Soft Comput (2012) 16:1393–1409
DOI 10.1007/s00500-012-0843-3

ORIGINAL PAPER

Random walk-based fuzzy linear discriminant analysis for dimensionality reduction

Mingbo Zhao · Tommy W. S. Chow · Zhao Zhang

Published online: 30 March 2012

Abstract Dealing with high-dimensional data has always been a major problem with the research of pattern recognition and machine learning, and linear discriminant analysis (LDA) is one of the most popular methods for dimensionality reduction. However, it suffers from the problem of being too sensitive to outliers. Hence to solve this problem, fuzzy membership can be introduced to enhance the performance of algorithms by reducing the effects of outliers. In this paper, we analyze the existing fuzzy strategies and propose a new effective one based on Markov random walks. The new fuzzy strategy can maintain high consistency of local and global discriminative information and preserve statistical properties of dataset. In addition, based on the proposed fuzzy strategy, we then derive an efficient fuzzy LDA algorithm by incorporating the fuzzy membership into learning. Theoretical analysis and extensive simulations show the effectiveness of our algorithm. The presented results demonstrate that our proposed algorithm can achieve significantly improved results compared with other existing algorithms.

Keywords Discriminative Learning · Dimensionality reduction · Fuzzy strategy · Markov random walks

M. Zhao (✉) · T. W. S. Chow · Z. Zhang
Electronic Engineering Department, City University of Hong Kong, Kowloon, Hong Kong
e-mail: mzhao4@student.cityu.edu.hk

T. W. S. Chow
e-mail: eetchow@cityu.edu.hk

Z. Zhang
e-mail: zhaozhang5@student.cityu.edu.hk

1 Introduction

Dealing with high-dimensional data has always been a major problem with the research of pattern recognition and machine learning. Typical applications of these include face recognition, document categorization, and image retrieval. Finding a low-dimensional representation of high-dimensional space, namely dimensionality reduction is thus of great practical importance. The goal of dimensionality reduction is to reduce the complexity of input space and embed high-dimensional space into a low-dimensional space while keeping most of the desired intrinsic information (Tenenbaum et al. 2000; Roweis and Saul 2000). Among all the dimensionality reduction techniques, Linear Discriminant Analysis (Belhumeur et al. 1997) is the most popular method and has been widely used in many classification applications. In LDA, it uses the within-class scatter matrix $S_w$to evaluate the aggregation within each class and between-class scatter matrix S_b to evaluate the separability between different classes. The objective of LDA is then to find the optimal projection that maximizes the between-class scatter matrix while minimizes the within-class scatter matrix. Given that the within-class scatter matrix is nonsingular, the optimization problem of LDA can be solved by generalized eigenvalue decomposition (GEVD), i.e. to find the d largest eigenvectors corresponding to the eigenvalues of $S_w^{-1}S_b$ (Fukunaga 1990). However, for many applications where the number of dimensionality is much larger than that of samples, the within-class scatter matrix tends to be singular. Hence, the optimal projection matrix may be found correctly. This is the so-called small sample problem (Fukunaga 1990). To solve this problem, many variants of LDA have been proposed which include null space LDA (Chen et al. 2000), direct LDA (Yu and Yang 2001),

LDA/GSVD (Howland and Park 2004), LDA/QR (Ye and Li 2005). Another drawback of LDA is that it is sensitive to the outlier samples as these samples may have adverse influences on the calculation of scatter matrixes hence causes the classification performance degraded (Song et al. 2009). Most existing LDA algorithms suffer from this drawback as they are based on a binary (0–1) class membership which cannot estimate the outlierness of a sample in each class (Kwak and Pedrycz 2005). Thus, to solve this problem, it is reasonable to take advantages of fuzzy memberships to enhance the performance as the fuzzy membership can evaluate the importance or representativeness of a sample in each class. In this paper, we will focus on the second issue.

One of the most important issues is how to develop a reasonable fuzzy strategy and how to redefine the subsequent scatter matrixes based on the fuzzy membership. To date, there are many algorithms that combine different fuzzy strategy with conventional LDA algorithm (Kwak and Pedrycz 2005; Song et al. 2009). These algorithms choose a fuzzy strategy based on the notion of fuzzy k neighborhood classifier (Kwak and Pedrycz 2005). Song et al. (2009) proposed another fuzzy strategy based on a relaxed normalized condition which is an extension of the former one. However, both of these two fuzzy strategies can only keep the local discriminative information but overlook the global discriminative information. In this paper, we further analyze the drawbacks of the above two fuzzy strategies and propose a new efficient algorithm based on Markov random walks (Moonesignhe and Tan 2006; Wang and Davidson 2009; Liu et al. 2010). The new fuzzy strategy starts with the construction of a neighborhood graph that represents the local structure of dataset. It then performs random walk along the graph to seek the global discriminative information. As a result, high consistency of local and global discriminative information can be preserved.

Another drawback of the recently proposed fuzzy LDA is that the scatter matrixes cannot satisfy the relationship of $S_t = S_b + S_w$, as these reformulated scatter matrixes do not exactly follow the definitions (Kwak and Pedrycz 2005; Song et al. 2009). To solve this problem, we propose a new formulation of scatter matrixes which can satisfy the above equation and can be viewed as an extension to the conventional LDA algorithm. In addition, it is worth noting that based on the notion of graph Laplacian matrix, the fuzzy scatter matrixes can be reformulated using a pairwise form of matrix (He et al. 2005). We also analyze our proposed algorithm under a least square framework (Howland and Park 2004; Ye and Li 2005). It can be concluded that given a certain class scatter indicator, the optimization problem of our algorithm can be equivalent to a weighted least square problem under a mild condition.

The main contribution of this paper is summarized as follows:

1. A new fuzzy strategy, which is based on Markov random walks, is proposed in this paper. Compared with the previous work (Keller et al. 1985; Kwak and Pedrycz 2005; Song et al. 2009; Sun and Chen 2007), our proposed fuzzy strategy can maintain high consistency between local and global discriminative information. The stationary distribution obtained by performing random walk can preserve the statistical properties of dataset, making the outlier detection possible.
2. We propose new definitions of scatter matrixes based on fuzzy membership. The new definitions of scatter matrixes satisfy the equality of $S_t = S_b + S_w$ and can be reformulated under the notion of graph Laplacian. In addition, given a certain class indicator, our proposed fuzzy LDA can be solved under a least square framework (Ye 2007; Zhang et al. 2009).
3. As an extended LDA algorithm, our proposed method can keep the statistical properties of dataset and eliminate the effects of outliers by incorporating fuzzy membership into learning. This is beneficial to the performance of classification. The proposed algorithm can also be extended to solving a nonlinear problem using kernel tricks (Muller et al. 2001; Yang et al. 2005).

This paper is organized as follows: in Sect. 2, we firstly review the basic idea of LDA and propose our extended method based on fuzzy membership. In Sect. 3, we further analyze the optimization problem of our algorithm under a graph Laplacian view and least square view. In Sect. 4, we review some previous work for determining the fuzzy strategies and propose our strategy based on Markov random walks. In Sect. 5, we extended our algorithm to solving a nonlinear problem using kernel tricks. Simulation results are shown in Sect. 6 and the final conclusions are drawn in Sect. 7.

2 A complete fuzzy linear discriminant analysis

2.1 Review of linear discriminant analysis

The goal of LDA is to seek the optimal projection that maximize between-class scatter matrix while minimize with-class scatter matrix. Let $X = \{x_1, x_2, \ldots, x_l\} \in R^{D\times l}$ be the matrix of training set, each sample x_i is associated with a class label c_i from $\{1, 2, \ldots, c\}$, the within-class, between-class and total-class scatter matrix S_t, S_w, S_b can be defined as

$$S_t = \sum_{i=1}^{c} \sum_{x \in c_i} (x - \mu)(x - \mu)^T$$
$$S_w = \sum_{i=1}^{c} \sum_{x \in c_i} (x - \mu_i)(x - \mu_i)^T \quad (1)$$
$$S_b = \sum_{i=1}^{c} l_i (\mu_i - \mu)(\mu_i - \mu)^T$$

where l_i is the number of labeled set in ith class, $\mu_i = \sum_{x \in c_i} x / l_i$ is the mean of labeled set in ith class, c is the number of classes, and $\mu = \sum_{i=1}^{c} \sum_{x \in c_i} x / l_i$ is the mean of all labeled set. The goal of LDA is to find the optimal projection matrix W^* by solving the following optimization problem:

$$V^* = \arg\max_{V \in R^{D \times d}} |V^T S_b V| / |V^T S_w V| \text{ or}$$
$$V^* = \arg\max_{V \in R^{D \times d}} |V^T S_b V| / |V^T S_t V| \quad (2)$$

where W is a linear transformation satisfying $V{:}R^D \rightarrow R^d$. The optimal projection V^* in Eq. (2) is then formed by eigenvectors corresponding to the d largest eigenvalues of $S_w^{-1} S_b$ or $S_t^{-1} S_b$. Since the rank of S_b has a critical limitation of $c - 1$, there are at most $c - 1$ eigenvectors corresponding to non-zero eigenvalues (Fukunaga 1990). The output set in the reduced space can be obtained by $Z = V^{*T} X$.

2.2 Official definition of fuzzy linear discriminant analysis

In this subsection, we propose a new fuzzy LDA algorithm by incorporating fuzzy membership into learning. This approach can be viewed as an extension to the classical LDA algorithm. Let $W \in R^{c \times l}$ be the matrix with each element $w_{ij} \in R^+$ representing the fuzzy membership of jth sample in ith class, we then define the fuzzy within-class, between-class and total-class scatter matrix $\widetilde{S_w}, \widetilde{S_b}$ and $\widetilde{S_t}$ as

$$\widetilde{S_t} = \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} (x_j - \widetilde{\mu})(x_j - \widetilde{\mu})^T$$
$$\widetilde{S_w} = \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} (x_j - \widetilde{\mu}_i)(x - \widetilde{\mu}_i)^T \quad (3)$$
$$\widetilde{S_b} = \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} (\widetilde{\mu}_i - \widetilde{\mu})(\widetilde{\mu}_i - \widetilde{\mu})^T$$

where $\widetilde{\mu}_i = \sum_{j=1}^{l} w_{ij} x_j / \sum_{j=1}^{l} w_{ij}$ and $\widetilde{\mu} = \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} x_j / \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij}$. Let $e = \{1, 1\ldots, 1\} \in R^{1 \times l}$ be the unit vector, $F \in R^{i \times i}$ be the diagonal matrix satisfying $F_{ii} = \sum_{j=1}^{l} w_{ij}$, we then have

$$\begin{aligned}\widetilde{S_t} &= \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} (x_j - \widetilde{\mu})(x_j - \widetilde{\mu})^T \\ &= \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} x_j x_j^T - \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} \widetilde{\mu} \widetilde{\mu}^T \\ &= \sum_{j=1}^{l} x_j x_j^T - \frac{1}{l} \sum_{j=1}^{l} \sum_{k=1}^{l} x_j x_k^T \\ &= XX^T - \frac{1}{l} Xe^T eX^T \end{aligned} \quad (4)$$

$$\begin{aligned}\widetilde{S_w} &= \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} (x_j - \widetilde{\mu}_i)(x_j - \widetilde{\mu}_i)^T \\ &= \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} x_j x_j^T - \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} \widetilde{\mu}_i \widetilde{\mu}_i^T \\ &= \sum_{j=1}^{l} x_j x_j^T - \sum_{j=1}^{l} \sum_{k=1}^{l} \left(\sum_{i=1}^{c} \frac{w_{ij} w_{ik}}{F_{ii}} \right) x_j x_k^T \\ &= XX^T - XW^T F^{-1} WX^T \end{aligned} \quad (5)$$

$$\begin{aligned}\widetilde{S_b} &= \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} (\widetilde{\mu}_i - \widetilde{\mu})(\widetilde{\mu}_i - \widetilde{\mu})^T \\ &= \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} \widetilde{\mu}_i \widetilde{\mu}_i^T - \sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} \widetilde{\mu} \widetilde{\mu}^T \\ &= \sum_{j=1}^{l} \sum_{k=1}^{l} \left(\sum_{i=1}^{c} \frac{w_{ij} w_{ik}}{F_{ii}} \right) x_j x_k - \frac{1}{l} \sum_{j=1}^{l} \sum_{k=1}^{l} x_j x_k^T \\ &= XW^T F^{-1} WX^T - \frac{1}{l} Xe^T eX^T \end{aligned} \quad (6)$$

