I$_2$.p-Norm Based Discriminant Subspace Clustering Algorithm

XIAOBIN ZHI$^{1}$, LONGTAO BI$^{2}$, AND JIULUN FAN$^{2}$

$^1$School of Science, Xi’an University of Posts and Telecommunications, Xi’an 710121, China
$^2$School of Communications and Information Engineering, Xi’an University of Posts and Telecommunications, Xi’an 710121, China

Corresponding author: Xiaobin Zhi (xbzhi@163.com)

This work was supported in part by the National Science Foundation of China under Grant 61671377, Grant 61571361, Grant 61901365, and Grant 61601362, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2020JM-580, in part by the Science Plan Foundation of the Education Bureau of Shaanxi Province of China under Grant 18JK0719, and in part by the New Star Team of Xi’an University of Posts and Telecommunications under Grant xyt2016-01.

ABSTRACT Discriminative subspace clustering (DSC) combines Linear Discriminant Analysis (LDA) with clustering algorithm, such as K-means (KM), to form a single framework to perform dimension reduction and clustering simultaneously. It has been verified to be effective for high-dimensional data. However, most existing DSC algorithms rigidly use the Frobenius norm (F-norm) to define model that may not always suitable for the given data. In this paper, DSC is extended in the sense of I$_2$.p-norm, which is a general form of the F-norm, to obtain a family of DSC algorithms which provide more alternative models for practical applications. In order to achieve this goal. Firstly, an efficient algorithm for the I$_p$-norm based KM (KM$_p$) clustering is proposed. Then, based on the equivalence of LDA and linear regression, a I$_2$.p-norm based LDA (I$_2$.p-LDA) is proposed, and an efficient Iteratively Reweighted Least Squares algorithm for I$_2$.p-LDA is presented. Finally, KM$_p$ and I$_2$.p-LDA are combined into a single framework to form an efficient generalized DSC algorithm: I$_2$.p-norm based DSC clustering (I$_2$.p-DSC). In addition, the effects of the parameters on the proposed algorithm are analyzed, and based on the theory of robust statistics, a special case of I$_2$.p-DSC, which can show better robustness on the data sets with noise and outlier, is studied. Extensive experiments are performed to verify the effectiveness of our proposed algorithm.

INDEX TERMS Subspace clustering, linear discriminant analysis, I$_2$.p-norm, iterative reweighted least squares, robustness.

I. INTRODUCTION

Cluster analysis is a basic method for multivariate statistical analysis, and it is also an important part of unsupervised pattern recognition. The nature of clustering is to assign the data with similar patterns to the same cluster by exploring the intrinsic class structure of the data. In the past decades, a great variety of clustering methods have been proposed. For example, the well known K-means (KM) algorithm [1], [2], which is simple in principle, easy to implement, and widely used; spectral clustering algorithm [3], [4], which can effectively cluster the data with manifold structure, obtain wide attention from many scholars, and has become a milestone in the development of cluster analysis; the k-centroid [5], the affine propagation [6] and the sparse subset selection [7] algorithms, which can not only cluster the data, but also find the representative elements of each class, are also indispensable clustering methods.

In recent years, with the rapid development of information technology, in many fields such as information retrieval, image processing and computational biology, a large number of high-dimensional data emerge every day. For example, in web text data mining, if we use a vector space model to describe each document, the dimensions tend to exceed 5000 dimensions because the word vocabulary is often large. In genomics, DNA microarray data measures the expression levels of thousands of genes in a single experiment. Gene expression data usually contain a large number of genes (dimensions), but a small number of samples [32]. The high dimension of data significantly increases the complexity of time and space of the clustering algorithm. Many dimensions in data are not helpful or may even worsen the performance of the subsequent clustering algorithms. How to effectively and efficiently cluster high-dimensional data has
obtained wide attention from many scholars, and many methods have been proposed to handle it. In supervised learning, an effective way to handle high dimension data is dimension reduction including feature selection and feature extraction. However, feature selection in clustering is more challenging for both effectiveness and efficiency since the class labels are unavailable to guide the search for discriminative feature subsets. A commonly alternative method is feature weighting clustering [8]–[10], in which, feature selection can be performed by ranking the obtained feature weight. Feature selection clustering has the advantages of interpretability of the selected features and easy implementation on a database [17]. However, the rigidity of original dimension do not have enough flexibility to handle clusters which extends along a mixture of directions [17]. Unlike feature selection clustering, feature extraction-based clustering transforms data to achieve dimension reduction, in which new features can be generated. Previous feature extraction-based clustering methods often projected the data onto a low-dimensional subspace by unsupervised dimension reduction method such as principal component analysis (PCA) [11] or manifold learning algorithm [12] before the clustering [3], [4], [13], [14]. Due to the inherent separation between subspace selection (via dimension reduction) and clustering, the projection may not necessarily improve the separability of the data for clustering [19].

Linear discriminant analysis (LDA) is a classical supervised method for feature extraction and dimension reduction [15]. It computes an optimal linear transformation matrix by minimizing the within-class distance of the data set while maximizing the between-class distance in the linearly transformed low dimensional space simultaneously. In some recent work [16]–[18], LDA (or its variants) was combined with clustering process to improve the performance of clustering. This method, which is called discriminative subspace clustering (DSC) in this paper, uses LDA to project data onto the optimal transformation subspace while completes the data clustering in the low dimension transformation subspace, and optimizes these two processes alternately to perform the clustering and dimension reduction of data simultaneously. The experimental results have shown the superiority of the DSC in comparison with other popular clustering algorithms in the same period [16]–[18]. DSC was first proposed by De la Torre in 2006 [16]. However, the LDA used in the DSC algorithm proposed by De la Torre only uses the between-class information in data, and is not a full LDA [17]. Ding and Li proposed combining another form of LDA, which involves not only between-class information but also within-class information in data, with the classical KM clustering algorithm to perform DSC [17]. The DSC algorithm proposed by Ding and Li has the advantage of simple principle, and is very easy to handle. In [18], Ye et al. interpreted DSC as a clustering algorithm that can adaptively perform clustering and distance metric learning simultaneously. In [19], Ye et al. also proposed discriminative k-means (DKM), which transforms DSC into a kernel KM clustering with a special kernel matrix.

The experimental results in [19] show the superiority of the DKM algorithm in comparison with some manifold learning based clustering algorithms. In addition, an outstanding feature of the DKM algorithm is its high efficiency for high-dimensional and small-sample data. However, the computation a kernel matrix of on a data set with a large sample number is very time-consuming. So DKM is inconvenient to use on the data set with a large sample number. In order to extend the scope of application of DSC, fuzzy versions of DSC were proposed by extending the membership function to fuzzy (or soft) case [20]–[22]. The fuzzy versions of DSC can use the fuzzy membership function to characterize the degree to which data belongs to some class expeditiously, and reflect the internal structure of the data [21]. Additionally, Hou et al. replaced the LDA in the DSC method with a general form of maximum margin criterion (MMC) [24], avoiding the singularity problem of the scatter matrix implicitly, and proposed a variant of DSC: discriminative embedded clustering (DEC) algorithm [23]. In [25], DC was shown equivalent to another version of discriminative clustering framework [26], [27], in which the cost function of support vector machine is used for linear classification as a clustering criterion, and extended to sparse version and multi-label version further. Seeing that when the data are large and high-dimensional, most DSC algorithms are time-consuming, recently, using QR decomposition-based Linear Discriminant Analysis, Zhi et al. proposed an efficient DSC algorithm, which makes DSC can be used on more large data set [28].

