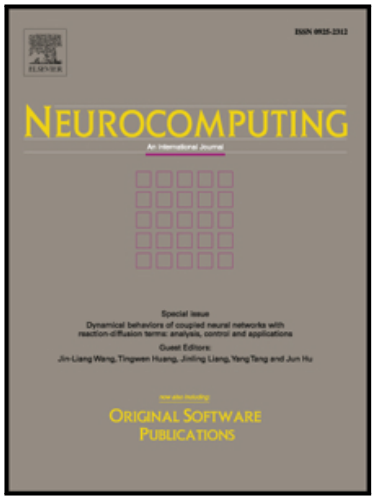
Unsupervised feature selection with adaptive residual preserving

Accepted Manuscript



Unsupervised Feature Selection with Adaptive Residual Preserving

Luyao Teng, Zhenye Feng, Xiaozhao Fang, Shaohua Teng,
Hua Wang, Peipei Kang, Yanchun Zhang

PII: S0925-2312(19)30898-7
DOI: https://doi.org/10.1016/j.neucom.2019.05.097
Reference: NEUCOM 20954

To appear in: *Neurocomputing*

Received date: 26 August 2018
Revised date: 25 November 2018
Accepted date: 10 May 2019

Please cite this article as: Luyao Teng, Zhenye Feng, Xiaozhao Fang, Shaohua Teng, Hua Wang, Peipei Kang, Yanchun Zhang, Unsupervised Feature Selection with Adaptive Residual Preserving, *Neurocomputing* (2019), doi: https://doi.org/10.1016/j.neucom.2019.05.097

Unsupervised Feature Selection with Adaptive Residual Preserving

Luyao Teng[a], Zhenye Feng[b], Xiaozhao Fang[b,*], Shaohua Teng[b], Hua Wang[a], Peipei Kang[b], Yanchun Zhang[a]

[a]*Institute for Sustainable Industries & Liveable Cities, VU Research, Victoria University, Ballarat Rd, Footscray VIC 3011, Australia*

[b]*School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, 510006, China*

Abstract

Many feature selection approaches are proposed in recent years. Most approaches utilize graph-based methods in studying the structure and relationship among data. However, many data relationships may loss during the graph construction, such as the residual relationships. To better preserve the relationships between data, in this paper, we propose a novel unified learning framework - unsupervised feature selection with adaptive residual preserving (UFSARP). The framework unifies feature selection, data reconstruction, and local residual preserving into one unified process, in which these tasks are completed simultaneously. We use the distance of projected data to learn the similarity matrix and simultaneously impose it on the data representation term to enforce that similar samples have similar reconstruction residuals. The use of such learning way has three-fold advantages: 1) The reconstruction residuals aim to maintain the residual relationships between data samples, namely, similar samples have similar residuals, and this helps to reconstruct the original data better; 2) Imposing the similarity matrix on the data representation term encourages similar samples not only have similar reconstruction residuals but also have similar reconstruction coefficients; 3) The similarity matrix and the reconstruction coefficient can be promoted by each other during the learning process. The experimental results show that the proposed algorithm is superior to other similar researches.

Keywords: Residual preserving, Unsupervised learning, Feature selection, Unified learning framework, and Sparse representation

1. Introduction

Nowadays, many real-world applications often confront data with high dimensions, such as image retrieval[52], pattern recognition[3], computer vision[51], and gene expression microarrays analysis[1]. As high dimensional data involves a large number of features, it brings considerable challenges in training time and storage resources. Moreover, the adverse impact of noisy and irrelevant features in all data may severely degrade the generalized performance. To address the difficulties, dimensionality reduction methods are proposed to learn a subset of informative features.

*Corresponding author

Email address: xzhfang168@126.com (Xiaozhao Fang)

These learnt features are representative and retain the salient characteristics of data. Therefore, using dimensionality reduction approaches helps to reduce data processing time and improves the performance.

There are two types of the dimensionality reduction: feature selection [5][13][18][19][42][63] and feature extraction [15][65][29]. Feature selection does not change the original representation of data and aims to obtain a subset of features to represent the original data. While, feature extraction forms new representations and commonly produces new features. In this paper, we focus on feature selection [41][40]. Depending on the availability of label information, feature selection is generally classified into supervised[22], semi-supervised[28] [48], and unsupervised[20][21][39][56][61]. Most available data in the real applications are unlabeled. Due to the lack of label knowledge, unsupervised learning is more challenging than the supervised and the semi-supervised learning. Thus it is necessary to develop an effective unsupervised feature selection method.

The studies on unsupervised feature selection can be mainly categorized into three groups: filters [62], wrappers, and embedding[16].

Filter methods select the optimal features based on a trace ratio score[62][33]. Typical filter methods are Laplacian score[18], spectral based feature selection[64], filter-based multivariate method[53], etc. However, filter methods do not fit a specific algorithm, thus they can only provide a generic selection of features and fail to select the most informative features for a particular learning task[16][27].

The approaches based on wrapper model require a predetermined learning algorithm and 'wrap' the feature selection to evaluate relevant features[27][49]. In past decades, Ma et al.[36], Guyon et al.[17], Maldonado et al.[37], and Dy et al.[12], etc. have dedicated a lot of researches on this topic, and experimental results show that algorithms based on wrapper model perform better than filter models. However, the wrapper models are usually computationally expensive, so they are not able to be applied in large-scale data sets. In addition, wrapper model is prone to the issue of overfitting.

Embedding-based methods unify the feature selection and model construction into one framework [56][66]. Zhu et al. [66][67] proposed an unsupervised spectral feature selection model by embedding a graph regularizer into one framework of joint sparse regression for preserving the local structures of data. Weston et al.[57] added the l_0-norm regularizer into an objective function to achieve sparse solution for performing feature selection and classifying object. Liu et al.[32] employed the $l_{2,1}$-norm regularizer to achieve the similar objective. Wang et al.[55] proposed an unsupervised feature selection algorithm that is built with a similarity matrix and a non-squared l_2-norm based sparsity. All these graph embedding methods process the classification in two separate steps: construct the data structure; select the target features. Thus the feature selection result is highly dependent on the built structure. Once the construction fails to represent the intrinsic data structure, the feature selection will also fail to represent the data characteristics. In recent years, Kang et al.[23][24][25][26] proposed multiple kernel-based learning methods that can not only learn both linear and nonlinear similarity information but also simultaneously learn cluster indicator matrix and similarity information in kernel spaces. In addition, there are many dimensionality reduction methods using the reconstruction coefficient matrix as a local similarity matrix to represent the data relationships[47][35][14]. In 2010, Qiao et al. [47] used the l_1-norm to minimize

the objective function. In the article, l_1-norm helps to maintain the local manifold structure and sparsity during the process of dimensionality reduction. Then, in 2016, Lu et al.[35] introduced a method that retained the low-rank information and global structure in dimensionality reduction process. In 2017, Fang et al.[14] proposed an algorithm that simultaneously obtained the feature representation and intrinsic similarity structure of data. Du et al.[11] introduced a unified learning framework which performed structure learning and feature selection simultaneously. In 2018, Li et al.[30] proposed a generalized uncorrelated regression with adaptive graph for unsupervised feature selection that can perform feature selection and spectral clustering simultaneously. As we all know, similar samples have similar properties. All these above methods only consider the local manifold structure but ignore the residual relationship among data.