It can be easily confirmed that $\widetilde{S_t} = \widetilde{S_w} + \widetilde{S_b}$. In addition, for $w_{ij} \in \{0, 1\}$, the fuzzy scatter matrixes defined in Eqs. (4, 5, 6) can become conventional scatter matrixes as in Eq. (3). The objective function and solution of the proposed fuzzy LDA are

$$J(V) = \max_{V \in R^{D \times d}} |V^T \widetilde{S_b} V| / |V^T \widetilde{S_w} V| \text{ or}$$
$$J(V) = \max_{V \in R^{D \times d}} |V^T \widetilde{S_b} V| / |V^T \widetilde{S_t} V| \quad (7)$$
$$V_F^* = \widetilde{S_w}^{-1} \widetilde{S_b} \text{ or } V_F^* = \widetilde{S_t}^{-1} \widetilde{S_b}$$

3 Analysis of fuzzy LDA

3.1 Graph Laplacian view of fuzzy LDA

Based on the notion of graph Laplacian, the conventional scatter matrixes can be reformulate in a pairwise form (He et al. 2005; Chung 1997). Here, we extend this idea and analyze the fuzzy scatter matrixes using a graph Laplacian view. Let $\widetilde{A_w}, \widetilde{A_b}, \widetilde{A_t}$ be the adjacent matrix of $\widetilde{S_w}, \widetilde{S_b}, \widetilde{S_t}$ satisfying

$$
\begin{aligned}
(\widetilde{A}_{\mathrm{t}})_{ij} &= \frac{1}{l} \\
(\widetilde{A_{\mathrm{w}}})_{ij} &= \sum_{k=1}^{c} \frac{w_{ki} w_{kj}}{F_{kk}} \\
(\widetilde{A_{\mathrm{b}}})_{ij} &= \frac{1}{l} - \sum_{k=1}^{c} \frac{w_{ki} w_{kj}}{F_{kk}}
\end{aligned} \tag{8}
$$

we then have:

$$
\begin{aligned}
\widetilde{S}_{\mathrm{t}} &= X \widetilde{L}_{\mathrm{t}} X^{\mathrm{T}} \\
\widetilde{S_{\mathrm{w}}} &= X \widetilde{L_{\mathrm{w}}} X^{\mathrm{T}} \\
\widetilde{S}_{\mathrm{b}} &= X \widetilde{L_{\mathrm{b}}} X^{\mathrm{T}}
\end{aligned} \tag{9}
$$

where $\widetilde{L_{\mathrm{w}}} = \widetilde{D_{\mathrm{w}}} - \widetilde{A_{\mathrm{w}}}, \widetilde{L_{\mathrm{b}}} = \widetilde{D_{\mathrm{b}}} - \widetilde{A_{\mathrm{b}}}, \widetilde{L_{\mathrm{t}}} = \widetilde{D_{\mathrm{t}}} - \widetilde{A_{\mathrm{t}}}$ are the graph Laplacian matrix of $\widetilde{S_{\mathrm{w}}}, \widetilde{S}_{\mathrm{b}}, \widetilde{S}_{\mathrm{t}}$, and $\widetilde{D_{\mathrm{w}}}, \widetilde{D_{\mathrm{b}}}, \widetilde{D_{\mathrm{t}}}$ are the diagonal matrix satisfying $(\widetilde{D_{\mathrm{w}}})_{ii} = \sum_{j=1}^{l} (\widetilde{A_{\mathrm{w}}})_{ij}$, $(\widetilde{D_{\mathrm{b}}})_{ii} = \sum_{j=1}^{l} (\widetilde{A_{\mathrm{b}}})_{ij}, (\widetilde{D_{\mathrm{t}}})_{ii} = \sum_{j=1}^{l} (\widetilde{A_{\mathrm{t}}})_{ij}$. Therefore, the objective function and solution of Fuzzy LDA can be rewritten as

$$
\begin{aligned}
&J(V) = \max_{V \in R^{D \times d}} |V^{\mathrm{T}} X \widetilde{L_{\mathrm{b}}} X^{\mathrm{T}} V| / |V^{\mathrm{T}} X \widetilde{L_{\mathrm{w}}} X^{\mathrm{T}} V| \text{ or} \\
&J(V) = \max_{V \in R^{D \times d}} |V^{\mathrm{T}} X \widetilde{L_{\mathrm{b}}} X^{\mathrm{T}} V| / |V^{\mathrm{T}} X \widetilde{L}_{\mathrm{t}} X^{\mathrm{T}} V| \\
&V_F^* = (X \widetilde{L_{\mathrm{w}}} X^{\mathrm{T}})^{-1} X \widetilde{L_{\mathrm{b}}} X^{\mathrm{T}} \text{ or } V_F^* = (X \widetilde{L}_{\mathrm{t}} X^{\mathrm{T}})^{-1} X \widetilde{L_{\mathrm{b}}} X^{\mathrm{T}}
\end{aligned} \tag{10}
$$

3.2 Least square view of fuzzy LDA

Least square is another popular technique which has been widely used for regression and classification (Hastie et al. 2001). Let $T \in R^{c \times l}$ be a class indicator matrix, the goal of *LS* is then to fix a linear model $T^{\mathrm{T}} = X^{\mathrm{T}} V + b$, where $b \in R^{l \times c}$ is the bias term. Assuming that both data matrix X and class indicator are centered, the bias term then becomes zero and can be ignored. Hence, the objective function and solution of LS is given by

$$
\begin{aligned}
&\min_{V} \|T^{\mathrm{T}} - X^{\mathrm{T}} V\|_F^2 \\
&V_{\mathrm{LS}}^* = (X X^{\mathrm{T}})^{-1} X T^{\mathrm{T}}.
\end{aligned} \tag{11}
$$

Actually, given a certain class indicator matrix T, the conventional LDA can be equivalent to LS under a mild condition (Ye 2007; Zhang et al. 2009). In the section, we further extend this relationship and analyze our proposed fuzzy LDA under a least square framework. Assuming the samples in dataset are centered by $\widetilde{\mu}$, i.e. $\widetilde{x_j} = x_j - \widetilde{\mu}$ satisfying $\sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} \widetilde{x_j} = 0$, we then have $\widetilde{X} \widetilde{X}^{\mathrm{T}} = \widetilde{S}_{\mathrm{t}}$. In addition, since the class indicator matrix is of great importance for multi-class classification, we choose a class indicator matrix as

$$
T_{ij} = \frac{w_{ij}}{\sqrt{F_{ii}}} - \frac{\sqrt{F_{ii}}}{l}. \tag{12}
$$

It can be easily proved that $T^{\mathrm{T}} T = \widetilde{L_{\mathrm{b}}}$ (Corollary 1, see proof in Appendix A). Thus given $\widetilde{H_{\mathrm{b}}} = \widetilde{X} T^{\mathrm{T}}$, we then have $\widetilde{H_{\mathrm{b}}} \widetilde{H_{\mathrm{b}}}^{\mathrm{T}} = \widetilde{S_{\mathrm{b}}}$. Therefore, the optimal solution in Eq. (11) can be rewritten as $V_{\mathrm{LS}}^* = \widetilde{S}_{\mathrm{t}}^{-1} \widetilde{H_{\mathrm{b}}}$. Next we show V_F^* and V_{LS}^* can be equivalent given the following condition

$$
\text{Condition 1 (C1)} : \mathrm{Rank}(\widetilde{S}_{\mathrm{t}}) = \mathrm{Rank}(\widetilde{S_{\mathrm{w}}}) + \mathrm{Rank}(\widetilde{S_{\mathrm{b}}}). \tag{13}
$$

We thus have the theorem as

Theorem 1: Assuming C1 holds, then $V_{\mathrm{LS}}^* = V_F^*$.

The proof of Theorem 1 is given in Appendix C. But in most cases we cannot guarantee that the condition of C1 is satisfied. Hence in these cases, the original optimization problem of the proposed fuzzy LDA can be solved by employing a two-stage approach (Sun et al. 2010). The proof is also given in Appendix D.

4 Strategy for determining the fuzzy membership

4.1 Previous work

The conventional LDA algorithm is based on 0–1 class membership which cannot reflect the statistical properties of samples in each class. Hence, fuzzy membership can be used instead to enhance the performance of conventional algorithm. To design a fuzzy strategy, the sum-to-one constraint should be satisfied as

$$
\begin{aligned}
&0 < w_{ij} < 1 \\
&0 < \sum_{j=1}^{l} w_{ij} < l \\
&\sum_{j=1}^{l} w_{ij} = 1 \\
&\sum_{i=1}^{c} \sum_{j=1}^{l} w_{ij} = l.
\end{aligned} \tag{14}
$$

In addition, since the class information of one sample is likely to be hidden in its neighborhoods, especially when the sample is close to the boundary of different classes, one can describe this information of one sample by observing the class memberships of its neighborhoods. Actually, this idea has been around for a long time and can be dated back to the work of Keller et al. (1985) which is based on the notion of a fuzzy k nearest neighbor classifier (FKNN). The algorithm of FKNN can estimate the statistical properties of a sample in each class hence can calculate the fuzzy

membership and incorporate them into learning (Kwak and Pedrycz 2005; Song et al. 2009; Sun and Chen 2007). The fuzzy membership of FKNN is defined as

$$w_{ij} = \begin{cases} 0.51 + 0.49(n_{ij}/k) & x_j \text{ belong to the } i\text{th class} \\ 0.49(n_{ij}/k) & \text{otherwise} \end{cases}, \tag{15}$$

where n_{ij} represents the number of neighborhoods of x_j belonging to the ith class. k is the selected number of neighborhoods. It can be easily verified that $\sum_{i=1}^{c} w_{ij} = 1$. But since misclassification often occurs due to the existence of outliers, it is unwise to achieve imprecise fuzzy membership. To solve this problem, Song et al. (2009) has proposed another fuzzy strategy (RFKNN) by reformulate FKNN under the restriction of condition. The fuzzy membership of RFKNN is defined as

$$w_{ij} = \begin{cases} (1-\alpha) + \alpha(n_{ij}/k) & x_j \text{ belong to the } i\text{th class} \\ \alpha(n_{ij}/k) & \text{otherwise} \end{cases}, \tag{16}$$

where α $(0 < \alpha < 1)$ are the parameter controlling the value of w_{ij}. It can also be verified that $\sum_{j=1}^{l} w_{ij} = 1$. The fuzzy membership of RFKNN is empirically effective than that of FKNN (Kwak and Pedrycz 2005). Given a sample is an outlier, a fuzzy membership close to 0.5 means that the outlier exhibits influentially to several classes. By carefully adjusting the parameter α in RFKNN, the relative large fuzzy membership can be achieved, which results in eliminating the effects of the outliers. But in FKNN, the parameter is fixed 0.49 which is not flexible and adaptive when handling datasets of different distribution. Therefore, RFKNN can be viewed as an extension of FKNN.

