International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-020-01188-2

ORIGINAL ARTICLE

A novel feature learning framework for high-dimensional data classification

Yanxia Li[2] · Yi Chai[1,2] · Hongpeng Yin[1,2] · Bo Chen[2]

Received: 19 November 2019 / Accepted: 19 August 2020

Abstract
Feature extraction is an essential component in many classification tasks. Popular feature extraction approaches especially deep learning-based methods, need large training samples to achieve satisfactory performance. Although dictionary learning-based methods are successfully used for feature extraction on both small and large datasets, however, when dealing with high-dimensional datasets, a large number of dimensions also mask the discriminative information embedded in the data. To address these issues, a novel feature learning framework for high-dimensional data classification is proposed in this paper. Specially, to discard the irrelevant parts that derail the dictionary learning process, the dictionary is adaptively learnt in the low-dimensional space parameterized by a transformation matrix. To ensure that the learned features are discriminative for the classifier, the classification results in turn are used to guide the dictionary and transformation matrix learning process. Compared with other methods, the proposed method simultaneously exploits the dimension reduction, dictionary learning and classifier learning in one optimization framework, which enables the method to extract low-dimensional and discriminative features. Experimental results on several benchmark datasets demonstrate the superior performance of the proposed method for high-dimensional data classification task, particularly when the number of training samples is small.

Keywords High-dimensional data classification · Feature extraction · Dimension reduction · Dictionary learning

1 Introduction

High-dimensional data is now ubiquitous in many domains, such as computer vision and bioinformatics [1, 2]. High dimensionality (usually several hundreds or thousands of dimensions) may produce the Hughes phenomenon, which can significantly reduce classification performance. Owing to accuracy consideration, numerous efforts have been made to produce good feature representation for high-dimensional classification task, by selecting [3–5] or extracting features [6, 7] from original high-dimensional data. Existing feature extraction methods mainly fall into two categories: designing features manually and learning features from data directly [8–10]. In general, designing features usually requires abundant engineering skills and domain expertise, which may limit their practical applications [11]. Learning features from data can overcome the limitations of hand-craft features and has been the focus of many recent researches.

A variety of feature learning methods have been proposed in recent years. Typical feature learning methods include, but not limited to, subspace learning [12, 13], dictionary learning [14–16], deep learning [17–19], hashing-based learning [20], and metric learning [21]. These feature learning methods has been successfully used in various applications. For example, Wu et al. [22] derived a low-dimensional subspace to reduce redundant information, so that document vectors can be grouped more reasonably. Liu et al. [23] proposed a distributed dictionary learning method for fault detection and fault isolation task. Li et al. [24] learned

✉ Hongpeng Yin
yinhongpeng@gmail.com

Yanxia Li
liyanxia106@gmail.com

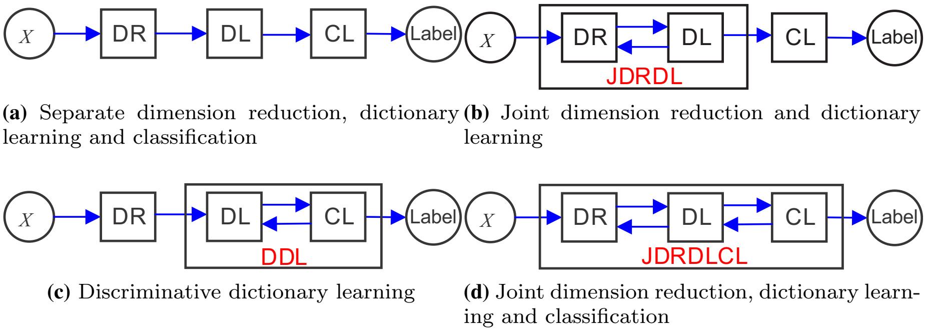
Yi Chai
chaiyi@cqu.edu.cn

Bo Chen
chenbocqu@163.com

[1] Power Transmission Equipment & System Security and New Technology, State Key Laboratory, Chongqing 400000, China

[2] College of Automation, Chongqing University, Chongqing 40000, China

Published online: 02 November 2020



(a) Separate dimension reduction, dictionary learning and classification

(b) Joint dimension reduction and dictionary learning

(c) Discriminative dictionary learning

(d) Joint dimension reduction, dictionary learning and classification

Fig. 1 The diagram of different dictionary learning-based high-dimensional classification framework

features via a convolutional neural network with mapping layers, which was efficient to maintain the spectral and spatial structures. Shen et al. [25] employed an effective unsupervised deep hashing framework to learn binary codes from raw images. The learned codes greatly improved the retrieval performance. Han et al. [26] presented a metric learning regularization term to learn a powerful feature representation for co-saliency detection. The aforementioned studies demonstrate that when compared with handcrafted feature-based methods, feature learning methods have obtained good performance.

Deep learning, also termed as deep neural networks, is one of the most popular feature learning methods. Representative deep learning models, such as convolutional neural network (CNN) [27], deep belief network (DBN) [28] and stacked auto-encoders (SAE) [29], have been in-depth researched and successfully employed to learn features for different classification tasks. These models exploit multiple hidden layers and a large scale of network parameters to discover abstract and hierarchical features. They are sufficient to produce state-of-the-art results but only with a large number of samples. For example, 4,000,000 labeled samples were used to train the deep convolutional neural networks proposed by Taigman et al. [30]. Parkhi et al. [31] introduced a deep convolutional neural networks by using 2.6M labeled training images. However, in practical scenarios, the high-dimensional datasets are characterized by a relatively small number of samples [32–35]. As a result, the application of deep learning in the above datasets may suffer from over-fitting problem and tends to degrade classification performance.

In the meantime, dictionary learning is emerged as a powerful tool for feature learning on both small and large datasets. Employing dictionary learning for classification is accomplished by considering the sparse coding coefficients as the learned features and utilizing a desired classifier to classify the data. Some known dictionary learning methods have been presented recently and performed well for different classification tasks [36, 37]. However, directly learning dictionary from original high-dimensional data is inefficient [38]. The reason is twofold. First, it is computationally intensive, since the complexity in dictionary learning increases dramatically along with the increase in dimensionality. Second, the presence of noisy, redundant information often obscures the discriminative structures within original data, leading to degraded classification performance.

To tackle with the challenges caused by high dimensionality, a dimension reduction technique is first adopted to project the original high-dimensional data to a low-dimensional latent space, and then a desirable dictionary is learned in the low-dimensional space. The whole process, which comprises of dimension reduction (DR), dictionary learning (DL) and classifier learning (CL), is summarized in Fig. 1a. Principal component analysis (PCA) [39] and its extensions [40] are the most representative algorithms employed for dimension reduction. Despite their easy implementation, they may not be powerful enough for dictionary learning [41]. This is mainly because most of these methods just perform dimension reduction, dictionary learning and classification as three individual stages, as shown in Fig. 1a. Hence some discriminative information which is essential for dictionary learning and classification may be lost in the individual dimension reduction stage.

Recent studies reveal that it is expected to jointly conduct the dimension reduction and dictionary learning processes. An overview of the joint dimension reduction and dictionary learning framework is illustrated in Fig.1b, from which one can note that the dimension reduction and dictionary learning stages are coupled into a unified framework for energy minimization, so a more effective dictionary can be obtained for better classification performance [42]. Only a few works have focused on Joint Dimension Reduction and Dictionary Learning (JDRDL). As an exploratory work, Feng et al. [43] jointly learned a projection matrix and a dictionary for face representation. Nguyen et al. [44] integrated the dimensionality reduction and dictionary learning together, which results in a simultaneous learning for optimal projection matrix and dictionary. More recently, Su et al. [45] carried out dimension reduction and dictionary learning jointly. Yang et al. [46] also formulated dimension reduction and dictionary learning as a whole optimization framework. Foroughi et al. [47] presented a novel scheme to optimize a transformation matrix and a desirable dictionary simultaneously. Sun et al. [48] proposed a jointly projection matrix and dictionary learning approach to find the most suitable projection matrix and dictionary. Despite their success, most of the existing JDLDR methods are just implemented by minimizing the reconstruction error of data while have no direct connection with the classification task. Thus they cannot fit the classifier well, since recent works have shown that better results can be obtained when the dictionary is tuned to the classification task it is intended for [49, 50].