In this paper, we proposed a novel unsupervised feature selection method, i.e., the unsupervised feature selection with adaptive residual preserving. In this framework, we not only unify the subspace learning and feature selection into one process, but also retain the local residuals in the data reconstruction process. Different from simply unifying the graph Laplacian and sparse representation to capture the data structure, the proposed method introduces the local residual, which assumes that similar samples have similar reconstructed residuals before and after transformation. Compared with traditional and aforementioned methods, our method could better capture the intrinsic data structure. With the refined reconstruction structure, a better feature selection result could be expected.

We emphasize three contributions of this paper as follows:

1) A novel unified learning framework - unsupervised feature selection with adaptive residual preserving is proposed in this paper. The framework unifies feature selection, data reconstruction, and local residual preserving into one process in which these tasks are completed simultaneously.
2) UFSARP holds the local manifold structure and intrinsic data relationship among data, so that similar samples not only have similar reconstruction coefficients and but also have similar reconstruction residuals.
3) The reconstruction residuals which preserve the residual relationships between data samples, help to reconstruct the original data better.

The rest of this paper is arranged as follows. We introduce the recent related works in Section 2. In Section 3, we propose our framework and the optimization. Section 4 discusses the computational complexity and convergence. In Section 5, comprehensive experiments are conducted to compare related methods. Last of all, we conclude our paper in Section 6.

2. Related work

This section overviews the literatures on the sparse representation, graph embedding, and unsupervised feature selection with adaptive structure learning (FSASL) proposed by Du et al.[11].

2.1. Notation

In this paper, we use bold uppercase letters to denote matrices, and bold lowercase letters to denote vectors. For an arbitrary matrix $\boldsymbol{F} \in R^{d\times k}$, $\boldsymbol{f}_i$ means the i-th row vector of $\boldsymbol{F}$ and $\boldsymbol{f}_j^T$ is

the j-th column vector of matrix $\boldsymbol{F}$. $\boldsymbol{F}_{ij}$ denotes the value of the i-th row and j-th column of $\boldsymbol{F}$. The $l_{2,1}$-norm is defined as $\|\boldsymbol{F}\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{k} \boldsymbol{F}_{ij}^2}$.

In our work, $\boldsymbol{X} \in R^{n \times d}$ denotes the sample matrix, where n is the number of samples, and d is the number of features.

2.2. Sparse Representation

The research on sparse representation has a long history, and recent literature proves that sparse representation is utilized not only in signal processing, but also in image processing and pattern recognition[9][44][45].

Sparse representation uses fewest samples to represent the randomly selected sample. Given a data set $\boldsymbol{X} = [x_1, x_2, ..., x_n] \in R^{d \times n}$, where $x_i \in R^d$ is a random sample in the data set. Then the objective function can be described as:

$$\min_{\boldsymbol{s}} ||\boldsymbol{s}||_0 \qquad s.t. \quad \boldsymbol{x_i} = \boldsymbol{X}\boldsymbol{s} \tag{1}$$

where $\boldsymbol{s} = [s_1, s_2, ..., s_n]^T \in R^n$ is the sparse coefficient vector, $||.||_0$ refers to the number of nonzero elements in the vector and is also viewed as the measure of sparsity. Although the sparse representation method with l_0-norm minimization can obtain the fundamental sparse solution of s^* over the matrix X, function (1) is still a non-deterministic polynomial-time hard (NP-hard) problem and the solution is difficult to approximate. Literatures[58][10][6][7] demonstrate that if the solution s^* is sparse enough, the solution of the above l_0 minimization problem is approximately equal to the solution of the following l_1 minimization problem:

$$\min_{\boldsymbol{s}} ||\boldsymbol{s}||_1 \qquad s.t. \quad \boldsymbol{x_i} = \boldsymbol{X}\boldsymbol{s} \tag{2}$$

Moreover, the l_1 minimization problem has an analytical solution and can be solved in polynomial time.

2.3. Graph Embedding

Graph embedding has been proved to be successful in preserving the manifold structure between data[60]. Many algorithms, including the isometric feature mapping (ISOMAP)[54], linear discriminant analysis (LDA)[8][38], locally linear embedding (LLE)[50], and laplacian eigenmaps (LE)[2], etc. can be classified into a graph embedding general framework[60].

These methods use k-nearest neighbor or ε-ball to determine the neighborhood relationship between samples. Then, the weight between neighbors is defined by different methods, including heat kernel, inverse Euclidean distance, and local linear reconstruction coefficient. These methods construct graphs and select features independently. Besides, the graph structure is derived from original data and remains invariant during the subsequent process, but real world data always contain lots of noisy samples and features, which make the graph structure unreliable.

2.4. FSASL

In 2014, Du et al. proposed a unified unsupervised feature selection framework FSASL[11]. The framework processes data reconstruction and feature selection simultaneously.

The objective of framework FSASL is as follows:

$$\begin{aligned} \min_{\boldsymbol{W},\boldsymbol{S},\boldsymbol{P}} \quad & (||\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X} - \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{S}||_{\boldsymbol{F}}^{\mathbf{2}} + \boldsymbol{\alpha}||\boldsymbol{S}||_{\mathbf{1}}) + \beta \sum_{i,j}^{n} (||\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{j}}||^2 \boldsymbol{P}_{\boldsymbol{ij}} + \mu \boldsymbol{P}_{\boldsymbol{ij}}^{\mathbf{2}}) \\ & + \gamma ||\boldsymbol{W}||_{2,1} \\ s.t. \quad & \boldsymbol{S}_{ii} = 0, \boldsymbol{P}\mathbf{1}_{\boldsymbol{n}} = \mathbf{1}_{\boldsymbol{n}}, \boldsymbol{P} \geq \mathbf{0}, \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W} = \boldsymbol{I} \end{aligned} \tag{3}$$

where $\mathbf{1}_n \in R^{n\times 1}$ is a vector with all elements valued 1.

In the framework, the first term of (3) is a sparse representation [58][59] that helps to determine the global structure of data. The coefficient regularization term $\boldsymbol{\alpha}||\boldsymbol{S}||_{\mathbf{1}}$ aims to balance the sparsity and the reconstruction error. In the second term of the framework, variable $\boldsymbol{P}$ helps to determine the local manifold structure of data. $||\boldsymbol{W}||_{2,1}$ encourages the rows of $\boldsymbol{W}$ to be zero. With the sparsity of W, the irrelevant and noisy features can be largely removed. $\beta(>0)$ and $\gamma(>0)$ are regularization parameters.

To draw the intrinsic structure of data, Du et al.[11] integrated the global with the local manifold structure into one framework, so that both structures were counted. In the framework, the structure determination and feature selection are conducted simultaneously. Thus it can find the intrinsic data structure and select the most relevant features simultaneously. In addition, since in conventional methods, the similarity matrix is predetermined and separated from feature selection, the process and result of the feature selection rely on this prebuilt similarity matrix. In FSASL, the problem is avoided because the local manifold structure and feature selection are conducted simultaneously. However, FSASL ignores the residual relationship among data, which can neither guarantee similar samples have similar reconstruction residuals, nor guarantee similar samples have similar reconstruction coefficient.

