Discriminative Graph Embedding Framework via Label-Free Marginal Fisher Analysis

Qianqian Wang[1], Mengping Jiang[1*], Wei Feng[2], Haixi Zhang[2], Bin Liu[2]

[1]School of Telecommunications Engineering, Xidian University, Xi'an, Shaanxi, China, 710071
[2]College of Information Engineering, Northwest A&F University, Yangling, China, 712100
qqwang@xidian.edu.cn, mpjiang@foxmail.com, wei.feng@nwafu.edu.cn, zh.haixi@nwafu.edu.cn, liubin0929@nwsuaf.edu.cn

Abstract

Marginal Fisher Analysis (MFA) is a classical dimensionality reduction (DR) method that leverages dual graphs to capture intra-class compactness and inter-class separability. However, MFA's reliance on high-quality labels limits its practical application. For another, existing unsupervised DR methods neglect data's local manifold relationship, resulting in poor discriminativeness. To address these limitations, we propose a novel DR method named **D**iscriminative **G**raph **E**mbedding **F**ramework (DGEF) via Label-Free Marginal Fisher Analysis. Our approach uses the adjacency matrix and cluster indicator matrix derived from centerless K-Means to construct intrinsic graph and penalty graph, which preserve the local manifold structure of the data. Additionally, we have derived the convertible relationship between centerless K-Means and Manifold learning and unified them within a graph embedding framework. By adopting the intrinsic graph and penalty graph, our DGEF avoids centroid initialization and ensures robustness and discriminativeness. This method achieves dimensionality reduction adaptively without relying on labeled data. Extensive experiments on benchmark datasets show that our approach outperforms conventional methods in clustering performance.

Introduction

Dimensionality reduction (DR) plays a critical role in various fields of machine learning and data analysis, aiming to transform high-dimensional data into a low-dimensional representation while capturing the essential characteristics of high-dimensional data (Ran et al. 2022; Wright and Ma 2022; Yi et al. 2020). Traditional DR methods, such as Principal Component Analysis (PCA) (Abdi and Williams 2010), Linear Discriminant Analysis (LDA) (Li et al. 2024), Locality Preserving Projections (LPP) (He and Niyogi 2003), and Marginal Fisher Analysis (MFA) (Yan et al. 2005, 2006), have achieved significant success in numerous applications (Nie et al. 2020a; Liu et al. 2023b).

Classical DR methods are roughly categorized into global methods and local methods, and PCA and LDA are two of the most popular global methods (Raju et al. 2022; Li et al. 2023). PCA extracts the most expressive features by minimizing the summation of the reconstruction error of each data point in least-squares sense (Turk and Pentland 1991). However, it is an unsupervised method and thus the extracted low-dimensional representation lacks discriminative information (Abdi and Williams 2010). LDA introduces labels to learn discriminative features by simultaneously maximizing the ratio of the trace of inter-class variance to the trace of intra-class variance. Nevertheless, LDA mainly captures the global structure and cannot well discover the intrinsic geometric structure of manifold on which data possibly reside (Chang et al. 2022; Fu et al. 2020). Hence, LDA cannot well obtain the low-dimensional representation with maximum margin, which is characterized by local geometric structure and is beneficial to the subsequent tasks (Ayesha, Hanif, and Talib 2020; Nie et al. 2020a).

To effectively capture the intrinsic geometric structure embedded within the high-dimensional data space, researchers have developed manifold learning techniques for the purpose of dimensionality reduction. For instance, LPP is able to effectively preserve the intrinsic geometry of data while producing an explicit linear mapping (He and Niyogi 2003). In order to further exploit the discriminant geometric structure, many discriminant manifold learning approaches (Wang et al. 2016, 2021b; Nie et al. 2022b; Yang, Ma, and Han 2023) were proposed. MFA (Yan et al. 2005), as one of the prominent discriminant manifold methods, leverages two adjacency graphs to capture both the discriminant structure and intrinsic geometric structure. By maximizing the distance between neighboring points with different class labels and minimizing the distance between neighboring points sharing the same class label, MFA derives the optimal projection matrix (Huang et al. 2018). Building upon these works, Yan et al. proposed a graph-embedding framework (GEF) that unifies various dimensionality reduction algorithms (Yan et al. 2006).

However, MFA and its variant methods (Fan et al. 2021; Hu, Zhang, and Dai 2021; Lu et al. 2020) heavily rely on labeled data, making them unsuitable for scenarios with limited or unavailable labels (Wang et al. 2021a). To overcome this limitation, unsupervised discriminative dimensionality reduction methods (He and Niyogi 2003; Lee and Verleysen 2010; Zhou et al. 2023) have received increasing attention. These methods (Liu et al. 2023a) take advantage of pseudo labels generated by K-Means to learn discriminative data representations for unsupervised dimensionality reduc-

*Corresponding author.

tion (Nie et al. 2020b; Kapoor and Singhal 2017), *e.g.*, LDA-Km (Ding and Li 2007) and Un-LDA (Wang et al. 2023). Nevertheless, the aforementioned unsupervised discriminative methods confront two problems. First, K-Means necessitates the initialization of cluster centroids, resulting in unstable performance (Pei et al. 2020; Lu et al. 2023, 2024; Gao et al. 2025). Second, in the low-dimensional space, none of the aforementioned approaches can achieve a better margin between different clusters, which is essential for both classification and clustering.

To tackle these challenges, we propose a novel discriminative graph embedding framework via label-free MFA and demonstrate the connection between manifold learning and K-Means. Our method effectively extracts features and generates a discrete cluster indicator matrix. Unlike MFA, we utilize the adjacency matrix of K-nearest neighbors and the learned cluster indicator matrix to generate a similarity matrix, thereby preserving the original manifold structure and cluster information. Leveraging a graph embedding framework for computing sample relationships and adopting centerless K-Means clustering, our approach eliminates the need for centroid computation and retains the local information structure of the original data. The main contributions of this paper are as follows:

- A novel DR method named discriminative graph embedding framework via label-free MFA is proposed, which unifies centerless K-Means and manifold learning within a graph embedding framework.
- We construct intrinsic graph and penalty graph guided by centerless K-Means, which better preserves the local structure of the data, reducing the impact of noise on clustering performance.
- Apart from Euclidean distance, we introduce various distance metrics to better handle nonlinearly separable data in our experiments on toy datasets and benchmark datasets, whose experimental results demonstrate the superiority of our proposed method.

Related Work

K-Means

The K-Means clustering is one of the classical clustering methods. It involves selecting K cluster centroids and then assigning data points to the nearest cluster centroids, resulting in the partitioning of the input data into K distinct clusters. Given a set of data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, the objective function of K-Means can be described as follows:

$$\min_{\mathbf{A}_c} \sum_{c=1}^{K} \sum_{\mathbf{x}_i \in \mathbf{A}_c} \|\mathbf{x}_i - \mathbf{u}_c\|_2^2 \tag{1}$$

where $\{\mathbf{A}_c\}_{c=1}^{K}$ represent K distinct clusters, $\mathbf{u}_c = \frac{1}{n_c} \sum_{\mathbf{x}_i \in \mathbf{A}_c} \mathbf{x}_i$ represents the cluster centroid of the c-th cluster, n_c is the number of samples in the c-th cluster.

