p^2 \mathbf{H}_p \mathbf{W}_p \right) \right) \\ \text{s.t.} \quad & \mathbf{H}^T \mathbf{H} = \mathbf{I}_c, \sum_{p=1}^m \gamma_p = 1, \gamma_p \geq 0, \mathbf{W}_p^T \mathbf{W}_p = \mathbf{I}_c, \forall p, \end{aligned} \quad (2)$$

which maximizes the perturbation matrices $\{\mathbf{W}_p\}_{p=1}^m$, minimizes view coefficients γ and maximizes the clustering partition matrix \mathbf{H} . The generalization clustering performance is upper bounded with the order $\mathcal{O}(\sqrt{k/n})$.

Notations and Preliminaries

Let \mathbb{P} be an unknown distribution on \mathcal{X} , and let $\mathbb{S} = \{\mathbf{x}_i\}_{i=1}^n \subseteq \mathcal{X}$ be a set of n i.i.d. samples drawn from \mathbb{P} . The empirical distribution \mathbb{P}_n is defined as $\mathbb{P}_n(\{\mathbf{x}_i\}) = \frac{1}{n}$ for each $\mathbf{x}_i \in \mathbb{S}$, and $\mathbb{P}_n(\mathbf{x}) = 0$ for $\mathbf{x} \notin \mathbb{S}$. Given a kernel function $K: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, the corresponding kernel matrix is defined as $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^n$. Let $\lambda_1(\mathbf{K}) \geq \lambda_2(\mathbf{K}) \geq \dots \geq \lambda_n(\mathbf{K}) \geq 0$ denote its eigenvalues. The Reproducing Kernel Hilbert Space (RKHS) \mathcal{H} is defined as the completion of the span of $\{K(\mathbf{x}, \cdot) : \mathbf{x} \in \mathcal{X}\}$.

In multiple kernel K-means, a common choice is the convex combination of m basic kernels:

$$\mathcal{K}^{\text{multiple}} = \left\{ \sum_{p=1}^m \gamma_p^2 \mathbf{K}_p \mid \sum_{p=1}^m \gamma_p = 1, \gamma_p \geq 0 \right\}, \quad (3)$$

where \mathbf{K}_p are the basic kernels.

Multiple Kernel K-Means for Late Fusion

In this section, we present to derive consensus clustering by integrating multiple embeddings $\{\mathbf{H}_p\}_{p=1}^m$ within a simplified linear case of the Multiple Kernel KMeans framework. The K-means clustering algorithm applied to the p -th partition $\mathbf{H}_p \in \mathbb{R}^{n \times r_p}$ can be formulated as follows:

$$\min_{\mathbf{Y} \in \text{Ind}} \text{tr}(\mathbf{H}_p \mathbf{H}_p^T) - \text{tr}(\mathbf{Y}(\mathbf{H}_p \mathbf{H}_p^T) \mathbf{Y}(\mathbf{Y}^T \mathbf{Y})^{-1}). \quad (4)$$

It is well established that K-means can be more effectively conducted in a kernel space, where data points are more easily separable. Given a set of orthogonal embeddings $\{\mathbf{H}_p\}_{p=1}^m$, we can construct a corresponding set of linear kernels $\{\mathbf{K}_p = \mathbf{H}_p \mathbf{H}_p^T\}_{p=1}^m$. For the LFMVC task, we propose to learn the consensus clustering using a consensus kernel \mathbf{K}_γ , which is obtained through a weighted aggregation of the individual kernels: $\mathbf{K}_\gamma = \sum_{p=1}^m \gamma_p^2 \mathbf{K}_p$. The MKKM-based LFMVC can then be formulated as follows:

$$\begin{aligned} \min_{\mathbf{Y}, \gamma} \quad & \text{tr} \left(\sum_{p=1}^m \gamma_p^2 \mathbf{H}_p \mathbf{H}_p^T \right) \\ & - \text{tr} \left(\mathbf{Y} \left(\sum_{p=1}^m \gamma_p^2 \mathbf{H}_p \mathbf{H}_p^T \right) \mathbf{Y}(\mathbf{Y}^T \mathbf{Y})^{-1} \right) \\ \text{s.t.} \quad & \mathbf{Y} \in \text{Ind}, \sum_{p=1}^m \gamma_p = 1, \gamma_p \geq 0, \forall i. \end{aligned} \quad (5)$$

Although applying the MKKM framework from Eq. (5) to the task of LFMVC may seem straightforward, several key insights deserve emphasis. First, MKKM demonstrates significant efficiency by operating in linear time when using linear kernels derived from candidate embeddings. This characteristic aligns with the inherent advantages of late

fusion-based approaches, making MKKM particularly suitable for large-scale datasets. Second, by employing the local Rademacher complexity technique, we establish that the generalization error for MKKM is of order $\mathcal{O}(1/n)$, which provides a much sharper bound than those achieved by previously discussed LFMVC methods. Furthermore, the generalization error bound can be further improved by enhancing the proportion of the principal eigenvalue.

Generalization Error Analysis of MKKM

The generalization error of the K-means algorithm is defined as the expected distance between unseen data and their corresponding cluster centers. To formalize this, we consider a class of functions to construct our hypothesis space:

$$\mathcal{L} = \left\{ \ell : \mathbf{x} \mapsto \min_{\mathbf{y} \in \{\mathbf{e}_1, \dots, \mathbf{e}_k\}} \|\Phi_\gamma(\mathbf{x}) - \mathbf{C}\mathbf{y}\|_{\mathcal{H}}^2 \mid \|\Phi_\gamma(\mathbf{x}_i)^T \Phi_\gamma(\mathbf{x}_j)\| \leq b, \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}, \mathbf{C} \in \mathcal{H}^k \right\}, \quad (6)$$

where $\Phi_\gamma(\mathbf{x}) = [\gamma_1 \Phi_1(\mathbf{x})^T, \dots, \gamma_m \Phi_m(\mathbf{x})^T]^T : \mathbb{R}^d \mapsto \mathcal{H}$. Suppose \mathbf{c}_j is the j -th clustering centroid and $f_{\mathbf{c}} = (f_{c_1}, \dots, f_{c_k})$, where $j \in \{1, \dots, k\}$. It follows that the mapping function f_j is bounded within $[0, 4b]$.