Many DSC algorithms have been proposed, and lots of experimental results have shown the effectiveness of DSC. However, the definitions of the loss functions in most of the existing DSC algorithms are rigidly in fact based on the F-norm of matrix (Corresponding to the vector form, it is the Euclidean norm.), which limits the use of DSC. Since the $l_{2,p}$-norm of a matrix is a general form of its F-norm, and it has been used in many machine learning problem [29]–[33] successfully recently, in this paper, a generalized DSC is proposed by extending the DSC algorithm [17] in the sense of $l_{2,p}$-norm. The main contributions of this paper are summarized as follows: (1) The Euclidean norm based KM clustering process is generalized in the sense of $l_{p}$-norm, and an efficient $l_{p}$-norm based on KM (KMP) algorithm is presented. (2) LDA is generalized in the sense of the $l_{2,p}$-norm based on the equivalence between LDA and linear regression (LR) [38], and an $l_{2,p}$-norm based LDA ($l_{2,p}$-LDA) algorithm is proposed. The main difference between $l_{2,p}$-LDA and LDA lies in the first stage. The first stage of $l_{2,p}$-LDA can be transformed into exactly an Iteratively Reweighted Least Squares (IRLS) problem [37], which can be solved efficiently using LSQR algorithm [38]. (3) Combining the proposed $l_{2,p}$-LDA and KMP algorithm into a single clustering framework, a generalized DSC algorithm, which is called $l_{2,p}$-norm based discriminative subspace clustering ($l_{2,p}$-DSC), is presented. (4) An intelligent version of KMP algorithm is used to initialize $l_{2,p}$-DSC to avoid falling into the local optimum. The presented theories and algorithms are evaluated
through experiments on twelve synthetic and real-world data sets.

The rest of this paper is organized as follows. The notations and definitions used in this paper are introduced in Section II. Some related works including classical KM, intelligent KM (IKM) based on anomalous cluster [41], [42], LDA, DSC [16], [17] and some variants and extensions of DSC are reviewed in Section III. IKM is extended in the sense of $l_p$-norm, LDA is extended in the sense of $l_2,p$-norm based on the equivalence between LDA and linear regression, and $l_2,p$-norm based DSC algorithm is proposed in Section IV. Section V includes the experimental results and the conclusions are given in Section VI.

II. NOTATIONS AND DEFINITIONS

This section summarizes the notations and definitions of the norms used in this paper. Matrices are written as boldface uppercase letters. Vectors are written as boldface lowercase letters. For a matrix $M \in \mathbb{R}^{n \times m}$, its $i$-th row and $j$-th column are denoted by $m^i$ and $m_j$ respectively. The $l_p$-norm of a vector $v \in \mathbb{R}^n$ is defined as:

$$\|v\|_p = \left( \sum_{i=1}^{n} |v_i|^p \right)^{1/p}, \quad (1)$$

where $v_i$ is the $i$-th component of the vector $v$. When $p = 1$ and 2, the $l_p$-norm of a vector $v$ will degenerate its $l_1$ -norm $\|v\|_1 = \sum_{i=1}^{n} |v_i|$ and $l_2$-norm (Euclidean norm) $\|v\|_2 = \left( \sum_{i=1}^{n} |v_i|^2 \right)^{1/2}$ respectively; and when $p \to \infty$, the $l_p$-norm of a vector $v$ will degenerate its $l_\infty$-norm $\|v\|_\infty = \max_i |v_i|$.

The $l_{r,p}$-norm of a matrix $M \in \mathbb{R}^{n \times m}$ is defined as:

$$\|M\|_{r,p} = \left( \sum_{i=1}^{n} \left( \sum_{j=1}^{m} |m_{ij}|^r \right)^{1/p} \right)^{1/p}. \quad (2)$$

When $r = p = 2$, the $l_{r,p}$-norm of the matrix $M$ will degenerate into its Frobenius norm ($F$-norm)

$$\|M\|_F = \left( \sum_{i=1}^{n} \sum_{j=1}^{m} m_{ij}^2 \right)^{1/2} = \left( \sum_{i=1}^{n} \|m^i\|_2^2 \right)^{1/2}. \quad (3)$$

When $r = 2$, the $l_{r,p}$-norm of the matrix $M$ will degenerate into its $l_{2,p}$-norm

$$\|M\|_{2,p} = \left( \sum_{i=1}^{n} \left( \sum_{j=1}^{m} |m_{ij}|^2 \right)^{1/2} \right)^{1/2} = \left( \sum_{i=1}^{n} \|m^i\|_2^{1/2} \right)^{1/2}. \quad (4)$$

Comparing the definition of $l_p$-norm of a vector with the one of the $l_{r,p}$-norm of a matrix, it can be seen that the $l_{2,p}$-norm of a matrix will degenerate into its $l_p$-norm when the matrix degenerates into a vector.

In the field of machine learning, $l_{2,1}$-norm and $l_{2,p}$-norm are mainly used in two aspects: one is to use the $l_{2,1}$-norm of the transformation matrix or the $p$-th power of $l_{2,p}$-norm as a penalty term to implement feature selection [29], [30], [32], [33]. The other is to replace the $F$-norm in the loss function using the $l_{2,1}$-norm or $p$-th power of $l_{2,p}$-norm to improve the adaptability of the algorithm to data [31]–[37], such as the robustness to noise points and outliers in the data. Of course, $l_{2,1}$-norm and $l_{2,p}$-norm are also used in two aspects at the same time: both as a penalty term to implement feature selection, and a loss function to improve the algorithm’s adaptability to data [32], [33].

In view of the generality of the $l_{2,p}$-norm and its successful use in machine learning, DSC is extended in the sense of the $l_{2,p}$-norm in this paper. For convenience, the notations used in this article are summarized in Table 1. For simplicity, in the following discussion, we assume that $x_1, x_2, \cdots, x_n$ form rows of a data matrix $X \in \mathbb{R}^{n \times m}$, and have been centered, so that the total mean data set $\bar{x} = X e / n = 0$, where $e \in [1, 1, \cdots, 1] \in \mathbb{R}^n$.

| Notation | Description | Notation | Description |
|----------|-------------|----------|-------------|
| $n$      | Number of data points | $G^2$ | Transformation matrix |
| $m$      | Number of features | $\text{trace}(A)$ | Trace of matrix $A$ |
| $k$      | Number of classes | $S_2$ | Within-class scatter matrix |
| $l$      | Number of retained dimension in DSC | $S_b$ | Within-class scatter matrix |
| $X$      | Data matrix of size $m$ by $n$ | $S_t$ | Total scatter matrix |
| $C_j$    | The set of consisting of $j$-th cluster | $U$ | Cluster membership matrix |
| $n_j$    | Size of the $j$-th cluster | $V$ | cluster center matrix |

III. RELATED WORK AND MOTIVATION

In this section, firstly, the classical KM [18], [19] and the intelligent KM (IKM) algorithms [41], [42] are reviewed. Then, linear discriminant analysis (LDA) and LDA based on least squares [31] are also reviewed. Finally, DSC [16], its variants and the motivation of the proposed work is presented.

A. KM AND IKM

Consider a data set consisting of $n$ data points $\{x_i\}_{i=1}^{n}$ and denote data matrix as $X = [x_1, x_2, \cdots, x_n]$ whose $i$-th column is given by $x_i$. KM clustering finds the partition of the data that minimizes the following objection function

$$\min J_{\text{KM}}(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij} \|x_i - v_j\|_2^2, \quad (5)$$
where the matrix $V = [v_1, v_2, \ldots, v_k]$ consisting of $k$ cluster centers, and $v_i$ is the center of the $j$-th cluster, and the matrix $U = [u_{ij}]_{n \times k}$ is a cluster membership matrix such that $u_{ij} = 1$ if $x_i$ belongs to the $j$-th cluster $X_i$, and $u_{ij} = 0$ otherwise; and the following conditions are satisfied:

$$\sum_{j=1}^{k} u_{ij} = 1, \ i = 1, 2, \ldots, n.$$  

The objective function (5) measures the total error of using $k$ cluster centers to represent $k$ class data. KM algorithm is a well known and widely used clustering algorithm because of its simplicity and efficiency, but it has mainly the following three defects: (1) The value of $k$ needs to be given in advance and belongs to the pre-knowledge. In many cases, the estimation of $k$ is very difficult. (2) The algorithm is sensitive to the initialization, the clustering results obtained by different random seed points may different, and it may converge to the local minimum. (3) Noise and outliers have a great impact on the algorithm.

The intelligent version of KM, intelligent k-means (IKM) [41], [42], finds the so-called anomaly cluster before running KM itself, by extracting the exception clusters one by one till there are no unclustered objects, after which the centers of the largest anomalous clusters are used to initialize KM. The detailed IKM algorithm can be found in [41], [42]. IKM algorithm solves the problem that KM algorithm is sensitive to initialization to a certain extent.