4.2 Our proposed fuzzy strategy

The above two fuzzy strategies are proved to be able to enhance the conventional LDA algorithms (Kwak and Pedrycz 2005; Song et al. 2009), but they have their own drawbacks. First, they cannot keep the consistency of local and global discriminative information. In FKNN and RFKNN, the determination of fuzzy membership only considers the label information of neighborhoods of each sample while neglecting the other samples. This means FKNN and RFKNN can only preserve the local discriminative information, but neglect the global discriminative information. Second, both of the above two fuzzy strategies cannot preserve the density distribution of samples in each class. However in some circumstances, the density distribution can directly reflect the importance or representativeness of each sample in class. The outliers can also be detected using density distribution. To solve these problems, Markov random walks, a method that has been widely used in a variety of pattern recognition and machine learning applications (Moonesignhe and Tan 2006; Wang and Davidson 2009; Liu et al. 2010), can fulfill this goal. This method represents the training samples as a stochastic graph matrix and performs random walk along the path on graph to assess the importance or representativeness of each sample. Given the stochastic graph matrix, i.e. the matrix with each element denoting the one-step transition probability form x_i to x_j, stationary distribution can represent the density distribution of each sample in class. Therefore, the importance of each sample can be evaluated and outliers can be detected. We next show our proposed fuzzy strategy and in this paper, we denote it as MFKNN.

Let we define a Markov random walks based on an adjacent matrix. The adjacent matrix can be approximated by a neighborhood graph associated with weights on the edges. Officially, let $\widehat{G} = (\widehat{V}, \widehat{E})$ denote this graph, where $\widehat{V}$ is the vertex set of $\widehat{G}$ representing the training samples, $\widehat{E}$ is the edge set of $\widehat{G}$ associated with a weight matrix containing the local information between two nearby samples. A natural method to define the weight matrix using Gaussian function as

$$A_{kj} = \begin{cases} \exp(-\|x_k - x_j\|^2/\sigma) & x_k \in N_k(x_j) \\ 0 & \text{otherwise} \end{cases}, \tag{17}$$

where $N_k(x_j)$ is the neighborhood set of x_j. Let S be the transition matrix, it can then be easily formed by normalizing the adjacent matrix as $S = AD^{-1}$, D is a diagonal matrix satisfying $D_{jj} = \sum_{k=1}^{l} A_{kj}$. We next consider an iterative process to calculate w_{ij}. In each iteration, the label information of x_j is partially received from its neighborhoods and the rest is received from its initial label. Hence, it is reasonable to let w_{ij} be

$$w_{ij}(t+1) = (1-\alpha) \sum_{x_k \in N_k(x_j)} w_{ik}(t) s_{kj} + \alpha y_{ij}, \tag{18}$$

where $Y \in R^{c \times l}$ is the matrix with each element $y_{ij} \in (0, 1)$ representing the exact membership of x_j belonging to ith class, $\alpha(0 < \alpha < 1)$ is a parameter balancing the tradeoff of label information received from x_j and its neighborhoods. We then reformulate Eq. (18) as

$$W(t+1) = (1-\alpha)W(t)S + \alpha Y \tag{19}$$

Based on Eq. (19) and let $W(0) = Y$, we then have

$$W(t+1) = Y((1-\alpha)S)^{t+1} + \alpha Y \sum_{i=0}^{t} ((1-\alpha)S)^i. \tag{20}$$

According to the properties of matrix, i.e. $\lim_{t\to\infty}((1-\alpha)S)^{t+1} = 0$, $\lim_{t\to\infty} \sum_{i=0}^{t}((1-\alpha)S)^i = (I - (1-\alpha)S)^{-1}$, the iteration process converges to

$$W = \lim_{t\to\infty} W(t) = \alpha Y(I - (1-\alpha)S)^{-1}, \tag{21}$$

Table 1 Algorithms of MF-LDA

1. Calculate the fuzzy membership by MFKNN
2. Formulate $\widetilde{S_b}, \widetilde{S_w}, \widetilde{S_t}$ according to Eqs. (4, 5, 6) or Eqs. (8, 9)
3. Obtain the optimal projection matrix V^* by solving generalized eigen-value decomposition (GEVD) of $\widetilde{S_w}^{-1}\widetilde{S_b}$ or $\widetilde{S_t}^{-1}\widetilde{S_b}$
4. Output V^*

where I is an identity matrix having the same size of S. Therefore, we can either iteratively calculate the fuzzy membership based on Eq. (20) or obtain it directly from Eq. (21). In addition, it can be easily proved that the sum of each column of W is equal to 1 (Corollary 2, see proof in Appendix B). This indicates that the elements in W are probability values, and w_{ij} can be seen as the posterior probability of x_j belonging to the ith class. Therefore, our proposed fuzzy strategy can reflect the importance or representativeness of each sample and the outliers can be detected.

4.3 Relation to FKNN and RFKNN

We next prove that MFKNN can be viewed as an extension to FKNN and RFKNN given a certain weight matrix. Recalling the iteration process in Eq. (18), if we simply define the weight of graph as $A_{kj} = 1$, $x_k \in N_k(x_j)$ and $A_{kj} = 0$, otherwise, the transition matrix S can then be formed as $S_{kj} = 1/k$, $x_k \in N_k(x_j)$ and $A_{kj} = 0$, otherwise. Let $W(0) = Y$ and we only consider the one-step results of MFKNN, we then have

$$w_{ij}(1) = \begin{cases} (1-\alpha) + \alpha \sum_{x \in N_k(x_j)} (1/k) y_{ij} = (1-\alpha) + \alpha\, n_{ij}/k \\ x_j \text{ belong to the } i\text{th class} \\ \alpha \sum_{x \in N_k(x_j)} (1/k) y_{ij} = \alpha\, n_{ij}/k \quad \text{otherwise} \end{cases}, \tag{22}$$

which is exactly the fuzzy strategies defined in RFKNN and FKNN ($\alpha = 0.49$). Therefore, FKNN and RFKNN can be viewed as one-step results of MFKNN.

Until here, we have reviewed some previous work for determining the fuzzy membership and have proposed a new effective strategy. By incorporating the fuzzy membership learned by MFKNN into Eqs. (4, 5, 6), we can then form our proposed fuzzy LDA algorithms (we denote it as MF-LDA). The basic steps of algorithms are shown in Table 1.

5 Kernelization

The proposed fuzzy LDA is a linear algorithm. In this section, we extend it to solve nonlinear problem using kernel trick (Muller et al. 2001; Yang et al. 2005). For convenience, we denote the kernel version of MF-LDA as MF-KDA.

Table 2 Algorithms of MF-KDA

1. Calculate the fuzzy membership by MFKNN
2. Calculate kernel induced scatter matrixes $K\widetilde{L_t}K$ or $K\widetilde{L_w}K$ and $K\widetilde{L_b}K$
3. Obtain the optimal projection matrix β^* by solving generalized eigen-value decomposition (GEVD) of $(K\widetilde{L_w}K)^{-1}K\widetilde{L_b}K$ or $(K\widetilde{L_t}K)^{-1}K\widetilde{L_b}K$
4. Output β^*

The basic idea of kernel trick is to map the original data space to a high-dimensional Hilbert space as $\phi : X \to F$, then perform linear dimensionality reduction on the new space. Denote $\phi(X) = \{\phi(x_1), \phi(x_2), \ldots \phi(x_l)\}$ be such high-dimensionality space, we assume the new space can be implicitly implemented in a kernel function as $K(x_i, x_j) = \phi(x_i)^{\mathrm{T}}\phi(x_i)$. The goal of MF-KDA is then to find an optimal projection $V^{\phi^*} \in R^{l\times d}$ satisfying

$$\begin{aligned} J(V^\phi) &= \max_{V^\phi \in R^{l\times d}} |V^{\mathrm{T}}\phi(X)\widetilde{L_b}\phi(X)^{\mathrm{T}}V| / |V^{\mathrm{T}}\phi(X)\widetilde{L_w}\phi(X)^{\mathrm{T}}V| \text{ or} \\ J(V^\phi) &= \max_{V^\phi \in R^{l\times d}} |V^{\mathrm{T}}\phi(X)\widetilde{L_b}\phi(X)^{\mathrm{T}}V| / |V^{\mathrm{T}}\phi(X)\widetilde{L_t}\phi(X)^{\mathrm{T}}V|. \end{aligned} \tag{23}$$

Note $\phi(X)$ is not available as it is only implicit. Hence, we cannot directly solve the problem in Eq. (23). To compute the optimal projection V^ϕ, we need to add some restrict to V^ϕ making the solution in Eq. (23) available. According to Representer theorem (Fukunaga 1990), we assume the projection lies in the span of $\phi(X)$

$$V^\phi = \phi(X)\beta, \tag{24}$$

where $\beta \in R^{l\times d}$ is the matrix representing the contribution of kernel-reduced space to the columns of V^ϕ, the objective function and optimal projection of MF-KDA can then be given as

$$\begin{aligned} J(\beta) &= \max_\beta |\beta^{\mathrm{T}}K\widetilde{L_b}K\beta| / |\beta^{\mathrm{T}}K\widetilde{L_w}K\beta| \text{ or} \\ J(\beta) &= \max_\beta |\beta^{\mathrm{T}}K\widetilde{L_b}K\beta| / |\beta^{\mathrm{T}}K\widetilde{L_t}K\beta| \\ \beta^*_{KS-LDA} &= (K\widetilde{L_w}K)^{-1}K\widetilde{L_b}K \text{ or} \\ \beta^*_{\mathrm{KS-LDA}} &= (K\widetilde{L_t}K)^{-1}K\widetilde{L_b}K \end{aligned} \tag{25}$$

Thus, the output in the reduced space can be given by $V^{\phi\mathrm{T}}\phi(X) = \beta^{*\mathrm{T}}_{\mathrm{KS-LDA}}K$. The basic steps of MF-LDA are in Table 2.

6 Simulations

In this section, we evaluate our algorithms with one synthetic dataset and several real-world datasets. For the

synthetic dataset, we evaluate our algorithm using a two-moon dataset. For real-world datasets, we focus on solving three pattern recognition problems (face recognition, object recognition and handwritten digit recognition) based on *UMIST* dataset (Graham and Allinson 1998), *COIL-20* dataset (Nene et al. 1996) and *USPS* dataset (Hull 1994). Furthermore, we compare our algorithm with other state-of-art algorithms such as PCA (Turk and Pentland 1991), LPP (He et al. 2005), LDA (Belhumeur et al. 1997), F-LDA (Kwak and Pedrycz 2005) and RF-LDA (Song et al. 2009).

6.1 Toy example for synthetic dataset

6.1.1 Two-moon dataset

In this toy example, we generate a dataset with two classes, each of which follows a half-moon distribution. Since the distribution of two-moon dataset is non-Gaussian, we only perform kernel version of our algorithm for evaluation. Figure 1 shows the gray images of decision surfaces learned by MF-KDA with different values of parameter α. The gray value of each pixel represents the distance difference from the pixel to its nearest samples in different classes after dimensionality reduction by MF-KDA. In this example, we set the dimensionality of projection as 1. The values of parameter α are set as 0.8, 0.5, 0.2 in Fig. 1b–d, respectively. From Fig. 1, we can see that for two-moon dataset, the decision surface learned by MF-KDA appropriately adjusts as the parameter α decreases from 0.8 to 0.2, and the best decision surface can be achieved when α is set 0.2. It can also be shown from Fig. 1a and d that MF-KDA is better than KDA as the dark area belonging to blue class learned by MF-KDA ($\alpha = 0.2$) is more distinctive than that learned by KDA. This indicates that the our proposed fuzzy KDA algorithm is more effective than KDA because the fuzzy membership can directly preserve the density distribution embedded in the dataset.