$$\begin{aligned} f_{c_j}(\mathbf{x}) &= \|\Phi_\gamma(\mathbf{x}) - \mathbf{c}_j\|_{\mathcal{H}}^2 = \left\| \Phi_\gamma(\mathbf{x}) - \frac{1}{|\mathbf{c}_j|} \sum_{i \in \mathbf{c}_j} \Phi_\gamma(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \\ &\leq 2\|\Phi_\gamma(\mathbf{x})\|^2 + \frac{2}{|\mathbf{c}_j|^2} \sum_{i_1, i_2 \in \mathbf{c}_j} \langle \Phi_\gamma(\mathbf{x}_{i_1}), \Phi_\gamma(\mathbf{x}_{i_2}) \rangle \\ &\leq 4b. \end{aligned} \quad (7)$$

The generalization error $R(f)$ is defined as

$$R(f) := \mathbb{E}_{\mathbf{x} \sim \mathbb{P}}[\ell_f(\mathbf{x})], \quad (8)$$

where ℓ_f is the loss function. Since \mathbb{P} is unknown, we estimate the empirical error using the observed data:

$$\hat{R}(f) := \frac{1}{n} \sum_{i=1}^n \ell_f(\mathbf{x}_i). \quad (9)$$

Lemma 1. *The loss function $\ell(f_{\mathbf{c}}(\mathbf{x})) = \min f_{\mathbf{c}}(\mathbf{x})$ is 1-Lipschitz with respect to the L_∞ norm, satisfying:*

$$\begin{aligned} \ell^2 &= \min(f_1, f_2, \dots, f_k)^2 \\ &\leq \max(f_1, f_2, \dots, f_k) \min(f_1, f_2, \dots, f_k) \\ &\leq 4b \min(f_1, f_2, \dots, f_k) = 4b\ell. \end{aligned} \quad (10)$$

Proof 1. Assuming that $\forall f, g \in \mathcal{H}$, $\ell(f_{\mathbf{c}}(\mathbf{x})) \geq \ell(g_{\mathbf{c}}(\mathbf{x}))$ and $\ell(g(\mathbf{x})) = g_{c_i}$, we have

$$\begin{aligned} |\ell(f_{\mathbf{c}}(\mathbf{x})) - \ell(g_{\mathbf{c}}(\mathbf{x}))| &= \min f_{\mathbf{c}}(\mathbf{x}) - g_{c_i}(\mathbf{x}) \\ &\leq f_{c_i}(\mathbf{x}) - g_{c_i}(\mathbf{x}) \\ &\leq \|f_{\mathbf{c}}(\mathbf{x}) - g_{\mathbf{c}}(\mathbf{x})\|_\infty, \end{aligned} \quad (11)$$

which shows that $\ell f(\mathbf{x})$ is a 1-Lipschitz continuous function.

Since the underlying probability distribution \mathbb{P} of the data is unknown, it is not possible to directly estimate the generalization error. To address this issue, a widely adopted strategy is to combine the empirical risk with a measure of model complexity, as shown in Eq. (12). For a hypothesis f , the generalization error is bounded as:

$$\mathcal{R}(f) \leq \mathcal{R}_{\text{emp}}(f) + \sup_{f \in \mathcal{H}} [\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)], \quad (12)$$

where the second term, $\sup_{f \in \mathcal{H}} [\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)]$, can be controlled using the Rademacher complexity $\mathfrak{R}(\mathcal{H})$:

$$\sup_{f \in \mathcal{H}} [\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)] \leq \mathfrak{R}(\mathcal{H}). \quad (13)$$

Compared to the global Rademacher complexity, the local Rademacher complexity (Bartlett, Bousquet, and Mendelson 2005) provides more refined bounds on the generalization error, avoiding the overly conservative estimates from global metrics. Its specific definition is given by:

$$\mathfrak{R}_n(\mathcal{H}, r) := \mathfrak{R}_n(\{f \in \mathcal{H} : \mathbb{E}[f^2] \leq r\}), \quad (14)$$

where r represents a tolerance for the complexity or error of the functions, and must satisfy $r > 0$.

Given the significant challenges in directly assessing local Rademacher complexity, an approximate estimation method based on the ratio of the eigenvalues to the secondary eigenvalues of the integral operator can be adopted (Mendelson 2003). Building on the concept of Principal Eigenvalue Proportion (PEP) introduced in (Liu et al. 2017), specifically Definitions 1 and 2, we utilize these to estimate the local Rademacher complexity of our model presented in Eq. (5).

Definition 1. (Liu et al. 2017) Assume K is a Mercer kernel with its eigenvalue decomposition given by $K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j(K) \phi_j(\mathbf{x}) \phi_j(\mathbf{x}')$, where the sequence $(\lambda_j(K))_{j=1}^{\infty}$ represents the eigenvalues of the kernel function arranged in decreasing order. The t -Principal Eigenvalue Proportion (t -PEP) of K , for $t \in \mathbb{N}_+$, is defined as follows:

$$\beta(K, t) = \frac{\sum_{i=1}^t \lambda_i(K)}{\sum_{i=1}^{\infty} \lambda_i(K)}. \quad (15)$$

Definition 2. (Liu et al. 2017) Assume K is a kernel function and let $\mathbf{K} = [\frac{1}{n} K(x_i, x_j)]_{i,j=1}^n$ denote its corresponding kernel matrix. The t -empirical principal eigenvalue proportion of \mathbf{K} , for $t \in \{1, \dots, n-1\}$, is given by

$$\hat{\beta}(\mathbf{K}, t) = \frac{\sum_{i=1}^t \lambda_i(\mathbf{K})}{\text{tr}(\mathbf{K})}, \quad (16)$$

where $\lambda_i(\mathbf{K})$ represents the i -th largest eigenvalue of \mathbf{K} , and $\text{tr}(\mathbf{K})$ denotes the trace of the matrix \mathbf{K} .