To enhance the discriminative capability of the dictionary, some discriminative dictionary learning (DDL) methods, which are efficient for classification task, have been proposed by incorporating some prediction loss into the dictionary learning process. Yang et al. [51] imposed the fisher discrimination on the coding vectors to enhance class discrimination. Jiang et al. [52] introduced a label consistent regularization to enforce the discrimination of coding vectors. Cai et al. [53] formulated the discrimination term as the weighted summation of the squared distances between all pairs of coding vectors. Yang et al. [54] incorporated the classification error on the discriminative analysis-synthesis dictionary learning procedure to make the dictionary discriminative. It is worth noting that incorporating the discrimination criteria into the objective function dose add discrimination property to the learned dictionary, however, the dictionary and dimension reduction are still two independent steps when applied to high-dimensional classification, as illustrated in Fig. 1c. The whole process may suffer from the same issues arising when applying PCA for dimension reduction in general. What lacks is a unified approach that optimizes dimension reduction transformation matrix with discriminative dictionary learning process.

To overcome the above limitations, in this paper a novel feature learning framework is proposed for high-dimensional data classification. The diagram of the proposed framework is illustrated in Fig. 1d. The basic idea of the proposed method is to perform dimension reduction, dictionary learning and classification in one optimization framework, so that in the transformed low-dimensional space, the learned features can have a better discriminability and suit the classifier well.

The major contribution of this paper is that a joint feature learning framework, which makes dimension reduction, dictionary learning and classification help each other in a coherent manner, is proposed for high dimensional data classification. Different from DDL methods which are separately optimized from the dimension reduction stage, the proposed method optimizes the transformation matrix and dictionary simultaneously, to perverse more vital information for dictionary learning and classification. Unlike JDRDL methods which have no direct connection with the classifier, in the proposed method, the classification results are used as the decision criteria to guide the dimension reduction and dictionary learning process, so that the transformation matrix and dictionary can be alternatively optimized according to the classification results. The whole framework simultaneously optimizes a dimension reduction transformation matrix, a dictionary and a classifier. It is efficient to learn effective and discriminative features, and shows more robust classification capability when the training sample size is small.

The rest of this paper is structured as follows. Section 2 highlights the proposed framework. The experimental results on multiple high-dimensional datasets are reported in Sect. 3. Section 4 presents the conclusion and potential future works.

2 The proposed methodology

The proposed methodology is described in detail in this section. Initially, a genetic framework of the proposed method is provided. Later, the objective function of the proposed method is reported. The detailed optimization steps of the proposed method is also explained, followed by the proposed classification algorithm.

2.1 The feature learning framework

The generic feature learning framework for high-dimensional data classification is shown in Fig. 2. In particular, a dictionary learning algorithm is proposed to extract abstract features due to its effective and efficient ability to represent high-dimensional data. To decrease the computation cost and make data more efficient for dictionary learning, the proposed method firstly projects the original high-dimensional to a low-dimensional space parameterized by a transformation matrix. Then, a dictionary is adaptively learnt in the reduced space. The sparse coding coefficients are used as the learned features inputted to the classifier (SVM). Further, to ensure the classification performance, the classification results (hinge loss of the SVM) in turn are used to guide the transformation matrix and dictionary learning process, which enables the proposed method to obtain more discriminative information. By integrating these process together, the dimension reduction transformation matrix, dictionary and classifier parameters are optimized simultaneously in one framework during the training stage. In the testing stage, the low-dimensional representation of high-dimensional data can be represented by the learned transformation matrix and dictionary. The label of the test data can be predicted simply with the learned classifiers.

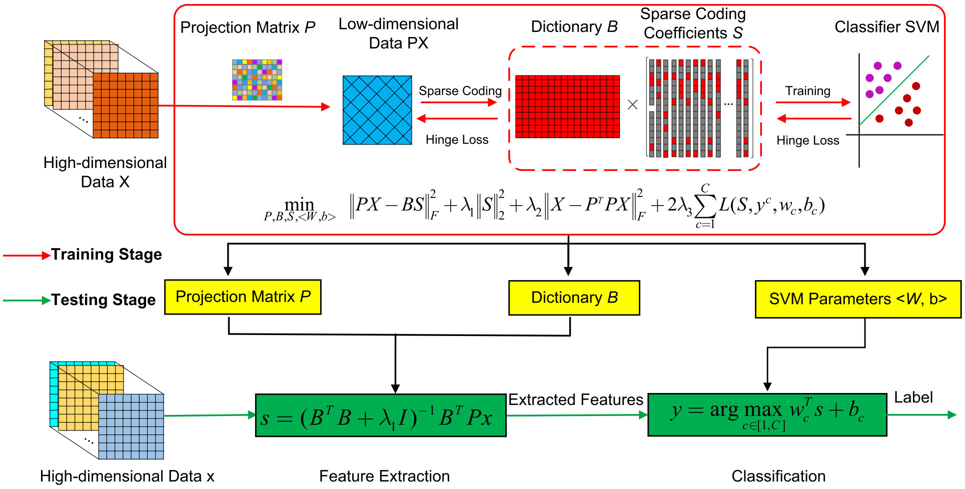


Fig. 2 The framework of the proposed method

2.2 The objective function

Let $X=[x_1,x_2,\dots,x_N]=[X_1,X_2,\dots,X_C]\in R^{D\times N}$ be the training samples of high-dimensional data, where C represents the number of classes. X_i denotes total training samples from class i, each sample is a D-dimensional vector. The optimization problem of dictionary learning can be formulated as

$$\min_{B,S}\|X-BS\|_F^2+\lambda_1\|S\|_p^p \tag{1}$$

where $\|\cdot\|_F^2$ is the Frobenious norm of a matrix. B is the desired dictionary, and K is the number of dictionary atoms. S is the sparse coding matrix of X over B. λ_1 is a Lagrange multiplier. The cost function in formulation (1) promotes a dictionary that can best represent X by minimizing the reconstruction error.

While for high-dimensional data, the dimensions of attributions are usually several hundreds or thousands. The dictionary learning algorithms are confronted with significant challenges when dealing with high-dimensional data, because the memory usage and computational complexity increase dramatically as the number of dimension increases. Moreover, some irrelevant attributes of the high-dimensional data may derail the dictionary learning process. To tackle with these issues, some dimension reduction methods are usually adopted to reduce the dimension of original high-dimensional data. Generally, it aims at learning an orthogonal transformation matrix and constructing a low-dimensional representation of the high-dimensional data as follows.

$$\min_{P}\left\|X-P^TPX\right\|_F^2 \quad s.t. PP^T=I \tag{2}$$

where $P\in R^{d\times D}$, d denotes the lower dimensionality of the reduced data and $d<D$. I is the identity matrix. Hence, the low-dimensional representation of original high-dimensional data can be represented as $\tilde{X}=PX\in R^{d\times N}$.

However, it is worth noting that the transformation matrix P is always pre-learned before dictionary learning in most previous works. Such a pre-learned transformation matrix limits the ability of dictionary learning. To address this shortcoming, the dimension reduction and dictionary learning process can be integrated together to jointly learn a transformation matrix and dictionary. Towards this end, the optimization problem can be exploited as follows.

$$\min_{P,B,S}\|PX-BS\|_F^2+\lambda_1\|S\|_p^p+\lambda_2\left\|X-P^TPX\right\|_F^2 s.t. PP^T=I \tag{3}$$

where λ_1 and λ_2 are two positive constants. Then the dictionary B can be learned in the low-dimensional space and $B\in R^{d\times K}$. Specially, the first term of (3) denotes dictionary learning in the reduced space. The second term ensures the coefficients matrix being sparse. The third term represents the difference between the reconstruction data and original high-dimensional data. By this way, the formulation (3) jointly conducts dimension reduction and dictionary learning.