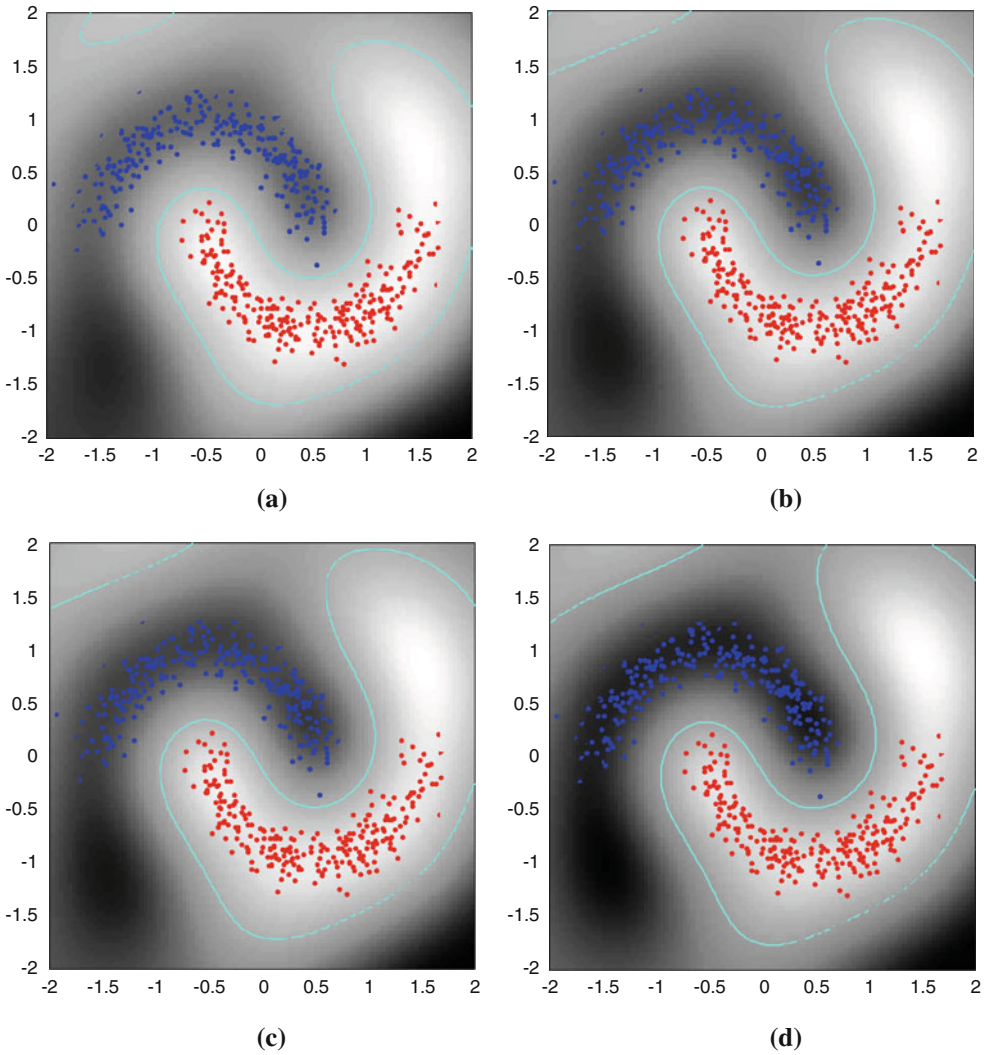


Fig. 1 Gray images of decision surfaces learned by MF-KDA with different values of parameter α: two-moon dataset. **a** $\alpha = 1$ (KDA), **b** $\alpha = 0.8$, **c** $\alpha = 0.5$, **d** $\alpha = 0.2$

Furthermore, to evaluate the effectiveness of our algorithm, we generate a sparse half-moon dataset, in which each class follows a sparse half-moon distribution and the boundary between two classes are more confused and overlapped. We then investigate the effectiveness of different algorithms based on such dataset. Figure 2 shows the gray images of decision surfaces learned by KDA, F-KDA, RF-KDA and MF-KDA. The parameter α in RF-KDA and MF-KDA is set to 0.2 in both. From the results, we can see that the performances of F-KDA, RF-KDA and MF-KDA are more effective than that of conventional KDA as the decision area belonging to different classes are more distinctive and accurate. The reason for it is that the fuzzy KDA can incorporate the discriminative information embedded in neighborhoods into learning, hence improve the classification performance. In addition, from Fig. 2b–d we can see that the result of MF-KDA is better than those of F-KDA and RF-KDA. The dark area belonging to blue class in MF-KDA is much more distinguished than that in F-KDA and RF-KDA. This is because F-KDA and RF-KDA only consider the local discriminative information of dataset, while MF-KDA can keep both local and global discriminative information by performing Markov random walks. Thus, the accuracy of classification can be significantly improved.

6.1.2 *Two-line dataset*

Finally, in the third toy example, we generate a two-line dataset, in which each class follows a line Gaussian distribution. We then add some outliers in the original dataset (seen in Fig. 3a) and investigate how our proposed algorithms can reduce the effects of outliers. Figure 3b shows the boundaries learned by PCA, LPP, LDA, F-LDA, RF-LDA and our proposed MF-LDA. In this simulation, the dimensionality of projections is set 1. From the simulation results, we can see that all supervised algorithms such as LDA, F-LDA, RF-LDA, MF-LDA are superior to the

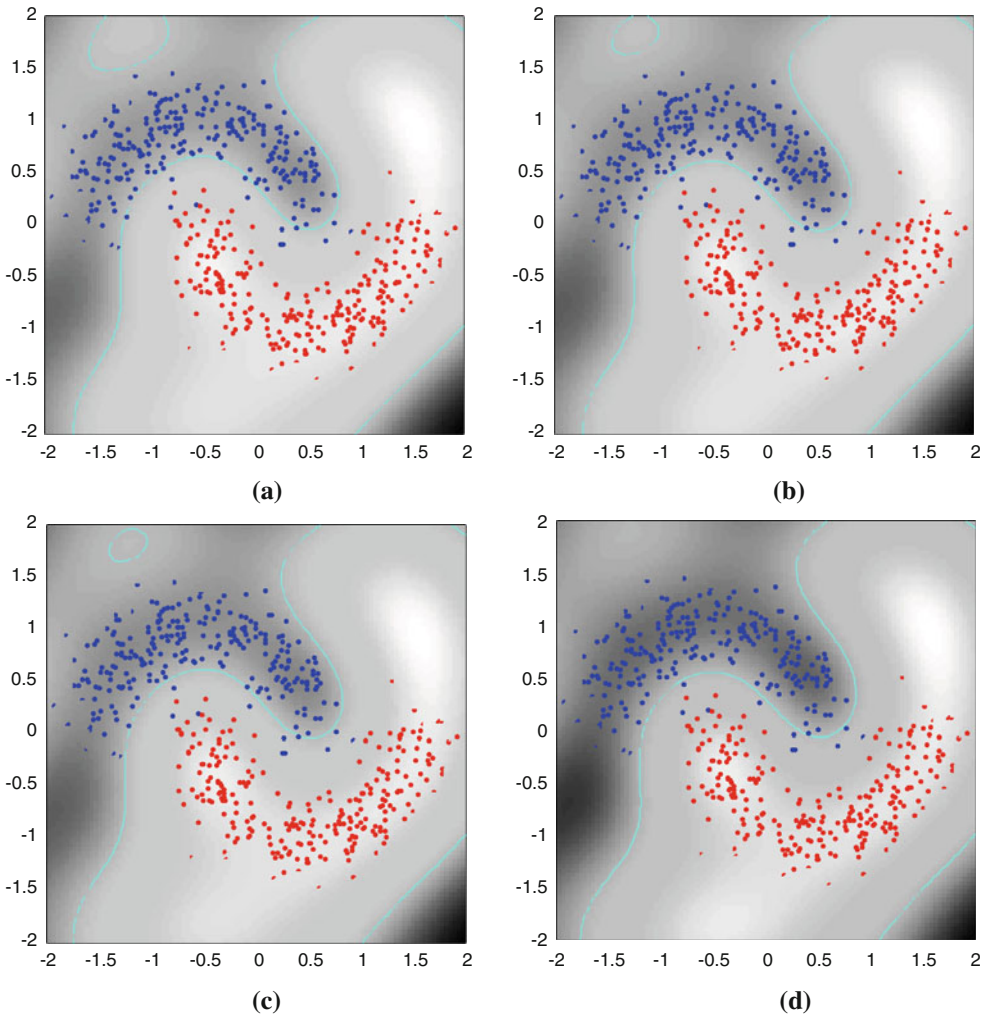


Fig. 2 Gray images of decision surfaces learned by KDA, F-KDA, RF-KDA and MF-KDA: sparse two-moon dataset. **a** KDA, **b** F-KDA, **c** RF-KDA ($\alpha = 0.2$), **d** MF-KDA ($\alpha = 0.2$)

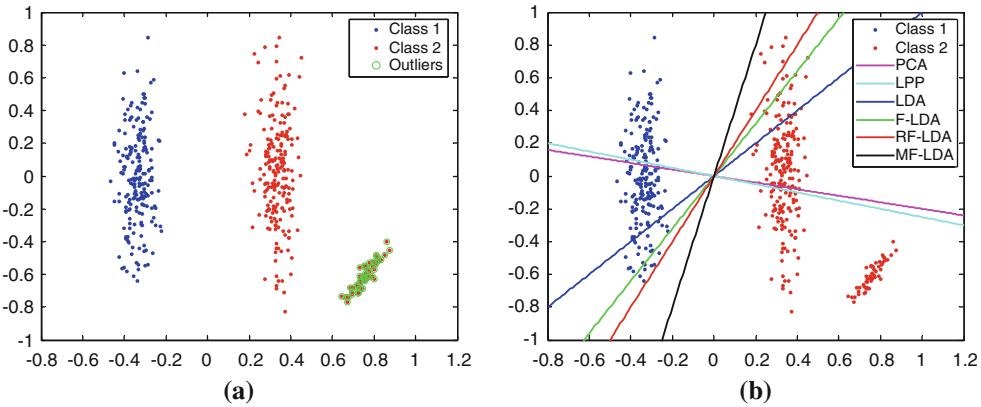


Fig. 3 Boundaries learned by PCA, LPP, LDA, F-LDA, RF-LDA and MF-LDA: two-line datasets with outliers. **a** Original dataset, **b** different boundaries learned by the algorithms

Table 3 Data information and simulation settings

Dataset	# Class (c)	# Images	# Feature	# Training	# Test
UMIST	20	564	32 × 32	10 × c	Remains
COIL-20	20	1,440	32 × 32	10 × c	40 × c
USPS	10	9,298	16 × 16	80 × c	100 × c

unsupervised algorithm such as PCA and LPP, as the boundaries learned by the supervised algorithms can better divide the samples into two classes, while for PCA and LPP, these samples may be seriously miss-classed. In addition, the proposed MF-LDA performs much better than other supervised algorithm. It can be seen in Fig. 3b that some of samples may be divided into false class for LDA, F-LDA, RF-LDA, but for the proposed MF-LDA it can precisely divide the samples into two classes. This indicates that our proposed algorithm can be more robust to outliers.

6.2 Classification

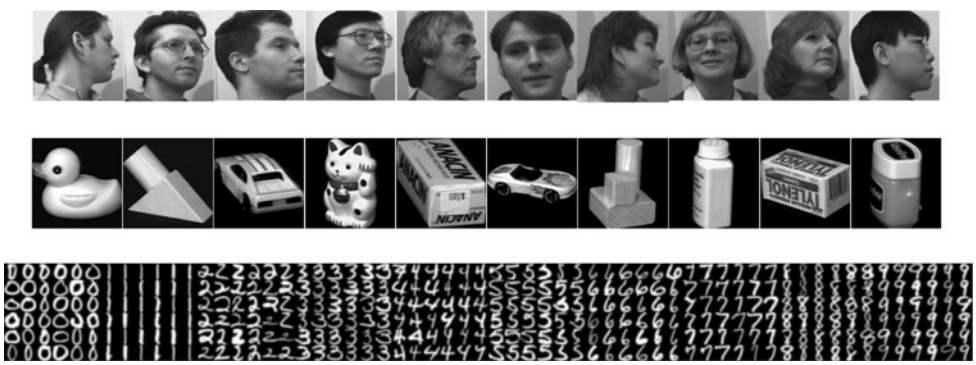
In this section, we used three datasets to evaluate the effectiveness of our algorithm and other algorithms such as PCA, LDA. The three datasets include the *UMIST* face dataset, *COIL-20* dataset and *USPS* handwritten digit dataset. The details of data information and samples are listed in Table 3 and Fig. 4.