Marginal Fisher Analysis

MFA was built upon the graph embedding framework by Yan et al. (2005; 2006). This method introduces two adjacency graphs, the intrinsic graph $\mathbf{S}^I$ and the penalty graph $\mathbf{S}^P$, which respectively represent the similarity information among samples of the same class and the separability between samples of different clusters. The process can be expressed as follows.

(a) The Intrinsic Graph S

Suppose l_i is the class $\mathbf{x}_i$ belongs to, $\mathbf{S}$ is defined by:

$$\mathbf{S}_{ij}^I = \begin{cases} 1, \mathbf{x}_i \in N_{k_1}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_{k_1}(\mathbf{x}_i), l_i = l_j \\ 0, \text{otherwise.} \end{cases} \tag{2}$$

That is, if $\mathbf{x}_i$ is one of the k_1-nearest neighbors of $\mathbf{x}_j$, or $\mathbf{x}_j$ is one of the k_1 nearest neighbors of $\mathbf{x}_i$, and $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class, then $\mathbf{S}_{ij}^I = \mathbf{S}_{ji}^I = 1$. Then, the objective of Intrinsic graph can be expressed by:

$$\min_{\mathbf{W}} \sum_i \sum_j \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^2 \mathbf{S}_{ij}^I \tag{3}$$

(b) The Penalty Graph $\mathbf{S}^p$

The penalty graph $\mathbf{S}^p$ can be defined by:

$$\mathbf{S}_{ij}^p = \begin{cases} 1, \mathbf{x}_i \in N_{k_2}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_{k_2}(\mathbf{x}_i), l_i \neq l_j \\ 0, \text{otherwise.} \end{cases} \tag{4}$$

Then, the objective about the penalty graph can be expressed by:

$$\max_{\mathbf{W}} \sum_i \sum_j \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^2 \mathbf{S}_{ij}^p, \tag{5}$$

The Laplacian matrix of the Intrinsic Graph $\mathbf{S}^I$ and the Penalty Graph $\mathbf{S}^p$ can be defined as $\mathbf{L}^I = \mathbf{D}^I - \mathbf{S}^I$ and $\mathbf{L}^p = \mathbf{D}^p - \mathbf{S}^p$; $\mathbf{D}^I$ and $\mathbf{D}^p$ are degree matrices, where $\mathbf{D}_{ii}^I = \sum_{i,j \neq i} \mathbf{S}_{ij}^I$ and $\mathbf{D}_{ii}^p = \sum_{i,j \neq i} \mathbf{S}_{ij}^p$.

According to (Yan et al. 2005, 2006), the overall objective function of MFA can be expressed as follows:

$$\max_{\mathbf{W}} \frac{\mathrm{tr}(\mathbf{W}^T \mathbf{X} \mathbf{L}^p \mathbf{X}^T \mathbf{W})}{\mathrm{tr}(\mathbf{W}^T \mathbf{X} \mathbf{L}^I \mathbf{X}^T \mathbf{W})} \tag{6}$$

By carefully examining the Intrinsic Graph representation of MFA (Eq. (3)) and the objective function of K-Means (Eq. (1)), we can observe that both of them aim to minimize the distance between samples within a cluster. *This indicates a common objective between MFA and K-Means, suggesting that they can be unified within a single framework. Achieving both MFA and K-Means can be accomplished through a single process.* Next, we will derive our objective function based on MFA and K-Means, allowing us to simultaneously obtain the cluster indicator matrix and the projection matrix within a unified framework without the need for labels.

Methodology

Motivations and Objective

To further illustrate the relationship between MFA and K-Means and eliminate MFA's reliance on data labels, we first reformulate K-Means into a centerless version through Theorem 1. Then, we discuss the relationship between the centerless K-Means and MFA and can be integrate them into a unified unsupervised MFA framework, which naturally leads to our objective function.

Theorem 1 *If the union of clusters as* $\mathbf{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K\}$, $\mathbf{A}_c \in \mathbf{A}(c = 1, 2, ..., K)$ *is a cluster, and* $\mathbf{u}_c$ *denotes the centroid of cluster* $\mathbf{A}_c$*, then*

$$\begin{aligned} &\min_{\mathbf{A}_c \in \mathbf{A}} \sum_{c=1}^{K} \sum_{\mathbf{x}_i \in \mathbf{A}_c} \|\mathbf{x}_i - \mathbf{u}_c\|_2^2 \\ &= \min_{\mathbf{A}_c \in \mathbf{A}} \sum_{c=1}^{K} \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{A}_c} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \end{aligned} \tag{7}$$

Proof 1 *For the c-th cluster* $\mathbf{A}_c$*, suppose there are* n_c *samples, and* $\mathbf{u}_c = \frac{1}{n_c} \sum_{\mathbf{x}_i \in \mathbf{A}_c} \mathbf{x}_i$*. Then, we have the following:*

$$\begin{aligned} &\min_{\mathbf{A}_c} \sum_{\mathbf{x}_i \in \mathbf{A}_c} \|\mathbf{x}_i - \mathbf{u}_c\|_2^2 \\ &= \min_{\mathbf{A}_c} \operatorname{tr} \sum_{\mathbf{x}_i \in \mathbf{A}_c} (\mathbf{x}_i - \mathbf{u}_c)^T (\mathbf{x}_i - \mathbf{u}_c) \\ &= \min_{\mathbf{A}_c} \operatorname{tr} \sum_{\mathbf{x}_i \in \mathbf{A}_c} (\mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{x}_i^T \mathbf{u}_c + \mathbf{u}_c^T \mathbf{u}_c) \\ &= \min_{\mathbf{A}_c} \operatorname{tr} (\sum_{\mathbf{x}_i \in \mathbf{A}_c} (\mathbf{x}_i^T \mathbf{x}_i) - 2 \sum_{\mathbf{x}_i \in \mathbf{A}_c} (\mathbf{x}_i^T \mathbf{u}_c) + n_c (\mathbf{u}_c^T \mathbf{u}_c)) \\ &= \min_{\mathbf{A}_c} \operatorname{tr} \sum_{\mathbf{x}_i \in \mathbf{A}_c} (\mathbf{x}_i^T \mathbf{x}_i - \mathbf{x}_i^T \mathbf{u}_c) \end{aligned} \tag{8}$$

and

$$\begin{aligned} &\min_{\mathbf{A}_c} \sum_{\mathbf{x}_i \in \mathbf{A}_c} \sum_{\mathbf{x}_j \in \mathbf{A}_c} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \\ &= \min_{\mathbf{A}_c} \operatorname{tr} \sum_{\mathbf{x}_i \in \mathbf{A}_c} \sum_{\mathbf{x}_j \in \mathbf{A}_c} (\mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{x}_i^T \mathbf{x}_j + \mathbf{x}_j^T \mathbf{x}_j) \\ &= \min_{\mathbf{A}_c} \operatorname{tr} (\sum_{\mathbf{x}_i \in \mathbf{A}_c} \sum_{\mathbf{x}_j \in \mathbf{A}_c} (\mathbf{x}_i^T \mathbf{x}_i + \mathbf{x}_j^T \mathbf{x}_j) \\ &- \sum_{\mathbf{x}_i \in \mathbf{A}_c} 2\mathbf{x}_i^T \sum_{\mathbf{x}_j \in \mathbf{A}_c} \mathbf{x}_j) \\ &= \min_{\mathbf{A}_c} \operatorname{tr} (\sum_{\mathbf{x}_i \in \mathbf{A}_c} 2n_c \mathbf{x}_i^T \mathbf{x}_i - \sum_{\mathbf{x}_i \in \mathbf{A}_c} 2\mathbf{x}_i^T (n_c \mathbf{u}_c)) \\ &= (2n_c) \min_{\mathbf{A}_c} \operatorname{tr} \sum_{\mathbf{x}_i \in \mathbf{A}_c} (\mathbf{x}_i^T \mathbf{x}_i - \mathbf{x}_i^T \mathbf{u}_c) \end{aligned} \tag{9}$$