**B. LINEAR DISCRIMINANT ANALYSIS BASED ON LEAST SQUARES**

LDA a classical supervised feature extraction and dimension reduction method [15]. In LDA, given membership matrix $U$, the weighted membership matrix [19] can be defined as follows:

$$L = [L_1, L_2, \ldots, L_c],$$  

where $L_j = \frac{1}{\sqrt{n_j}} U_j$, $U_j$ is the $j$-th column of $U$, and $n_j$ is the number of points in the $j$-th class. In LDA, the within-class, between-class, and total scatter matrices are defined as

$$S_w = \sum_{j=1}^{k} \sum_{x \in C_j} (x_i - v_j) (x_i - v_j)^T,$$

$$S_b = \frac{1}{n} \sum_{j=1}^{k} n_j v_j v_j^T = X L L^T X^T,$$

and

$$S_t = \frac{1}{n} X X^T,$$

where $v_j = \frac{1}{n_j} \sum_{x \in C_j} x$ is the means of the data points in $j$-th class. It can be shown that $S_t = S_b + S_w$. The sizes of trace ($S_b$) and trace ($S_w$) measure the separation between classes and the compactness of the data points within the class respectively. The goal of LDA is to find an optimal linear transformation $G$ that maps $m$-dimensional data into a low-dimensional space: $G^T : x_j \in R^m \rightarrow y_j \in R^t$ such that the class structure of data set in the original high-dimensional space is optimally preserved. An common used objective function of LDA is

$$\max_{G} \text{trace} \left( (G^T S_w G)^{-1} (G^T S_b G) \right).$$  

(10)

The optimal $G$ can be obtained by computing the eigen-decomposition on the matrix $S_w^{-1} S_b$ if $S_w$ is nonsingular. However, in many applications, the data are so high-dimensional that scatter matrix often does not satisfy nonsingularity, so the classical LDA cannot be used. The regularization method [43] is usually used to improve the singularity of the scatter matrix, i.e., using $\tilde{S} = S + \lambda I_m$ replace the singular scatter $S$, where $I_m$ is the unit matrix of size $m$, $\lambda > 0$ is a regularization parameter.

The solution of LDA can be transformed into a generalized eigenvalue problem, and the equivalence of this generalized eigenvalue problem and the least squares (LS) based linear regression (LR) problem has been strictly proved in [38]. Based on this equivalence, an efficient algorithm, which is called least squares based LDA (LSLDA), is proposed for LDA [38]. The LSLDA algorithm can be divided into two stages. The first stage aims to solve the following LS problem:

$$\min_{G_1} \left\| G_1^T X - L^T \right\|^2_F + \gamma \left\| G_1 \right\|^2_F.$$  

(11)

The second stage use the transformation matrix $G_1$ obtained in the first stage to project the original data onto a low-dimensional space, and perform a LDA procedure in the dimension reduced subspace [38]. Since the LS problem in LSLDA can be efficiently solved using LSQR algorithm [40], LSLDA can be used on large data sets.

**C. DSC**

The DSC [16], [17] algorithms can be seen unsupervised procedures of LDA. In DSC [16], [17], the transformation matrix $G$ and the cluster membership matrix $U$ are computed by maximizing the objective function as

$$\max_{G, U} \text{trace} \left( (G^T S_w G)^{-1} (G^T S_b G) \right),$$

(12)

or

$$\max_{G, U} \text{trace} \left( (G^T S_w G)^{-1} (G^T S_b G) \right).$$

(13)

respectively. The algorithms work in an intertwined and iterative fashion. More specifically, for a given $U$, the optimal $G$ can be obtain by the standard LDA procedure. For a given $G$, the optimal $U$ can be computed by applying the gradient descent strategy [16] or by solving KM problem in the lower-dimensional space resulting from the transformation $G$ [17].

Once DSC [16] was proposed, it has attracted the attention of many scholars. In [17], Ding and Li pointed out the LDA used in the DSC algorithm [16] only used the between-class information in data, and is not a full LDA [17]. They proposed combining another form of LDA, which involves not only
between-class information but also within-class information in data, with the classical KM clustering algorithm to perform DSC [17]. In [18], Ye et al. interpreted DSC as a clustering algorithm that can adaptively performs clustering and distance metric learning simultaneously. In [19], Ye et al. proposed discriminative KM (DKM), which transformed DSC into a kernel KM with the special kernel matrix. However, the computation of a kernel matrix on a data set with large sample number is more time-consumes. So DKM is not fit large data set. In [23], Hou et al. proposed Discriminative embedded clustering (DEC) algorithm, which can be formulated as following optimization problem

$$\max \text{trace} (G^T S b G) + (1 - \eta) \text{trace} (G^T S_n G)$$

where \(\eta\) is a balance parameter. DEC replaces the LDA in DSC with a general form of MMC [24], avoiding the singularity problem of the scatter matrix implicitly. Seeing that when the data are large and high-dimensional, most DSC algorithm are time-consumes, recently, Zhi et al. proposed an efficient algorithm for DSC by using QR decomposition-based LDA, which makes DSC can be used on more larger data set [28].

**D. MOTIVATION**

Many DSC algorithms have been proposed, and lots of experimental results have shown the effectiveness of DSC [16]–[28]. However, the definitions of the objective function in most of the existing DSC algorithms are based on the F-norm (Corresponding to the vector form, it is the Euclidean norm.). In fact, based on the equivalence between LDA and LS based LR [38], it can be seen that the optimization of DSC can be divided into two parts, i.e.

$$\min_{G_1, L} \left\| G_1^T X - L^T \right\|_F^2 + \gamma \left\| G_1 \right\|_F^2$$

and

$$\max \text{trace} (G_2^T S b G_2)$$

respectively, where \(S b = G_1^T S b G_1\). Eq.(15) constitutes the main body of the optimization of DSC since the optimization (16) does not change the class label. So, the objective function of DSC can be thought to be defined based on the F-norm. In [31], [32], [36], it was pointed out that the loss function based on the mean square F-norm is sensitive to noise and outliers in the data set. So the used square F-norm based loss function in the objective function of DSC may limits the application of DSC. Especially, if the data is with noise and outlier, the performance DSC may be greatly degraded. On the other hand, compared to squared F-norm, \(l_{2,1}\)-norm based loss function is more robust to noise and outlier [31], [36]. In [31], The \(l_{2,1}\)-norm of a matrix was first introduced in Principal Component Analysis to improve its robustness to noise and outlier. In [32], \(l_{2,1}\)-norm is applied to the construction of the loss function and penalty term of the linear regression to obtain a noise-resistant supervised feature selection algorithm. In [36], \(l_{2,1}\)-norm was applied to the construction of the loss function of linear discriminant analysis (LDA) to obtain a noise-resistant LDA algorithm. In [33], the algorithm in [32] was extended to the sense of \(l_{2,p}\)-norm to further improve the adaptability of the algorithm to data.

It is certain that the rigidly used F-norm to define loss functions in DSC would limit its application. Existing lots of work shows that replacing the F-norm in the loss function with \(l_{2,p}\)-norm can general the original algorithm, which provides more alternative models for practical applications, and can improve the flexibility of the original algorithm greatly. So, in this paper, the loss function of the DSC objective function will be extended in the sense of \(l_{2,p}\)-norm to improve the flexibility of the DSC algorithm.

**IV. DISCRIMINANT SUBSPACE CLUSTERING BASED ON \(l_{2,p}\)-NORM**

Since DSC is a framework of the combination of LDA and clustering, a direct method of extending DSC in the sense of \(l_{2,p}\)-norm is to extend LDA and clustering simultaneously. In order to achieve this goal, in this section, firstly, efficient algorithms for \(l_p\)-norm based KM (KM\(_p\)) and its intelligent version are presented. Then, based on the equivalence of FDA and linear regression, an \(l_{2,p}\)-norm based LDA \((l_{2,p}\)-LDA\) algorithm is proposed. Finally, KM\(_p\) and \(l_{2,p}\)-LDA are combined into a single clustering framework, and obtain an \(l_{2,p}\)-norm based generalized DSC algorithm: \(l_{2,p}\)-DSC. In addition, a special case of the proposed \(l_{2,p}\)-LSDC algorithm, which shows the robustness to noise and outlier in data set, is studied.