3. Algorithm

An informative graph is critical for graph based learning models. Different from classic graph constructions, the graph of UFSARP is constructed by using sparse representation and local residual preserving. In this section, we briefly introduce residual preserving.

3.1. Similarity Matrix Learning

The similarity matrix can preserve the local manifold structure of data. Instead of using the graph Laplacian with the determined neighborhood relationship, a local similarity matrix P is used to solve the following problem according to the FSASL:

$$\min_{\boldsymbol{P}} \quad \sum ||\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{j}}||^2 \boldsymbol{P}_{\boldsymbol{ij}} + \mu \boldsymbol{P}_{\boldsymbol{ij}}^{\mathbf{2}} \qquad s.t. \quad \boldsymbol{P}\mathbf{1}_n = \mathbf{1}_n, \boldsymbol{P} \geq \mathbf{0} \tag{4}$$

This is a typical graph construction function where variable $\boldsymbol{P}$ reflects the local manifold structure. It can be found that a large distance of $||\boldsymbol{x}_i - \boldsymbol{x}_j||^2$ will lead to a small probability $\boldsymbol{P}_{\boldsymbol{ij}}$. μ is a regularization parameter. The regularization term is used to avoid the trivial solution and can be seen as a prior of uniform distribution.

3.2. Residual Preserving Learning

As introduced in the previous section, sparse representation can capture the global representation structure of data. Besides, sparse representation is used to eliminate noise and reduce the impact of outliers. So we choose sparse representation to reconstruct data. In sparse representation, the i-th sample is represented by a linear combination of other samples, which can be formulated as the following function:

$$\min_{\boldsymbol{S}} \quad \sum_{i=1}^{n} ||\boldsymbol{x}_i - \boldsymbol{X}\boldsymbol{s}_i||^2 + \alpha||\boldsymbol{S}||_1 \qquad s.t. \quad \boldsymbol{S}_{ii} = 0 \tag{5}$$

where $X = [x_1, x_2, ..., x_n]$ refers to the data set, and s_i is the reconstruction coefficients. By employing the l_1 norm constraint, s_i is sparse. Sample x_i can be sparsely represented by other samples. However, the sparse representation only considers the global structure but ignores the local structure and residual relationship between samples. Therefore, we introduce the graph $\boldsymbol{P}$ and modify the above function, so that similar samples not only have similar reconstruction coefficient, but also have similar reconstruction residuals. Thus the above idea can be formulated as follows:

$$\min_{\boldsymbol{P},\boldsymbol{S}} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} ||\boldsymbol{x}_i - \boldsymbol{X}\boldsymbol{s}_j||^2\boldsymbol{P}_{\boldsymbol{ij}} + \alpha||\boldsymbol{S}||_1 \qquad s.t. \quad \boldsymbol{S}_{ii} = 0 \tag{6}$$

where α is a parameter of $\boldsymbol{S}$, the sparsity variable. $\boldsymbol{X}s_j$ is the reconstructed sample of original sample x_i.

Lemma 1: For samples x_s and x_t, if x_s and x_t are similar at the measure of Euclidean distance, then $||x_s - x_t||^2$ is small.

Proof: The distance measure between two samples can be defined as dissimilarity between them. The smaller the distance measure is, the more similar they are. When samples x_s and x_t are similar at the measure of Euclidean distance, $||x_s - x_t||^2$ is small.

Lemma 2: For samples x_s and x_t, if x_s is similar to x_t at the measure of Euclidean distance, and $\sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{s}_i$ and $\sum_{j=1}^{n} \boldsymbol{X}\boldsymbol{s}_j$ approximate to x_s and x_t, respectively, then $\sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{s}_i$ and $\sum_{j=1}^{n} \boldsymbol{X}\boldsymbol{s}_j$ are also similar at the measure of Euclidean distance.

Proof: According to the trigonometric inequality, we have

$$\begin{aligned} ||\sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{s}_i - \sum_{j=1}^{n} \boldsymbol{X}\boldsymbol{s}_j||^2 &= ||\sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{s}_i - \sum_{j=1}^{n} \boldsymbol{X}\boldsymbol{s}_j + \boldsymbol{x}_s - \boldsymbol{x}_s + \boldsymbol{x}_t - \boldsymbol{x}_t||^2 \\ &\leq ||\boldsymbol{x}_s - \sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{s}_i||^2 + ||\boldsymbol{x}_t - \sum_{j=1}^{n} \boldsymbol{X}\boldsymbol{s}_j||^2 + ||\boldsymbol{x}_s - \boldsymbol{x}_t||^2 \end{aligned} \tag{7}$$

Since $\sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{s}_i$ and $\sum_{j=1}^{n} \boldsymbol{X}\boldsymbol{s}_j$ approximate to x_s and x_t, respectively, $||x_s - \sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{s}_i||^2$ and $||x_t - \sum_{j=1}^{n} \boldsymbol{X}\boldsymbol{s}_j||^2$ are small. According to *Lemma 1*, $||x_s - x_t||^2$ is small too. Therefore $\sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{s}_i$ and $\sum_{j=1}^{n} \boldsymbol{X}\boldsymbol{s}_j$ are similar at the measure of Euclidean distance.

Theorem 1: For samples x_s and x_t, if x_s is similar to x_t at the measure of Euclidean distance,

and $\sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{s}_i$ and $\sum_{j=1}^{n} \boldsymbol{X}\boldsymbol{s}_j$ approximate to x_s and x_t, respectively, then reconstruction coefficient matrix S not only makes similar samples be still similar, but also preserve residuals between the similar samples.

Proof: According to *Lemma 1* and *Lemma 2*, $\sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{s}_i$ and $\sum_{j=1}^{n} \boldsymbol{X}\boldsymbol{s}_j$ are similar at the measure of Euclidean distance. Therefore, $||\sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{s}_i - \sum_{j=1}^{n} \boldsymbol{X}\boldsymbol{s}_j||^2$ is small. So reconstruction coefficient matrix S not only makes similar samples be still similar, but also preserve residuals between the similar samples.

Furthermore, we hope that dimensionality reduction can maintain above characteristics which aim to select useful and discriminative features. Thus, we propose the following formulation:

$$\min_{\boldsymbol{S},\boldsymbol{P},\boldsymbol{W}} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} ||\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{X}\boldsymbol{s}_j||^2\boldsymbol{P}_{ij} + \alpha||\boldsymbol{S}||_1 \qquad s.t. \quad \boldsymbol{S}_{ii} = 0 \tag{8}$$

Accordingly, a data reconstruction function with preserving the residual relationship according to local manifold structure is proposed as follows:

$$\begin{aligned} \min_{\boldsymbol{S},\boldsymbol{P},\boldsymbol{W}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} ||\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{X}\boldsymbol{s}_j||^2\boldsymbol{P}_{ij} + \alpha||\boldsymbol{S}||_1 + \beta(\sum_{i=1}^{n}\sum_{j=1}^{n} ||\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j||^2\boldsymbol{P_{ij}} + \mu\boldsymbol{P}_{ij}^2) \\ s.t. \quad & \boldsymbol{P}\boldsymbol{1}_n = \boldsymbol{1}_n, \boldsymbol{P} \geq \boldsymbol{0}, \boldsymbol{S}_{ii} = 0 \end{aligned} \tag{9}$$

where β is to regularize the local similarity matrix $\boldsymbol{P}$.