From the above proof, when $n_1 = n_2 = \cdots = n_K$*, we can conclude that Theorem 1 holds true.*

According to Theorem 1, the optimization objective function Eq.(1) of K-Means can be rewritten as follows:

$$\min_{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K} \sum_{c=1}^{K} \sum_{\mathbf{x}_i \in \mathbf{A}_c} \sum_{\mathbf{x}_j \in \mathbf{A}_c} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \tag{10}$$

By introducing the cluster indicator matrix $\mathbf{F} \in \mathrm{Ind}$, which is a discrete matrix, each row representing a sample, where if the i-th sample $\mathbf{x}_i$ belongs to the c-th cluster $\mathbf{A}_c$, then $\mathbf{F}_{ic} = 1$, otherwise $\mathbf{F}_{ic} = 0$. Since the result of K-Means clustering is that each sample belongs to only one cluster, resulting in hard labels, the matrix $\mathbf{F} \in \mathrm{Ind}$ has only one element equal to 1 in each row, with the rest of the elements being 0. Therefore, the inner product of the cluster labels for two samples $\langle \mathbf{f}_i, \mathbf{f}_j \rangle = 1$ only when the two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same cluster, and $\langle \mathbf{f}_i, \mathbf{f}_j \rangle = 0$ when the samples $\mathbf{x}_i$ and $\mathbf{x}_j$ do not belong to the same cluster. Denoting the union of clusters as $\mathbf{A} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K\}$, the Eq.(10) can be rewritten as:

$$\min_{\mathbf{A}, \mathbf{F}} \sum_{i=1}^{n} \sum_{j=1}^{n} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \langle \mathbf{f}_i, \mathbf{f}_j \rangle \text{ s.t. } \mathbf{F} \in \mathrm{Ind}. \tag{11}$$

The above formula shows that this is another form of K-Means clustering that does not require initializing cluster centroids. This reduces the impact of initial cluster centroids on the results. Instead of calculating the distance between samples and cluster centroids, it uses the distance between two samples. By avoiding the computation of cluster means, it mitigates the influence of outliers on the cluster centroids, resulting in more robust clustering results. If K-Means clustering is performed in the subspace $\mathbf{Y} = \mathbf{W}^T \mathbf{X}$, then Formula (11) can be transformed as follows:

$$\min_{\substack{\mathbf{F} \in \mathrm{Ind}, \\ \mathbf{W}^T \mathbf{W} = \mathbf{I}}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j) \right\|_2^2 \langle \mathbf{f}_i, \mathbf{f}_j \rangle \tag{12}$$

$$\langle \mathbf{f}_i, \mathbf{f}_j \rangle = \begin{cases} 1 & , \quad \mathbf{x}_i \in \mathbf{A}_c \text{ and } \mathbf{x}_j \in \mathbf{A}_c \\ 0 & , \quad \text{others} \end{cases}$$

Subsequently, we search for the k nearest neighbors of each sample within the obtained clusters. When sample $\mathbf{x}_i$ and $\mathbf{x}_j$ are k nearest neighbors of of each other, we set $\mathbf{S}_{ij} = 1$; otherwise, $\mathbf{S}_{ij} = 0$. Accordingly, $\mathbf{S}_{ij} \cdot \langle \mathbf{f}_i, \mathbf{f}_j \rangle = 1$ indicates that sample $\mathbf{x}_i$ is both a k nearest neighbor of sample $\mathbf{x}_j$ and belongs to the same cluster. This leads us to the following expression:

$$\min_{\substack{\mathbf{F} \in \mathrm{Ind}, \\ \mathbf{W}^T \mathbf{W} = \mathbf{I}}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j) \right\|_2^2 \cdot \mathbf{S}_{ij} \cdot \langle \mathbf{f}_i, \mathbf{f}_j \rangle \tag{13}$$

$$\mathbf{S}_{ij} \cdot \langle \mathbf{f}_i, \mathbf{f}_j \rangle = \begin{cases} 1 & , \quad \begin{array}{l} \mathbf{x}_i, \mathbf{x}_j \in \mathbf{A}_c, \mathbf{x}_j \in N_k(\mathbf{x}_i) \\ \mathbf{x}_i \in N_k(\mathbf{x}_j) \end{array} \\ 0 & , \quad \text{otherwise} \end{cases}$$

Let $\mathbf{S}_{ij} \cdot \langle \mathbf{f}_i, \mathbf{f}_j \rangle = \tilde{\mathbf{S}}_{ij}$. The equation above represents the objective function (3) for optimizing the eigenmaps in MFA. Without using true labels, we can generate pseudo-labels through K-Means clustering and perform dimensionality reduction, obtaining both the cluster indicator matrix $\mathbf{F} \in \mathrm{Ind}$ and the projection matrix $\mathbf{W}$. Similarly, $\mathbf{S}_{ij} \cdot (1 - \langle \mathbf{f}_i, \mathbf{f}_j \rangle) = 1$ indicates that sample $\mathbf{x}_i$ is a k nearest neighbor of sample $\mathbf{x}_j$, but they do not belong to the same cluster. Let $\mathbf{S}_{ij} \cdot (1 - \langle \mathbf{f}_i, \mathbf{f}_j \rangle) = \tilde{\mathbf{S}}_{ij}^p$, and we obtain the objective function for penalizing the graph in MFA:

$$\max_{\substack{\mathbf{F} \in \mathrm{Ind}, \\ \mathbf{W}^T \mathbf{W} = \mathbf{I}}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j) \right\|_2^2 \cdot \mathbf{S}_{ij} \cdot (1 - \langle \mathbf{f}_i, \mathbf{f}_j \rangle) \tag{14}$$

$$\mathbf{S}_{ij} \cdot (1 - \langle \mathbf{f}_i, \mathbf{f}_j \rangle) = \begin{cases} 1 & , \quad \begin{array}{l} \mathbf{x}_i \in \mathbf{A}_c, \mathbf{x}_j \notin \mathbf{A}_c, \\ \mathbf{x}_i \in N_k(\mathbf{x}_j), \mathbf{x}_j \in N_k(\mathbf{x}_i) \end{array} \\ 0 & , \quad \text{others} \end{cases}$$