Lemma 2. *Assume $\kappa = \sup_{p \in \{1, 2, \dots, m\}} \text{tr}(K_p) < \infty$, the Theorem 6 of (Kloft and Blanchard 2012) and the Corollary 1 of (Liu et al. 2017) show that for $\forall K \in \mathcal{K}^{\text{multiple}}, D > 0$, denote*

$$\mathcal{H}_K^{\text{multiple}} = \{f_w : x \mapsto \langle \mathbf{w}, \Phi_K(\mathbf{x}) \rangle : \|\mathbf{w}\| \leq D\},$$

where $\mathcal{K}^{\text{multiple}} = \sum_{p=1}^m \gamma_p^2 K_p$. The local Rademacher complexity of the hypothesis space $\mathcal{H}_K^{\text{multiple}}$ is bounded by:

$$\begin{aligned} R_n(\mathcal{H}_K^{\text{multiple}}, r) &\leq \sqrt{\frac{2}{n} \min_{\theta \geq 0} \left(\theta r + D^2 m^{\frac{z}{2}-1} \sum_{j>\theta} \lambda_j(K) \right)} \\ &\leq \sqrt{\frac{2}{n} \left(\theta r + D^2 m^{\frac{z}{2}-1} \kappa \sum_{p=1}^m \frac{1}{\beta(K_p, t)} \right)}, \end{aligned}$$

Where z is the optimal norm value corresponding to the proof of Proposition 2 in (Kloft and Blanchard 2012).

Generalization Bounds with PEP

Within the purview of this section, an estimation of the generalization error margin is conducted by leveraging the ratio of the dominant eigenvalues.

Theorem 1. Let ℓ defined in Eq. (6) be the multiple kernel k -means loss function associated with the score function f defined in Eq (7). Then, $\forall \delta > 0$, with probability at least $1 - \delta$ over the choice of a sample $\mathbb{S} = \{\mathbf{x}_i\}_{i=1}^n$ drawn i.i.d according to \mathbb{P} , the following inequality holds: $\forall \theta > 1$ and $\forall f \in \mathcal{H}_K^{\text{multiple}}$,

$$R(f) \leq \frac{\theta}{\theta - 1} \hat{R}(f) + \frac{c_1}{\sqrt{nz}} + \frac{c_2 + c_3}{n},$$

where $z = \sum_{p=1}^m \beta(K_p, t)$, $c_1 = 6\theta b D \sqrt{2\kappa m^{\frac{z}{2}-1}}$, $c_2 = 48\theta b t$, $c_3 = (44b + 20b\theta) \log \frac{1}{\delta}$.

Proof 2. It follows from Lemma 1 that $\ell_f(x)^2 \leq 4b\ell_f(x)$, we get $\mathbb{E}\ell_f^2(x) \leq 4b\mathbb{E}\ell_f(x)$. According to Theorem 3.3 of (Bartlett, Bousquet, and Mendelson 2005), we can know with at least $1 - \delta$ probability that:

$$\forall \theta > 1, R(f) \leq \frac{\theta}{\theta - 1} R_{\text{emp}}(f) + \frac{6\theta r^*}{4b} + \frac{e}{n},$$

where $e = \log \frac{1}{\delta} (44b + 20b\theta)$ and r^* is the fixed point of $4bR_n(\mathcal{L}, r)$. According to Lemma 1, we get ℓ_f is a 1-Lipschitz with respect to the L_∞ norm. According to Lemma 5 of (Cortes et al. 2016), we get $R_n(\mathcal{L}, r) \leq R_n(\mathcal{H}_K^{\text{multiple}}, r)$. According to Lemma 2, we know

$$R_n(\mathcal{H}_K^{\text{multiple}}, r) \leq \sqrt{\frac{2}{n} \left(\theta r + D^2 m^{\frac{z}{2}-1} \kappa \sum_{p=1}^m \frac{1}{\beta(K_p, t)} \right)}.$$

Then we have

$$\begin{aligned} 4bR_n(\mathcal{L}, r) &\leq 4bR_n(\mathcal{H}_K^{\text{multiple}}, r) \\ &\leq 4b \sqrt{\frac{2}{n} \left(t \cdot r + D^2 m^{\frac{z}{2}-1} \kappa \sum_{p=1}^m \frac{1}{\beta(K_p, t)} \right)}. \end{aligned}$$

Let $a = \frac{32b^2 t}{n}$, $c = \frac{32\kappa b^2 D^2 m^{\frac{z}{2}-1}}{n} \sum_{p=1}^m \frac{1}{\beta(K_p, t)}$, and by

solving $r^2 - ar - c = 0$, we can obtain:

$$\begin{aligned} r^* &= \frac{a + \sqrt{a^2 + 4c}}{2} \leq \frac{a + \sqrt{a^2} + \sqrt{4c}}{2} = a + \sqrt{c} \\ &= \frac{32b^2 t}{n} + \sqrt{\frac{32\kappa b^2 D^2 m^{\frac{z}{2}-1}}{n} \sum_{p=1}^m \frac{1}{\beta(K_p, t)}} \end{aligned}$$

then bring r^* back to Eq. (17) and complete this proof.

Assumption 1. Assume that the kernel function spectrum satisfies the polynomial spectrum decay rate, that is

$$\exists \gamma > 1 : \lambda_i(K) = O(i^{-\gamma}).$$

The polynomial spectral decay rate is a common assumption, which is satisfied by both finite-rank kernels and convolution kernels. Assuming that the kernel function K satisfies the polynomial spectral decay and we assume that the t -PEP value corresponding to the v -th kernel is the smallest, we can get

$$\begin{aligned} \sum_{i=t+1}^{\infty} \lambda_i^v(K) &= \sum_{i=t+1}^{\infty} O(i^{-\gamma}) \leq \int_t^{\infty} O(x^{-\gamma}) dx \\ &= O\left(\frac{x^{1-\gamma}}{1-\gamma} \Big|_t^{\infty}\right) = O\left(\frac{t^{1-\gamma}}{\gamma-1}\right). \end{aligned} \quad (17)$$

Therefore, there is

$$\begin{aligned} \sum_{i=1}^t \lambda_i^v(K) &= \text{tr}(K_v) - \sum_{i=t+1}^{\infty} \lambda_i^v(K) \\ &\geq \text{tr}(K_v) - O\left(\frac{t^{1-\gamma}}{\gamma-1}\right), \\ \frac{\sum_{i=1}^t \lambda_i^v(K)}{\sum_{i=1}^{\infty} \lambda_i^v(K)} &\geq \Omega\left(\frac{t^{1-\gamma}}{\gamma-1}\right). \end{aligned}$$

In this case, the generalization error bound converges to

$$O\left(\frac{1}{\sqrt{n} \sum_{p=1}^m \beta(K_p, t)} + \frac{1}{n}\right).$$

If $t^{\gamma-1}(\gamma-1) \geq \frac{\sqrt{n}}{m}$ means $t \geq \left(\frac{\sqrt{n}}{m(\gamma-1)}\right)^{\frac{1}{\gamma-1}}$, the convergence rate can reach $\mathcal{O}\left(\frac{1}{n}\right)$.