Although the joint dimension reduction and dictionary learning method can preserve the inherent structure in the low-dimensional space, the feature extraction approaches have no direct connection with the classifier thus they cannot fit the classifier perfectly. Intuitively, the discriminative capability of extracted features can be improved when the dictionary is tuned to the classification task it is intended for. Following the idea of popular used classifier SVM, the classification loss (hinge loss of SVM) is further incorporated on the joint dimension reduction and dictionary learning objective function, to guide the transformation matrix and dictionary learning process. In this spirit, a transformation matrix, a dictionary and classifier parameters can be simultaneously learned in one framework by solving the following optimization problem.

$$\min_{P,B,S,W,b} \|PX - BS\|_F^2 + \lambda_1 \|S\|_p^p + \lambda_2 \left\|X - P^T PX\right\|_F^2 + 2\lambda_3 \sum_{c=1}^{C} L(S, y^c, w_c, b_c) \quad s.t. PP^T = I \tag{4}$$

where $W = \left[w_1, w_2, \dots w_C\right]$ and $b = \left[b_1, b_2, \dots b_C\right]$ denote C hyperplanes and the corresponding biases, respectively. y is the label vector, and $L(S, y, w, b)$ is defined as:

$$L(S, y, w, b) = \|w\|_2^2 + \theta \sum_{i=1}^{N} l(s_i, y_i, w, b) \tag{5}$$

where θ is a fixed constant, $l(s_i, y_i, w, b)$ is a quadratic hinge loss function, which can be written as follows:

$$l(s_i, y_i, w, b) = \{\max(0, 1 - y_i(w^T s_i + b))\}^2 \tag{6}$$

According to [33], l_2-norm regularizer is more computational efficient than l_1-norm regularizer, then (7) can be simplified as:

$$\min_{P,B,S,W,b} \|PX - BS\|_F^2 + \lambda_1 \|S\|_F^2 + \lambda_2 \left\|X - P^T PX\right\|_F^2 + 2\lambda_3 \sum_{c=1}^{C} L(S, y^c, w_c, b_c) \quad s.t. PP^T = I \tag{7}$$

The formulation (7) consists of four parts. The first item is designed to minimize the reconstruction error between PX and BS. The second item enforces the coefficients being sparse. The third term requires the data in the high-dimensional space can be well reconstructed by P. The last item is the classification loss, which is exploited to guide the dimension reduction and dictionary learning process. By introducing the classification loss term, the model (7) has the potential to catch discriminative information. From the objective function, it can be noted that the proposed method can simultaneously learn a transformation matrix P, a dictionary B and SVM parameters $< W, b >$ in one optimization framework, which is efficient to extract low-dimensional and discriminative features for high-dimensional data classification task.

2.3 Optimization

The objective function in (7) is not convex for P, B, S and $< W, b >$ simultaneously, it can be partitioned into three sub-problems and optimized iteratively by updating one variable with other ones fixed. The optimization procedures are detailed in the following subsections. Although the optimization method yields local minimum solution, this local minimum has been experimentally shown in the literature to be good enough for the classification task [55].

2.3.1 Learn S with fixed P, B and $< W, b >$

When P, B and $< W, b >$ are fixed, the coefficient matrix S can be optimized as follows:

$$\min_S \|PX - BS\|_F^2 + \lambda_1 \|S\|_F^2 + 2\lambda_3 \theta \sum_{i=1}^{N} \sum_{c=1}^{C} l(s_i, y_i^c, w_c, b_c). \tag{8}$$

In each iteration, for each c, if $y_i^c\left(w_c^T s_i + b_c\right) - 1 > 0$ in the previous iteration, the squared hinge loss is replaced by $\left\|y_i^c\left(w_c^T s_i + b_c\right) - 1\right\|^2$, else it is replaced by 0.

Let $\tilde{X} = PX$ denote the data after dimension reduction. When $y_i^c\left(w_c^T s_i + b_c\right) - 1 > 0$, for each s_i, the optimization problem reduces to:

$$\begin{cases} \min\limits_{s_i} f(s_i) \\ f(s_i) = \|\tilde{x}_i - Bs_i\|_F^2 + \lambda_1 \|s_i\|_2^2 + 2\lambda_3 \theta \sum\limits_{c=1}^{C} \left\|1 - y_i^c(w_c^T s_i + b_c)\right\|_2^2 \end{cases} \tag{9}$$

Further, the $f(s_i)$ can be simplified to :

$$\tilde{f}(s_i) = Tr\{-2s_i \tilde{x}_i^T B + s_i^T B^T B s_i + \lambda_1 s_i^T s_i\} + 2\lambda_3 \theta \sum_{c=1}^{C} Tr\{-2s_i y_i^c w_c^T + s_i^T w_c w_c^T s_i + 2s_i b_c^T w_c^T\} \tag{10}$$

Take the deviation of the objective function in (10) with respect to s_i, yields

$$\frac{\partial \tilde{f}(s_i)}{\partial s_i} = 2B^T B s_i + 2\lambda_1 s_i - 2B^T \tilde{x}_i + 4\lambda_3 \theta \sum_{c=1}^{C} w_c w_c^T s_i + w_c b_c - w_c y_i^{cT} \tag{11}$$

Let the above equation equal zero, and it can be got that

$$s_i = \left(B^T B + \lambda_1 I + 2\lambda_3 \theta W W^T\right)^{-1} \left[B^T \tilde{x}_i + 2\lambda_3 \theta \left(w_c y_i^{cT} - w_c b_c\right)\right] \tag{12}$$

where $W = [w_1, w_2, \dots, w_C]$.

When $y_i^c\left(w_c^T s_i + b_c\right) - 1 \le 0$, the optimization problem is equivalent to

$$\min_S \|\tilde{X} - BS\|_F^2 + \lambda_1 \|S\|_F^2 \tag{13}$$

Taking the derivatives of (13) with respect to S and letting the equation equal zero, we can get the closed form of S mathematically as follows:

$$S = \left(B^T B + \lambda_1 I\right)^{-1} B^T \tilde{X} \tag{14}$$

2.3.2 Learn B with fixed P, S and $< W, b >$

When P, S and $< W, b >$ is fixed, (7) is converted to the following problem:

$$\min_B \|PX - BS\|_F^2 \quad s.t. \sum_{i=1}^{r} B_{i,j}^2 \leq 1, \forall j = 1, \ldots, K \tag{15}$$

It can be solved effectively by the Lagrange dual method [56]. Consider the Lagrangian:

$$L(B, \lambda) = Tr((PX - BS)^T (PX - BS)) + \sum_{j=1}^{K} \lambda_j (\sum_{i=1}^{r} B_{i,j}^2 - c) \tag{16}$$

where $\lambda = \left(\lambda_1, \ldots, \lambda_K\right)$, $\lambda_i > 0$ denotes the Lagrange multiplier of the ith equality constraint. $\Lambda \in R^{K \times K}$ is a diagonal matrix and its elements are defined as $\Lambda_{ii} = \lambda_i (i = 1, \ldots, K)$ for all i. The optimal bases B can be obtained as follows:

$$B^T = (SS^T + \Lambda)^{-1} (\tilde{X} S^T)^T. \tag{17}$$

2.3.3 Learn $< W, b >$ with fixed P, B and S

When P, B and S are fixed, (7) can be rewritten as:

$$\min_{<W,b>} \sum_{c=1}^{C} L(S, y^c, w_c, b_c), \tag{18}$$

where $L(S, y^c, w_c, b_c) = \left\|w_c\right\|_2^2 + \theta \sum_{i=1}^{N} l(s_i, y_i^c, w_c, b_c)$ and $l(s_i, y_i^c, w_c, b_c)$ is a quadratic hinge loss:

$$l(s_i, y_i^c, w_c, b_c) = \begin{cases} 0, if \ 1 - y_i^c (w_c^T s_i + b_c) \leq 0 \\ \left\| y_i^c (w_c^T s_i + b_c) - 1 \right\|_2^2, \ else \end{cases} \tag{19}$$

Equation (18) is an unconstrained optimization problem. In this work, the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) is adopted to solve the problem as suggested in [57].

2.3.4 Learn P with fixed B, S and $< W, b >$

When B, S and $< W, b >$ are fixed, (7) can be rewritten as:

$$\begin{cases} f(P) = \|PX - BS\|_F^2 + \lambda_2 \left\|X - P^{\mathrm{T}} PX\right\|_F^2 \\ \min_P f(P) \quad s.t. PP^{\mathrm{T}} = I \end{cases} \tag{20}$$

The cost function can be expanded as follows:

$$\begin{aligned} f(P) &= \mathrm{Tr}\{\lambda_2 (X^T - X^T P^T P)(X - P^T PX) \\ &\quad + (X^T P^T - S^T B^T)(PX - BS)\} \\ &= \mathrm{Tr}\{(1 - \lambda_2) PXX^T P^T - 2S^T B^T PX + W\} \end{aligned} \tag{21}$$

where $W = \lambda_2 X^{\mathrm{T}} X + S^{\mathrm{T}} B^{\mathrm{T}} BS$. Note that W is a constant when B, S are fixed, the cost function can be further simplified as:

$$\tilde{f}(P) = Tr\{(1 - \lambda_2) PXX^T P^T - 2S^T B^T PX\} \tag{22}$$

Thus the optimization function can be simplified to a more elegant form:

$$\min_P Tr\{(1 - \lambda_2) PXX^T P^T - 2S^T B^T PX\} \quad s.t. \quad PP^T = I \tag{23}$$

The above model can be efficiently solved by the algorithm exploited in [58]. These steps are repeated until the convergence is met or maximum number of iteration is reached. In summary, the procedure of the proposed method is listed in Algorithm 1.