Hence, the objective function for our Label-Free MFA can be defined as follows:

$$\max_{\substack{\mathbf{W}^T \mathbf{W} = \mathbf{I} \\ \mathbf{F} \in \mathrm{Ind}}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} ||\mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j)||_F^2 \cdot \mathbf{S}_{ij} \cdot (1 - \langle \mathbf{f}_i, \mathbf{f}_j \rangle)}{\sum_{i=1}^{n} \sum_{j=1}^{n} ||\mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j)||_F^2 \cdot \mathbf{S}_{ij} \cdot \langle \mathbf{f}_i, \mathbf{f}_j \rangle} \tag{15}$$

Optimization

We optimize and solve the variables $\mathbf{F} \in \mathrm{Ind}$ and $\mathbf{W}$ iteratively. The specific solving process can be divided into the following two steps:

Fix W and solve F: When $\mathbf{W}$ is fixed, the distances between samples in the subspace are determined. To optimiz $\mathbf{F}$, it is necessary to obtain the pairwise distance between all the samples. Therefore, we construct $\mathbf{S}$ as a fully-connected graph by setting $\mathbf{S}_{ij} = 1$ for $1 \leq i, j \leq n$. Thus, for the intra-class distances, equation (13) transforms into:

$$\min_{\mathbf{F}} \sum_{i=1}^{n}\sum_{j=1}^{n} \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^2 \cdot \langle \mathbf{f}_i, \mathbf{f}_j\rangle \quad \text{s.t. } \mathbf{F} \in \mathrm{Ind}. \tag{16}$$

Theorem 2 *If we define each element of the distance matrix* $\mathbf{Q}$ *as* $\mathbf{Q}_{ij} = \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^2$*, then*

$$\begin{aligned} &\min_{\mathbf{F}\in\mathrm{Ind}} \sum_{i=1}^{n}\sum_{j=1}^{n} \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^2 \cdot \langle \mathbf{f}_i, \mathbf{f}_j\rangle \\ &= \min_{\mathbf{F}\in\mathrm{Ind}} \mathrm{tr}(\mathbf{F}^T\mathbf{Q}\mathbf{F}). \end{aligned} \tag{17}$$

Proof 2 *Expanding the left side of Eq.(17) as follows:*

$$\begin{aligned} &\sum_{i=1}^{n}\sum_{j=1}^{n} \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^2 \langle \mathbf{f}_i, \mathbf{f}_j\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{tr}(\mathbf{Q}_{ij} \langle \mathbf{f}_i, \mathbf{f}_j\rangle) \\ &= \sum_{i=1}^{n} \mathrm{tr}(\mathbf{Q}_{ij}\mathbf{F}^T\mathbf{f}_i) = \mathrm{tr}(\mathbf{F}^T\mathbf{Q}\mathbf{F}) \end{aligned} \tag{18}$$

Hence, equation (17) holds.

According to Theorem 2, solving (16) is equivalent to solving:

$$\min_{\mathbf{F}} \mathrm{tr}(\mathbf{F}^T\mathbf{Q}\mathbf{F}) \quad \text{s.t. } \mathbf{F} \in \mathrm{Ind} \tag{19}$$

For inter-class distances, similarly, when $\mathbf{W}$ is fixed, Eq.(14) becomes:

$$\begin{aligned} &\max_{\mathbf{F}\in\mathrm{Ind}} \sum_{i=1}^{n}\sum_{j=1}^{n} \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^2 \cdot (1 - \langle \mathbf{f}_i, \mathbf{f}_j\rangle) \\ &\Rightarrow \min_{\mathbf{F}\in\mathrm{Ind}} \sum_{i=1}^{n}\sum_{j=1}^{n} \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^2 \langle \mathbf{f}_i, \mathbf{f}_j\rangle \end{aligned} \tag{20}$$

From the above derivation, according to Theorem 2, it can be concluded that solving the variable $\mathbf{F} \in \mathrm{Ind}$ is equivalent to minimizing $\mathrm{tr}(\mathbf{F}^T\mathbf{Q}\mathbf{F})$.

Theorem 3 *Since the cluster indicator matrix* $\mathbf{F} \in \mathrm{Ind}$ *is a discrete matrix with only one element as 1 in each row and the remaining elements as 0, and each row's elements are mutually independent, it is challenging to solve it directly. Therefore, we need to fix the other rows and solve it row by row. For the distance matrix* $\mathbf{G}$*, where each element* $\mathbf{Q}_{ij} = \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^2$*, the optimal solution for the elements* $\mathbf{F}_{ic}$ *of* $\mathbf{F} \in \mathrm{Ind}$ *in minimizing* $\mathrm{tr}(\mathbf{F}^T\mathbf{Q}\mathbf{F})$ *is:*

$$\mathbf{F}_{ic} = \begin{cases} 1, & c = \arg\min\limits_{c} (\mathbf{F}^T\mathbf{q}_i)_c \\ 0, & \text{others} \end{cases} \tag{21}$$

Algorithm 1: Discriminative Graph Embedding Framework via Label-Free Marginal Fisher Analysis

Input: A set of input data $\mathbf{X}\in\mathbb{R}^{d\times n}$; cluster number K, λ.
Output: Projection matrix $\mathbf{W} \in \mathbb{R}^{d\times t}$, cluster indication matrix $\mathbf{F} \in \mathbb{R}^{n\times K}$.

1: Initialize $\mathbf{W} \in \mathbb{R}^{d\times t}$, $\mathbf{F} \in \mathbb{R}^{n\times K}$ and adjancency matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$.
2: **repeat**
3: Update the distance matrix $\mathbf{Q}$;
4: Update $\mathbf{F}$ by solving Eq. (21);
5: Update $\mathbf{S}_{ij} = \begin{cases} 1 & , \ \mathbf{x}_i \in N_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_k(\mathbf{x}_i) \\ 0 & , \ \text{others} \end{cases}$;
6: Update $\tilde{\mathbf{S}}_{ij}$ according to Eq. (13);
7: Update $\tilde{\mathbf{S}}^p_{ij}$ according to Eq. (14);
8: Update $\mathbf{W}$ by performing eigenvalue decomposition on $\mathbf{X}(\tilde{\mathbf{L}}^p - \lambda\tilde{\mathbf{L}})\mathbf{X}^T$;
9: **until** $||\mathbf{F} - \mathbf{F}_{\text{old}}||_F^2 \leq 10^{-5}$
10: **return** $\mathbf{W}$ and $\mathbf{F}$

Proof 3 *When fixing the other rows and solving for the* i*-th row* $\mathbf{f}_i$ *of* $\mathbf{F} \in \mathrm{Ind}$*, we have:*

$$\min_{\mathbf{f}_i\in\mathrm{Ind}} \sum_{i,j} \mathbf{Q}_{ij}\mathrm{tr}(\mathbf{f}_i^T\mathbf{f}_j) \Leftrightarrow \min_{\mathbf{f}_i\in\mathrm{Ind}} \mathbf{f}_i(\sum_{i,j} \mathbf{Q}_{ij}\mathbf{f}_j^T) \tag{22}$$