Remark 1. γ_p represents the weight for each kernel function K_p , while PEP measures the complexity of the kernel functions. When the proportion of the principal eigenvalue is low, $\beta(K_p, t)$ is also low, indicating a higher complexity of the kernel function. This conclusion further supports the strategy in multiple kernel learning of assigning appropriate weights to kernel functions of varying complexities to achieve models with better generalization performance.

Generalization Bounds with Empirical PEP

Since it is not easy to compute the value of PEP, we have to use the empirical PEP for practical kernel learning. Fortunately, we can also use it to derive sharp bound.

Theorem 2. Let ℓ defined in Eq. (6) be the multiple kernel k -means loss function associated with the score function f defined in Eq. (7). Then, with $1 - \delta, \forall \theta \geq 1, f \in \mathcal{H}_K^{\text{multiple}}$, we have

$$R(f) \leq \frac{\theta}{\theta - 1} \hat{R}(f) + \frac{c_3}{\sqrt{n\hat{z}}} + \frac{c_4 + c_5}{n} \quad (18)$$

where $\hat{z} = \sum_{p=1}^m \hat{\beta}(\mathbf{K}_p, t)$, $c_3 = 12\theta D \sqrt{2\kappa m^{\frac{z}{2}-1}}$, $c_4 = 384\theta bt$, $c_5 = (11 + 20b\theta + \frac{39\theta}{b}) \log(\frac{3}{\delta})$.

Proof 3. To prove the theorem, we introduce the empirical localized Rademacher complexity of \mathcal{H} , which is defined as:

$$\hat{R}_n(\mathcal{H}, r) := R_n \left(\left\{ f \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i))^2 \leq r \right\} \right).$$

From (Bartlett, Bousquet, and Mendelson 2005), it is known that with at least probability $1 - \delta$, for all $\theta > 1$, the following formula holds:

$$R(f) \leq \frac{\theta}{\theta - 1} R_{\text{emp}}(f) + \frac{6\theta \hat{r}^*}{4b} + \frac{\log(\frac{3}{\delta})(44b + 20b\theta)}{n}, \quad (19)$$

where \hat{r}^* is the fixed point of $8b\hat{R}_n(\mathcal{L}, 2r) + \frac{13 \log(3/\delta)}{n}$ and the empirical localized Rademacher complexity of \mathcal{H} is defined as:

$$\hat{R}_n(\mathcal{H}, r) := R_n \left(\left\{ f \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i))^2 \leq r \right\} \right).$$

According to the Lemma 2, we have

$$\begin{aligned} & 8b\hat{R}_n(\mathcal{L}, 2r) + \frac{13 \log(3/\delta)}{n} \\ & \leq 8b\hat{R}_n(\mathcal{H}_K^{\text{multiple}}, 2r) + \frac{13 \log(3/\delta)}{n} \\ & \leq 8b \sqrt{\frac{2}{n} \left(t \cdot 2r + D^2 m^{\frac{z}{2}-1} \kappa \sum_{p=1}^m \frac{1}{\hat{\beta}(\mathbf{K}_p, t)} \right)} \\ & \quad + \frac{13 \log(3/\delta)}{n}. \end{aligned}$$

Similar to the proof of the Theorem 1, we need to solve for

$$\begin{aligned} r &= 8b \sqrt{\frac{2}{n} \left(t \cdot 2r + D^2 m^{\frac{z}{2}-1} \kappa \sum_{p=1}^m \frac{1}{\hat{\beta}(\mathbf{K}_p, t)} \right)} \\ & \quad + \frac{13 \log(\frac{3}{\delta})}{n} \end{aligned}$$

Defining $a = \frac{256tb^2}{n}$, $c = \frac{128b^2 D^2 m^{\frac{z}{2}-1} \kappa}{n} \sum_{p=1}^m \frac{1}{\hat{\beta}(\mathbf{K}_p, t)}$ and $d = \frac{13 \log(\frac{3}{\delta})}{n}$. The above is equivalent to solving

$$r^2 - (2d + a)r - (c - d^2) = 0$$

for a positive root. It is easy to verify that

$$\begin{aligned} \hat{r}^* &\leq \frac{2d + a + \sqrt{a^2 + 4da + 4c}}{2} \\ &\leq \frac{2d + a + (a + 2d + 2\sqrt{c})}{2} \\ &= a + 2d + \sqrt{c} = \frac{256tb^2}{n} + \frac{26 \log(\frac{3}{\delta})}{n} \\ & \quad + 8bD \sqrt{\frac{2\kappa m^{\frac{z}{2}-1} \sum_{p=1}^m \frac{1}{\hat{\beta}(\mathbf{K}_p, t)}}{n}} \end{aligned}$$

Then we bring \hat{r}^* back to Eq. (19) and complete this proof. Same as the assumption 1, when $t^{\gamma-1}(\gamma-1) \geq \frac{\sqrt{n}}{m}$ means $t \geq \left(\frac{\sqrt{n}}{m(\gamma-1)}\right)^{\frac{1}{\gamma-1}}$, the convergence rate can reach $\mathcal{O}\left(\frac{1}{n}\right)$.

Graph Filter Enhanced Multiple Linear K-means

In this section, we introduce a strategy within the linear case of the MKKM framework to denoise base partitions by leveraging smooth representations obtained through an optimal graph filter, which aims to enhance the principal eigenvalue proportions.

In graph signal processing, natural signals are expected to be smooth across adjacent nodes, aligning with the underlying graph structure. Smoother signals often lead to clearer clustering, consistent with the cluster and manifold assumption (Zhou, Du, and Li 2023b). To achieve this, it is crucial to smooth raw features, reducing high-frequency noise while preserving key graph-based properties.