Algorithm 1 The Proposed Method

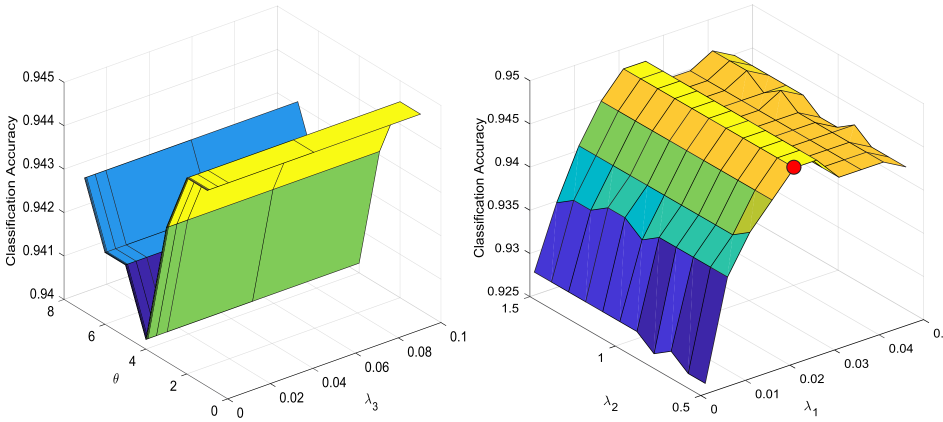
```
Input: Training set $X = [x_1, x_2, \ldots, x_N] \in R^{D \times N}$,
    Labels $y = [y_1, y_2, \cdots, y_N]$where $y_i \in \{1, 2, \cdots, C\}$ ,
    Parameters $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\theta$, reduced dimensionality $d$, and max iteration number $T$.
Output: Dimension reduction transformation matrix $P$, dictionary $B$ and one-against-all
    SVM parameters $< W, b >$
Step 1. Initialization
    1.1 Initialize $P$ as the PCA projection matrix;
    1.2 initialize the atoms of $B$ with DCT;
Step 2. Fix $< W, b >$ , $P$ and $B$, update $S$
    If $y_i^c \left(w_c^T s_i + b_c\right) - 1 > 0$ , compute $S$ according to (12), else, compute $S$ according to (14);
Step 3. Fix $< W, b >$ , $P$ and $S$, update $B$
    Compute Lagrange dual $\Lambda$ using Newton's method or conjugate gradient. Then, induce
    the $\Lambda$ in (17) to compute dictionary $B$;
Step 4. Fix $< W, b >$ , $B$ and $S$, update $P$
    Compute $P$ as an optimization problem with orthogonal constrained method by the method
    presented in [57] ;
Step 5. Fix $B$, $S$ and $P$, update $< W, b >$
    Compute $< W, b >$ with LBFGS algorithm presented in [58] ;
Step 6. Go to Step 2 until the convergence is met or maximum number of iteration $T$ is
    reached.
```

2.4 Classification

Once the dimension reduction matrix P, dictionary B and the classifier parameterized by $< W, b >$ are learned, the classification of high-dimensional data can be performed as follows. For a test sample x, it can be coded as $s = \tilde{P}x$, where $\tilde{P} = (B^T B + \lambda_1 I)^{-1} B^T P$. Then the label of x can be predicted simply with linear classifiers $< w_c, b_c >$, where $c = 1, 2, \ldots, C$ via (24).

$$y = \arg \max_{c \in [1, C]} w_c^T s + b_c \tag{24}$$

Fig. 3 Classification accuracies on the AR Face dataset versus different parameters



(a) λ_3 and θ

(b) λ_1 and λ_2

3 Experimental results

To evaluate the classification performance, the proposed method is applied on several publicly available high-dimensional datasets, including the AR Face dataset[1], Isolet dataset[2], and Colon dataset[3]. The details of these datasets are presented later. The proposed method is first compared with several related dictionary learning methods.

Individual dimension reduction and dictionary learning (IDRDL): this category first uses PCA technology to pre-reduce the dimension, then learns a desirable dictionary via the K-means Single Value Decomposition (KSVD) algorithm [59]. The projection matrix, dictionary and classifier are learned individually. This category is termed as PCA + DL + SVM.

Joint dimension reduction and dictionary learning (JDRDL) [45]: this category conducts dimension reduction and dictionary learning in a united scheme to jointly optimize the projection matrix and dictionary. However, the classifier is separately learned from the JDRDL stage. Unlike the proposed method using classification loss to optimize the dimension reduction and dictionary learning process, no criterion is exploited to guide the transformation matrix and dictionary learning process.

Discriminative dictionary learning (DDL): this category considers discriminant power of extracted features in the dictionary learning stage. However, the dimension reduction stage is still separate from the dictionary learning stage. This category includes PCA + label consistent K-SVD (LC-KSVD) [52], PCA + Fisher Discrimination dictionary learning (FDDL) [51]and PCA + Support vector guided dictionary learning (SVGDL) [53].

3.1 Parameters selection

As shown in the formulation (7), there are four important parameters $\lambda_1, \lambda_2, \lambda_3, \theta$ needed to be decided in advance in the proposed method. λ_1 is related with sparsity of the coefficients, λ_2 is added to ensure that the origin high-dimensional data can be well constructed in the low-dimensional space, λ_3 controls the weight of classification loss, θ is the hyper parameter of SVM. These parameters are selected and determined by the grid search method.

The effect of parameters λ_3 and θ is first studied. Specially, λ_1 and λ_2 are fixed as 0.03 and 0.5 while λ_3 and θ are turned from $[1e-4, 5e-4, 1e-3, 2e-3, 2e-3, 5e-2, 1e-1]$ and $[1, 2, 3, 4, 5, 6, 7]$, respectively. Figure 3a shows the classification accuracies versus different λ_3 and θ. From Fig. 3a, it can be observed that good performance can be obtained in a wide range of λ_3 and θ. Hence, the parameters λ_3 and θ can be set at the middle interval of the tuned range. Therefore, λ_3 and θ are set as 0.002 and 4, respectively.

λ_1 and λ_2 are then selected and determined. In detail, the suitable λ_1 and λ_2 are investigated by assigning $\lambda_1 \in [0.001, 0.01, \ldots, 0.05]$, $\lambda_2 \in [0.1, 0.2, \ldots, 0.5]$. Figure 3b reports the classification accuracies versus different parameters on the AR Face dataset. It is obvious that when $\lambda_1 = 0.03$ and $\lambda_2 = 0.5$, the classification accuracy reaches the highest point. Therefore, the parameters λ_1 and λ_2 are set as 0.03 and 0.5 on the AR Face dataset in the experiment.

3.2 Convergence study

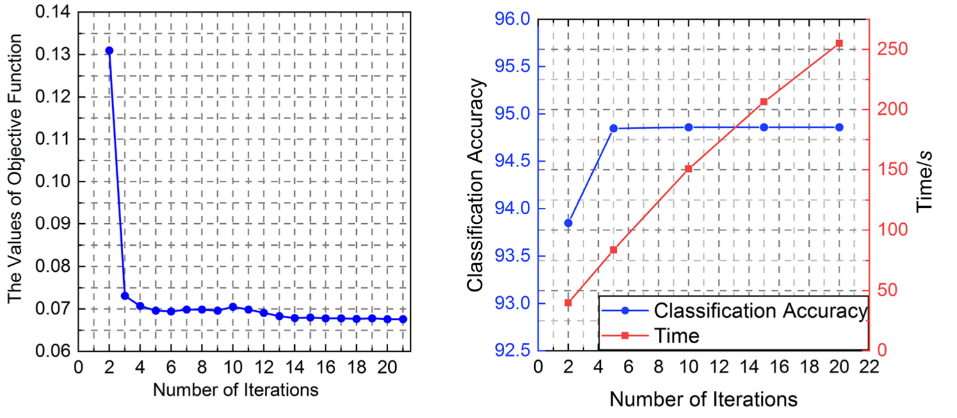
The convergence characteristic of the proposed method is investigated by providing some numerical results. The resulted objective function values of (7) between iterations are plotted in Fig. 4a. As observed, the objective function value of the proposed method converges to a fixed value when the number of iterations arrives at around 15. In addition, Fig. 4b depicts classification accuracies on the AR

[1] http://www2.ece.ohio-state.edu/~aleix/ARdatabase/.