In Eq. (9), $\boldsymbol{P}$, preserves the local manifold structure of data. In the function, $||\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j||^2$ represents the reconstruction residuals relationships between samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. As we have taken the local residual relationships into account, the local residual relationship will be well preserved, which encourages the similar samples have similar reconstruction residuals during data reconstruction process.

3.3. UFSARP

Finally, in order to maintain the characteristics of residuals and achieve the purpose of feature selection, the $l_{2,1}$ norm constraint is added to $\boldsymbol{W}$. Due to the good row-sparsity property of $l_{2,1}$ norm, we impose it on the selection matrix as follows:

$$\begin{aligned} \min_{\boldsymbol{S},\boldsymbol{P},\boldsymbol{W}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} ||\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{X}\boldsymbol{s}_j||^2\boldsymbol{P}_{ij} + \alpha||\boldsymbol{S}||_1 + \beta(\sum_{i=1}^{n}\sum_{j=1}^{n} ||\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j||^2\boldsymbol{P_{ij}} + \mu\boldsymbol{P}_{ij}^2) \\ & + \gamma||\boldsymbol{W}||_{2,1} \\ s.t. \quad & \boldsymbol{P}\boldsymbol{1}_n = \boldsymbol{1}_n, \boldsymbol{P} \geq \boldsymbol{0}, \boldsymbol{S}_{ii} = 0, \boldsymbol{W}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W} = \boldsymbol{I} \end{aligned} \tag{10}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\mu}$ in the framework are the regularization parameters that are used to balance the the importance of the corresponding terms. $||\boldsymbol{W}||_{2,1}$ encourages $\boldsymbol{W}$ to be row sparse, which can be used to select features.

In this framework, we integrate the sparse representation with local residuals. Term $||\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j||^2\boldsymbol{P}_{ij}$ learns the local manifold structure. Given two similar samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, if the value of $||\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j||^2$ is small, $\boldsymbol{P}_{ij}$ will be relatively large. Then $\boldsymbol{P}$ can be adaptively learnt. Besides, the local similarity matrix $\boldsymbol{P}$ encourages similar samples to have similar reconstruction residuals after transformation, so the reconstruction residual can be preserved. In addition, since in conventional methods, the similarity matrix is predetermined and separated from feature selection, the process and result of the feature selection rely on this prebuilt similarity matrix. However, in UFSARP, the problem is avoided because the local manifold structure and feature selection are conducted simultaneously. And UFSARP considers the residual relationship between samples during the process of data reconstruction. Therefore, UFSARP can better preserve the relationships between data, which helps in achieving the efficiency and effectiveness of the framework.

3.4. Optimization

Since $\boldsymbol{W}$, $\boldsymbol{P}$ and $\boldsymbol{S}$ are unknown variables, it is challenging to find the optimial solution of our objective function directly, particularly in calculating the derivative of the function in term of $\boldsymbol{S}$. Therefore, to solve the problem, we attempt to minimize the following augmented Lagrange multiplier (ALM) formula Γ:

$$
\begin{aligned}
&\Gamma(\boldsymbol{W},\boldsymbol{S},\boldsymbol{P},\boldsymbol{Q},\boldsymbol{T},\boldsymbol{Y}_1,\boldsymbol{Y}_2,\boldsymbol{Y}_3,\mu_1)\\
&\quad=\sum_{i=1}^{n}\sum_{j=1}^{n}||\boldsymbol{W}^T\boldsymbol{x}_i-\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{s}_j||^2\boldsymbol{P}_{ij}+\alpha||\boldsymbol{T}||_1+\beta(\sum_{i=1}^{n}\sum_{j=1}^{n}||\boldsymbol{W}^T\boldsymbol{x}_i-\boldsymbol{W}^T\boldsymbol{x}_j||^2\boldsymbol{P}_{ij}+\mu\boldsymbol{P}_{ij}^2)\\
&\quad+\gamma||\boldsymbol{W}||_{2,1}+\langle\boldsymbol{Y}_1,\boldsymbol{P}\boldsymbol{1}_n-\boldsymbol{1}_n\rangle+\langle\boldsymbol{Y}_2,\boldsymbol{P}-\boldsymbol{Q}\rangle+\langle\boldsymbol{Y}_3,\boldsymbol{S}-\boldsymbol{T}\rangle\\
&\quad+\frac{\mu_1}{2}(||\boldsymbol{P}\boldsymbol{1}_n-\boldsymbol{1}_n||_F^2+||\boldsymbol{P}-\boldsymbol{Q}||_F^2)+\frac{\mu_1}{2}||\boldsymbol{S}-\boldsymbol{T}||_F^2\\
&s.t.\quad \boldsymbol{W}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W}=\boldsymbol{I},\boldsymbol{S}_{\boldsymbol{ii}}=\boldsymbol{0}
\end{aligned}
\tag{11}
$$

where $\langle\boldsymbol{A},\boldsymbol{B}\rangle = Tr(\boldsymbol{A}^T\boldsymbol{B})$. $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$ and $\boldsymbol{Y}_3$ are Lagrange multipliers, $\mu_1 > 0$ is a penalty parameter. Then, we employ the Alternating Direction Method of multipliers (ADMM) algorithm to problem (11) [4].

ADMM updates each of these variables with all other variables fixed in each iteration by minimizing Γ. The main steps of solving the term (11) are shown as follows:

3.4.1. Updating variable $\boldsymbol{W}$

Firstly, we fix the value of other variables in order to update the $\boldsymbol{W}$. The optimization problem can be rewritten as:

$$\begin{aligned}\Gamma(\boldsymbol{W}) = &\min_{\boldsymbol{W}} \sum_{i=1}^{n}\sum_{j=1}^{n} ||\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{X}\boldsymbol{s}_j||^2\boldsymbol{P}_{ij} + \beta(\sum_{i=1}^{n}\sum_{j=1}^{n} |||\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j||^2\boldsymbol{P}_{ij} + \mu\boldsymbol{P}_{ij}) \\ &+ \gamma||\boldsymbol{W}||_{2,1} \\ &s.t. \quad \boldsymbol{W}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W} = \boldsymbol{I}\end{aligned} \tag{12}$$