We start by constructing a set of affinity graphs $\{\mathbf{A}_p\}_{p=1}^m$, with each $\mathbf{A}_p \in \mathbb{R}^{n \times n}$. For the p -th data view, the normalized adjacency matrix is $\mathbf{D}_p^{-\frac{1}{2}} \mathbf{A}_p \mathbf{D}_p^{-\frac{1}{2}}$, where \mathbf{D}_p is the diagonal degree matrix. The corresponding normalized graph Laplacian is:

$$\mathbf{L}_p = \mathbf{I} - \mathbf{D}_p^{-\frac{1}{2}} \mathbf{A}_p \mathbf{D}_p^{-\frac{1}{2}}. \quad (20)$$

A common technique is to apply a low-pass graph filter, typically expressed as:

$$\hat{\mathbf{x}} = \sum_{o=1}^{\bar{o}} \left(\frac{\mathbf{I} + \mathbf{D}_p^{-\frac{1}{2}} \mathbf{A}_p \mathbf{D}_p^{-\frac{1}{2}}}{2} \right)^o \mathbf{x} = \mathbf{G}_p \mathbf{x}, \quad (21)$$

where o represents the o -hop neighborhood.

Traditional methods, such as those in (Lin et al. 2023), typically use fixed pre-defined graph filters, which may not adapt well to complex multi-view data. To address this limitation, we propose learning an optimal low-pass graph filter to enhance smoothness across embeddings. This is achieved by expressing the smoother embedding $\hat{\mathbf{H}}_p$ as:

$$\hat{\mathbf{H}}_p = \left(\sum_{i=1}^m \mu_i \mathbf{G}_i \right) \mathbf{H}_p, \quad (22)$$

where $\{\mu_i\}_{i=1}^m$ are learnable combination coefficients, and $\{\mathbf{G}_i\}$ are graph-based filters associated with different views.

Algorithm 1: Algorithm for the problem in Eq. (23).

Require: Base partitions $\{\mathbf{H}_p\}_{p=1}^m$, the cluster number k , the neighborhood size is set to 5.

- 1: Compute the probabilistic affinity graph $\{\mathbf{S}_r\}_{r=1}^m$ according to (Nie, Wang, and Huang 2014);
- 2: Initialize γ, μ, \mathbf{Y} ;
- 3: **repeat**
- 4: Update \mathbf{Y} by CD algorithm (Nie et al. 2022);
- 5: Update γ according to Eq. (26);
- 6: Update μ according to Eq. (27);
- 7: **until** Converges

Ensure: The discrete clustering result \mathbf{Y} .

This formulation ensures that the derived embedding is not only smoother but also incorporates complementary information from multiple views, optimizing the clustering process. Based on this, we propose the Graph filter-enhanced Multiple Linear K-Means (GMLKM):

$$\begin{aligned} \min \quad & \text{tr} \left(\sum_{p=1}^m \gamma_p^2 \tilde{\mathbf{G}} \mathbf{H}_p \mathbf{H}_p^T \tilde{\mathbf{G}} \right) \\ & - \text{tr} \left(\mathbf{Y}^T \left(\sum_{p=1}^m \gamma_p^2 \tilde{\mathbf{G}} \mathbf{H}_p \mathbf{H}_p^T \tilde{\mathbf{G}} \right) \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{-1} \right) \\ \text{s.t.} \quad & \tilde{\mathbf{G}} = \sum_{i=1}^m \mu_i \mathbf{G}_i, \mathbf{Y} \in \text{Ind}, \\ & \sum_{p=1}^m \gamma_p = 1, \sum_{i=1}^m \mu_i = 1, \gamma_p \geq 0, \mu_i \geq 0, \forall p, i. \end{aligned} \quad (23)$$

This filtering approach is preferred as it captures high-order relationships while maintaining linear computational complexity through sparse matrix multiplication.

Remark 2. PEP quantifies the proportion of variance captured by the leading eigenvalues of the kernel matrix. Higher PEP values suggest that most of the variance is explained by the dominant eigenvalues, enabling effective clustering as the data can be well-represented in a lower-dimensional space. The adaptive graph filter enhances PEP by increasing the smoothness and relevance of the data representation, resulting in a more concentrated and informative eigenvalue distribution.

Alternate Optimization

The objective in Eq. (23) involves optimizing four variables. We propose a three-step alternating optimization procedure by optimizing one variable while keeping the others fixed.

Optimization \mathbf{Y} : With γ and μ fixed, the optimization problem in Eq. (23) with respect to \mathbf{Y} reduces to:

$$\max_{\mathbf{Y} \in \text{Ind}} \text{tr} \left(\mathbf{Y}^T (\tilde{\mathbf{H}} \tilde{\mathbf{H}}^T) \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{-1} \right), \quad (24)$$

where $\tilde{\mathbf{H}} = \tilde{\mathbf{G}} [\gamma_1 \mathbf{H}_1, \gamma_2 \mathbf{H}_2, \dots, \gamma_m \mathbf{H}_m]$. This formulation resembles the K-means algorithm applied to $\tilde{\mathbf{H}}$. Hence, the optimal \mathbf{Y} can be efficiently obtained using the Coordinate Descent (CD) method (Nie et al. 2022).

Optimization γ : With \mathbf{Y} and μ fixed, the optimization problem in Eq. (23) with respect to γ simplifies to:

$$\min \sum_{p=1}^m \gamma_p^2 \alpha_p, \quad \text{s.t.} \quad \sum_{p=1}^m \gamma_p = 1, \gamma_p \geq 0, \forall p, \quad (25)$$

where $\alpha_p = -\text{tr}(\mathbf{Y}^T \tilde{\mathbf{G}} \mathbf{H}_p \mathbf{H}_p^T \tilde{\mathbf{G}} \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{-1}) + \text{tr}(\tilde{\mathbf{G}} \mathbf{H}_p \mathbf{H}_p^T \tilde{\mathbf{G}})$. The optimal solution (Du, Zhou et al. 2015) for γ_p is given by:

$$\gamma_p = \frac{1}{\sum_{p'=1}^m \frac{\alpha_p}{\alpha_{p'}}}. \quad (26)$$

Optimization μ : With \mathbf{Y} and γ fixed, the optimization problem in Eq. (23) with respect to μ is reformulated as:

$$\min \mu^T \mathbf{M} \mu, \quad \text{s.t.} \quad \sum_{i=1}^m \mu_i = 1, \mu_i \geq 0, \forall i, \quad (27)$$

where the matrix $\mathbf{M} \in \mathbb{R}^{m \times m}$ is defined as $\mathbf{M}_{ij} = -\text{tr}(\mathbf{Y}^T \mathbf{G}_i (\sum_{p=1}^m \gamma_p \mathbf{H}_p \mathbf{H}_p^T) \mathbf{G}_j \mathbf{Y} (\mathbf{Y}^T \mathbf{Y})^{-1}) + \text{tr}(\mathbf{G}_i (\sum_{p=1}^m \gamma_p \mathbf{H}_p \mathbf{H}_p^T) \mathbf{G}_j)$. This is a standard quadratic programming problem. The whole optimization procedure in solving Eq. (23) is outlined in Algorithm 1.