**A. INTELLIGENT KM BASED ON \(l_p\)-NORM**

The KM algorithm based on \(l_1\)-norm has been proposed in [43]. In [42], the feature weighted KM algorithm based on the more general \(l_p\)-norm has been proposed. The fuzzy \(k\)-means algorithm based on \(l_p\)-norm has been introduced in [44]. However, the KM algorithm based on \(l_p\)-norm (KM\(_p\)) has not been studied directly. In this section, an efficient algorithm for KM\(_p\) is presented.

Given the data matrix \(X = [x_1, x_2, \ldots, x_n]\), the objective function of the KM\(_p\) algorithm is:

$$\min J_{\text{KM}_p}(U, V) = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij} \left\| x_i - v_j \right\|_p$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij} \sum_{l=1}^{m} |x_{il} - v_{jl}|^p.$$  (17)

Similar as KM, KM\(_p\) can be solved by using alternating optimization method. Given the clustering center matrix \(V\), the optimal membership function matrix \(U\) can be obtained by computing the following equation

$$u_{ij} = \begin{cases} 1, & \text{if } \left\| x_i - v_j \right\|_p = \min_{1 \leq l < c} \left\| x_i - v_l \right\|_p \\ 0, & \text{else} \end{cases}.$$  (18)
Given the membership function matrix \( U \), for \( p = 1 \), the optimal clustering center of the \( j \)-th cluster is the median of the data point in the \( j \)-th cluster [43]. For \( p \neq 1 \), it can be shown that the objective function of \( \text{KM}_p \) can be reformulated as

\[
\min J_{\text{KM}_p}(U, V) = \sum_{j=1}^{c} \sum_{i=1}^{n} \sum_{l=1}^{m} u_{ij} \left( x_{ij} - v_{il} \right)^2 
\]

Differentiating Eq. (19) with respect to \( v_{il} \) and setting the derivative to zero, the \( v_{il} \) can be computed as

\[
v_{il} = \frac{\sum_{x \in X_{jl}} w_{xjl} x}{\sum_{x \in X_{jl}} w_{xjl},} 
\]

where \( X_{jl} \) is the set of consisting of the data points in the \( j \)-th clustering on the \( k \)-th feature, and \( \sum_{x \in X_{jl}} w_{xjl} = \left( x - v_{il} \right)^2 \right)^{p-2}/2 \). So given the membership function matrix \( U \), for \( p \neq 1 \), the optimal \( V \) can be obtained by computing Eq. (20). As a summary, the detailed algorithm flow of \( \text{KM}_p \) can be described as the following **Algorithm 1**.

**Algorithm 1 KMp Algorithm**

**Input:** Data matrix \( X \), the number of clusters \( k \), membership matrix \( U \), the clustering center matrix \( V \), parameter \( p \) and the convergence threshold \( \epsilon \).

1. Compute the \( i \)-th feature of the cluster center of the \( j \)-th cluster according the following rules: \( v_{il} = \begin{cases} \text{median}(X_{jl}), & \text{if } p=1 \\ \frac{\sum_{x \in X_{jl}} w_{xjl} x}{\sum_{x \in X_{jl}} w_{xjl}}, & \text{else} \end{cases} \) where \( X_{jl} \) is the set of consisting of the data points in the \( j \)-th clustering on the \( k \)-th feature and \( \sum_{x \in X_{jl}} w_{xjl} = \left( x - v_{il} \right)^2 \right)^{p-2}/2 \).

2. Update membership matrix \( U \) by computing Eq. (14). If \( \| U^T - U \| < \epsilon \), then stop iteration; else, let \( U \leftarrow U' \), return to Step 1.

**Output:** Membership matrix \( U \).

Based on the proposed \( \text{KM}_p \) algorithm in section IV.A, and the IKM algorithm [41], [42], an \( l_p \)-norm based intelligent K-means (IKMp) algorithm can be proposed. The following **Algorithm 2** presents the detailed IKMp algorithm.

**B. LDA BASED ON \( l_p \)-NORM**

Recall that the loss term in the objective function of the first stage of the LSLDA algorithm [38] is \( \| G_1^T X - L^T \|_{F}^2 \), which can be rewritten as

\[
\| X^T G_1 - L \|_{F}^2.
\]

Extending the square F-norm used in the Eq.(21) to the \( p \)-th power of \( l_p \)-norm, and adding the square F-norm penalty of \( G_1 \) to it, a more general objective function can be obtained as follows:

\[
\min J(G_1) = \| X^T G_1 - L \|_{l_p}^p + \gamma \| G_1 \|_{F}^2.
\]

The objective function in Eq.(22) is a general form of the one in Eq.(11). So replacing the optimization (11) by (22) in LSLDA algorithm [38], a general LDA algorithm in the sense of the \( l_{p,2} \)-norm can be obtained. It can be attempted to find a numerical solution of the optimization problem (22) since to find its analytical solution is very difficult. In fact, the optimization problem (22) can be rewritten as

\[
\min J(G_1) = \sum_{i=1}^{n} \left\| G_1^T x_i \right\|_{l_p}^p + \gamma \| G_1 \|_{F}^2,
\]

where \( L^i \) is the \( i \)-th line of \( L \). Let \( w_i = \left( \left\| G_1^T x_i \right\|_{l_p}^p + \gamma \right)^{-1} \). Then differentiating (23) with respect to \( G_1 \) and setting the derivative to zero, an equation that the optimal \( G_1 \) must satisfies can be obtained as

\[
\left( \sum_{i=1}^{n} w_i^2 x_i x_i^T + 2 \gamma I \right) G_1 = \sum_{i=1}^{n} w_i^2 x_i L^i.
\]

It follows that

\[
G_1 = \left( \sum_{i=1}^{n} w_i^2 x_i x_i^T + 2 \gamma I \right)^{-1} \left( \sum_{i=1}^{n} w_i^2 x_i L^i \right).
\]

Let \( \tilde{x}_i = w_i x_i, \tilde{L}^i = w_i L^i \), and let \( \tilde{X} = [\tilde{x}_1, \ldots, \tilde{x}_n], \tilde{L} = [	ilde{L}^1, \ldots, (\tilde{L}^p)^T]^T \). Then the \( G_1 \) can be rewritten as

\[
G_1 = (\tilde{X}^T \tilde{X} + \gamma I)^{-1} \tilde{X} \tilde{L}.
\]
From Eq.(27), it can be known that the problem (23) can be solved by using an iterative algorithm: Given \( w_i (i = 1, 2, \cdots, n) \), \( G_1 \) can be updated by computing Eq. (27) which can be computed using LSQR algorithm [40] efficiently; and given \( G_1, w_i (i = 1, 2, \cdots, n) \) can be updated by computing Eq. (24). This algorithm is in essence an Iterative Weighted Least Squares (IRLS) [39], and it can be initialized by using the transformation matrix obtained by classical linear regression (corresponding to the proposed algorithm with \( p = 2 \)).

Based on the LSLDA [38] and the above proposed IRLS algorithms, a \( l_{2,p} \)-norm based LDA (\( l_{2,p} \)-LDA) algorithm can be obtained as the following Algorithm 3.

**Algorithm 3 \( l_{2,p} \)-LDA Algorithm**

**Input:** Data matrix \( X \), the weighted membership matrix \( L \), regularization \( \gamma \);

1. Solve the following \( l_{2,p} \)-norm regression problem
   \[
   \min J (G_1) = \| X^T G_1 - L \|_{l_{2,p}}^p + \gamma \| G_1 \|_F^2
   \]
   by using IRLS.
2. Compute \( \tilde{X} = G_1^T X, D = \tilde{X} L \); Compute the compact SVD \( D = U \Sigma V^T \), \( G_2 = U \Sigma_r \).
3. Compute \( G = G_1 G_2 \) as the final transformation matrix.