Denote $\boldsymbol{L_{SP}} = (\boldsymbol{D_P} + \boldsymbol{S}\boldsymbol{D_{P^T}}\boldsymbol{S}^T - 2\boldsymbol{P}\boldsymbol{S}^T + \beta(\boldsymbol{D_P} + \boldsymbol{D_{P^T}} - 2\boldsymbol{P}))$, where $\boldsymbol{D_P}$ and $\boldsymbol{D_{P^T}}$ are two diagonal matrices, $\boldsymbol{D_{Pii}} = \sum_j P_{ij}$ and $\boldsymbol{D_{P^T}} = \sum_i P_{ij}$. Then we can transfer $\Gamma(\boldsymbol{W})$ into:

$$\begin{aligned}\Gamma(\boldsymbol{W}) = &\min_{\boldsymbol{W}} Tr(\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{L_{SP}}\boldsymbol{X}^T\boldsymbol{W}) + \gamma||\boldsymbol{W}||_{2,1} \\ &s.t. \quad \boldsymbol{W}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W} = \boldsymbol{I}\end{aligned} \tag{13}$$

Eq. (13) is equivalent to:

$$\begin{aligned}\Gamma(\boldsymbol{W}) = &\min_{\boldsymbol{W}} Tr(\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{L_{SP}}\boldsymbol{X}^T\boldsymbol{W}) + \gamma(\boldsymbol{W}^T\boldsymbol{D_W}\boldsymbol{W}) \\ &s.t. \quad \boldsymbol{W}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W} = \boldsymbol{I}\end{aligned} \tag{14}$$

where $\boldsymbol{D_W} \in R^{d\times d}$ is a diagonal matrix whose $\boldsymbol{i}$-th diagonal element is:

$$\boldsymbol{D_{W_{i,i}}} = \frac{1}{2||\boldsymbol{W}_i||_2} \tag{15}$$

Consequently, we can obtain the solution of problem (12) by solving the following problem:

$$\min_{\boldsymbol{W}} Tr(\boldsymbol{W}^T(\boldsymbol{X}\boldsymbol{L_{SP}}\boldsymbol{X^T} + \gamma\boldsymbol{D_W})^T\boldsymbol{W}) \qquad s.t. \quad \boldsymbol{W}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W} = \boldsymbol{I} \tag{16}$$

and it is clear that the optimal solutions of $\boldsymbol{W}$ are the eigenvectors corresponding to the c smallest eigenvalues of the eigenvalue decomposition problem:

$$(\boldsymbol{X}\boldsymbol{L_{SP}}\boldsymbol{X^T} + \gamma\boldsymbol{D_W})^T\boldsymbol{W} = \Lambda\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W} \tag{17}$$

3.4.2. Updating variable $\boldsymbol{P}$

Secondly, in order to update variable $\boldsymbol{P}$, we fix other variables, and the optimization objective function can be rewritten as the following form.

$$\begin{aligned}\Gamma(\boldsymbol{P}) = &\min_{\boldsymbol{P}} \sum_{j=1}^{n} ||\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{X}\boldsymbol{s}_j||^2\boldsymbol{P}_{ij} + \beta(\sum_{i=1}^{n}\sum_{j=1}^{n} ||\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j||^2\boldsymbol{P}_{ij} + \mu\boldsymbol{P}_{ij}^2) \\ &+ \frac{\mu_1}{2}||\boldsymbol{P}\boldsymbol{1}_n - \boldsymbol{1}_n + \frac{\boldsymbol{Y}_1}{\mu_1}||_F^2 + \frac{\mu_1}{2}||\boldsymbol{P} - \boldsymbol{Q} + \frac{\boldsymbol{Y}_2}{\mu_1}||_F^2\end{aligned} \tag{18}$$

Set the partial derivative of $\Gamma(\boldsymbol{P})$ with respect to $\boldsymbol{P}$ as zero, then, we have the following equation:

$$\frac{(2\mu+\mu_1)}{\mu_1}\boldsymbol{P}+\boldsymbol{P}\mathbf{1}_n\mathbf{1}_n^T-\frac{\boldsymbol{E}}{\mu_1}=\mathbf{0} \tag{19}$$

or equivalent

$$\boldsymbol{P}=\frac{\boldsymbol{E}}{2\mu+\mu_1}(\boldsymbol{I}+\frac{\mu_1}{2\mu+\mu_1}\mathbf{1}_n\mathbf{1}_n^T)^{-1} \tag{20}$$

with $\boldsymbol{E}=(\mu_1\mathbf{1}_n\mathbf{1}_n^T+\mu_1\boldsymbol{Q}-\boldsymbol{A}-\beta\boldsymbol{B}-\boldsymbol{Y}_1\mathbf{1}_n^T-\boldsymbol{Y}_2)$, where $\boldsymbol{A}_{ij}=||\boldsymbol{W}^T\boldsymbol{x}_i-\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{s}_j||^2$, and $\boldsymbol{B}_{ij}=||\boldsymbol{W}^T\boldsymbol{x}_i-\boldsymbol{W}^T\boldsymbol{x}_j||^2$. We can update $\boldsymbol{P}$ by Eq. (20).

3.4.3. Updating variable $\boldsymbol{S}$

Thirdly, we fix other variables, and the optimization objective function can be transferred into:

$$\begin{aligned}\Gamma(\boldsymbol{S})=&\min_{\boldsymbol{S}}\sum_{j=1}^{n}||\boldsymbol{W}^T\boldsymbol{x}_i-\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{s}_j||^2\boldsymbol{P}_{ij}+\alpha||\boldsymbol{T}||_1+\frac{\mu_1}{2}||\boldsymbol{S}-\boldsymbol{T}+\frac{\boldsymbol{Y}_3}{\mu_1}||_F^2\\ &s.t.\quad \boldsymbol{S}_{\boldsymbol{ii}}=\mathbf{0}\end{aligned} \tag{21}$$

In order to find the optimal solution of $\Gamma(\boldsymbol{S})$, we take the derivative of the function with respect to $\boldsymbol{S}$, and set it as zero ($\frac{\partial\Gamma}{\partial\boldsymbol{S}}=\mathbf{0}$):

$$\frac{\boldsymbol{C}}{\mu_1}\boldsymbol{S}+\boldsymbol{S}\boldsymbol{D}_{\boldsymbol{P}^T}^{-1}+\frac{(\boldsymbol{C}\boldsymbol{P}-\mu_1\boldsymbol{T}+\boldsymbol{Y}_3)\boldsymbol{D}_{\boldsymbol{P}^T}^{-1}}{\mu_1}=\mathbf{0} \tag{22}$$

where $\boldsymbol{C}=2\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}$. Then $\boldsymbol{S}$ can be obtained by solving a Sylvester equation.

3.4.4. Updating variables $\boldsymbol{Q}$, $\boldsymbol{T}$ and Lagrange multipliers

$\boldsymbol{Q}$ can be updated by solving the following optimization function:

$$\begin{aligned}\boldsymbol{Q}^*=arg\quad&\min_{\boldsymbol{Q}}\frac{\mu_1}{2}||\boldsymbol{P}-\boldsymbol{Q}+\frac{\boldsymbol{Y}_2}{\mu_1}||_F^2\\ &s.t.\quad \boldsymbol{Q}>\mathbf{0}\end{aligned} \tag{23}$$

and the solution is:

$$\boldsymbol{Q}^*=max(\boldsymbol{P}+\frac{\boldsymbol{Y}_2}{\mu_1},\mathbf{0}) \tag{24}$$

$\boldsymbol{T}$ can be updated by solving the following optimization function:

$$\boldsymbol{T}^*=arg\quad\min_{\boldsymbol{T}}\alpha||\boldsymbol{T}||_1+\frac{\mu_1}{2}||\boldsymbol{S}-\boldsymbol{T}+\frac{\boldsymbol{Y}_3}{\mu_1}||_F^2 \tag{25}$$

by utilizing the shrinkage operator, we have:

$$\boldsymbol{T}^* = shrink(\boldsymbol{S} + \frac{\boldsymbol{Y}_3}{\mu_1}, \frac{\alpha}{\mu_1}) \tag{26}$$

where $shrink(x, a) = sign(x)max(|x| - a, 0)$.