By substituting $\mathbf{Q}_{ii} = 0$ *and letting* $\mathbf{q}_i = (\mathbf{Q}_{i1}, \mathbf{Q}_{i2}, \cdots, \mathbf{Q}_{in})^T$*, we have:*

$$\min_{\mathbf{f}_i\in\mathrm{Ind}} \mathbf{f}_i(\sum_{j\neq i} \mathbf{Q}_{ij}\mathbf{f}_j^T) \Leftrightarrow \min_{\mathbf{f}_i\in\mathrm{Ind}} \mathbf{f}_i(\mathbf{F}^T\mathbf{q}_i) \tag{23}$$

$$\Rightarrow \mathbf{F}_{ic} = \begin{cases} 1, & c = \arg\min\limits_{c} (\mathbf{F}^T\mathbf{q}_i)_c \\ 0, & \text{others} \end{cases} \tag{24}$$

Consequently, Theorem 3 is proved. The optimal solution for the cluster indicator matrix $\mathbf{F} \in \mathrm{Ind}$ *is given by Eq. (21).*

Fix F and solve for W:

According to the principles of graph embedding framework, for computational convenience, let $\tilde{\mathbf{S}}_{ij} = \mathbf{S}_{ij} \cdot \langle \mathbf{f}_i, \mathbf{f}_j\rangle$ and $\tilde{\mathbf{S}}^p_{ij} = \mathbf{S}_{ij} \cdot (1 - \langle \mathbf{f}_i, \mathbf{f}_j\rangle)$, then

$$\min_{\mathbf{W}} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{Q}_{ij}\tilde{\mathbf{S}}_{ij} = \min_{\mathbf{W}} \mathrm{tr}(\mathbf{W}^T\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T\mathbf{W}) \tag{25}$$

$$\max_{\mathbf{W}} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{Q}_{ij}\tilde{\mathbf{S}}^p_{ij} = \max_{\mathbf{W}} \mathrm{tr}(\mathbf{W}^T\mathbf{X}\tilde{\mathbf{L}}^p\mathbf{X}^T\mathbf{W}) \tag{26}$$

As a result, solving for $\mathbf{W}$ is equivalent to solving:

$$\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{\mathrm{tr}(\mathbf{W}^T\mathbf{X}\tilde{\mathbf{L}}^p\mathbf{X}^T\mathbf{W})}{\mathrm{tr}(\mathbf{W}^T\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T\mathbf{W})} \tag{27}$$

This is a classic trace ratio (TR) problem, Guo et al. (2003) proposed a trace difference method that converts the TR problem into $\max\limits_{\mathbf{W}} \mathrm{tr}[\mathbf{W}^T\mathbf{X}(\tilde{\mathbf{L}}^p - \lambda\tilde{\mathbf{L}})\mathbf{X}^T\mathbf{W}]$.

Then, we perform an eigenvalue decomposition on $\mathbf{X}(\tilde{\mathbf{L}}^p - \lambda\tilde{\mathbf{L}})\mathbf{X}^T$, and the obtained eigenvalues and their corresponding eigenvectors are sorted in descending order. We select the corresponding eigenvectors of the top t largest eigenvalues to construct the projection matrix $\mathbf{W}$.

Calculation of the Distance Matrix Q

The squared Euclidean distance between the samples $\mathbf{x}_i$ and $\mathbf{x}_j$ is calculated as $(\mathbf{Q}_{ij})_{\text{euc}} = \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|_2^2$.

1. Gaussian kernel distance: To handle nonlinear data, we map the original data $\mathbf{x}$ to a higher-dimensional space $\phi(\mathbf{x})$ to achieve linear separability. The Gaussian kernel function is defined as $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right)$. The calculation of the Gaussian kernel distance in the subspace $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$ is given by $(\mathbf{Q}_{ij})_{\text{ker}} = \|\phi(\mathbf{y}_i) - \phi(\mathbf{y}_j)\|_2^2 = K(\mathbf{y}_i, \mathbf{y}_i) - 2K(\mathbf{y}_i, \mathbf{y}_j) + K(\mathbf{y}_j, \mathbf{y}_j)$.

2. K-nearest neighbor distance:

$$(\mathbf{Q}_{ij})_{\text{knn}} = \begin{cases} \|\mathbf{y}_i - \mathbf{y}_j\|_2^2, & \mathbf{y}_i \in N_k(\mathbf{y}_j) \text{ or } \mathbf{y}_j \in N_k(\mathbf{y}_i) \\ \sigma, & \text{otherwise} \end{cases}$$

where σ is a large constant.

3. Butterworth distance:

By utilizing the system function of Butterworth filter on the input similarity matrix $\mathbf{T}$, we transform it into Butterworth distance: $(\mathbf{Q}_{ij})_{\text{btw}} = \sqrt{\frac{1}{1+\left(\frac{\mathbf{T}_{ij}}{\Omega}\right)^4}}$, where Ω is a hyperparameter. It brings similar data points closer and increases the distance between dissimilar data points.

Experiments

In this section, we conducted comprehensive experiments on two toy datasets and six benchmark datasets to compare the clustering performance of our method with that of ten compared approaches. "Ours(euc)", "Ours(ker)", "Ours(knn)", and "Ours(btw)" respectively represent our methods based on squared Euclidean distance, Gaussian kernel distance, K-nearest neighbor (KNN) distance, and Butterworth distance. These experiments are implemented on a Windows 11 desktop computer with a 2.60GHz 13th Gen Intel Core i7-13650HX CPU, 16 GB RAM, and MATLAB R2023a.

Experiments on the Toy Datasets

Noise Dataset: This dataset contains 43 samples belonging to two clusters. As shown in Figure 1(a), Cluster 1 contains 20 samples, while Cluster 2 consists of 23 samples, including 3 noise points. To evaluate the robustness of our method to noise, we compare it with the K-Means algorithm. The clustering result of K-Means is shown in Figure 1(c), and that of our method is shown in Figure 1(e).

It can be observed that K-Means produces incorrect clustering results due to its sensitivity to outliers and reliance on the Euclidean distance. In contrast, our method achieves accurate clustering. This is attributed to the use of the Butterworth distance, which assigns more appropriate distances to both similar and dissimilar data points, thereby effectively reducing the impact of noise on the clustering results.

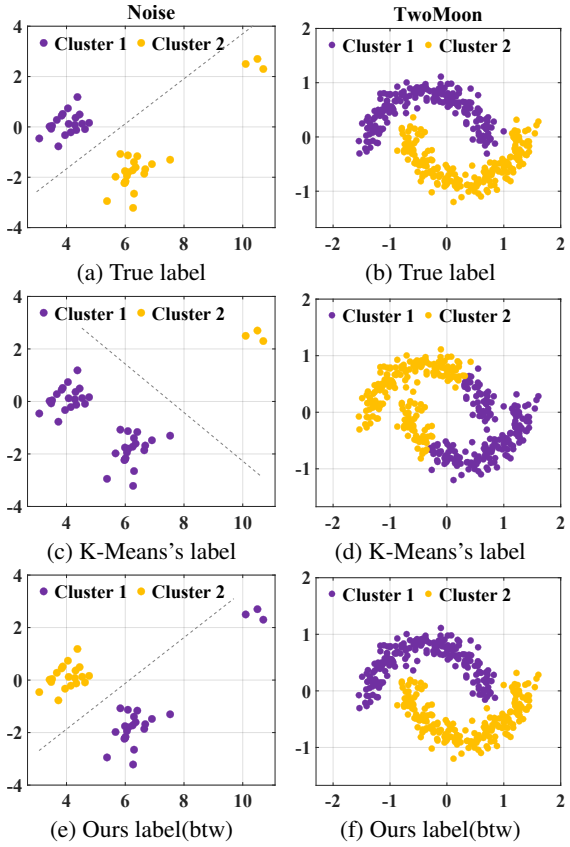


Figure 1: Clustering results of K-Means and our method on the Noise and TwoMoon Datasets.