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

[3] http://microarray.princeton.edu/oncology/.

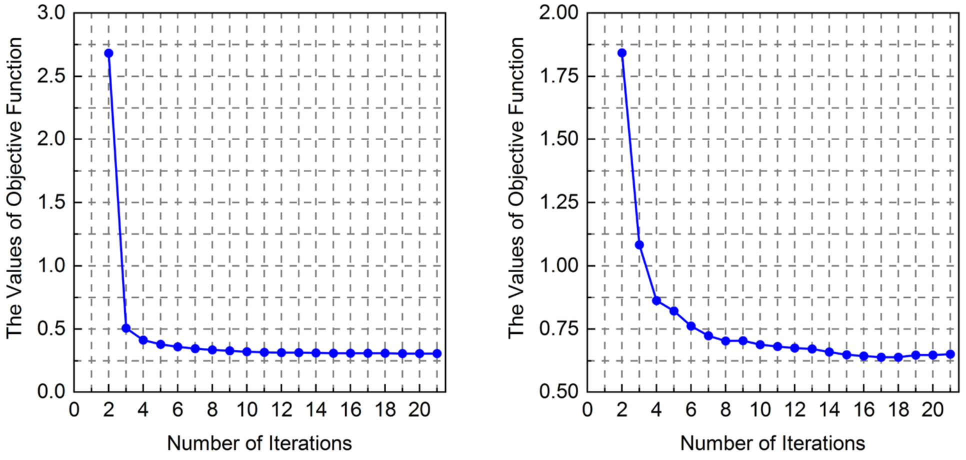
Fig. 4 Convergence behavior of the proposed method on the AR Face dataset



(a) The objective function values between iterations.

(b) Classification accuracies and computation costs between iterations.

Fig. 5 Convergence behavior of the proposed method on the Isolet and Colon dataset



(a) The Isolet dataset.

(b) The Colon dataset.

dataset, when the maxima number of iterations are ranged from 2 to 20. It can be noted classification accuracies quickly arrives at a stable value, while the corresponding running time keeps increasing linearly. Similar observations can be found from the Isolet and Colon dataset, as shown in Fig. 5a, b, respectively. Therefore, the number of iterations is set at 15 in all the experiments. These observations demonstrate that the proposed is highly efficient for high-dimensional data classification task in practice because of the fast converse speed.

3.3 Experimental results on the AR Face dataset

The AR Face dataset consists of over 4000 images of 126 people. A subset selected from the database are used in this experiment. The subset consists of 100 subjects with 6 illumination and 8 expression variations. The resolution of each image is 60 × 43 pixels, hence the vectorization of a face image is a 2580-dimensional vector. The training data and test data used in this experiment both are 700 samples. The projected dimension and dictionary size are set as 300 and 500, respectively.

Table 1 Classification accuracies on the AR Face dataset

Method	Accuracy (%)	Method	Accuracy(%)
PCA + DL + SVM	87.86	PCA + SVGDL	94.29
PCA + LCKSVD2	89.71	JDRDL + SVM	88.86
PCA + FDDL	93.14	The proposed method	**94.86**

The evaluation index of classification performance is defined by the classification accuracy. A higher value of accuracy means that the classification method has led to a better performance. Table 1 tabulates the average classification accuracies of all competing methods. The best result is emphasized in bold. According to the results, the proposed method significantly outperforms the PCA + DL + SVM method. This is mainly because some useful information

Table 2 Classification accuracies (%) on the AR Face dataset with different number of training samples

No. of training samples	3	4	5	6
PCA + DL + SVM	70.86	79.43	84.43	85.57
PCA + LCKSVD2	79.71	80.29	82.71	88.14
PCA + FDDL	71.57	86.14	87.86	89.71
PCA + SVGDL	73.29	81.57	89.14	92.29
JDRDL + SVM	75.86	81.86	89.71	92.43
The proposed method	**86.14**	**90.43**	**91.43**	**93.71**

may be lost in the separate dimension reduction and dictionary learning stage. When it comes to the DDL methods, which mainly includes PCA + LCKSVD2, PCA + FDDL, and PCA + SVGDL, the proposed method can significantly improve classification accuracies, which proves the effectiveness of joint dimension reduction and dictionary learning. Also, it is clear that the proposed method consistently performs better, when compared with JDRDL methods. It indicates that embedding classification results in the cost function indeed improve the classification performance. All these results validate that the proposed method is effective in dealing with the high-dimensional classification task.

Then the classification results of the proposed method against the number of training samples is investigated. 3–6 samples each class are randomly selected for training. Table 2 depicts the classification accuracies versus different number of training samples. The best result is indicated in bold. It can be noted that the proposed method is less sensitive to the number of training samples. In particular, when the training number is 3, the proposed method gets 86.14% in accuracy while the best value obtained by other competing methods only reaches 79.71%, which is far below than the proposed method's. This observation demonstrates that the proposed method can achieve a relatively stable performance when there are just a few training samples.

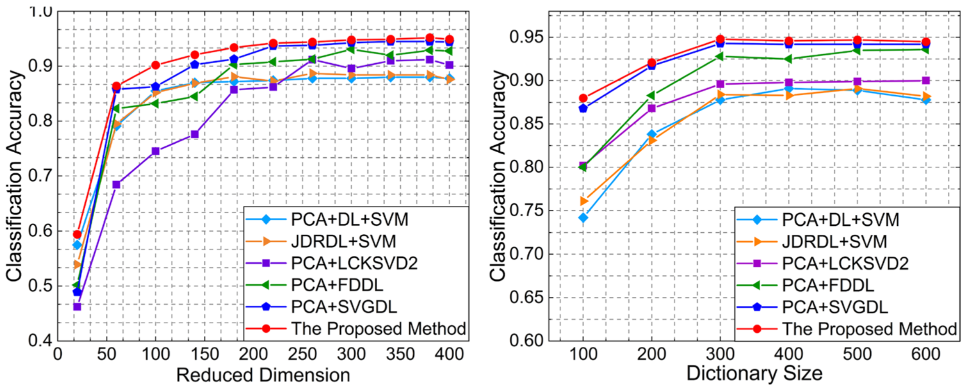
The effects of the different reduced dimensions and dictionary sizes are also evaluated. Figure 6a shows the classification accuracies of the proposed method versus different reduced dimensions on the AR Face dataset. It can be noted that the proposed method achieves stable performance when the feature dimension reaches around 300. In particular, the proposed method can guarantee better classification performance, even with fewer number of dimensions. Moreover, the classification accuracies of all the methods using different dictionary size are displayed in Fig. 6b. It can be seen that the proposed method achieves stable performance when the dictionary size reaches around 500. On the other hand, when the dictionary size exceeds 500, the classification accuracies are not improved significantly or decreased. It is interesting to observe that the classification results of the proposed method and SVGDL are nearly equal. This is because they both use the classification loss to determine the dictionary. This also indicates that the classification results can play an important role in learning a discriminative dictionary. Since the proposed method implement dimension reduction, dictionary and classification in a united framework, it can further improve classification performance.

3.4 Experimental results on the Isolet dataset

The Isolet dataset contains 1560 recognition data of 26 spoken letters. Each data is composed of 617 attributes. 12 samples per category are randomly selected as the training samples and the remaining for testing. The experiments are performed in the same way as the AR Face dataset. All the samples are projected into a 150 dimensional subspace. The dictionary size is set as 182. The optimal parameters selected by the grid search method are $\lambda_1 = 0.041$, $\lambda_2 = 0.02$, $\lambda_3 = 0.002$ and $\theta = 2$. Table 3 lists the classification accuracies. The best result is indicated in bold.

As shown in Table 3, it is easy to see that when compared with the PCA+DL+SVM method, the proposed method

Fig. 6 Classification results versus different reduced dimensions and different dictionary sizes on the AR Face dataset



(a) Different reduced dimensions.

(b) Different dictionary sizes.