Finally, we can update Lagrange multipliers $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3$, and the penalty parameter μ_1 by Eq. (27):

$$\begin{cases} \boldsymbol{Y}_1 = \boldsymbol{Y}_1 + \mu_1(\boldsymbol{P}\mathbf{1}_n - \mathbf{1}_n) \\ \boldsymbol{Y}_2 = \boldsymbol{Y}_2 + \mu_1(\boldsymbol{P} - \boldsymbol{Q}) \\ \boldsymbol{Y}_3 = \boldsymbol{Y}_3 + \mu_1(\boldsymbol{S} - \boldsymbol{T}) \\ \mu_1 = min(\mu_{max}, \rho\mu_1) \end{cases} \tag{27}$$

where $\rho > 0$ is the iteration step-size, and μ_{max} is a constant.

In summary, the procedure of optimization process is listed in Algorithm 1.

Algorithm 1 Solving problem (11) by ADMM

Input: $\boldsymbol{X} \in \mathbb{R}^{d\times n}$, α, β, γ, μ, c;

Initialization: $\boldsymbol{P} = \boldsymbol{P}_{knn}$, $\boldsymbol{Q} = \boldsymbol{P}$, $\boldsymbol{S} = \frac{1}{n}\mathbf{1}$, $\boldsymbol{T} = \boldsymbol{S}$, $\mu_{max} = 10^7$, $\boldsymbol{Y}_1 = 0$, $\boldsymbol{Y}_2 = \boldsymbol{Y}_3 = \mathbf{0}$, $\mu_1 = 0.1$, $\rho = 1.01$, $\varepsilon = 10^{-6}$;

1: **while** not converged **do**
2: Fix other variables and update $\boldsymbol{W}$ by solving Eq. (17)
3: Fix other variables and update $\boldsymbol{P}$ by Eq. (20)
4: Fix other variables and update $\boldsymbol{S}$ by solving Eq. (22)
5: Fix other variables and update $\boldsymbol{Q}$ by Eq. (24)
6: Fix other variables and update $\boldsymbol{T}$ by Eq. (26)
7: Update the multipliers and the penalty parameter by solving Eq. (27)
8: Check the convergence condition $||\boldsymbol{P}\mathbf{1}_n - \mathbf{1}_n||_F < \varepsilon$ AND $||\boldsymbol{P} - \boldsymbol{Q}||_F < \varepsilon$ AND $||\boldsymbol{S} - \boldsymbol{T}||_F < \varepsilon$
9: **end while**

Output: $\boldsymbol{W}$

4. Discussion

4.1. Computational Complexity

The major computational burden of our algorithm is in Eq. (17), Eq. (20), and Eq. (22), because eigenvalue decomposition (EVD) and Sylvester equation problem are involved. Especially in Eq. (17), the EVD is operated on a $d \times d$ matrix with the computational complexity $o(d^3)$. In addition, the Eq. (20) and Eq. (22) are operated on $n \times n$ matrices with the complexity about $o(2n^3)$. Thus,the main computational complexity of Algorithm 1 is $o(\tau(d^3 + 2n^3))$ where τ is the number of iterations.

4.2. Convergence Analysis

In this section, we would like to represent the convergence behavior of our proposed framework. We operate the framework with the k-means clustering on four image data sets, including TOX, YALE, UMIST, and ORL. We deem that the algorithm tends to converge if all variables and the recognition accuracies are stable. Therefore we define an objective function in formula (28), which represents the sum of changes of all variables. We operate our method for 300 steps of iterations to show the numerical value of the objective function and clustering accuracy in Fig.1.

$$obj = ||\boldsymbol{P} - \boldsymbol{Q}||_F + ||\boldsymbol{S} - \boldsymbol{T}||_F < \varepsilon \tag{28}$$

Although UFSARP is a nonconvex optimization problem, we can obtain its local optimal solution by using the ADMM algorithm. In Fig.1, it is clear that the value of objective function decreases along with the increase of iterations. The convergence of the objective function is not smooth, which may be caused by regularization terms. The main influence may come from the term $||S||_1$ and term $||W||_{2,1}$. Variable S and variable W are not convex in each iteration. There are two possible reasons: 1) It is difficult to ensure that matrices X and X^T (in Eq. (17)) are nonsingular. So the pseudo inverse of XX^T may cause the fluctuation; 2) The Sylvester equation (in Eq. (22)) may also cause the fluctuation of curves. But the value of the function will eventually converge by ADMM. For the accuracy curves, there still exist some waves. The reason may be that k-means based clustering depends on the initialization.

4.3. The Determination of Parameter μ

It is clear that the parameter μ is used to control the trade off between the trivial solutions ($\mu = 0$) and the uniform distribution ($\mu = \infty$).

The determination of parameter μ has been researched in some literatures[11][43]. Based on the previous research, we define μ as follows,

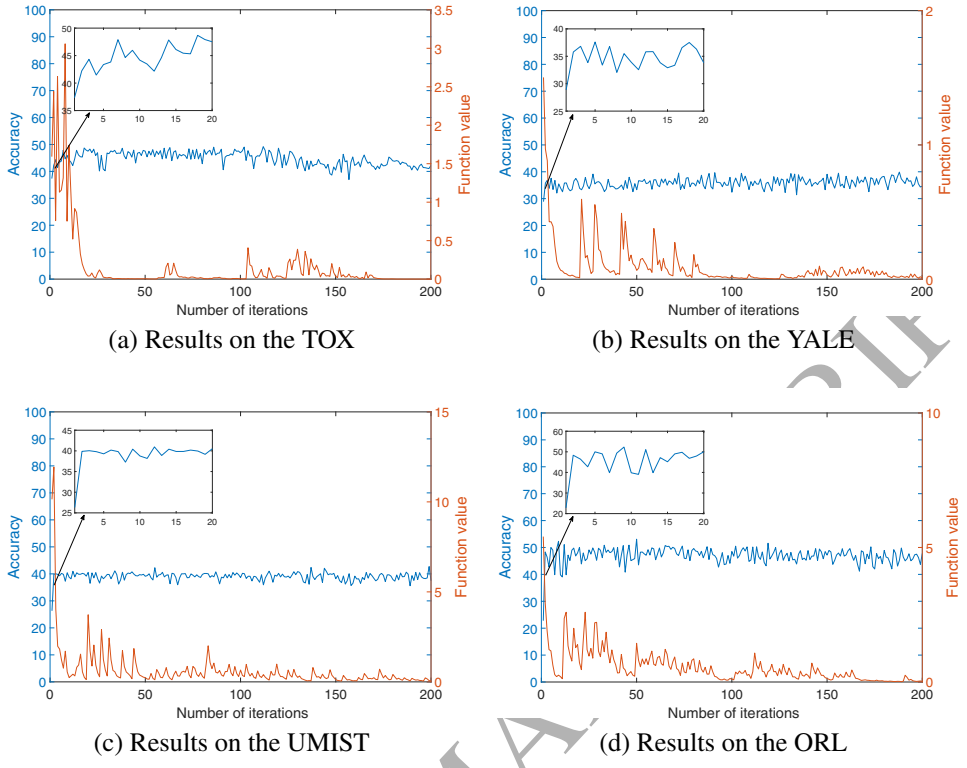
$$\mu = \frac{1}{n}\sum_{i}^{n}(\frac{k}{2}d_{i,k'+1}^{W} - \frac{1}{2}d_{ik'}^{W}) \tag{29}$$