In this comparative study, we randomly split each dataset into training set and test set. The training set in all datasets are preliminarily processed with a PCA operator to eliminate the null space before performing dimensionality reduction. All algorithms used the training set in the output-reduced space to train a nearest neighborhood (*1NN*) classifier with Euclidean distance for evaluating the accuracies of test set (Demsar 2006; Garcia and Herrera 2008).

6.2.1 *Face recognition*

For face recognition, we use the *UMIST* dataset to evaluate the performance of algorithms. The simulation settings are as follows: We randomly select 4, 7, 10 samples per class as training set and the remaining as test set. The regularized parameter α is set 0.1 in RF-LDA and the same as MF-LDA. For F-LDA, RF-LDA and MF-LDA the number of neighborhoods is set 8. We then explore the Gaussian function to construct the similarity matrix in MF-LDA. The parameter σ in the Gaussian function is determined as follows: we first calculated all the pairwise distances among data points of the whole training set. We then set σ equivalent to half the median of those distances. As a result, a reasonable estimation for the value of σcan be obtained.

Fig. 4 Some samples of dataset



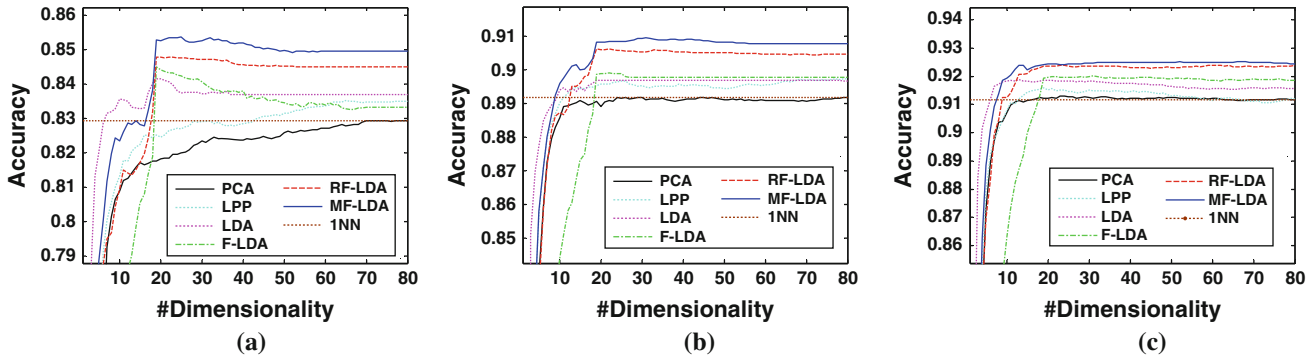


Fig. 5 Average accuracy (20 random splits) under different dimensionality: *UMIST* dataset **a** 4 labels, **b** 7 labels, **c** 10 labels

Table 4 Average accuracy (20 random splits) on the test set: *UMIST* dataset

Dataset	Method	4 labeled			7 labeled			10 labeled		
		Mean	Var	Dim	Mean	Var	Dim	Mean	Var	Dim
UMIST	*1NN*	82.95	2.87	–	89.25	1.38	–	91.22	0.91	–
	PCA	82.95	2.87	79	89.18	1.27	45	91.32	0.80	23
	LPP	83.52	2.59	65	89.73	1.25	79	91.60	1.10	19
	LDA	84.16	2.39	19	89.68	1.26	19	91.88	0.85	19
	F-LDA	84.50	2.26	19	89.89	1.75	19	92.01	1.30	19
	RF-LDA	84.80	2.66	19	90.61	1.66	19	92.38	1.06	19
	MF-LDA	85.38	2.67	19	90.94	1.45	19	92.53	1.08	19

The average accuracies over 20 random splits with the above parameters under different dimensionality are shown in Fig. 5. Table 4 shows the average accuracy with the best dimensionality for the *UMIST* dataset. The number of training set is fixed to 4, 7 and 10. From the results displayed in Fig. 5 and Table 4, we can observe that the fuzzy LDA outperforms conventional algorithm, such as PCA, LPP and LDA by about 2–3 %. This indicates that by incorporating the fuzzy membership into learning, the classification accuracies can be markedly improved as the fuzzy membership preserves the local discriminative information in each class. In addition, MF-LDA can deliver slightly better result than F-LDA and RF-LDA by about 1 and 2 %, respectively. This is mainly due to the fact that MF-LDA can maintain the consistency of local and global discriminative information by performing Markov random walks while F-LDA and RF-LDA only consider the local discriminative information of the training samples. In Table 3, it is noticed that the classification accuracy of all algorithms change when the number of training set increases. For instance, the accuracy of MF-LDA increased from about 83–91 % when the number of training samples increased from 4 to 10. We can also observe from Fig. 4 that the accuracy of all algorithms varies when the number of reduced dimensionality increased. For LDA, F-LDA, RF-LDA and MF-LDA, their accuracy remained unchanged beyond the bound of $c - 1$.

6.2.2 *Object recognition*

For object recognition, we use the *COIL-20* dataset to evaluate the performance of algorithms. The simulation settings are the same as the *UMIST* dataset. In the simulation, we randomly selected 4, 7, 10 samples per class as training set and 20 samples as test set. The average accuracies over 20 random split under different dimensionality are shown in Fig. 6 and the best results with optimal dimensionality are listed in Table 5. From Fig. 6 and Table 5, the following observation can be obtained: (1) the fuzzy LDA algorithms outperform conventional algorithms such as PCA, LPP and LDA by about 3–4 %. (2) MF-LDA can deliver about 2 and 3 % improvements compared with F-LDA and RF-LDA. (3) The accuracies of all algorithms change significantly when the labeled number increased, i.e. the accuracy of MF-LDA increases from 83 to 93 % when the number of labeled samples increased from 4 to 10. (4) The accuracies of all algorithms vary when the reduced dimensionality increases. For example, the best results can be achieved at the dimensionality of 19, 13, 11 for F-LDA, RF-LDA and MF-LDA, respectively. When the dimensionality increases beyond these bounds, the accuracies of above algorithms degrade gradually. (5) MF-LDA can reach the highest accuracy using fewest number of dimensionality. This shows a great superiority of MF-LDA over other algorithms.

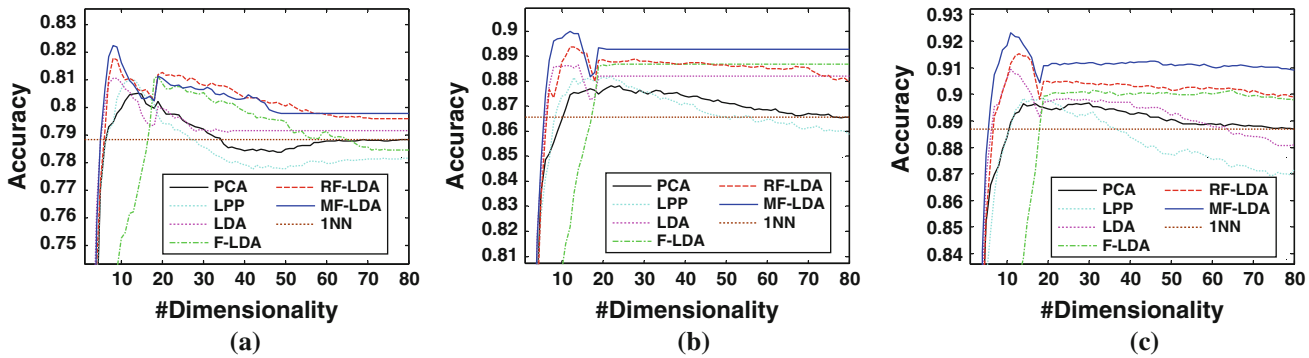


Fig. 6 Average accuracy (20 random splits) under different dimensionality: *COIL-20* dataset **a** 4 labels, **b** 7 labels, **c** 10 labels

Table 5 Average accuracy (20 random splits) on the test set: *COIL-20* dataset

Dataset	Method	4 labeled			7 labeled			10 labeled		
		Mean	Var	Dim	Mean	Var	Dim	Mean	Var	Dim
COIL20	1NN	78.80	2.76	–	86.51	1.52	–	88.49	1.48	–
	PCA	80.51	2.84	14	87.82	1.60	22	89.66	1.39	30
	LPP	81.03	2.68	13	88.19	2.05	22	89.87	1.68	18
	LDA	81.06	3.39	9	88.14	2.14	14	90.90	1.24	11
	F-LDA	81.14	2.81	19	89.19	2.18	19	90.15	1.92	19
	RF-LDA	81.79	3.08	8	89.38	1.97	12	91.51	1.76	13
	MF-LDA	82.76	3.17	8	90.00	1.82	12	92.30	1.40	11

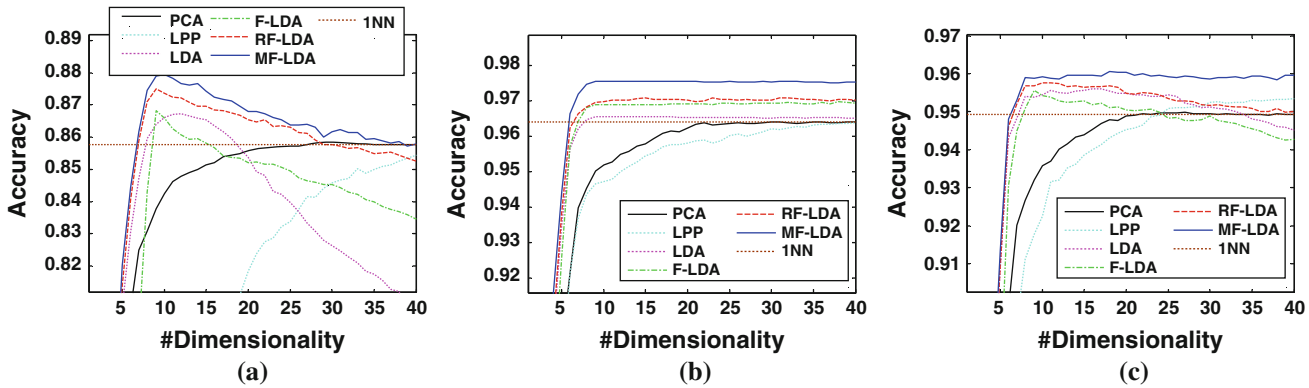


Fig. 7 Average accuracy (20 random splits) under different dimensionality: *USPS* dataset **a** 20 labels, **b** 50 labels, **c** 80 labels

6.2.3 Handwritten digit recognition

For handwritten digit recognition, we used the *USPS* dataset to evaluate the performance of algorithms. The simulation settings are as follows: We randomly selected 100 samples per class as training set and 100 samples as test set. The regularized parameter α is set to 0.1 in RF-LDA and MF-LDA. For F-LDA, RF-LDA and MF-LDA the number of neighborhoods is set to 16. The Gaussian function is used to construct the similarity matrix in MF-LDA and the parameter σ in Gaussian function is determined using the same strategy as *UMIST* and *COIL* dataset.