Table 3 Classification accuracies on the Isolet dataset

Method	Accuracy (%)	Method	Accuracy (%)
PCA + DL + SVM	84.62	PCA + SVGDL	93.59
PCA + LCKSVD2	90.71	JDRDL + SVM	91.99
PCA + FDDL	93.27	The proposed method	**95.51**

Table 4 Classification accuracies (%) on the Isolet dataset with different number of training samples

No. of training samples	5	6	7	8	9
PCA + DL + SVM	71.79	76.28	77.56	78.21	78.53
PCA + LCKSVD2	74.47	80.77	81.45	83.01	85.58
PCA + FDDL	75.96	79.49	83.01	83.33	84.29
PCA + SVGDL	78.20	81.41	83.01	84.29	86.54
JDRDL + SVM	75.32	80.45	82.69	83.65	86.54
The proposed method	**78.85**	**82.37**	**84.29**	**85.90**	**88.46**

consistently performs better, which proves the effectiveness of the proposed joint optimization framework. Compared with existing DDL methods such as PCA + LCKSVD2, PCA + FDDL and PCA + SVGDL, the proposed method achieves higher classification accuracy. This is because the joint dimension reduction and DDL can characterize more efficient information than these using two independent steps. Meanwhile, the proposed method achieves comparative performance than the JDRDL methods, which indicates that the classification results indeed improve the final performance. In summary, as displayed in Table 3, the proposed method achieves the highest classification accuracy, demonstrating that optimizing transformation matrix, dictionary and classifier parameters simultaneously could bring benefit to the final classification performance.

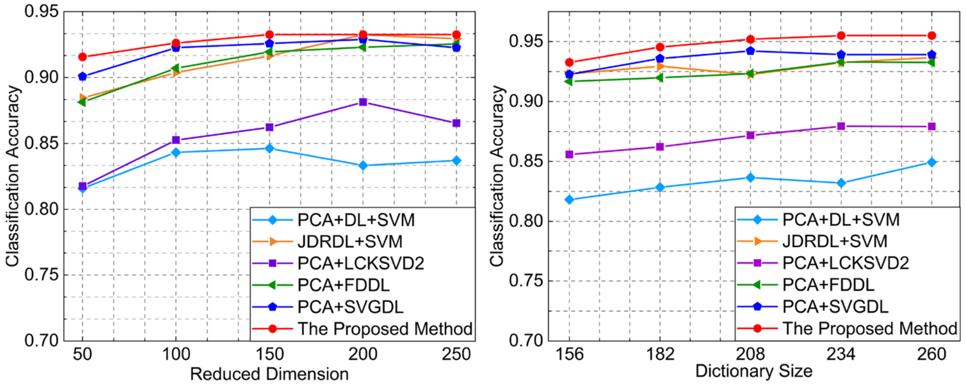
Similarly, different numbers of training samples, including 5, 6, 7, 8, 9 samples per category, are randomly selected for training. The classification accuracies versus different training numbers per class are reported in Table 4. The best result is labeled in bold. As observed, although the classification accuracies of all methods drop with the decrease of number of training samples, the proposed method still maintains satisfactory performance when the number of training samples is small. This observation clearly validates that the proposed method is robust to the number of training samples.

The effect of reduced dimension is evaluated on the classification performance. As Fig.7a demonstrates, the classification accuracies of the proposed method are gradually improved with the number of dimensionality increases. Moreover, compared with other methods, the proposed method achieves a better performance and maintains a relatively stable performance in lower dimensions. Next, the proposed method is compared with other methods with respect to different dictionary sizes. The classification results are depicted in Fig.7b, from which it can be noted that as the number of dictionary size increased, the classification performance of the proposed method can be improved. When the dictionary size reaches a fixed number, the performance of the proposed method becomes stable. Moreover, the proposed method performs better than other methods as the number of dictionary size varies. This shows that both the reduced dimension and dictionary size can significantly affect final classification results. Fortunately, the proposed method can maintain a remarkable performance across all dimensions and dictionary sizes in comparison to other methods.

3.5 Experimental results on the Colon dataset

The Colon dataset deal with the problem of cancer detection. This dataset contains 62 microarray samples of the

Fig. 7 Classification results versus different reduced dimensions and different dictionary sizes on the Isolet dataset



(a) Different reduced dimensions.

(b) Different dictionary sizes.

Table 5 Classification accuracies on the Colon dataset

Method	Accuracy (%)	Method	Accuracy (%)
PCA + DL + SVM	75.00	PCA + SVGDL	84.38
PCA + LCKSVD2	81.25	JDRDL + SVM	81.25
PCA + FDDL	81.25	The proposed method	**87.50**

Table 6 Classification accuracy(%) versus different dimensions on the Colon dataset

Methods	Dimensions				
	12	14	16	18	20
PCA + DL + SVM	68.75	71.88	75.00	75.00	71.88
PCA + LCKSVD2	71.88	78.13	81.25	81.25	75.00
PCA + FDDL	75.00	78.13	81.25	81.25	81.25
PCA + SVGDL	75.00	78.13	84.38	81.25	81.25
JDRDL + SVM	75.00	78.13	81.25	81.25	78.13
The proposed method	76.50	81.25	87.50	84.38	81.25

colon cancer. Every sample is composed of 2000 attributes. Obviously, it is characterized by a large number of attributes and a relatively small number of samples. The experiments are conducted with 30 randomly chosen training samples and 32 testing samples. The reduced dimension and dictionary size are set as 15 and 24, respectively. The parameters searched by the grid search method are $\lambda_1 = 0.021$, $\lambda_2 = 0.1$, $\lambda_3 = 0.002$ and $\theta = 2$. The classification accuracies are summarized in Table 5, the best value is shown in bold.

Similar to the results in other datasets, the proposed method is comparable to the IDRDL and DDL methods owing to the learned transformation matrix, which is more powerful than a pre-defined one. When compared with the JDRDL method, the proposed method achieves an obvious improvement. This is mainly because that the proposed method exploits the classification results to optimize dimension reduction and dictionary learning stage, which is beneficial for extracting discriminative information among original high-dimensional dataset and leading to better classification performance. All these results verify that the proposed method can effectively learn discriminative representations when the number of training samples is small.

Note that the Colon dataset itself is a typical high dimension small sample size (HDSSS) dataset, thus the classification performance versus different number of training samples is not specially investigated. Referring to the classification accuracies under different reduced dimensions, one can see that the proposed method achieves the best performance compared to the competing methods at each individual reduced dimension, and maintains a relatively remarkable performance in lower dimensions, as shown in Table 6. In addition, the classification results with respect to different dictionary sizes are listed in Table 7. It also can be noted that with the increasing number of dictionary sizes, the proposed method gets higher values at most time in comparison to other methods, which indicates that the proposed method is more efficient for high-dimensional data classification with small sample size.

3.6 Comparison to deep learning-based methods

Recently, deep neural networks have drawn significant attention and achieved promising performance in classification tasks. However, if the size of the training dataset is small, the classification performance of deep learning methods would drop significantly [60]. To show the superiority over the deep learning based-methods when the number of training samples is relatively limited, the proposed method is compared with stack auto-encoders (SAE) method and PCA + SAE [61] method. SAE means the features are directly learned from high-dimensional data using the stack auto-encoders, which is one of the representative deep learning models. PCA + SAE means the high-dimensional data is first projected to the low-dimensional space by PCA, and then the SAE rather than dictionary learning is adopted to learn features form the reduced space. Specially, a SAE model with two hidden layers is considered. The details of the experiments are listed in Table 8, in which h_1 and

Table 7 Classification accuracies(%) versus dictionary sizes on the Colon dataset

Methods	Dictionary sizes					
	16	18	20	22	24	26
PCA + DL + SVM	68.75	68.75	71.88	71.88	75.00	68.75
PCA + LCKSVD2	71.88	78.13	78.13	81.25	81.25	71.88
PCA + FDDL	71.88	78.13	78.13	78.13	81.25	78.13
PCA + SVGDL	81.25	81.25	81.25	84.38	84.38	81.25
JDRDL + SVM	78.13	81.25	81.25	81.25	81.25	78.13
The proposed method	81.25	81.25	84.38	84.38	87.50	81.25

Table 8 The implementation details of the deep learning-based methods

Methods	SAE	PCA + SAE
AR Face	$h_1 = 100, h_2 = 80$	$d = 300, h_1 = 100, h_2 = 80$
Isolet	$h_1 = 100, h_2 = 80$	$d = 150, h_1 = 100, h_2 = 80$
Colon	$h_1 = 12, h_2 = 10$	$d = 15, h_1 = 12, h_2 = 10$

Table 9 Classification accuracies(%) on the AR Face and Isolet dataset

Methods	SAE	PCA + SAE	The proposed method
AR Face	24.71	58.14	94.86
Isolet	33.01	60.51	95.24
Colon	68.75	88.33	85.70

h_2 denote the node numbers of the first hidden layer and the second hidden layer, respectively, and d represents the dimensionality of the reduced data.