**Output:** transformation matrix \( G \).

The time complexity of the \( l_{2,p} \)-LDA algorithm can be analyzed as follows: Line 1 takes \( t_1 (t_2 k (3n + 5m + 2mn)) \) time for using IRLS to solve \( l_{2,p} \)-norm regression problem, where \( t_1 \) is the number of the iterations of IRLS, and \( t_2 \) is the number of the iterations of LSQR algorithm. Line 2 takes \( O (k mn) \) time for computing \( \tilde{X}, O (k^2 n) \) time for computing the matrix \( D \) and \( O (k^3) \) for computing the matrix the matrix \( G_2 \). So the total computational complexity of Line 2 is \( O (mkn + k^2 n + k^3) \). Line 3 takes \( O (mk^2) \) for combining \( G_1 \) and \( G_2 \). Therefore, the total computational complexity of the \( l_{2,p} \)-LDA algorithm is

\[
O (t_1 (t_2 k (3n + 5m + 2mn)) + mnk + k^2 n + k^3 + mk^2)
\]

It can be simplified as

\[
O (t_1 (t_2 k (3n + 5m + 2mn)) + mnk)
\]

when \( k \ll m \) and \( k \ll n \).

**C. DSC BASED ON \( l_{2,p} \)-NORM**

In this section, the proposed \( l_{2,p} \)-LDA algorithm is introduced into DSC, and combined with KM_\( p \) into a single clustering framework. Thus, a generalized DSC algorithm, which is called \( l_{2,p} \)-norm based discriminant subspace clustering (\( l_{2,p} \)-DSC) algorithm, is presented. In \( l_{2,p} \)-DSC, \( l_{2,p} \)-LDA is used to reduce the dimension of data, and KM_\( p \) is used to cluster the dimension-reduced data. The dimension reduction and the clustering processes are mutually optimized until the optimal clustering and dimensional reduction of the data are finally obtained. The following Algorithm 4 presents the detailed \( l_{2,p} \)-DSC algorithm. In Algorithm 4, for distinction, The parameters \( p \) used in KM_\( p \) and \( l_{2,p} \)-LDA are noted as \( p_1 \) and \( p_2 \) respectively.

The time complexity of the \( l_{2,p} \)-DSC algorithm can be analyzed as follows: Line 1 takes \( O (nm (2k_0 t_0)) \) time for KM_\( p \) clustering on the initial data, where \( k_0 \) is the number of the abnormal clusters found finally, \( t_0 \) is the maximum number of iterations of the two-class KM_\( p \) in IKM_\( p \). Line 2 takes \( O (t_1 (t_2 k (3n + 5m + 2mn)) + mnk) \) time complexity for applying \( l_{2,p} \)-LDA algorithm to compute the transformation matrix \( G \). Line 3 takes \( O (mk) \) for the dimension reduction of data as using the transformation matrix \( G \). Line 4 takes \( O (nk^2 t_3) \) time for KM_\( p \) clustering on the dimension reduced data, where \( t_3 \) is the number of iterations of KM_\( p \) on the dimension reduced data. So the total time complexity of the \( l_{2,p} \)-DSC algorithm is

\[
O \left( nm (2k_0 t_0) + t_1 (t_2 k (3n + 5m + 2mn)) + mnk + nk^2 t_3 \right).
\]

**D. A SPECIAL CASE OF DSC: ROBUST DSC**

Parameter \( p_1 \) and \( p_2 \) in the \( l_{2,p} \)-DSC algorithm have potent effect on the performance of the \( l_{2,p} \)-LSDC algorithm. In this section, based on the theories of Robust Statistics [46], [47], the robustness to noise and outlier of the proposed \( l_{2,p} \)-DSC algorithm is studied.

In fact, let \( [x_1, x_2, \cdots, x_n] \) be an observed data set and \( \theta \) is an unknown parameter to be estimated. In robust statistical theory [46], [47], an M-estimate of \( \theta \) can be generated by minimizing the following form

\[
\sum_{i=1}^{n} \rho (x_i - \theta).
\]

Here \( \rho \) is an arbitrary function that can measure the loss of \( x_i \) and \( \theta \). In a location estimate that minimizes \( \sum_{i=1}^{n} \rho (x_i - \theta) \), the M-estimator is generated by solving the equation

\[
\sum_{i=1}^{n} \psi (x_i - \theta) = 0,
\]

where \( \psi (x_i - \theta) = \partial \rho (x_i - \theta) / \partial \theta \). If \( \rho (x - \theta) \) is taken as \( |x - \theta|^{p_1} = \left( |x - \theta|^2 \right)^{p_1/2} \), it follows that the estimate

\[
Recall that the loss term in the regression (23) is the sum of a less rapid increasing function of the residuals. Eq. (30) shows the estimate as a weighted mean of the observed data points. Since $W(x)$ is a non-increasing function of $|x|$ when $p_1 < 2$, outlying observations will receive smaller weights [47]. So when $p_1 < 2$, the estimate (30) is robust to noise and outlier.

The object function of KM$_p$ can be rewritten as

$$ J_{KM_p}(U, V) = \sum_{j=1}^{k} \sum_{l=1}^{m} \left[ \sum_{i=1}^{n} u_{ij} \left| x_{il} - v_{jl} \right|^{p_1} \right] $$

(31)

Given $U$, the part $\sum_{i=1}^{n} u_{ij} \left| x_{il} - v_{jl} \right|^{p_1}$ of the Eq. (31) can be seen as an M-estimate of the $j$-th clustering center using the $j$-th cluster as the observed data set on the $l$-th feature, and $\left| x_{il} - v_{jl} \right|^{p_1}$ as the loss function. Based on the above analysis, it can be known that this estimate is robust to noise when $p_1 < 2$. So KM$_p$ is a noise-robust clustering algorithm when $p_1 < 2$.

The use of robust statistical theory in regression leads to robust regression method [46], [47]. The classical linear regression can be robustified in a straightforward way. Instead of minimizing a sum of square, one can minimize a sum of less rapid increasing function of the residuals [46]. Recall that the loss term in the regression (23) is

$$ \sum_{i=1}^{n} \left[ G_i^* x_i - (L_i) \right]^T \left[ G_i^* x_i - (L_i) \right]^{p_2} $$

(32)

Compared with $\|x\|_2^{p_2}$, $\rho(x) = \|x\|_2^{p_2}$ is a less rapid increasing function of the residuals when $p_2 < 2$. So the regression (23), which uses Eq. (32) as loss function, is a more robust regression model than the classical linear regression when $p_2 < 2$. Since the regression (23) constitutes the main part of $l_2,p$-LDA (Algorithm 3), it follows that $l_2,p$-LDA is a robust version of classical LDA when $p_2 < 2$.

Both KM$_p$ (with $p_1 < 2$) and $l_2,p$-LDA (with $p_2 < 2$) are robust to noise and outlier, and the proposed $l_2,p$-DSC algorithm is composed of KM$_p$ and $l_2,p$-LDA. Therefore, with $p_1 < 2$ and $p_2 < 2$, the proposed $l_2,p$-DSC algorithm is more robust to noise and outlier than the DSC [17] and its variant algorithms [19], [23].

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the performance the posposed $l_2,p$-DSC algorithm is empirically studied in comparison with the IKM [41], DC [17], DKM [19] and DEC [23] algorithms.