where k is the neighborhood size. In this way, the search of μ can be better handled by searching k, which is more intuitive and easy to tune.

4.4. Comparison with FSASL

Du et al. proposed FSASL method in 2015, and the method can be formulated as Eq. (3). Comparing Eq. (3) with our proposed method Eq. (11), the main differences between UFSARP and FSASL are as follows:

UFSARP not only unifies the subspace learning and feature selection into one process, but also retains the local residuals during the data reconstruction process, while FSASL does not consider the local residuals. More specifically, we capture the local structure and residual relationships between samples during reconstruction process, which encourages similar samples not only have similar reconstruction coefficients, but also have similar reconstruction residuals. Therefore, our method can help to better represent the intrinsic characteristics of the data set.



(a) Results on the TOX (b) Results on the YALE

(c) Results on the UMIST (d) Results on the ORL

Figure 1: The objective function value and clustering accuracy versus the iterations of the the proposed method on (a) TOX biomedical data set, (b) YALE face image data set, (c) UMIST face image data set, and (d) ORL biomedical data set.

5. Experiments

In this section, we conduct extensive experiments to evaluate the performance of the proposed UFSARP for the task of unsupervised feature selection.

5.1. Data Sets

The experiments will be tested based on ten open source data sets, including handwritten and spoken digit/letter recognition data sets (i.e., MFEA from UCI repository[1] and USPS[2] data sets), five face image data sets (i.e., UMIST[3], JAFFE[4], AR[5], YALE[6], and ORL[7]), one object data set

[1]http://www.ics.uci.edu/∼mlearn/MLRepository.html
[2]http://www-stat-class.stanford.edu/∼tibs/ElemStatLearn/data.html
[3]http://images.ee.umist.ac.uk/danny/database.html
[4]http://www.kasrl.org/jaffe.html
[5]http://www2.ece.ohio-state.edu/∼aleix/ARdatabase.html
[6]http://cvc.yale.edu/projects/yalefaces/yalefaces.html
[7]www.zjucadcg.cn/dengcai/Sata/FaceData.html

(i.e., COIL[8]) and two biomedical data sets (i.e., LUNG[9] and TOX[10]). In this paper, we conduct experiments based on all samples in the data sets except for MFEA and USPS. For these two data sets, we randomly select 200 images from each category, and the brief of these benchmark data sets are summarized in Table 1.

- MFEA200 contains series of handwritten numerals $(0-9)$ which are extracted from a collection of Dutch utility maps. The size of one image has 15×16 pixels. In our experiments, we randomly selected 200 images from 10 projects with 20 samples for each project.
- USPS200 has 9298 samples in total, in which there are 7291 samples for the training set and 2007 samples for the test set. The samples are handwritten images in digits $4\ versus\ 9$, and each image is in 16×16 pixels. In our experiments, we randomly selected 200 samples from 2 projects with 100 samples for each project.
- UMIST is a data set that contains 564 images of 20 people. Each person conducts a range of poses from profile to frontal views, and the persons are from different races, sexes, and appearances. The files are all in PGM format with 220×220 pixels in 256 shades of grey.
- JAFFE is a data set that contains 213 images of 7 facial expressions, including six basic facial expressions and one neutral, posed by 10 Japanese female models.
- AR is a data set about color face images of 126 people and has over 4000 images in total. The images are taken from frontal views with various facial expressions, lighting conditions and occlusions.
- COIL is a data set that consists 1440 samples of 20 objects with 72 samples for each object. The samples are captured from varying angles at intervals of 5 degrees. All the images are cropped and resized to 32×32 pixels.
- ORL is consisted of 400 images of 40 distinct subjects with ten images for each subject. The images are taken at different time, lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (wearing/not wearing glasses). All the images are taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement), and are cropped and then resized to 32×32 pixels.
- YALE is a data set that has 165 grayscaling images in GIF format of 15 individuals, 11 images per person. The images are in different facial expression or configuration: center/right/left lighting, wearing/not wearing glasses, happy/normal/sad/sleepy/surprised/wink. The images are cropped and then resized to 32×32 pixels.

[8]http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
[9]https://archive.ics.uci.edu/ml/datasets/lung+cancer
[10]http://featureselection.asu.edu/datasets.php

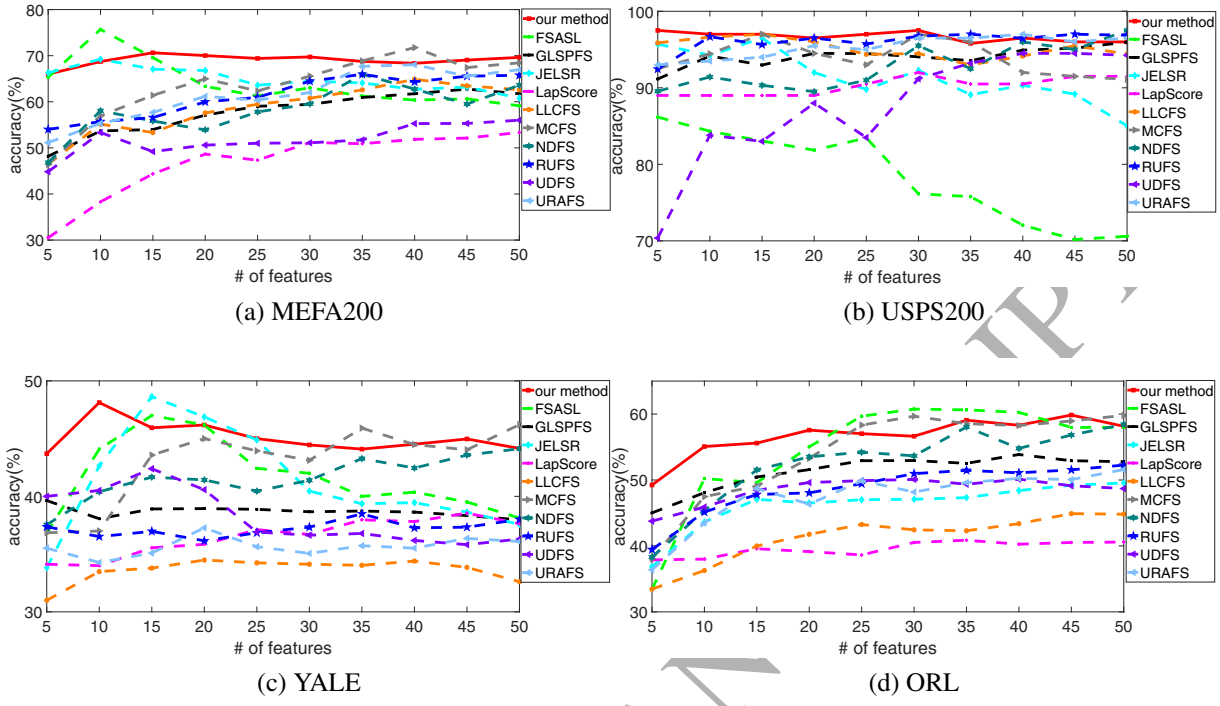
- LUNG has 203 images from 5 classes, and each class has 139, 21, 20, 6, 17 samples respectively. Each sample has 12600 genes. We removed those genes with standard deviations smaller than 50 expression units and then obtained a data set with 203 samples and 3312 genes in our experiments.
- TOX has 171 samples from 4 classes in total, and we have obtained a data set with 171 samples and 5748 genes in our experiments.