Convergence Analysis

The optimization problem in Eq. (23) is evidently lower-bounded, and the variable updates for updating \mathbf{Y} , γ , and μ monotonically decrease the objective function. Thus, convergence of Algorithm 1 is guaranteed.

Experiments

Datasets

In Table 1, we evaluate our method on ten benchmark datasets: Lung (D1), JAFFE (D2), CSTR (D3), ISOLET(D4), WAP (D5), Mfeat (D6), K1b (D7), MouseBladder (D8), PBMC (D9), and CBMC (D10). These datasets cover a range of domains, including image datasets (JAFFE, ISOLET, Mfeat), biological datasets (Lung, MouseBladder, PBMC, CBMC), and text corpora (CSTR, WAP, K1b).

ID	Dataset	Samples	Features	Classes
D1	Lung	203	3312	5
D2	JAFFE	213	676	10
D3	CSTR	476	1000	4
D4	ISOLET	1560	617	26
D5	WAP	1560	8460	20
D6	Mfeat	2000	240	10
D7	K1b	2340	21839	6
D8	MouseBladder	2746	11829	16
D9	PBMC	4271	12206	8
D10	CBMC	8617	1703	15

Table 1: Summary of ten datasets

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
ACC(%)										
avgH	79.5 ± 7.3	87.4 ± 7.9	76.5 ± 14.9	55.4 ± 2.9	40.3 ± 6.2	85.0 ± 5.6	81.2 ± 15.8	53.7 ± 3.7	58.8 ± 2.1	62.1 ± 3.1
AWP	64.0 ± 0.0	62.0 ± 0.0	66.7 ± 0.0	45.2 ± 0.0	33.7 ± 0.0	64.2 ± 0.0	85.4 ± 0.0	53.4 ± 0.0	53.7 ± 0.0	59.7 ± 0.0
LFMVC	81.5 ± 7.1	89.7 ± 7.2	80.8 ± 7.0	55.7 ± 2.7	42.2 ± 3.2	84.6 ± 9.0	83.4 ± 8.5	54.4 ± 2.2	59.4 ± 2.3	62.2 ± 2.9
OPLF	86.7 ± 0.0	97.3 ± 2.7	90.3 ± 0.0	59.3 ± 1.1	47.2 ± 1.4	94.9 ± 2.8	92.0 ± 0.3	61.6 ± 0.5	64.1 ± 0.7	66.0 ± 0.2
LFLKA	81.8 ± 9.1	89.5 ± 7.2	80.8 ± 7.1	55.7 ± 2.7	42.9 ± 3.4	85.8 ± 5.2	86.0 ± 8.0	54.8 ± 2.4	59.1 ± 2.0	62.9 ± 2.6
ALMVC	79.0 ± 12.1	83.5 ± 9.9	80.6 ± 7.0	56.8 ± 3.4	41.9 ± 3.6	84.6 ± 9.0	83.1 ± 8.5	54.1 ± 2.3	58.9 ± 1.8	62.2 ± 3.0
M3LF	79.7 ± 7.1	84.2 ± 7.9	76.1 ± 12.6	56.3 ± 2.7	42.0 ± 3.1	82.7 ± 8.3	82.1 ± 4.5	53.8 ± 2.7	60.0 ± 2.6	62.2 ± 2.6
sLGm	79.6 ± 7.3	84.2 ± 7.4	80.7 ± 7.1	55.8 ± 2.7	41.6 ± 2.7	84.6 ± 9.0	84.0 ± 8.5	54.6 ± 2.0	59.0 ± 2.1	62.2 ± 2.9
RIWLF	79.5 ± 8.0	86.6 ± 8.2	80.6 ± 7.0	57.1 ± 2.6	42.2 ± 2.7	84.6 ± 9.0	84.0 ± 8.5	54.5 ± 2.4	59.0 ± 2.1	62.2 ± 3.0
GMLKM	87.2 ± 0.0	97.3 ± 2.7	91.8 ± 0.0	60.4 ± 0.6	47.7 ± 0.5	94.7 ± 2.7	93.0 ± 0.0	62.3 ± 0.5	66.6 ± 0.6	68.5 ± 0.3
NMI(%)										
avgH	56.3 ± 6.8	91.8 ± 4.1	64.8 ± 18.6	73.9 ± 1.6	48.2 ± 6.6	86.4 ± 2.9	68.4 ± 18.7	61.7 ± 1.8	59.1 ± 1.5	67.5 ± 0.9
AWP	42.3 ± 0.0	69.8 ± 0.0	47.4 ± 0.0	65.0 ± 0.0	42.6 ± 0.0	68.7 ± 0.0	66.9 ± 0.0	61.6 ± 0.0	53.1 ± 0.0	65.8 ± 0.0
LFMVC	59.5 ± 7.4	92.7 ± 3.8	69.5 ± 4.0	74.2 ± 1.3	50.3 ± 1.3	86.3 ± 4.2	80.8 ± 7.8	61.6 ± 1.7	60.5 ± 1.3	67.8 ± 0.5
OPLF	66.6 ± 0.0	96.8 ± 1.7	75.3 ± 0.0	77.0 ± 0.5	52.0 ± 0.2	90.7 ± 1.8	80.7 ± 1.2	68.6 ± 0.1	61.7 ± 0.6	68.2 ± 0.3
LFLKA	60.5 ± 9.3	92.7 ± 3.8	69.6 ± 3.9	74.2 ± 1.3	50.5 ± 1.0	86.6 ± 2.8	73.9 ± 7.2	61.9 ± 1.9	60.5 ± 1.4	67.5 ± 0.6
ALMVC	59.4 ± 9.1	88.4 ± 5.1	69.2 ± 3.8	74.5 ± 1.3	50.0 ± 1.2	86.3 ± 4.2	70.9 ± 7.5	61.3 ± 1.9	60.1 ± 1.6	67.6 ± 0.7
M3LF	56.3 ± 6.8	89.7 ± 4.5	66.5 ± 8.2	74.6 ± 1.4	50.0 ± 1.0	85.1 ± 3.9	65.4 ± 4.1	61.4 ± 2.1	59.5 ± 1.7	67.0 ± 0.7
sLGm	56.6 ± 7.2	90.2 ± 3.8	69.4 ± 4.2	74.2 ± 1.3	50.2 ± 0.9	86.3 ± 4.2	72.6 ± 8.1	61.8 ± 1.8	59.4 ± 1.3	67.9 ± 0.3
RIWLF	56.6 ± 7.4	91.0 ± 4.5	69.2 ± 3.8	74.7 ± 1.3	50.7 ± 1.0	86.3 ± 4.2	72.6 ± 8.0	61.3 ± 2.1	59.4 ± 1.3	67.6 ± 0.6
GMLKM	66.7 ± 0.0	96.9 ± 1.3	79.4 ± 0.0	77.0 ± 0.4	47.9 ± 0.8	90.7 ± 1.8	82.7 ± 0.0	68.9 ± 0.5	60.1 ± 0.6	67.7 ± 0.3
ARI(%)										
avgH	59.9 ± 10.8	85.0 ± 8.4	64.2 ± 20.8	50.4 ± 2.6	31.8 ± 10.6	81.5 ± 5.4	71.1 ± 21.3	40.9 ± 3.3	45.8 ± 1.8	51.0 ± 2.7
AWP	31.9 ± 0.0	54.1 ± 0.0	35.5 ± 0.0	38.5 ± 0.0	21.3 ± 0.0	58.0 ± 0.0	70.9 ± 0.0	41.0 ± 0.0	35.7 ± 0.0	51.8 ± 0.0
LFMVC	65.4 ± 11.9	87.0 ± 8.0	69.0 ± 6.7	52.3 ± 2.6	33.7 ± 3.9	81.6 ± 7.3	71.7 ± 13.2	41.2 ± 2.3	46.8 ± 2.3	51.2 ± 2.5
OPLF	73.6 ± 0.0	95.0 ± 2.9	80.2 ± 0.0	55.3 ± 0.6	40.2 ± 1.7	90.1 ± 2.7	86.5 ± 0.6	44.1 ± 0.3	50.4 ± 0.1	54.8 ± 0.1
LFLKA	67.4 ± 14.5	87.0 ± 8.0	69.0 ± 6.7	52.3 ± 2.6	34.9 ± 4.5	82.3 ± 4.8	76.2 ± 12.6	41.4 ± 2.6	46.5 ± 2.4	51.7 ± 2.5
ALMVC	63.3 ± 16.2	78.9 ± 9.4	68.6 ± 6.6	51.2 ± 2.3	33.4 ± 4.5	81.6 ± 7.3	71.3 ± 12.8	40.9 ± 2.5	46.2 ± 2.3	51.1 ± 2.6
M3LF	61.7 ± 11.2	81.3 ± 9.1	64.2 ± 14.0	51.6 ± 1.8	34.8 ± 4.3	79.8 ± 6.7	69.3 ± 8.1	41.0 ± 2.8	47.3 ± 1.9	51.3 ± 2.4
sLGm	60.1 ± 11.6	81.6 ± 8.2	68.9 ± 6.9	52.3 ± 2.6	33.4 ± 4.0	81.6 ± 7.3	73.6 ± 13.4	41.3 ± 2.5	46.2 ± 1.8	51.2 ± 2.5
RIWLF	60.5 ± 11.7	83.8 ± 9.3	68.6 ± 6.6	51.5 ± 2.0	34.3 ± 3.7	81.6 ± 7.3	73.7 ± 13.5	40.8 ± 2.7	46.2 ± 1.8	51.1 ± 2.6
GMLKM	74.7 ± 0.0	95.2 ± 2.4	84.2 ± 0.0	54.5 ± 0.7	35.2 ± 0.7	89.7 ± 2.8	89.4 ± 0.0	43.8 ± 0.1	48.2 ± 0.1	56.0 ± 0.1