Table 9 shows the classification performance on different datasets. According to the results, it can be found that the proposed method outperforms the deep learning-based methods and thus demonstrates its superiority to enhance the classification performance, when the data are high-dimensional with relatively small sample size. This is mainly because the deep learning-based methods exploit multilayer neural networks to learn hierarchical features, optimizing the parameters of these neural networks requires plenty of training numbers. However, since the AR Face, Isolet and Colon datasets do not consist of a lager number of training samples, the neural networks may be over-fit, so that the power of deep learning-based methods on such dataset is not so strong. In the proposed method, the dimension reduction transformation matrix, dictionary and classifier are simultaneously optimized to construct discriminative terms. In this way, the discrimination of the learned dictionary can be enhanced, resulting in superior classification performance.

3.7 Regularization terms effect

The contributions of different regularization terms in the proposed method are also investigated. Three alternative baselines are defined to study the importance of different terms in the proposed method. $\lambda_2 = 0$, learning the model without reconstruction term $\|X - P^{\mathrm{T}}PX\|_F^2$. $\lambda_3 = 0$, learning the model without classification loss term $\sum_{c=1}^{C} L(S, y^c, w_c, b_c)$. $\lambda_2 = \lambda_2 = 0$, learning the model without both reconstruction term and classification loss term. Note that λ_1 term is a basic regularizer of the proposed method, thus it is not considered in this subsection.

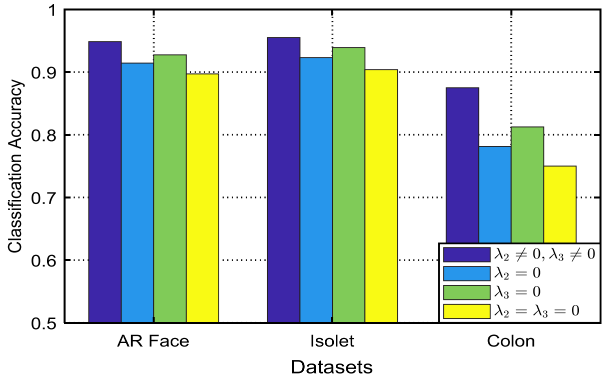


Fig. 8 Classification results on different datasets with/without λ_2 term and λ_3 term

Figure 8 illustrates the classification results with different regularization terms. When $\lambda_2 = \lambda_3 = 0$, the lowest classification performance is obtained, demonstrating that it is essential to minimize the reconstruction error and preserve label information in dimension reduction and dictionary learning stages. Obviously, the best performance can be obtained when both reconstruction term and classification loss term are used together to learn the model. This observation indicates that both of the two terms are helpful for final classification performance. Moreover, it can be noted that the reconstruction error term contributes more than the classification loss term in terms of the final performance. This is mainly because the reconstruction term can ensure that the data representation in the reduced space is presentative to the original high-dimensional data. Preserving the vital information from the original high-dimensional data is helpful for discrimination preservation in the later dictionary learning stage.

3.8 Discussion

The reasons why the proposed method achieves better performance on all classification tasks are discussed. It can be noted that the competing method can be divided into three types. The first type includes PCA + DL + SVM, which conducts dimension reduction, dictionary learning and classification individually. The second type includes PCA + LCKSVD2, PCA + FDDL, PCA + SVGDL that all exploit discriminative information for dictionary learning process, however the dimension reduction is separated from the dictionary learning process. The last type includes JDRDL method which jointly conducts dimension reduction and dictionary learning, but the classification error cannot be minimized simultaneously with the dimension reduction and dictionary learning process. Compared with these competing methods, the proposed method implements dimension reduction, dictionary learning and classifier learning in one optimization framework, so the learned features and classification errors can be simultaneously minimized. The classification results in turn can guide the dimension reduction and dictionary learning process, which is efficient to capture discriminative features from original high-dimensional data. These are the major reasons why the proposed method outperforms its competing methods by delivering higher classification accuracy. In addition, since the proposed method is a general feature learning framework, it can also be potentially applied to other tasks, such as high-dimensional data regression [62, 63] and clustering [64], by exploring alternative regression or clustering loss functions in their objective functions. The adaptation of the proposed framework to other problems involving high-dimensional data regression and clustering deserves a further study.

4 Conclusion

A novel feature learning framework is proposed for high-dimensional data classification. Different from existing methods which focus on individual dimension reduction, dictionary learning and classification, the proposed method considers the interaction of dimension reduction, dictionary learning and classification. By integrating them to a joint framework, the proposed method can extract low-dimensional and discriminative features, even when the training samples per class are limited. Extensive experiments demonstrate the effectiveness and superiority of the proposed method in different high-dimensional data classification tasks.

However, the proposed method can still be revised in some aspects. For instance, the proposed method is indeed a linear method, which fails to consider the complex nonlinear relationships within high-dimensional data. Also, the proposed method assumes that the sizes of different classes are similar. However, in real-word applications the collected data often exhibit imbalanced class distribution. Therefore, the future work will focus on nonlinear extension of the proposed method and considering imbalanced data classification problem.

Acknowledgements The authors would like to thank the support by National Natural Science Foundation of China (61773080 and 61633005), Fundamental Research Funds for the Central Universities (2019CDYGZD001), Scientific Reserve Talent Programs of Chongqing University (cqu2018CDHB1B04) and Graduate Scientific Research and Innovation Foundation of Chongqing (CYB20065).

References

1. Yamada M, Tang J, Lugo-Martinez J (2018) Ultra high-dimensional nonlinear feature selection for big biological data. IEEE Trans Knowl Data Eng 30(7):1352–1365
2. Sun W, Xie S, Han N (2019) Robust discriminant analysis with adaptive locality preserving. Int J Mach Learn Cybern 10:2791–2804
3. Wu Y, Hoi SCH, Mei T, Yu N (2017) Large-scale online feature selection for ultra-high dimensional sparse data. ACM Trans Knowl Discov Data 11(4):48.1–48.22
4. Tan M, Tsang IW, Wang L (2013) Minimax sparse logistic regression for very high-dimensional feature selection. IEEE Trans Neural Netw Learn Syst 24(10):1609–1622
5. Tan M, Wang L, Tsang IW (2010) Learning sparse SVM for feature selection on very high dimensional datasets. Proc Int Conf Mach Learn 2010:1047–1054
6. Zhang M, Li W, Du Q, Gao L, Zhang B (2020) Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN. IEEE Trans Cybern 50(1):100–111
7. Zhao W, Du S (2016) Spectral-spatial feature extraction for hyperspectral image classification: a dimension reduction and deep learning approach. IEEE Trans Geosci Remote Sens 54(8):4544–4554
8. Fei L, Lu G, Jia W, Teng S, Zhang D (2018) Feature extraction methods for palmprint recognition: a survey and evaluation. IEEE Trans Syst Man Cybern Syst 49(2):346–363
9. Wei Z, Peipei K, Xiaozhao F, Luyao T, Nan H (2019) Joint sparse representation and locality preserving projection for feature extraction. Int J Mach Learn Cybern 10:1731–1745
10. Kurup AR, Ajith M, Ramón MM (2019) Semi-supervised facial expression recognition using reduced spatial features and deep belief networks. Neurocomputing 367:188–197
11. Wang X, Zhang B, Yang M, Ke KY, Zheng WS (2019) Robust joint representation with triple local feature for face recognition with single sample per person. Knowl Based Syst 181:104790
12. Li S, Fu Y (2016) Learning robust and discriminative subspace with low-rank constraints. IEEE Trans Neural Netw 27(11):2160–2173
13. Xu N, Guo Y, Wang J, Luo X, Kong X (2017) Multi-view clustering via simultaneously learning shared subspace and affinity matrix. Int J Adv Robot Syst 14(6):1–8
14. Wang H, Wang P, Song L, Ren B, Cui L (2019) A novel feature enhancement method based on improved constraint model of online dictionary learning. IEEE Access 7:17599–17607
15. Zhang G, Porikli F, Sun H, Sun Q, Zheng Y (2020) Cost-sensitive joint feature and dictionary learning for face recognition. Neurocomputing 391:177–188
16. Song P, Weizman L, Mota JF, Eldar YC, Rodrigues MRD (2020) Coupled dictionary learning for multi-contrast MRI reconstruction. IEEE Trans Med Imaging 39(3):621–633
17. Cabrera D, Sancho F, Cerrada M, Sanchez R, Li C (2020) Knowledge extraction from deep convolutional neural networks applied to cyclo-stationary time-series classification. Inf Sci 524:1–14
18. Dimitriou N, Leontaris L, Vafeiadis T (2020) Fault diagnosis in microelectronics attachment via deep learning analysis of 3-D laser scans. IEEE Trans Ind Electron 67(7):5748–5757
19. Cao Z, Wan C, Zhang Z (2019) Hybrid ensemble deep learning for deterministic and probabilistic low-voltage load forecasting. IEEE Trans Power Syst 35(3):1881–1897
20. Wu G, Han J, Guo Y (2019) Unsupervised deep video hashing via balanced code for large-scale video retrieval. IEEE Trans Image Process 28(4):1993–2007