A. EXPERIMENTAL SETUP

A total of twelve data sets are collected to verify the effectiveness of the proposed $l_2,p$-DSC algorithm in the experiments. They are eight UCI data sets [48] including Wine, Post-Operative Patient, Statlog, page_blocks, Vehicle and Letter (abc), two synthetic data constructed of Wine: Wine1 and Wine2 (here Wine1 and Wine2 are constructed by adding one and five outliers to Wine data set respectively), two gene expression data set including Brain [49] and Leukemia (http://featureselection.asu.edu/datasets.php), and two text data set including Cora_OS and WebKB_wisconsin (http://www.escience.cn/people/fmnie/index.html). See Table 2 for more details. All of them are normalized using the following method:

$$ x_{ji}' = \begin{cases} 0.5, & \text{if } x_{j1} = x_{j2} = \cdots = x_{jm}, \\ \frac{x_{ji} - \min\{x_{ji}\}}{\max\{x_{ji}\} - \min\{x_{ji}\}}, & \text{else}, \end{cases} $$

(33)

Clustering accuracy is used to measure the performance of the algorithms. It is defined as

$$ Acc = \frac{\sum_{j=1}^{k} n_j}{n}, $$

(34)

where $n_j$ represents the number of the data points clustered correctly in $j$-th class. The regularization parameters in DSC [17], DKM [19] and the $l_2,p$-DSC algorithm are all set to $10^{-6}$. The balance parameter in DEC [23] is set to 2. DSC [17], DKM [19] and DEC [23] algorithms are all initialized using IKM [41] algorithm. The maximum number of iterations of the $l_2,p$-DSC algorithm is set to 100, and the iteration threshold is set to $10^{-3}$. All experiments are performed on Matlab2014a under Inter(R) Core (TM) i5-6200 CPU @ 2.4G and 8GB RAM.

| Data set             | Size | Dimension | Class |
|----------------------|------|-----------|-------|
| Post-Operative Patient | 87   | 8         | 3     |
| Wine                 | 178  | 13        | 3     |
| Wine1                | 179  | 13        | 3     |
| Wine2                | 183  | 13        | 3     |
| Statlog              | 270  | 13        | 2     |
| Vehicle              | 946  | 18        | 4     |
| Letter               | 2291 | 16        | 3     |
| Brain                | 42   | 5597      | 5     |
| page_blocks          | 5473 | 10        | 5     |
| Leukemia             | 72   | 7129      | 5     |
| Cora_OS              | 1246 | 6737      | 4     |
| WebKB_wisconsin      | 1210 | 4189      | 7     |
B. CONVERGENCE STUDY

To analyze the overall convergence of the proposed $l_{2,p}$-DSC theoretically is difficult. Since the proposed $l_{2,p}$-DSC algorithm can be decomposed into two stages (the stage of dimension reduction and the stage of clustering on the dimension reduced data), in this section, the convergence of each stage of $l_{2,p}$-DSC algorithm is studied empirically.

Firstly, the convergence of the algorithm (Algorithm 1) of clustering stage is studied. Fig.1 depicts the evolution of objective function (OF) of KM$_P$ (Eq.(19)) as a function of iterations (N) on all twelve data sets, where parameters $p_1$ and $p_2$ are chosen as 1.5 and 2 respectively. From Fig.1, it can be observed that the values of objective function of KM$_P$ show downward trends with the iteration, and KM$_P$ converges for no more than dozens of iterations on each data set. These show the KM$_P$ algorithm (Algorithm 1) can minimize the objective function of KM$_P$ effectively and efficiently.

Then the convergence of the algorithm (Algorithm 3) of dimension reduction stage is studied. Fig.2 depicts the evolution of objective function (OF) of $l_{2,p}$-norm based regression (Eq.(23)) as a function of iterations (N) on all twelve data sets, where parameters $p_1$ and $p_2$ are chosen as 2 and
1.2 respectively. From Fig.2, it is observed that the values of objective function of $l_{2,p}$-norm based regression (Eq. (23)) show downward trends with the iteration, and $l_{2,p}$-norm based regression converges for no more than dozens of iterations for all data sets. These show IRLS can minimize the objective function of $l_{2,p}$-norm based regression effectively and efficiently.

C. STUDY ON THE EFFECTS OF PARAMETER $p_1$ AND $p_2$

$p_1$ and $p_2$ are important parameters involved in the proposed $l_{2,p}$-DSC algorithm. $p_1$ is used in the KM$_p$ algorithm for clustering initialization in $l_{2,p}$-DSC algorithm and clustering on the dimension-reduced data. $p_2$ is used in the $l_{2,p}$-norm regression for the dimension reduction process in $l_{2,p}$-DSC algorithm. In this section, the effects of parameter $p_1$ and $p_2$ on the performance of the proposed $l_{2,p}$-DSC algorithm are studied empirically.

Firstly, the effect of parameter $p_1$ on the performance of the $l_{2,p}$-DSC algorithm is studied. To this end, $p_2$ is set to 2 in the experiments. At this point, the $l_{2,p}$-norm based regression used in the $l_{2,p}$-DSC algorithm degenerates the commonly used F-norm based regression. Fig.3 shows the clustering accuracies (y-axis) of $l_{2,p}$-DSC on twelve data sets for different $p_1$ values (x-axis), where the clustering accuracies of IKM [41] and DSC [17] are recorded as baselines. From the Fig.3, it can be observed that the clustering accuracies of DSC are higher than the ones of IKM on eight data sets including Wine, Wine1, Wine2, Letter(abc), Brain and Leukemia. This show DSC is commonly more effective than IKM. It can also be observed that, with proper $p_1$ values, the clustering accuracies of $l_{2,p}$-DSC are higher than the ones of IKM and DSC on almost all data sets ($l_{2,p}$-DSC and DSC achieve the same accuracy on Wine). On Wine1, Wine2, Letter(abc), Brain, page_blocks, Leukemia and WebKB_wisconsin, the clustering accuracies of $l_{2,p}$-DSC are much higher than the ones of IKM and DSC. At the same time, comparing the clustering results of $l_{2,p}$-DSC and DSC on Wine1 and Wine2, it can be seen that when the value of the parameter $p_1$ is chosen from the interval [1-2], the algorithm $l_{2,p}$-DSC shows a relatively higher clustering precision. This shows that the $l_{2,p}$-DSC algorithm has better noise-robustness when the value of the parameter $p_1$ is between 1 and 2 (excluding 2). So it can be suggested that, given $p_2 = 2$, commonly, the $p_1$ in the $l_{2,p}$-DSC algorithm can be selected in the range [1-2.5]; and, if the data is noisy, the $p_1$ in the $l_{2,p}$-DSC algorithm can be selected in the smaller range [1-2].

Then, the effect of parameter $p_2$ on the performance of the $l_{2,p}$-DSC algorithm is studied. In order to study the the effect of parameter $p_2$ on the performance of the $l_{2,p}$-DSC algorithm, the $p_1$ in the $l_{2,p}$-DSC algorithm is fixed to 2 in the experiment. At this point, the KM$_p$ clustering degenerates into the KM clustering accordingly. Fig.4 shows the clustering accuracies (y-axis) of $l_{2,p}$-DSC on twelve data sets for different $p_2$ values (x-axis), and the clustering accuracies of IKM [36] and DSC [14] are recorded as baselines. From the Fig.4, it can be observed that, with proper $p_2$ values, the clustering accuracies of $l_{2,p}$-DSC are higher than the ones of IKM and DSC on all data sets except for Post-Operative Patient. compared with IKM and DSC, $l_{2,p}$-DSC shows much higher clustering accuracy On Letter(abc), page_blocks, Leukemia and WebKB_wisconsin. At the same time, it can be also observed that for too or large values of $p_2$, the $l_{2,p}$-DSC algorithm [17] may show poor clustering accuracy. So it can be suggested that, given $p_1 = 2$, the $p_2$ in the $l_{2,p}$-DSC algorithm can be selected in the range [1-2.5]. At the same
FIGURE 4. The clustering accuracies (y-axis) of $l_{2,p}$-DCS on twelve data sets for different $p_2$ values (x-axis), and the clustering accuracies of IKM and DSC are recorded as baselines.

Figure 4 shows the clustering accuracies (y-axis) of $l_{2,p}$-DCS on twelve data sets for different $p_2$ values (x-axis), and the clustering accuracies of IKM and DSC are recorded as baselines.