Table 2: Clustering results in terms of ACC/NMI/ARI

Experimental Setup and Results

We evaluated the proposed GMLKM method against recently developed LFMVC approaches, including AWP (Nie, Tian, and Li 2018), LFMVC (Wang et al. 2019), OPLF (Liu et al. 2021), LFLKA (Zhang et al. 2023), ALMVC (Zhang et al. 2022), M3LF (Li et al. 2024b), sLGm (Wu et al. 2024), and RIWLF (Zhang et al. 2024). Additionally, the average performance of a single embedding, denoted as AvgH, was reported as a baseline. The related codes is available at <https://github.com/csliangdu/GMLKM>.

Experimental settings followed those in (Wu et al. 2024), where 12 shifted Laplacian kernels were generated using kernel parameters specified in (Du, Zhou et al. 2015). Kernel K-means was employed to derive base partitions, and the dimensionality for all methods was selected from $\{c, 2c, 3c, 4c\}$, as recommended in (Wang et al. 2019; Zhang et al. 2022). Baseline algorithms were executed using their recommended parameter settings. To mitigate the influence of randomness, each method was executed 10 times, and the average performance was reported.

For the proposed GMLKM method, the neighborhood size in Eq. (23) was fixed at 5 across all datasets without parameter tuning, while the order \bar{o} was chosen from $\{1, 2, 3\}$. The evaluation metrics included Accuracy (ACC), Normal-

ized Mutual Information (NMI), and Adjusted Rand Index (ARI), and the results are summarized in Table 2.

The results demonstrate that the proposed method consistently outperforms other approaches in terms of ACC, NMI, and ARI across a variety of datasets.

Conclusion

This work presents a novel analysis using the principal eigenvalue proportion for Multiple Kernel K-means achieving a generalization error bound of $O(1/n)$. By incorporating low pass graph filtering techniques it is expected that the principal eigenvalue proportion could be improved during for multiple kernel clustering task. We take the graph filter enhanced multiple linear K-means on base partitions in the late fusion scenarios to demonstrate the improved results. In future, we plan to conduct a thorough investigation into the theoretical and practical implications of enhancing the principal eigenvalue proportion using graph filters.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376146,62176001), the Shanxi Province Central Guidance for Local Science and Technology Development Special Project (YDZJSX20231D003).