21. Goel A, Banerjee B, Pizurica A (2019) Hierarchical metric learning for optical remote sensing scene categorization. IEEE Geosci Remote Sens Lett 16(6):952–956
22. Wu X, Chen X, Li X (2014) Adaptive subspace learning: an iterative approach for document clustering. Neural Comput Appl 25(2):333–342
23. Huang K, Wu Y, Wen H (2020) Distributed dictionary learning for high-dimensional process monitoring. Control Eng Pract 98:104386
24. Li R, Pan Z, Wang Y (2019) A convolutional neural network with mapping layers for hyperspectral image classification. IEEE Trans Geosci Remote Sens 58(5):3136–3147
25. Shen F, Xu Y, Liu L (2018) Unsupervised deep hashing with similarity-adaptive and discrete optimization. IEEE Trans Pattern Anal Mach Intell 40(12):3034–3044
26. Han J, Cheng G, Li Z (2018) A unified metric learning-based framework for co-saliency detection. IEEE Trans Circ Syst Video Technol 28(10):2473–2483
27. Yu Y, Liu F, Mao S (2018) Fingerprint extraction and classification of wireless channels based on deep convolutional neural networks. Neural Process Lett 48(3):1767–1775
28. Sari CT, Gunduz-Demir C (2018) Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images. IEEE Trans Med Imaging 38(5):1139–1149
29. Gogna A, Majumdar A (2019) Discriminative autoencoder for feature extraction: application to character recognition. Neural Process Lett 49(3):1723–1735
30. Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1701–1708
31. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: Proceedings of British machine vision conference, pp 1–12
32. Pappus V, Panagopoulos OP, Xanthopoulos P, Pardalos PM (2015) Sparse proximal support vector machines for feature selection in high dimensional datasets. Expert Syst Appl 42:9183–9191
33. Zhu L, Zhang C, Zhang C, Zhang Z, Nie X, Zhou X, Wang X (2019) Forming a new small sample deep learning model to predict total organic carbon content by combining unsupervised learning with semisupervised learning. Appl Soft Comput 83:105596
34. Liaghat S, Mansoori EG (2019) Filter-based unsupervised feature selection using Hilbert Schmidt independence criterion. Int J Mach Learn Cybern 10(9):2313–2328
35. Vinyals O, Blundell C, Lillicrap T, Wierstra D (2016) Matching networks for one shot learning. In: Advances in neural information processing systems, pp 3630–3638
36. Li ZM, Lai ZH, Xu Y, Yang J, Zhang D (2017) A locality-constrained and label embedding dictionary learning algorithm for image classification. IEEE Trans Neural Netw Learn Syst 28(2):278–293
37. Sun Z, Hu Z, Wang M, Zhao S (2019) Dictionary learning feature space via sparse representation classification for facial expression recognition. Artif Intell Rev 51:1–18
38. Qi N, Shi Y, Sun X, Wang J, Yin B, Gao J (2018) Multi-dimensional sparse models. IEEE Trans Pattern Ana Mach Intell 40:163–178
39. Thomas M, Brabanter KD, Moor BD (2014) New bandwidth selection criterion for Kernel PCA: approach to dimensionality reduction and classification problems. Bmc Bioinform 15:1–12
40. Liu LT, Dobriban E, Singer A (2018) ePCA: high dimensional exponential family PCA. Ann Appl Stat 12:2121–2150
41. Wang A, Lu J, Cai J, Wang G, Cham TJ (2015) Unsupervised joint feature learning and encoding for RGB-D scene labeling. IEEE Trans Image Process 24:4459–4473
42. Lu J, Liong VE, Zhou J (2018) Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition. IEEE Trans Pattern Ana Mach Intell 40:1979–1993
43. Feng Z, Yang M, Zhang L, Liu Y, Zhang D (2013) Joint discriminative dimensionality reduction and dictionary learning for face recognition. Pattern Recognit 46(8):2134–2143
44. Nguyen HV, Patel VM, Nasrabadi NM, Chellappa R (2012) Sparse embedding: a framework for sparsity promoting dimensionality reduction. In: European conference on computer vision, pp 414–427
45. Chen Y, Su J (2017) Sparse embedded dictionary learning on face recognition. Pattern Recognit 64:51–59
46. Yang BQ, Gu CC, Wu KJ, Zhang T, Guan XP (2017) Simultaneous dimensionality reduction and dictionary learning for sparse representation based classification. Multimed Tools Appl 76:8969–8990
47. Foroughi H, Ray N, Zhang H (2018) Object classification with joint projection and low-rank dictionary learning. IEEE Trans Image Process 27:806–821
48. Zheng Z, Sun H (2019) Jointly discriminative projection and dictionary learning for domain adaptive collaborative representation-based classification. Pattern Recognit 90:325–336
49. Cheng M, Wu G, Yuan M, Wan H (2016) Semi-supervised software defect prediction using task-driven dictionary learning. Chin J Electron 25(6):1089–1096
50. Mairal J, Bach F, Ponce J (2012) Task-driven dictionary learning. IEEE Trans Pattern Anal Mach Intell 34(4):791–804
51. Yang M, Zhang L, Feng X, Zhang D (2014) Sparse representation based fisher discrimination dictionary learning for image classification. Int J Comput Vis 109:209–232
52. Jiang Z, Lin Z, Davis LS (2013) Label consistent K-SVD: learning a discriminative dictionary for recognition. IEEE Trans Pattern Anal Mach Intell 35:2651–2664
53. Cai S, Zuo W, Zhang L, Feng X, Wang P (2014) Support vector guided dictionary learning. In: European conference on computer vision, pp 624–639
54. Yang M, Chang H, Luo W (2017) Discriminative analysis-synthesis dictionary learning for image classification. Neurocomputing 219:404–411
55. Abdi A, Rahmati M, Ebadzadeh MM (2019) Dictionary learning enhancement framework: learning a non-linear mapping model to enhance discriminative dictionary learning methods. Neurocomputing 357:135–150
56. Lee, H, Battle A, Raina R, Ng AY (2007) Efficient sparse coding algorithms. In: Advances in neural information processing systems, pp 801–808
57. Yang, J, Yu K, Gong Y, Huang T (2009) Linear spatial pyramid matching using sparse coding for image classification. In: IEEE conference on computer vision and pattern recognition, pp 1794–1801
58. Wen Z, Yin W (2013) A feasible method for optimization with orthogonality constraints. Math Program 142:397–434
59. Aharon M, Elad M, Bruckstein A (2006) K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Signal Process 54(11):4311–4322
60. Foroughi H, Ray N, Zhang H (2018) Object classification with joint projection and low-rank dictionary learning. IEEE Trans Image Process 27(2):806–821
61. Fakoor R, Ladhak F, Nazi A (2013) Using deep learning to enhance cancer diagnosis and classification. In: Proceedings of the 30th International conference on machine learning, pp 1–7
62. Zhang W, Wang W, Wang J (2018) User-guided hierarchical attention network for multi-modal social image popularity prediction. In: Proceedings of the 2018 world wide web conference, pp 1277–1286

63. Mohammadi MR, Fatemizadeh E, Mahoor MH (2017) A joint dictionary learning and regression model for intensity estimation of facial AUs. J Vis Commun Image Represent 47:1–6

64. Ji M, Rao H, Li Z (2019) Partial multi-view clustering based on sparse embedding framework. IEEE Access 7:29332–29343

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.