TwoMoon Dataset: This dataset contains 400 samples with two clusters, as shown in Figure 1(b). Figure 1(d) shows the clustering results of K-Means on the moon-shaped dataset, and Figure 1(f) presents the clustering results of our method on the same dataset. As can be seen, in the case of complex and non-linearly separable data distribution, by combining the similarity graph and Butterworth distance, our algorithm can more effectively identify and extract the intrinsic structure of the data, leading to more accurate clustering. In contrast, K-Means struggles to achieve good clustering results due to its inability to handle non-linearly separable cases.

Experiments Settings on Benchmark Datasets

FaceV5 (Team 2009) contains 2,500 images across 500 face classes, utilizing 256 features. **JAFFE** (Lyons, Budynek, and Akamatsu 1999) includes 213 images representing various facial expressions, categorized into 10 classes, and employs 676 features. **MSRCV2** (Ali and Zafar 2018) comprises 210 images covering 7 object classes, with 576 features used. **ORL** (Cai, Zhang, and He 2010) consists of facial images from 400 samples across 40 individuals, utilizing 1,024 features. **USPS** (Hull 1994) consists of 3,000 handwritten digit images, categorized into 10 classes, and employs 256 features. **Yaleface** (Georghiades, Belhumeur,

Dataset	FaceV5			JAFFE			MSRCV2		
Metric	**ACC**	**NMI**	**Purity**	**ACC**	**NMI**	**Purity**	**ACC**	**NMI**	**Purity**
RKM	0.8540	0.9563	0.8640	0.8310	0.8159	0.8310	0.6286	0.5612	0.6333
Ksums	0.9676	0.9893	0.9709	0.8789	0.8764	0.8789	0.7524	0.6110	0.7524
Ksums-x	0.9620	0.9857	0.9659	0.8930	0.9013	0.8977	0.6852	0.5753	0.6910
K-Means	0.8422	0.9624	0.8796	0.7108	0.7981	0.7441	0.6657	0.5693	0.6795
CDKM	0.7633	0.9369	0.8114	0.7085	0.8010	0.7455	0.6052	0.5280	0.6276
PCA	0.7792	0.9366	0.8192	0.8873	0.9135	0.8873	0.6905	0.6026	0.7190
LPP	0.7640	0.9393	0.8116	0.9765	0.9740	0.9765	0.7238	0.6245	0.7238
LDA-Km	0.8096	0.9516	0.8404	0.9577	0.9484	0.9577	0.7286	0.5841	0.7286
Un-RTLDA	0.7644	0.9247	0.7864	0.9671	0.9625	0.9671	0.6714	0.5482	0.6905
Un-TRLDA	0.8196	0.9556	0.8512	0.9155	0.9225	0.9155	0.7524	0.6383	0.7524
Ours(euc)	0.9564	0.9827	0.9608	**0.9859**	**0.9816**	**0.9859**	**0.8143**	**0.6721**	**0.8143**
Ours(ker)	0.9624	0.9869	0.9644	**1.0000**	**1.0000**	**1.0000**	**0.8714**	**0.7572**	**0.8714**
Ours(knn)	0.9404	0.9793	0.9428	**1.0000**	**1.0000**	**1.0000**	**0.8667**	**0.7401**	**0.8667**
Ours(btw)	**0.9696**	**0.9903**	**0.9716**	**1.0000**	**1.0000**	**1.0000**	**0.8524**	**0.7273**	**0.8524**

Table 1: Clustering performance on FaceV5, JAFFE and MSRCV2 datasets.

Dataset	ORL			USPS			Yaleface		
Metric	**ACC**	**NMI**	**Purity**	**ACC**	**NMI**	**Purity**	**ACC**	**NMI**	**Purity**
RKM	0.5000	0.7143	0.5200	0.6673	0.5865	0.6827	0.4485	0.5099	0.4848
Ksums	0.6337	0.7940	0.6562	0.7664	0.6615	0.7677	0.4339	0.4975	0.4733
Ksums-x	0.5877	0.7693	0.6060	0.7240	0.6078	0.7240	0.4418	0.5018	0.4915
K-Means	0.5507	0.7529	0.6090	0.6491	0.6147	0.6803	0.3964	0.4779	0.4188
CDKM	0.5198	0.7234	0.5705	0.6416	0.6059	0.6746	0.3812	0.4389	0.4030
PCA	0.5525	0.7334	0.6025	0.7780	0.6670	0.7780	0.4727	0.5082	0.4727
LPP	0.5500	0.6917	0.5875	0.6857	0.6529	0.7223	0.4000	0.4298	0.4242
LDA-Km	0.5675	0.7463	0.6100	0.7590	0.6555	0.7590	0.4545	0.5084	0.4727
Un-RTLDA	0.6625	0.8159	0.6975	0.7437	0.6417	0.7437	0.4727	0.5210	0.4970
Un-TRLDA	0.6075	0.7723	0.6450	0.7913	0.6760	0.7913	0.4848	0.5265	0.5030
Ours(euc)	0.6475	0.7908	0.6725	0.7710	0.6346	0.7710	**0.5152**	**0.5670**	**0.5152**
Ours(ker)	0.6500	0.7985	0.6650	0.6637	0.5796	0.6637	**0.5212**	**0.5319**	**0.5212**
Ours(knn)	0.6600	0.8027	0.6775	**0.8843**	**0.8015**	**0.8843**	**0.5333**	**0.5427**	**0.5333**
Ours(btw)	**0.7250**	**0.8328**	**0.7375**	**0.8207**	**0.7631**	**0.8207**	**0.5636**	**0.5671**	**0.5697**

Table 2: Clustering performance on ORL, USPS, and Yaleface datasets.

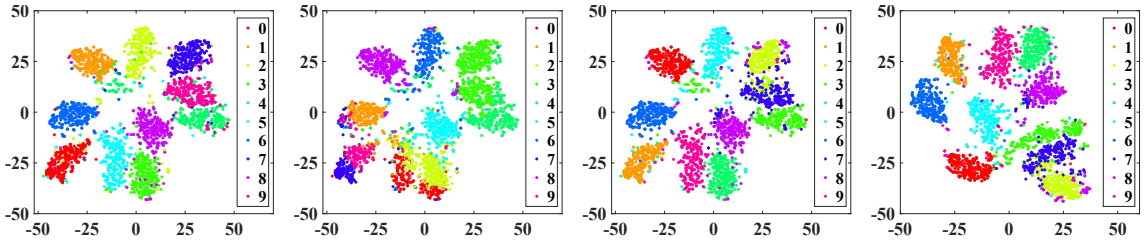
and Kriegman 1997) contains 165 grayscale images from 15 individuals, using 1,024 features.