We first fixed the number of training samples to 20, 50 and 80 to train the learner. The average accuracies over 20 random splits with the above parameters under different dimensionality are shown in Fig. 7 and the best results with optimal dimensionality are listed in Table 6. From the results in Fig. 7 and Table 6, we can observe that for the *USPS* dataset, RF-LDA and MF-LDA outperform other algorithms by about 2–3 %. F-LDA does not seem to be

Table 6 Average accuracy (20 random splits) on the test set: *USPS* dataset

Dataset	Method	20 labeled			50 labeled			80 labeled		
		Mean	Var	Dim	Mean	Var	Dim	Mean	Var	Dim
USPS	1NN	85.72	0.92	–	94.90	0.32	–	96.39	0.37	–
	PCA	85.83	0.87	31	94.98	0.31	27	96.41	0.35	39
	LPP	85.54	2.12	39	95.35	0.65	39	96.39	0.52	39
	LDA	86.73	1.63	9	95.61	0.71	9	96.55	0.50	9
	F-LDA	86.82	2.09	9	95.55	0.78	9	96.98	0.59	9
	RF-LDA	87.50	1.75	9	95.77	0.66	9	97.09	0.58	9
	MF-LDA	87.95	1.83	9	96.06	0.87	9	97.56	0.51	9

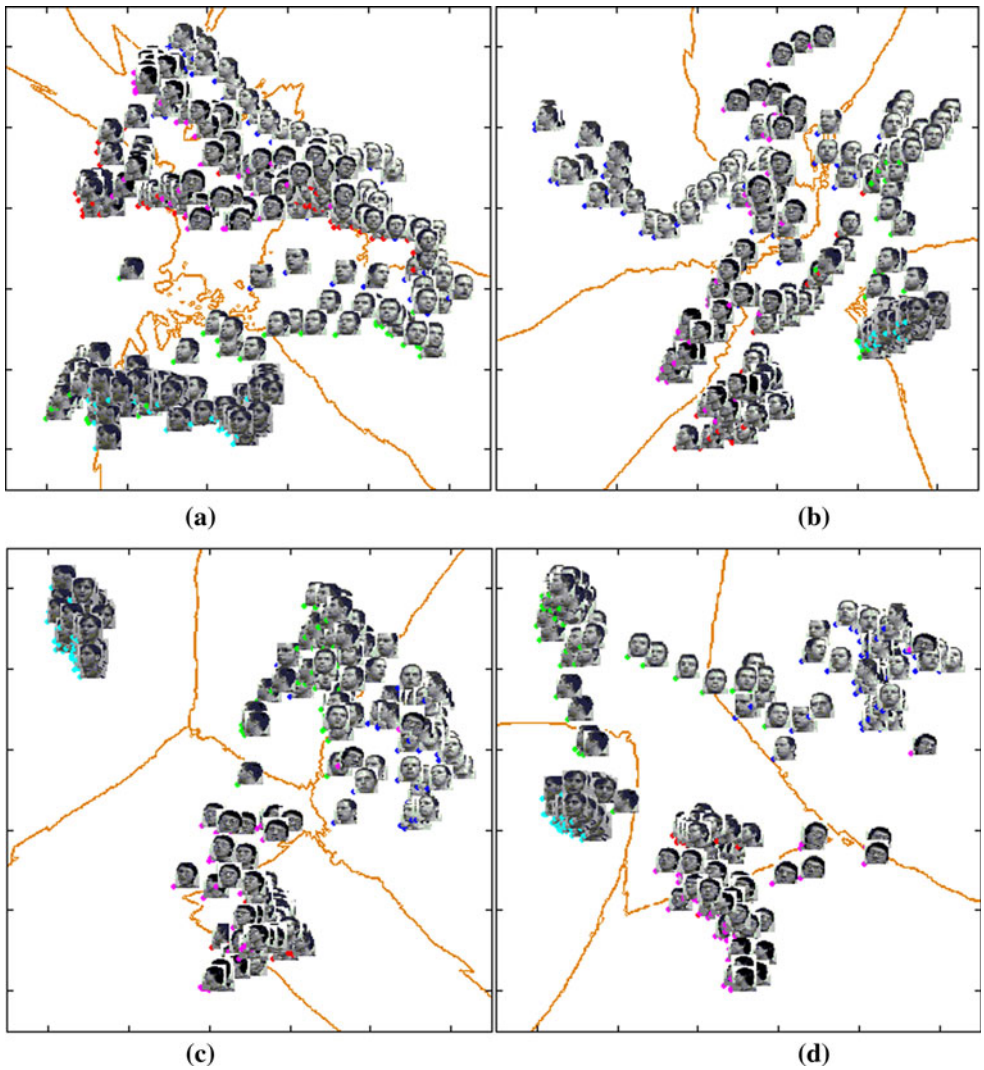


Fig. 8 2D visualization performed by PCA, LPP, LDA and MF-LDA: five individuals of *UMIST* dataset **a** PCA, **b** LPP, **c** LDA **d** MF-LDA (each color represents the faces of one person)

able to deliver better performance compared to LDA. The reason for it is that the regularized parameter α is fixed to 0.51 for F-LDA. But in most case, the handwritten digit dataset is usually sparse with lots of outliers, the fuzzy membership close to 0.5 such as in F-LDA means that the outliers exhibit influentially to several classes. Hence, the accuracy of classification may be degraded. On the other hand, by adjusting regularized parameter α in RF-LDA and MF-LDA, the relative large fuzzy membership can be achieved which results in eliminating the negative effects of the outliers to several classes. In addition, it can also be observed from Fig. 7 and Table 6 that the classification

accuracies of all algorithms change when the number of training set increases. For instance the accuracy of MF-LDA increased from about 88–98 % when the number of training samples increased from 20 to 80. Another phenomenon shown in Fig. 7 is that the accuracies of all algorithms vary when the reduced dimensionality increases and MF-LDA can reach the highest accuracy using the fewest number of dimensionality compared with other algorithms.

6.3 Visualization

We demonstrate the visualization of our proposed MF-LDA and compare it with PCA, LPP and LDA. In this study, two real-world datasets including the *UNIST* and *USPS* dataset are used.

In the *UMIST* face dataset, we selected five individuals to illustrate the visualization of dataset. In each individual, we selected four samples as labeled set and the remaining as test set. Fig. 8 shows the 2D visualization of test set performed by PCA, LPP, LDA and MF-LDA. From the results in Fig. 8a and b, we can see that in the unsupervised methods such as PCA and LPP, the ordering of face poses from profile to frontal views can be well preserved. LPP can deliver better visualization performance than PCA as the ordering lines of face poses are much smoother. This improved performance is mainly due to the characteristics of LPP that local information of dataset is embedded. But we can also observe from Fig. 8a and b that the decision boundaries between different classes are heavily overlapped and confused, which indicate that both PCA and LPP cannot preserve the discriminative information. On the other hand, from Fig. 8c and d, we can see that in supervised methods such as LDA and MF-LDA, such information can be well preserved because the boundaries learned by LDA and MF-LDA are more accurate and distinctive. It is clear that our proposed MF-LDA is able to deliver better performance compared to LDA. It can be observed from

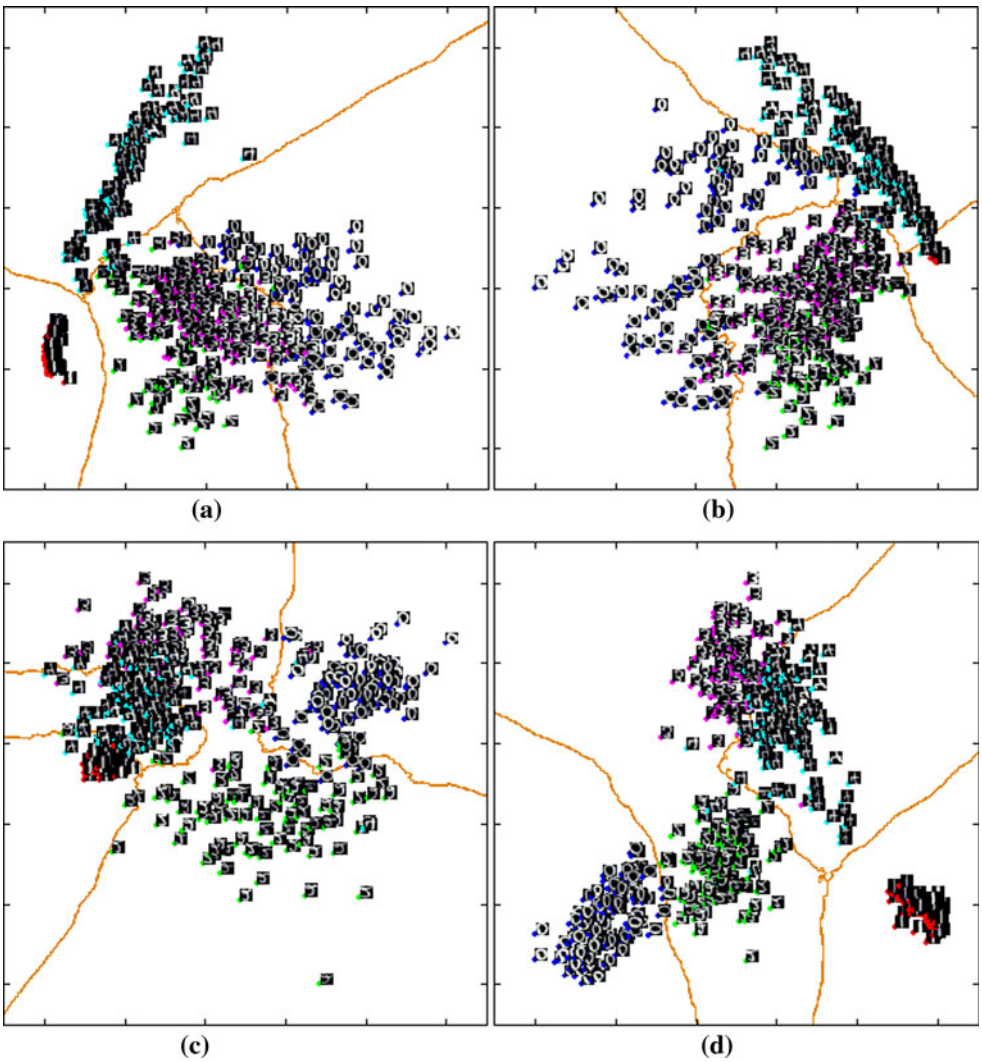


Fig. 9 2D visualization performed by PCA, LPP, LDA and MF-LDA: handwritten digits *0-4* of *USPS* dataset **a** PCA, **b** LPP, **c** LDA, **d** MF-LDA (*blue* digit 0; *red* digit 1; *green* digit 2; *magenta* digit 3; *cyan* digit 4) (color figure online)

Table 7 *UCI* datasets descriptions ("Balance" is defined as the ratio between the number of samples in the smallest class to the number of samples in the largest class)

Datasets	#Class	#Data	#Feature	"Balance"
Heart-Statlog	2	270	13	0.8
Ionosphere	2	351	34	0.56
Iris	3	150	4	1
Wine	3	178	13	0.1455
Balance	3	625	4	0.1701
Waveform	3	5,000	40	0.9770
SCCTS	6	600	60	1

Fig. 8d that samples in the upper left corner and upper right corner can be well separated in MF-LDA. But in Fig. 8c, these samples are heavily overlapped and conglomerated in the upper right corner.

In the *USPS* handwritten digit dataset, we selected five digits 0–4 to illustrate the visualization of dataset. For each digit, we randomly selected 50 samples as labeled set and another 100 samples as test set. Figure 9 shows the 2D visualization of test set performed by PCA, LPP, LDA and MF-LDA. From the results in Fig. 9 we can see that all algorithms can achieve satisfactory visualization performance and the boundaries learned by different algorithm can separate the test set nicely. The supervised algorithms (LDA and MF-LDA) are better than the unsupervised algorithms (PCA and LPP). For instance in Fig. 9a and b, we can notice that the boundaries between samples of number '2' and number '3' are not clear and seriously confused. But in Fig. 9c and d, these samples of two classes can be separated as there are distinctive boundaries across the two classes. In Fig. 9c and d, it demonstrates that our proposed MF-LDA can outperform LDA in a way that samples in each class are closely conglomerated, while those belonging to different classes are separated far apart. The reason for it is that by incorporating the fuzzy membership into the learning, the visualization performance of our proposed MF-LDA can be markedly enhanced as the fuzzy membership preserves the statistical properties of dataset and eliminates the effects of outliers.