D. CLUSTERING ACCURACY EVALUATION

Table 3 presents the accuracies achieved by all five algorithms on all twelve data sets, where the accuracy achieved by each algorithm is recorded as the y-axis value for each data set.

| Data Sets          | IKM   | DSC   | DKM   | DEC   | $l_{2,p}$-DSC |
|--------------------|-------|-------|-------|-------|---------------|
| Post-Operative Patient | 0.5037 | 0.4805 | 0.5057 | **0.5517** | **0.5517** |
| Wine               | 0.9494 | 0.9775 | 0.9775 | 0.7023 | **0.9944**   |
| Wine1              | 0.6760 | 0.7151 | 0.6872 | 0.7039 | **0.9326**   |
| Wine2              | 0.6503 | 0.6557 | 0.6503 | 0.6885 | **0.9438**   |
| Statlog            | 0.5037 | 0.5185 | 0.5185 | 0.5037 | **0.5407**   |
| Vehicle            | 0.3700 | 0.4090 | 0.4078 | 0.4479 | **0.4551**   |
| Letter             | 0.8254 | 0.8316 | 0.8850 | **0.8961** | 0.8926       |
| Brain              | 0.5000 | 0.5238 | 0.5238 | 0.5               | **0.6429**   |
| Pageblocks         | 0.4000 | 0.3873 | 0.5388 | 0.4486 | **0.9012**   |
| Leukemia           | 0.6528 | 0.6805 | 0.7500 | 0.7361 | **0.7639**   |
| CoraOS             | 0.3941 | 0.3467 | 0.3820 | 0.3860 | **0.4406**   |
| WebKBWisconsin     | 0.5678 | 0.4835 | 0.5455 | 0.5744 | **0.7339**   |

Table 4 presents the running times of five algorithms on eight data sets.

| Data Sets          | IKM   | DSC   | DKM   | DEC   | $l_{2,p}$-LSDC |
|--------------------|-------|-------|-------|-------|--------------|
| Post-Operative Patient | 0.0494 | 0.0180 | 0.7252 | 1.3119 | 0.9280       |
| Wine               | 0.1678 | 0.0290 | 0.9456 | 1.8056 | 0.9286       |
| Wine1              | 0.0802 | 0.0424 | 0.8576 | 1.6614 | 2.0330       |
| Wine2              | 0.0893 | 0.0247 | 0.9615 | 1.7129 | 1.5597       |
| Statlog            | 0.0949 | 0.0245 | 0.9449 | 2.0792 | 0.6622       |
| Vehicle            | 0.4238 | 0.5194 | 4.0613 | 6.7821 | 22.5402      |
| Letter             | 1.7077 | 0.8999 | 14.8476 | 12.0322 | 5.4225       |
| Brain              | 0.3649 | 401.5352 | 1.5179 | 26.2133 | 34.7076     |
| pageblocks         | 1.9442 | 58.5612 | 71.6063 | 27.6055 | 197.1766    |
| Leukemia           | 0.5240 | 658.2372 | 0.6052 | 161.1307 | 31.6549     |
| CoraOS             | 7.9839 | 740.7129 | 12.1053 | 259.6040 | 170.2408    |
| WebKBWisconsin     | 5.9565 | 3321.351 | 10.3073 | 177.9539 | 54.736      |

Table 3 presents the accuracies achieved by all five algorithms on all twelve data sets, where the accuracy achieved by each algorithm is recorded as the y-axis value for each data set.

Table 4 presents the running times of five algorithms on eight data sets, where the running time achieved by each algorithm is recorded as the y-axis value for each data set.
by $l_{2,p}$-DSC is the the optimal accuracy by using $p_1$ from the range of $[1.5, 2.5]$ and $p_2$ from the range of $[1, 2.5]$. Table 4 presents the corresponding running time. From Table 3, it can be seen that, among five algorithms, $l_{2,p}$-DSC achieves the best on all data sets except for Letter in terms optimal accuracy. $l_{2,p}$-DSC yields improvements over DSC on all twelve data sets in terms of optimal accuracy. These appearances show the extension of DSC including clustering and dimension reduction processes from F-norm to $l_{2,p}$-norm is beneficial as a whole.

E. CLUSTERING EFFICIENCY EVALUATION

In this section, the efficiency of the $l_{2,p}$-DSC algorithm in terms of running time is studied. The running times of $l_{2,p}$-DSC are compared with the ones of IKM [36], DSC [17], DKM [19] and DEC [23] on twelve data sets. The running times of five algorithms on twelve data sets are presented in Table 4. From Table 4, it can be observed that, the running time of the proposed $l_{2,p}$-DSC algorithm generally increases with the increase of the data size. This phenomenon is more obvious on the following five data sets: Brain, page-blocks, Leukemia, Cora_OS and WebKB_wisconsin. In terms of running time, the efficiency of $l_{2,p}$-DSC is competitive with the DSC [17] and DEC [23] algorithms for high-dimensional data sets including Brain, Leukemia, Cora_OS and WebKB_wisconsin.

VI. CONCLUSION

In this paper, the DSC algorithm [17] is generalized in the sense of $l_{2,p}$-norm and $l_{2,p}$-DSC algorithm is proposed. This generalization provides more alternative models for practical applications, and improves the flexibility of the original DSC algorithm. Specially, theoretical analysis and experimental results show, compared with the DSC algorithm, with $p_1 < 2$ and $p_2 < 2$, $l_{2,p}$-DSC is more robust to the noise and outlier in data. In addition, the efficient LSQR algorithm based solution to $l_{2,p}$-LDA dimension reduction and iteration procedure based solution to $KMP$ clustering in $l_{2,p}$-DSC offer a possibility to use $l_{2,p}$-DSC on large data sets. The selection of parameter is commonly an open problem in an unsupervised learning algorithm due to the deficiency of class label. In semi-supervised case, devising $l_{2,p}$-DSC algorithm, especially considering the selection of parameter $P_1$ and $P_2$, is interesting, that will be carried out in our next work. In addition, multi-view learning has become a hotpot of machine learning since the features of data may come from different sources and multi-view learning can handle this kind of data effectively. DEC algorithm [23], which is a variant of DSC, has been extended to multi-view case, and the corresponding algorithm has show encouraging effect on the data with multiple source features [50]. So another further research will include the generalization of $l_{2,p}$-DSC in the sense of multi-view case.

REFERENCES

[1] E. W. Forgy, “Cluster analysis of multivariate data: Efficiency versus interpretability of classification,” Biometrics, vol. 21, no. 3, pp. 768–769, 1965.
[2] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proc. 5th Berkeley Symp. Math. Statist. Probab. Berkeley, CA, USA: Univ. California Press, 1967, pp. 281–297.
[3] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in Proc. 14th Adv. Neural Inf. Process. Syst. Cambridge, MA: MIT Press, 2002, pp. 849–856.
[4] J. Shi and J. Malik, “Normalized cuts and image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 888–905, Aug. 2000.
[5] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. New York, NY, USA: Wiley, 1990.
[6] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” Science, vol. 315, no. 5814, pp. 972–976, Feb. 2007.
[7] E. Elhamifar, G. Sapiro, and S. S. Sastry, “Dissimilarity-based sparse subset selection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 11, pp. 2182–2197, Nov. 2016.
[8] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, “Automated variable weighting in K-means type clustering,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 5, pp. 657–668, May 2005.
[9] H. Parvin and B. Minaei-Bidgoli, “A clustering ensemble framework based on elite selection of weighted clusters,” Adv. Data Anal. Classification, vol. 7, no. 2, pp. 181–208, Jun. 2013.
[10] H. Parvin and B. Minaei-Bidgoli, “A clustering ensemble framework based on selection of fuzzy weighted clusters in a locally adaptive clustering algorithm,” Pattern Anal. Appl., vol. 18, no. 1, pp. 87–112, Feb. 2015.
[11] I. T. Jolliffe, Principal Component Analysis, 2nd ed. New York, NY, USA: Springer, 2002.
[12] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduc tion and data representation,” in Proc. NIPS, 2003, pp. 1–7.
[13] A. Nasser, D. Hamad, and C. Nasr, “K-means clustering algorithm in projected spaces,” in Proc. 9th Int. Conf. Inf. Fusion, 2006, pp. 1–6.
[14] Q. Xu, C. Ding, J. Liu, and B. Luo, “PCA-guided search for K-means,” Pattern Recognit. Lett., vol. 54, pp. 50–55, Mar. 2015.
[15] K. Fukunaga, Introduction to Statistical Pattern Recognition. New York, NY, USA: Academic, 1990.
[16] F. De la Torre and T. Kanade, “Discriminative cluster analysis,” in Proc. 23rd Int. Conf. Mach. Learn. (ICML), 2006, pp. 241–248.
[17] C. Ding and T. Li, “Adaptive dimension reduction using discriminant analysis and K-means clustering,” in Proc. ICML, 2007, pp. 521–528.
[18] J. Ye, Z. Zhao, and H. Liu, “Adaptive distance metric learning for clustering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2007, pp. 1–7.
[19] J. Ye, Z. Zhao, and M. Wu, “Discriminative K-means for clustering,” in Proc. Int. Conf. Inf. Process. Syst. Red Hook, NY, USA: Curran Associates, 2007, pp. 1649–1656.
[20] C. Bouveyron and C. Brunet, “Simultaneous model-based clustering and visualization in the Fisher discriminative subspace,” Statist. Comput., vol. 22, no. 1, pp. 301–324, Jan. 2012.
[21] X.-B. Zhi, J.-L. Fan, and F. Zhao, “Fuzzy linear discriminant analysis-guided maximum entropy fuzzy clustering algorithm,” Pattern Recognit., vol. 46, no. 6, pp. 1604–1615, Jun. 2013.
[22] C.-L. Liu, W.-H. Hsiao, C.-H. Lee, and F.-S. Gou, “Semi-supervised linear discriminant clustering,” IEEE Trans. Cybern., vol. 44, no. 7, pp. 989–1000, Jul. 2014.
[23] C. Hou, F. Nie, D. Yi, and D. Tao, “Discriminative embedded clustering: A framework for grouping high-dimensional data,” IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 6, pp. 1287–1299, Jun. 2015.
[24] H. Li, T. Jiang, and K. Zhang, “Efficient and robust feature extraction by maximum margin criterion,” IEEE Trans. Neural Netw., vol. 17, no. 1, pp. 157–165, Jan. 2006.
[25] N. Flammarion, B. Palaniappan, and F. Bach, “Robust discriminative clustering with sparse regularizers,” J. Mach. Learn. Res., vol. 18, no. 1, pp. 2764–2813, 2017.
[26] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, “Maximum margin clustering,” in Proc. NIPS, 2005, pp. 1537–1544.
[27] T. De Bie and N. Cristianini, “Fast SDP relaxations of graph cut clustering, transduction, and other combinatorial problems,” J. Mach. Learn. Res., vol. 7, pp. 1409–1436, Jul. 2006.
[28] X. Zhi, H. Yan, J. Fan, and S. Zheng, “Efficient discriminative clustering via QR decomposition-based linear discriminant analysis,” Knowl.-Based Syst., vol. 153, pp. 117–132, Aug. 2018.
XIAOBIN ZHI received the B.Sc. degree in mathematics from Shaanxi Normal University, China, in 2000, the M.Sc. degree in applied mathematics from Shantou University, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, in 2013. He is currently an Associate Professor with the School of Science, Xi’an University of Posts and Telecommunications. His research interests include pattern recognition and machine learning.