Cpmparasion Methods: Five variants of K-Means: **K-Means** (Hartigan and Wong 1979), Regularized K-Means (**RKM**) (Lin, He, and Xiao 2019), **Ksums**(Pei et al. 2023), **Ksums-x** (Pei et al. 2023), Coordinate Descent Method for K-Means (**CDKM**) (Nie et al. 2022a); two methods that apply dimensionality reduction followed by K-Means: **PCA** (Turk and Pentland 1991) and **LPP** (He and Niyogi 2003); and three unsupervised LDA methods: **LDA-Km** (Ding and Li 2007), **Un-RTLDA** and **Un-TRLDA** (Wang et al. 2023).

Clustering Performance

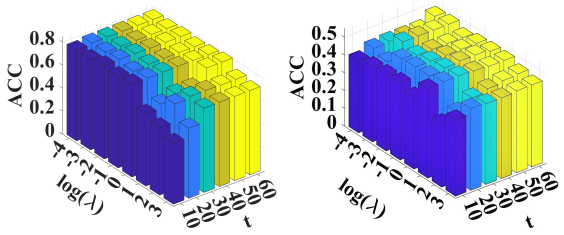
The clustering results of our method and the comparative approaches on the benchmark datasets are presented in Table 1 and Table 2. Among the five K-Means variants, Ksum and Ksum-x, which integrate K-Means with spectral clustering and avoid initialization and cluster centroids computation, achieve relatively strong performance, but are still affected by high-dimensional data.

Applying PCA or LPP before clustering helps improve performance on high-dimensional datasets. For instance, on the JAFFE dataset, LPP+K-Means achieves an ACC 10% higher than Ksums. However, these methods are limited by the decoupling of dimensionality reduction and clustering. LDA-Km, Un-RTLDA, and Un-TRLDA are unsupervised LDA-based methods that perform joint clustering and dimensionality reduction without labels. As shown in Table 2, they achieve better clustering performance but remain sensitive to cluster initialization and noise.



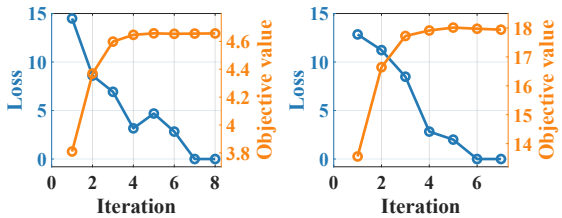
(a) True label (b) K-Means's label (c) Our label on original data (d) Our label on subspace

Figure 2: Clustering performance of our method and K-Means on the USPS dataset.



(a) MSRCV2 (b) Yaleface

Figure 3: ACC with parameter λ and dimension t on MSRCV2 and Yaleface datasets



(a) MSRCV2 (b) Yaleface

Figure 4: Objective function value and loss with the number of iterations on MSRCV2 and Yaleface datasets

Our method addresses these issues by replacing cluster centroids computation with a predefined distance matrix, unifying MFA and K-Means into a single framework. In particular, the Butterworth distance, when applied to a similarity graph, effectively handles non-linearly separable data.

T-SNE Visualization

The T-SNE visualizations of the clustering results of K-Means and our method on the USPS dataset are shown in Figure 2. Figure 2(a) presents the ground-truth labels of the USPS dataset, which contains 10 distinct clusters. Figure 2(b) shows the clustering result of K-Means, where multiple nearby but different clusters are mixed together. Figure 2(c) displays the clustering result of our method visualized in the original space, and Figure 2(d) shows the result in the embedding space. It can be observed that our method clearly separates different clusters, closely aligning with Figure 2(a). This demonstrates the effectiveness of our approach.

Parameter Analysis

In Figure3, we investigate the impact of the parameters λ and the reduced dimensionality on clustering performance ACC. The range of λ values is 0.0001, 0.001, 0.01, 0.05, 0.1, 1, 10, 100, 1000, while the dimensionality is selected as t = 10, 20, 30, 40, 50, 60. For the MSRCV2 and Yaleface datasets, under the same dimensionality t, the ACC exhibits an upward trend followed by a decline. The highest accuracy is achieved at λ = 0.05. However, for each dataset, different dimensions need to be selected to achieve the best clustering performance under the same λ parameter. This indicates that dimensionality reduction can improve clustering performance.

Convergence Analysis

The curve depicting the variation of the objective function value and loss $||\mathbf{F}-\mathbf{F}_{\text{old}}||_F^2$ with the number of iterations for our method is shown in Figure 4. As the number of iterations increases, the objective function value also increases while the loss decreases and eventually reaches a stable value. This indicates that our method is convergent and achieves convergence within a small number of iterations.

Conclusion

In this paper, we propose a novel discriminative graph embedding framework via MFA that can extract discriminative features without the need for data labels, while preserving the geometric manifold structure. Moreover, we establish the equivalence between MFA and K-Means, and integrate them into a unified graph embedding framework that directly yields the cluster indicator matrix and the projection matrix. Additionally, we generate the similarity graph required for MFA using the KNN graph and the cluster indicator matrix, thus eliminating the need for true labels and improving the discriminability of MFA. To further enhance robustness and clustering performance, we adopt a centerless K-Means approach, replacing the computation of cluster centroids with a predefined distance matrix. This avoids the influence of centroid initialization. Furthermore, to handle complex nonlinear separability data, we introduce four different distance matrix computation methods. Experimental results demonstrate the effectiveness and superiority of our method.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 62176203, the Fundamental Research Funds for the Central Universities (ZYTS25267, QTZX25004), and the Science and Technology Project of Xi'an (Grant 2022JH-JSYF-0009), Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (No. MMC202416), Selected Support Project for Scientific and Technological Activities of Returned Overseas Chinese Scholars in Shaanxi Province 2023-02.

References

Abdi, H.; and Williams, L. J. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4): 433–459.

Ali, N.; and Zafar, B. 2018. MSRC-v2 image dataset.

Ayesha, S.; Hanif, M. K.; and Talib, R. 2020. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, 59: 44–58.

Cai, D.; Zhang, C.; and He, X. 2010. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 333–342.

Chang, W.; Nie, F.; Wang, Z.; Wang, R.; and Li, X. 2022. Self-weighted learning framework for adaptive locality discriminant analysis. *Pattern Recognition*, 129: 108778.

Ding, C.; and Li, T. 2007. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the 24th international conference on Machine learning*, 521–528.

Fan, Y.; Liu, J.; Liu, P.; Du, Y.; Lan, W.; and Wu, S. 2021. Manifold learning with structured subspace for multi-label feature selection. *Pattern Recognition*, 120: 108169.

Fu, L.; Li, Z.; Ye, Q.; Yin, H.; Liu, Q.; Chen, X.; Fan, X.; Yang, W.; and Yang, G. 2020. Learning Robust Discriminant Subspace Based on Joint L2,p-and L2,s-Norm Distance Metrics. *IEEE transactions on neural networks and learning systems*, 33(1): 130–144.

Gao, Q.; Li, F.; Wang, Q.; Gao, X.; and Tao, D. 2025. Manifold Based Multi-View K-Means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 3175–3182.