References

- Bartlett, P. L.; Bousquet, O.; and Mendelson, S. 2005. Local Rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537.
- Cortes, C.; Kuznetsov, V.; Mohri, M.; and Yang, S. 2016. Structured prediction theory based on factor graph complexity. *Advances in Neural Information Processing Systems*, 29.
- Du, L.; Zhou, P.; et al. 2015. Robust multiple kernel k-means using l₂₁-norm. In *Twenty-fourth international joint conference on artificial intelligence*, 3476–3482.
- Kloft, M.; and Blanchard, G. 2012. On the convergence rate of lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 13(1): 2465–2502.
- Li, M.; Liu, X.; Zhang, Y.; and Liang, W. 2024a. Late fusion multiview clustering via min-max optimization. *EEE TNNLS*, 35(7): 9417–9427.
- Li, M.; Liu, X.; Zhang, Y.; and Liang, W. 2024b. Late Fusion Multiview Clustering via Min-Max Optimization. *EEE TNNLS*, 35(7): 9417–9427.
- Li, Z.; Tang, C.; Liu, X.; Zheng, X.; Zhang, W.; and Zhu, E. 2021. Consensus graph learning for multi-view clustering. *IEEE Transactions on Multimedia*, 24: 2461–2472.
- Lin, Z.; Kang, Z.; Zhang, L.; and Tian, L. 2023. Multi-View Attributed Graph Clustering. *IEEE Transactions on Knowledge Data Engineering*, 35(02): 1872–1880.
- Liu, J.; Liu, X.; Xiong, J.; Liao, Q.; Zhou, S.; Wang, S.; and Yang, Y. 2020. Optimal neighborhood multiple kernel clustering with adaptive local kernels. *EEE TNNLS*, 34(6): 2872–2885.
- Liu, S.; Wang, S.; Zhang, P.; Xu, K.; Liu, X.; Zhang, C.; and Gao, F. 2022. Efficient one-pass multi-view subspace clustering with consensus anchors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7576–7584.
- Liu, X. 2023. Simplemkkm: Simple multiple kernel k-means. *EEE TPAMI*, 45(4): 5174–5186.
- Liu, X.; Liu, L.; Liao, Q.; Wang, S.; Zhang, Y.; Tu, W.; Tang, C.; Liu, J.; and Zhu, E. 2021. One pass late fusion multi-view clustering. In *International Conference on Machine Learning*, 6850–6859.
- Liu, Y.; Liao, S.; Lin, H.; Yue, Y.; and Wang, W. 2017. Infinite kernel learning: generalization bounds and algorithms. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 31.
- Liu, Y.; Yang, X.; Zhou, S.; Liu, X.; Wang, S.; Liang, K.; Tu, W.; and Li, L. 2023. Simple contrastive graph clustering. *EEE TNNLS*.
- Mendelson, S. 2003. On the performance of kernel classes. *Journal of Machine Learning Research*, 4(Oct): 759–771.
- Nie, F.; Tian, L.; and Li, X. 2018. Multiview clustering via adaptively weighted procrustes. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2022–2030.
- Nie, F.; Wang, X.; and Huang, H. 2014. Clustering and projected clustering with adaptive neighbors. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 977–986.
- Nie, F.; Xue, J.; Wu, D.; Wang, R.; Li, H.; and Li, X. 2022. Coordinate Descent Method for k-means. *IEEE TPAMI*, 44(5): 2371–2385.
- Tang, C.; Li, Z.; Wang, J.; Liu, X.; Zhang, W.; and Zhu, E. 2022. Unified one-step multi-view spectral clustering. *EEE TNNLS*, 35(6): 6449–6460.
- Tang, C.; Liu, X.; Zhu, X.; Zhu, E.; Luo, Z.; Wang, L.; and Gao, W. 2020. CGD: Multi-view clustering via cross-view graph diffusion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 5924–5931.
- Wang, S.; Liu, X.; Liu, L.; Zhou, S.; and Zhu, E. 2023. Late fusion multiple kernel clustering with proxy graph refinement. *EEE TNNLS*, 34(8): 4359–4370.
- Wang, S.; Liu, X.; Zhu, E.; Tang, C.; Liu, J.; Hu, J.; Xia, J.; and Yin, J. 2019. Multi-view Clustering via Late Fusion Alignment Maximization. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 3778–3784.
- Wang, S.; Liu, X.; Zhu, X.; Zhang, P.; Zhang, Y.; Gao, F.; and Zhu, E. 2021. Fast parameter-free multi-view subspace clustering with consensus anchor guidance. *IEEE Transactions on Image Processing*, 31: 556–568.
- Wu, X.; Huang, C.; Liu, X.; Zhou, F.; and Ren, Z. 2024. Multiple Kernel Clustering with Shifted Laplacian on Grassmann Manifold. In *Proceedings of the ACM International Conference on Multimedia (MM)*.
- Zhang, T.; Liu, X.; Gong, L.; Wang, S.; Niu, X.; and Shen, L. 2023. Late fusion multiple kernel clustering with local kernel alignment maximization. *IEEE Transactions on Multimedia*, 25: 993–1007.
- Zhang, T.; Liu, X.; Zhu, E.; Zhou, S.; and Dong, Z. 2022. Efficient anchor learning-based multi-view clustering—a late fusion method. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3685–3693.
- Zhang, Y.; Tian, F.; Ma, C.; Li, M.; Yang, H.; Liu, Z.; Zhu, E.; and Liu, X. 2024. Regularized Instance Weighting Multiview Clustering via Late Fusion Alignment. *EEE TNNLS*, 1–13.
- Zhou, P.; and Du, L. 2023. Learnable graph filter for multi-view clustering. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3089–3098.
- Zhou, P.; Du, L.; and Li, X. 2023a. Adaptive consensus clustering for multiple K-means via base results refining. *EEE TNNLS*, 35(10): 10251–10264.
- Zhou, P.; Du, L.; and Li, X. 2023b. Adaptive Consensus Clustering for Multiple K-Means Via Base Results Refining. *EEE TNNLS*, 35(10): 10251–10264.
- Zhou, P.; Hu, B.; Yan, D.; and Du, L. 2024. Clustering Ensemble Via Diffusion on Adaptive Multiplex. *EEE TNNLS*, 36(4): 1463–1474.
- Zhou, S.; Ou, Q.; Liu, X.; Wang, S.; Liu, L.; Wang, S.; Zhu, E.; Yin, J.; and Xu, X. 2023. Multiple Kernel Clustering With Compressed Subspace Alignment. *EEE TNNLS*, 34(1): 252–263.