6.4 *UCI* datasets

In this section, we choose seven *UCI* datasets to evaluate our algorithms and compare them with other algorithms. The datasets include *Heart-Statlog*, *Ionosphere*, *Iris*, *Wine*, *Balance*, *Waveform* and Synthetic *Control Chart Times Seires* (*SCCTS*) dataset. The detailed data information is listed in Table 7.

For comparative study, we randomly chose 30 % samples from each dataset as training set and the rest 70 % as test set. All algorithms used training set in the output-reduced space to train a nearest neighborhood classifier for evaluating the accuracies of test set. In these simulations, we simply set the reduced dimensionality as $c - 1$. The average accuracies over 20 random splits are shown in Table 8 for different UCI datasets. From the results in Table 8, we can observe that (1) supervised algorithms such as LDA, F-LDA, RF-LDA and the proposed MF-LDA are superior to the unsupervised algorithms such as PCA and LPP, e.g. MF-LDA can achieve 24, 14, 4, 20, 20 and 10 % improvements in the above datasets compared with PCA and LPP. (2) The fuzzy LDA algorithms outperform the conventional algorithms such as PCA, LPP and LDA by about 3–4 % in different datasets. (3) Our proposed algorithm can deliver about 2–3 % improvements compared with F-LDA and RF-LDA.

7 Conclusions

To reduce the effects of outliers, we propose a new efficient variant of LDA algorithm by incorporating the fuzzy membership into the learning. The fuzzy membership can be used to enhance the conventional LDA algorithm by detecting the outliers in each class. In this paper, we further analyze the existing fuzzy strategies and propose a new effective one based on Markov random walks. Our results

Table 8 Average accuracy (20 random splits) on the test set for 30 % labeled UCI datasets

Datasets	Methods						
	1NN	PCA	LPP	LDA	F-LDA	RF-LDA	MF-LDA
Heart-Statlog	59.78	59.78	60.43	78.30	79.54	80.36	82.30
Ionosphere	69.63	69.63	70.73	80.82	81.45	81.89	82.82
Iris	91.23	91.23	91.23	94.19	94.64	95.19	95.19
Wine	64.80	64.80	71.20	83.60	83.80	84.40	84.60
Balance	65.06	65.06	66.55	81.44	82.27	83.36	84.21
Waveform	80.94	80.94	82.85	84.45	84.85	85.40	85.45
SCCTS	81.90	81.90	84.76	90.95	91.19	91.66	92.95

show that the proposed fuzzy strategy can provide a remarkable effect on maintaining the consistency of local and global discriminative information embedded in datasets. It can also reflect the statistical properties of dataset by performing random walk along the neighborhood graph, which results in detecting the outliers. Finally, theoretical analysis and extensive simulations show the effectiveness of our algorithm. The results in simulations demonstrate that our proposed algorithm can achieve great superiority compared with other existing algorithms.

Appendix

Proof of Corollary 1

According to Eq. (12), we have

$$\begin{aligned}(T^{\mathrm{T}}T)_{ij} &= \sum_{k=1}^{c}\left(\frac{w_{ki}}{\sqrt{F_{kk}}}-\frac{\sqrt{F_{kk}}}{l}\right)\left(\frac{w_{kj}}{\sqrt{F_{kk}}}-\frac{\sqrt{F_{kk}}}{l}\right)\\ &= \sum_{k=1}^{c}\frac{w_{ki}w_{kj}}{F_{kk}}-\sum_{k=1}^{c}\frac{w_{ki}}{l}-\sum_{k=1}^{c}\frac{w_{kj}}{l}+\sum_{k=1}^{c}\frac{F_{kk}}{l^2}\\ &= \sum_{k=1}^{c}\frac{w_{ki}w_{kj}}{F_{kk}}-\frac{1}{l}=-(\widetilde{A_{\mathrm{b}}})_{ij}.\end{aligned} \tag{26}$$

According to Eq. (8), we have

$$\begin{aligned}(\widetilde{D_{\mathrm{b}}})_{ii} = \sum_{j=1}^{l}(\widetilde{A_{\mathrm{b}}})_{ij} &= \sum_{j=1}^{l}\left(\frac{1}{l}-\sum_{k=1}^{c}\frac{w_{ki}w_{kj}}{F_{kk}}\right)\\ &= 1-\sum_{k=1}^{c}\frac{1}{F_{kk}}\sum_{j=1}^{l}w_{ki}w_{kj}\\ &= 1-\sum_{k=1}^{c}w_{ki}=1-1=0\end{aligned} \tag{27}$$

The second equality holds as $\sum_{j=1}^{l} w_{kj} = F_{kk}$ and the third equality holds as $\sum_{k=1}^{c} w_{ki} = 1$. We then have $(\widetilde{L_{\mathrm{b}}})_{ij} = (\widetilde{D_{\mathrm{b}}})_{ii} - (\widetilde{A_{\mathrm{b}}})_{ij} = -(\widetilde{A_{\mathrm{b}}})_{ij}$, hence we prove $T^{\mathrm{T}}T = \widetilde{L_{\mathrm{b}}}$.

Proof of Corollary 2

We prove Corollary 2 using a recursive algorithm. According to Eq. (18) and $W(0) = Y$, have

$$\begin{aligned}\sum_{i=1}^{c}w_{ij}(1) &= \alpha\sum_{i=1}^{c}\sum_{x_k\in N_k(x_j)}y_{ik}s_{kj}+(1-\alpha)\sum_{i=1}^{c}y_{ij}\\ &= \alpha\sum_{x_k\in N_k(x_i)}s_{kj}\sum_{i=1}^{c}y_{ik}+(1-\alpha)\sum_{i=1}^{c}y_{ij}\\ &= \alpha\sum_{x_k\in N_k(x_i)}s_{kj}+(1-\alpha)=\alpha+(1-\alpha)=1\end{aligned} \tag{28}$$

The third equality holds as $\sum_{i=1}^{c} y_{ik} = 1 \sum_{i=1}^{c} y_{ik} = 1$ and the fourth equality holds as $\sum_{x_k\in N_k(x_i)} s_{kj} = 1$. Hence, Eq. (28) indicates that the sum of each column of $W(1)$ is equivalent to 1. We next assume that the sum of each column of $W(t)$ is equivalent to 1, i.e. $\sum_{i=1}^{c} w_{ij}(t) = 1$ for any iteration t, we then have

$$\begin{aligned}\sum_{i=1}^{c}w_{ij}(t+1) &= \alpha\sum_{i=1}^{c}\sum_{x_k\in N_k(x_i)}w_{ik}(t)s_{kj}+(1-\alpha)\sum_{i=1}^{c}y_{ij}\\ &= \alpha\sum_{x_k\in N_k(x_i)}s_{kj}\sum_{i=1}^{c}w_{ik}(t)+(1-\alpha)\sum_{i=1}^{c}y_{ij}\\ &= \alpha\sum_{x_k\in N_k(x_i)}s_{kj}+(1-\alpha)=\alpha+(1-\alpha)=1\end{aligned} \tag{29}$$

This indicates that the sum of each column of $W(t+1)$ is also equivalent to 1. Thus, we prove that $\sum_{i=1}^{c} w_{ij} = \sum_{i=1}^{c} \lim_{t\to\infty} w_{ij}(t) = 1$.

Proof of Theorem 1

1. Computing V_F^* via eigen-decomposition (Ye 2005)

Let t be the rank of $\widetilde{S}_{\mathrm{t}}$, by performing SVD to $\widetilde{S}_{\mathrm{t}}$, we have

$$\widetilde{S}_{\mathrm{t}} = U\begin{pmatrix}\Sigma_t^2 & 0\\ 0 & 0\end{pmatrix}U^{\mathrm{T}}, \tag{30}$$

where U is an orthogonal matrix, Σ_t^2 is a diagonal matrix with rank t. Let $U = [U_1, U_2]$ be a partition of U such that $U_1 \in \mathbb{R}^{d\times t}$ and $U_2 \in \mathbb{R}^{d\times(d-t)}$, where U_2 lies in the null space of $\widetilde{S}_{\mathrm{t}}$ satisfying $U_2^{\mathrm{T}}\widetilde{S}_{\mathrm{t}}U_2 = 0$, we then have $S_{\mathrm{t}} = U_1\sum_{\mathrm{t}}^2 U_1^{\mathrm{T}}$. Since $\widetilde{S}_{\mathrm{t}} = \widetilde{S}_{\mathrm{b}} + \widetilde{S_{\mathrm{w}}}$, we have

$$U^{\mathrm{T}}\widetilde{S}_{\mathrm{b}}U = \begin{pmatrix}U_1^{\mathrm{T}}\widetilde{S}_{\mathrm{b}}U_1 & 0\\ 0 & 0\end{pmatrix}, U^{\mathrm{T}}\widetilde{S_{\mathrm{w}}}U = \begin{pmatrix}U_1^{\mathrm{T}}\widetilde{S_{\mathrm{w}}}U_1 & 0\\ 0 & 0\end{pmatrix} \tag{31}$$

From Eqs. (30, 31), it follows

$$\begin{aligned}I_{\mathrm{t}} &= \sum_{\mathrm{t}}^{-1}U_1^{T}\widetilde{S}_{\mathrm{t}}U_1\sum_{\mathrm{t}}^{-1}\\ &= \sum_{\mathrm{t}}^{-1}U_1^{\mathrm{T}}\widetilde{S}_{\mathrm{b}}U_1\sum_{\mathrm{t}}^{-1}+\sum_{\mathrm{t}}^{-1}U_1^{T}\widetilde{S_{\mathrm{w}}}U_1\sum_{\mathrm{t}}^{-1},\end{aligned} \tag{32}$$

where $I_{\mathrm{t}} \in \mathbb{R}^{t\times t}$ is an identity matrix. Recall that $\widetilde{S}_{\mathrm{b}} = \widetilde{H}_{\mathrm{b}}\widetilde{H}_{\mathrm{b}}^{\mathrm{T}}$, if we let $G = \sum_{\mathrm{t}}^{-1}U_1^{\mathrm{T}}\widetilde{H}_{\mathrm{b}}$ and its *SVD* be $G = P\sum_{\mathrm{b}}Q$, where $P \in \mathbb{R}^{t\times t}$ and $Q \in \mathbb{R}^{t\times c}$ are two orthogonal matrixes and $\sum_b \in \mathbb{R}^{t\times t}$ is a diagonal matrix, we then have

$$\sum_{t}^{-1}U_1^{\mathrm{T}}\widetilde{S}_bU_1\sum_{t}^{-1} = GG^{\mathrm{T}} = P\sum_{b}^{2}P^{\mathrm{T}}. \tag{33}$$

Therefore, according to Eqs. (30, 31, 33), we rewrite $V_F^* = \widetilde{S}_{\mathrm{t}}^{-1}\widetilde{S}_{\mathrm{b}}$ as

$$\begin{aligned}\widetilde{S}_{\mathrm{t}}^{-1}\widetilde{S}_{\mathrm{b}} &= U_1\begin{pmatrix}\sum_t^{-1}\sum_t^{-1} & 0\\ 0 & 0\end{pmatrix}U^{\mathrm{T}}\widetilde{S}_{\mathrm{b}}U\begin{pmatrix}\sum_t^{-1}\sum_{\mathrm{t}} & 0\\ 0 & 0\end{pmatrix}U^{\mathrm{T}}\\ &= U\begin{pmatrix}\sum_t^{-1} & 0\\ 0 & 0\end{pmatrix}P\sum_b^2 P^{\mathrm{T}}\begin{pmatrix}\sum_t & 0\\ 0 & 0\end{pmatrix}U^{\mathrm{T}}\\ &= U\begin{pmatrix}\sum_t^{-1}P & 0\\ 0 & I^{D-t}\end{pmatrix}\begin{pmatrix}\sum_b^2 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}P^{\mathrm{T}}\sum_t & 0\\ 0 & I^{D-t}\end{pmatrix}U^{\mathrm{T}}\end{aligned} \tag{34}$$

The first equality holds as Eq. (30), the second equality holds as Eq. (33). From Eq. (34), if we let $V_F^* = U_1\sum_t^{-1}P$, we have $\widetilde{S}_{\mathrm{t}}^{-1}\widetilde{S}_{\mathrm{b}}V_F^* = \sum_{\mathrm{b}}^2 V_F^*$, which indicate the column vectors of V_F^* are eigenvectors of $\widetilde{S}_{\mathrm{t}}^{-1}\widetilde{S}_{\mathrm{b}}$.