Georghiades, A.; Belhumeur, P.; and Kriegman, D. 1997. Yale Face Database. http://cvc.yale.edu/projects/yale-face-database.

Guo, Y.-F.; Li, S.-J.; Yang, J.-Y.; Shu, T.-T.; and Wu, L.-D. 2003. A generalized Foley–Sammon transform based on generalized Fisher discriminant criterion and its application to face recognition. *pattern recognition letters*, 24(1-3): 147–158.

Hartigan, J. A.; and Wong, M. A. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1): 100–108.

He, X.; and Niyogi, P. 2003. Locality preserving projections. *Advances in neural information processing systems*, 16.

Hu, L.; Zhang, W.; and Dai, Z. 2021. Joint sparse locality-aware regression for robust discriminative learning. *IEEE Transactions on Cybernetics*, 52(11): 12245–12258.

Huang, Z.; Zhu, H.; Zhou, J. T.; and Peng, X. 2018. Multiple marginal fisher analysis. *IEEE Transactions on Industrial Electronics*, 66(12): 9798–9807.

Hull, J. J. 1994. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5): 550–554.

Kapoor, A.; and Singhal, A. 2017. A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms. In *international conference on computational intelligence & communication technology*, 1–6.

Lee, J. A.; and Verleysen, M. 2010. Unsupervised dimensionality reduction: Overview and recent advances. In *The International Joint Conference on Neural Networks*, 1–8.

Li, Z.; Nie, F.; Wu, D.; Hu, Z.; and Li, X. 2023. Unsupervised Feature Selection With Weighted and Projected Adaptive Neighbors. *IEEE Transactions on Cybernetics*, 53(2): 1260–1271.

Li, Z.; Nie, F.; Wu, D.; Wang, Z.; and Li, X. 2024. Sparse Trace Ratio LDA for Supervised Feature Selection. *IEEE Transactions on Cybernetics*, 54(4): 2420–2433.

Lin, W.; He, Z.; and Xiao, M. 2019. Balanced Clustering: A Uniform Model and Fast Algorithm. In *IJCAI*, 2987–2993.

Liu, M.; Feng, W.; Pei, C.; Wang, Q.; and Gao, Q. 2023a. Capped L2p-Norm Based 2DPCA with Adaptive Capped Threshold Learning for Image Processing. In *IEEE International Conference on Communication Technology*, 665–672.

Liu, X.; Wang, S.; Lu, S.; Yin, Z.; Li, X.; Yin, L.; Tian, J.; and Zheng, W. 2023b. Adapting feature selection algorithms for the classification of Chinese texts. *Systems*, 11(9): 483.

Lu, H.; Gao, Q.; Wang, Q.; Yang, M.; and Xia, W. 2023. Centerless multi-view K-means based on the adjacency matrix. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8949–8956.

Lu, H.; Xu, H.; Wang, Q.; Gao, Q.; Yang, M.; and Gao, X. 2024. Efficient Multi-View -Means for Image Clustering. *IEEE Transactions on Image Processing*, 33: 273–284.

Lu, J.; Lai, Z.; Wang, H.; Chen, Y.; Zhou, J.; and Shen, L. 2020. Generalized embedding regression: A framework for supervised feature extraction. *IEEE transactions on neural networks and learning systems*, 33(1): 185–199.

Lyons, M. J.; Budynek, J.; and Akamatsu, S. 1999. Automatic classification of single facial images. *IEEE transactions on pattern analysis and machine intelligence*, 21(12): 1357–1362.

Nie, F.; Wang, Z.; Wang, R.; Wang, Z.; and Li, X. 2020a. Adaptive local linear discriminant analysis. *ACM Transactions on Knowledge Discovery from Data*, 14(1): 1–19.

Nie, F.; Xue, J.; Wu, D.; Wang, R.; Li, H.; and Li, X. 2022a. Coordinate Descent Method for k-means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5): 2371–2385.

Nie, F.; Zhao, X.; Wang, R.; and Li, X. 2022b. Fast locality discriminant analysis with adaptive manifold embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9315–9330.

Nie, F.; Zhao, X.; Wang, R.; Li, X.; and Li, Z. 2020b. Fuzzy K-means clustering with discriminative embedding. *IEEE Transactions on Knowledge and Data Engineering*, 34(3): 1221–1230.

Pei, S.; Chen, H.; Nie, F.; Wang, R.; and Li, X. 2023. Centerless Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 167–181.

Pei, S.; Nie, F.; Wang, R.; and Li, X. 2020. Efficient clustering based on a unified view of k-means and ratio-cut. *Advances in Neural Information Processing Systems*, 33: 14855–14866.

Raju, K.; Chinna Rao, B.; Saikumar, K.; and Lakshman Pratap, N. 2022. An optimal hybrid solution to local and global facial recognition through machine learning. *A fusion of artificial intelligence and internet of things for emerging cyber systems*, 203–226.

Ran, R.; Feng, J.; Zhang, S.; and Fang, B. 2022. A General Matrix Function Dimensionality Reduction Framework and Extension for Manifold Learning. *IEEE Transactions on Cybernetics*, 52(4): 2137–2148.

Team, C. F. I. D. S. 2009. CAS Institute of Automation. http://biometrics.idealtest.org/.

Turk, M.; and Pentland, A. 1991. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1): 71–86.

Wang, J.; Xie, F.; Nie, F.; and Li, X. 2021a. Unsupervised adaptive embedding for dimensionality reduction. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11): 6844–6855.

Wang, Q.; Gao, Q.; Xie, D.; Gao, X.; and Wang, Y. 2016. Robust DLPP With Nongreedy ℓ_1-Norm Minimization and Maximization. *IEEE transactions on neural networks and learning systems*, 29(3): 738–743.

Wang, Q.; Wang, F.; Ren, F.; Li, Z.; and Nie, F. 2023. An Effective Clustering Optimization Method for Unsupervised Linear Discriminant Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35(4): 3444–3457.

Wang, Q.; Xia, W.; Tao, Z.; Gao, Q.; and Cao, X. 2021b. Deep self-supervised t-SNE for multi-modal subspace clustering. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1748–1755.

Wright, J.; and Ma, Y. 2022. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press.

Yan, S.; Xu, D.; Zhang, B.; and Zhang, H.-J. 2005. Graph embedding: A general framework for dimensionality reduction. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, 830–837. IEEE.

Yan, S.; Xu, D.; Zhang, B.; Zhang, H.-J.; Yang, Q.; and Lin, S. 2006. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1): 40–51.

Yang, C.; Ma, S.; and Han, Q. 2023. Unified discriminant manifold learning for rotating machinery fault diagnosis. *Journal of Intelligent Manufacturing*, 34(8): 3483–3494.

Yi, Z.; Shang, W.; Xu, T.; Guo, S.; and Wu, X. 2020. Local discriminant subspace learning for gas sensor drift problem. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(1): 247–259.

Zhou, Q.; Gao, Q.; Wang, Q.; Yang, M.; and Gao, X. 2023. Sparse discriminant PCA based on contrastive learning and class-specificity distribution. *Neural Networks*, 167: 775–786.