LONGTAO BI received the B.S. degree in electronic and information engineering from Shangluo University, in 2017. He is currently pursuing the M.S. degree with the Xi’an University of Posts and Telecommunications. His research interests include pattern recognition and machine learning.

JIULUN FAN received the B.S. and M.S. degrees in fundamental mathematics from Shaanxi Normal University, in 1984 and 1988, respectively, and the Ph.D. degree in signal and information processing from Xidian University, in 1998. He is currently the President and a Professor with the Xi’an University of Posts and Telecommunications, Xi’an, Shaanxi, China. He is the author of five books and more than 200 articles. His research interests include fuzzy set theory, pattern recognition, and image processing.

* * *

References:
[29] A. Argyriou, T. Evgeniou, and M. Pontil, “Multi-task feature learning,” in Proc. NIPS, 2007, pp. 41–48.
[30] H. Tao, C. Hou, F. Nie, Y. Jiao, and D. Yi, “Effective discriminative feature selection with nontrivial solution,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 4, pp. 796–808, Apr. 2016.
[31] C. Ding, D. Zhou, X. He, and H. Zha, “R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization,” in Proc. 23rd Int. Conf. Mach. Learn. (ICML), Jun. 2006, pp. 281–288.
[32] F. Nie, H. Huang, X. Cai, and C. H. Ding, “Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization,” in Proc. Int. Conf. Neural Inf. Process. Syst., Vancouver, BC, Canada, Dec. 2010, pp. 1813–1821.
[33] L. Wang, S. Chen, and Y. Wang, “A unified algorithm for mixed $\ell_{p,q}$-minimizations and its application in feature selection,” Comput. Optim. Appl., vol. 58, no. 2, pp. 409–421, 2014.
[34] S. Liao, Q. Gao, Z. Yang, F. Chen, F. Nie, and J. Han, “Discriminant analysis via joint Euler transform and $\ell_{2,1}$-norm,” IEEE Trans. Image Process., vol. 27, no. 6, pp. 5668–5682, Nov. 2018.
[35] M. Zhao, H. Zhang, W. Cheng, and Z. Zhang, “Joint $\ell_p$ and $\ell_{2,q}$-norm minimization for subspace clustering with outlier pursuit,” in Proc. Int. Joint Conf. Neural Netw. (IJCNN), 2016, pp. 3658–3665.
[36] H. Zhao, Z. Wang, and F. Nie, “A new formulation of linear discriminant analysis for robust dimensionality reduction,” IEEE Trans. Knowl. Data Eng., vol. 31, no. 4, pp. 629–640, Apr. 2019.
[37] H. Du, G. Li, S. Wang, and F. Zhang, “Discriminant locality preserving projections based on $\ell_{2,q}$-norm for image feature extraction and recognition,” J. Vis. Commun. Image Represent., vol. 58, pp. 166–177, Jan. 2019.
[38] L. Sun, B. Cenar, and J. Ye, “A scalable two-stage approach for a class of dimensionality reduction techniques,” in Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2010, pp. 313–322.
[39] P. W. Holland and R. E. Welsch, “Robust regression using iteratively reweighted least-squares,” Commun. Statist., vol. 6, no. 9, pp. 813–827, 1977.
[40] C. C. Paige and M. A. Saunders, “LSQR: An algorithm for sparse linear equations and sparse least squares,” ACM Trans. Math. Softw., vol. 8, no. 1, pp. 43–71, Mar. 1982.
[41] M. M.-T. Chiang and B. Mirkin, “Intelligent choice of the number of clusters in K-means clustering: An experimental study with different cluster spreads,” J. Classification, vol. 27, no. 1, pp. 3–40, Mar. 2010.
[42] R. C. de Amorim and B. Mirkin, “Minkowski metric, feature weighting and anomalous cluster initializing in K-means clustering,” Pattern Recognit., vol. 45, no. 3, pp. 1061–1075, Mar. 2012.
[43] P. S. Bradley, O. L. Mangasarian, and W. N. Street, “Clustering via concave minimization,” in Advances in Neural Information Processing System. Cambridge, MA, USA: MIT Press, 1997, pp. 368–374.
[44] R. J. Hathaway, J. C. Bezdek, and Y. Hu, “Generalized fuzzy C-means clustering strategies using $L_p$ norm distances,” IEEE Trans. Fuzzy Syst., vol. 8, no. 5, pp. 576–582, Oct. 2000.
[45] J. H. Friedman, “Regularized discriminant analysis,” J. Amer. Statist. Assoc., vol. 84, no. 405, pp. 165–175, 1989.
[46] P. J. Huber, Robust Statistics. New York, NY, USA: Wiley, 1981.
[47] R. A. Maronna, D. R. Martin, and V. J. Yohai, Robust Statistics: Theory and Methods. New York, NY, USA: Wiley, 2006.
[48] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. (1998). UCI Repository of Machine Learning Databases. [Online]. Available: http://www.ics.uci.edu/mlrelearn/MLRepository.html
[49] S. Pomeroy et al., “Prediction of central nervous system embryonal tumour outcome based on gene expression,” Nature, vol. 415, pp. 436–442, Jan. 2002.
[50] J. Xu, J. Han, F. Nie, and X. Li, “Re-weighted discriminatively embedded K-means for multi-view clustering,” IEEE Trans. Image Process., vol. 26, no. 6, pp. 3016–3027, Jun. 2017.
[51] F. Nie, S. Yang, R. Zhang, and X. Li, “A general framework for auto-weighted feature selection via global redundancy minimization,” IEEE Trans. Image Process., vol. 28, no. 5, pp. 2428–2438, May 2019.