2. Equivalence relationship to least square

Recall V_{LS}^* in Eq. (11), it can be rewritten as

$$\begin{aligned}\widetilde{S}_{\mathrm{t}}^{-1}\widetilde{H_{\mathrm{b}}} &= U\begin{pmatrix}\sum_t^{-1}\sum_t^{-1} & 0\\ 0 & 0\end{pmatrix}U^{\mathrm{T}}\widetilde{H_{\mathrm{b}}}\\ &= U_1\sum_t^{-1}\left(\sum_t^{-1}U_1^{\mathrm{T}}\widetilde{H_{\mathrm{b}}}\right) = U_1\sum_t^{-1}G\\ &= U_1\sum_t^{-1}P\sum_{\mathrm{b}}Q^{\mathrm{T}} = V_F^*\sum_{\mathrm{b}}Q^{\mathrm{T}}\end{aligned} \tag{35}$$

From Eq. (35), we can neglect Q as it is orthogonal. Thus, the main difference between V_F^* and V_{LS}^* is the diagonal matrix $\sum_b$. We next show that given the condition in Eq. (13), $\sum_b$ is an identity matrix hence resulting in $V_{\mathrm{LS}}^* = V_F^*$. Let $H \in \mathbb{R}^{D\times D}$ be a non-degenerate matrix defined as:

$$H = U\begin{pmatrix}\sum_t^{-1}P & 0\\ 0 & I_{D-t}\end{pmatrix} \tag{36}$$

According to Eq. (31) (32) and $\widetilde{S_{\mathrm{w}}} = \widetilde{S}_{\mathrm{t}} - \widetilde{S}_{\mathrm{b}}$, we have

$$\begin{aligned}&H^T\widetilde{S}_{\mathrm{t}}H = \begin{pmatrix}I_{\mathrm{t}} & 0\\ 0 & 0\end{pmatrix}, H^{\mathrm{T}}\widetilde{S_{\mathrm{w}}}H = \begin{pmatrix}\sum_w^2 & 0\\ 0 & 0\end{pmatrix},\\ &H^{\mathrm{T}}\widetilde{S}_{\mathrm{b}}H = \begin{pmatrix}\sum_b^2 & 0\\ 0 & 0\end{pmatrix}\end{aligned} \tag{37}$$

where $\sum_b^2 = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_t^2, 0, \ldots, 0)$ and $\sum_w^2 = \mathrm{diag}(\tau_1^2, \tau_2^2, \ldots, \tau_t^2, 0, \ldots, 0)$ are two diagonal matrixes satisfying $\sigma_i^2 + \tau_i^2 = 1, \forall i$. This indicates that there is at least one of σ_i and τ_i to be nonzero, $\forall i$. Since $\mathrm{rank}(A) + \mathrm{rank}(B) \geq \mathrm{rank}(A + B)$ (Hull 1994), we have $\mathrm{rank}(\widetilde{S}_{\mathrm{b}}) + \mathrm{rank}(\widetilde{S_{\mathrm{w}}}) \geq \mathrm{rank}(\widetilde{S}_{\mathrm{t}})$ According to the Sylvester's law of interia (Hull 1994), it follows $\mathrm{rank}(\sum_b^2) + \mathrm{rank}(\sum_w^2) \geq \mathrm{rank}(I_t)$. Let b be the rank of $\sum_b^2$ and assume $\mathrm{rank}(\sum_b^2) + \mathrm{rank}(\sum_w^2) = \mathrm{rank}(I_t) + s$, to satisfy this rank equality, we have

$$\begin{aligned}&1 = \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_{b-s}^2 > \sigma_{b-s+1}^2 > \cdots \sigma_b^2 > \sigma_{b+1}^2\\ &= \cdots \sigma_t^2 = 0\\ &0 = \tau_1^2 = \tau_2^2 = \cdots \tau_{b-s}^2 < \tau_{b-s+1}^2 < \cdots \tau_b^2 < \tau_{b+1}^2 = \cdots \tau_t^2 = 1\end{aligned} \tag{38}$$

Since C1 holds, we have $s = 0$ and $1 = \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_b^2 > \sigma_{b+1}^2 = \cdots = \sigma_t^2 = 0$, which indicates that $\sum_b$ is an identity matrix.

Appendix D

1. Computing V_F^* via two-stage approach

In the two-stage approach, we first solve a least square problem by regressing X on T, i.e. projecting the original high-dimensional dataset into a low-dimensional subspace, we then calculate a auxiliary matrix $M \in R^{d\times d}$ and its *SVD*. Finally, the optimal projection matrix can be obtained from the *SVD* of M. Since the size of M is very small, the cost for calculating the *SVD* of M is relatively low. The basic steps of two-stage approach are listed as follows:

1. Solve the least square problem $\min_V \|T^T - X^T V\|_F^2$ and obtain the optimal solution V_{LS}^*.
2. Let $\widetilde{X} = V_{\mathrm{LS}}^{*\mathrm{T}}X$ and calculate the auxiliary matrix as $M = V_{\mathrm{LS}}^{*\mathrm{T}}XT^{\mathrm{T}}$.
3. Perform *SVD* to M as $M = U_M\Sigma_M U_M^T$ and obtain $V_M^* = U_M\Sigma_M^{-1/2}$.
4. The optimal solution can be given by $V_{\mathrm{T}}^* = V_{\mathrm{LS}}^*V_M^*$.

2. Equivalent relationship

We next prove the optimal solution V_T^* obtained by two-stage approach is equivalent to that in Eq. (34). By solving least square problem in Eq. (11), we have $V_{LS}^* = (XX^T)^{-1}XT^T$. Hence, $\widetilde{X} = V_{\mathrm{LS}}^{*T}X = TX^{\mathrm{T}}(XX^{\mathrm{T}})^{-1}X$ The auxiliary matrix M can then be given by

$$M = \widetilde{X}Y^{\mathrm{T}} = TX^{\mathrm{T}}(XX^{\mathrm{T}})^{-1}XT^{\mathrm{T}} = \widetilde{H_{\mathrm{b}}}^{\mathrm{T}}U_1\sum_t^{-1}\sum_t^{-1}U_1^{\mathrm{T}}\widetilde{H_{\mathrm{b}}}. \tag{39}$$

The third equation holds as $\widetilde{H_{\mathrm{b}}} = XT^{\mathrm{T}}$ and $XX^{\mathrm{T}} = S_t = U_1\sum_t^2 U_1^{\mathrm{T}}$. Since $G = \sum_t^{-1}U_1^{\mathrm{T}}\widetilde{H_{\mathrm{b}}}$ and its SVD is $G = P\sum_b Q^{\mathrm{T}}$, we have $M = G^{\mathrm{T}}G = Q\sum_b^2 Q^{\mathrm{T}}$. This indicates that $Q\sum_b^2 Q^{\mathrm{T}}$ is a SVD of M, we thus have $V_M^* = Q\sum_b^{-1}$ and the optimal solution of two-stage approach can be given by:

$$
\begin{aligned}
V_{\mathrm{T}}^{*} &= V_{\mathrm{LS}}^{*} V_{M}^{*} = (XX^{\mathrm{T}})^{-1} XY^{\mathrm{T}} Q \sum_{b}^{-1} \\
&= U_1 \sum_{1}^{-1} \left(\sum_{1}^{-1} U_1^{\mathrm{T}} \widetilde{H_{\mathrm{b}}} \right) Q \sum_{b}^{-1} \\
&= U_1 \sum_{1}^{-1} P \sum_{b} Q^{\mathrm{T}} Q \sum_{b}^{-1} \\
&= U_1 \sum_{1}^{-1} P, \qquad (40)
\end{aligned}
$$

which is equivalent to V_F^* in Eq. (34).

References

Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720

Chen L, Liao H, Ko M, Lin J, Yu G (2000) A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recogn 33(10):1713–1726

Chung FRK (1997) Spectral graph theory. American Mathematical Society, Providence

Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

Fukunaga K (1990) Introduction to statistical pattern classification. Academic Press, New York

Garcia S, Herrera F (2008) An extension on "Statistical Comparisons of Classifiers over Multiple Datasets" for all Pairwise Comparisons. J Mach Learn Res 9:2677–2694

Graham DB, Allinson NM (1998) Characterizing virtual eigensignatures for general purpose face recognition in face recognition: from theory to application. NATO ASI Series F, Computer and Systems Sciences, vol 163, pp 446–456

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning data mining, inference and prediction. Springer, Berlin

He X, Yan S, Hu Y, Niyogi P, Zhang H (2005) Face recognition using Laplacianfaces. IEEE Trans Pattern Anal Mach Intell 27(3): 328–340

Howland P, Park H (2004) Generalizing discriminant analysis using the generalized singular value decomposition. IEEE Trans Pattern Anal Mach Intell 26(8):995–1005

Hull J (1994) A database for handwritten text recognition research. IEEE Trans Pattern Recogn Mach Intell 16(5):550–554

Keller JM, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbor algorithm. IEEE Trans Syst Man Cybern 15(4):580–585

Kwak KC, Pedrycz W (2005) Face recognition using a fuzzy Fisherface classifier. Pattern Recogn 38(10):1717–1732

Liu X, Lu C, Chen F (2010) Spatial outlier detection: random walk based approaches. In: ACM SIG SPATIAL Proceedings of GIS

Moonesignhe HDK, Tan P (2006) Outlier detection using random walks. In: Proceedings of ICTAI

Muller KR, Mika S, Ratsch G, Tsuda K, Scholkopf B (2001) An introduction to kernel-based learning algorithms. IEEE Trans Neural Netw 12(2):181–201

Nene SA, Nayar SK, Murase H (1996) Columbia object image library (COIL-20). Technical report CUCS-005-96, Columbia University

Roweis S, Saul L (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326

Song X, Zheng Y, Wu X, Yang X, Yang J (2009) A complete fuzzy discriminant analysis approach for face recognition. Appl Soft Comput 10(1):208–214

Sun T, Chen S (2007) Class label versus sample label-based *CCA*. Appl Math Comput 185(1):272–283

Sun L, Ceran B, Ye J (2010) A scalable two-stage approach for a class of dimensionality reduction techniques. In: Proceedings of KDD

Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323

Turk M, Pentland A (1991) Face recognition using Eigenfaces. In: Proceedings of CVPR

Wang X, Davidson I (2009) Discovering contexts and contextual outliers using random walks in graphs. In: Proceedings of ICDM

Yang J, Frangi AF, Yang J, Zhang D, Jin Z (2005) *KPCA* plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition. IEEE Trans Pattern Anal Mach Intell 27(2):230–244

Ye J (2005) Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. J Mach Learn Res 6:483–502

Ye J (2007) Least square linear discriminant analysis. In: Proceedings of ICML

Ye J, Li Q (2005) A two-stage linear discriminant analysis via *QR*-decomposition. IEEE Trans Pattern Anal Mach Intell 27(6): 929–941

Yu H, Yang J (2001) A direct LDA algorithm for high-dimensional data with application to face recognition. Pattern Recogn 34(10): 2067–2070

Zhang Z, Dai G, Jordan MI (2009) A flexible and efficient algorithm for regularized Fisher discriminant analysis. In: Proceedings of ECML PKDD