Table 1: Summary of the benchmark data sets and the number of selected features.

Data Sets.	sample	feature	class	selected features
MFEA200	200	240	10	[5,10,...,50]
USPS200	200	256	2	[5,10,...,50]
UMIST	575	644	20	[5,10,...,50]
JAFFE	213	676	10	[5,10,...,50]
AR	840	768	120	[5,10,...,50]
COIL	1440	1024	20	[5,10,...,50]
ORL	400	1024	40	[5,10,...,50]
YALE	165	1024	15	[5,10,...,50]
LUNG	203	3312	5	[10,20,...,150]
TOX	171	5748	4	[10,20,...,150]

Table 2: Aggregated clustering results measured by Accuracy (%) of the compared methods

DataSets	UMIST	JAFFE	AR	COIL	LUNG	TOX	Average
AllFea	42.40	71.57	30.26	59.17	72.46	43.65	53.25
LapScore	36.73 ± 1.18	67.62 ± 8.49	25.29 ± 2.89	45.60 ± 6.16	58.97 ± 5.24	40.25 ± 0.65	45.74
MCFS	44.46 ± 3.26	73.56 ± 4.83	29.05 ± 1.19	51.50 ± 5.38	70.42 ± 3.41	43.10 ± 1.86	52.02
LLCFS	47.31 ± 0.83	64.79 ± 4.08	34.22 ± 2.70	50.84 ± 3.76	71.58 ± 5.85	39.28 ± 0.49	51.34
UDFS	48.04 ± 1.92	75.48 ± 1.63	30.87 ± 0.35	48.40 ± 16.89	65.46 ± 3.88	47.14 ± 0.75	52.57
NDFS	52.80 ± 2.26	74.98 ± 2.15	32.34 ± 1.52	52.22 ± 6.33	75.52 ± 1.57	38.28 ± 1.64	54.36
URAFS	45.77 ± 2.89	79.86 ± 8.63	**40.67 ± 1.30**	56.68 ± 3.84	66.85 ± 7.65	49.80 ± 1.68	56.61
RUFS	50.87 ± 1.95	75.75 ± 2.53	34.84 ± 1.90	59.20 ± 3.28	77.35 ± 2.62	49.17 ± 0.83	57.86
JELSR	53.52 ± 1.54	77.77 ± 1.87	34.19 ± 2.52	59.53 ± 4.01	77.86 ± 3.12	43.96 ± 1.56	57.81
GLSPFS	50.53 ± 0.59	75.46 ± 1.61	34.12 ± 1.60	57.96 ± 2.27	77.83 ± 2.70	47.38 ± 1.93	57.21
FSASL	**54.92 ± 1.89**	79.29 ± 2.24	36.11 ± 0.75	60.93 ± 2.50	81.93 ± 1.63	50.12 ± 0.67	60.55
UFSARP	53.99 ± 4.14	**81.39 ± 9.11**	39.32 ± 0.87	**61.87 ± 5.91**	**83.33 ± 2.58**	**52.57 ± 3.36**	**62.08**



(a) MEFA200 (b) USPS200

(c) YALE (d) ORL

Figure 2: The clustering accuracy versus the number of selected features of the the comparison methods on (a) MEFA200, (b) USPS200, (c) YALE, and (d) ORL.

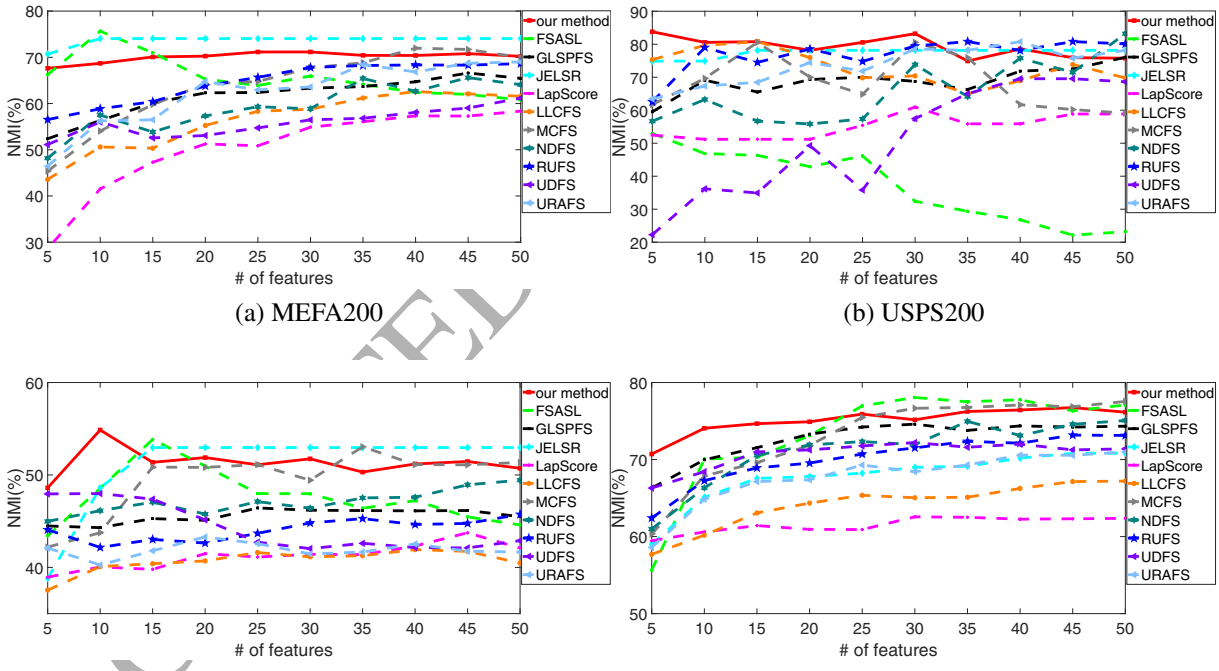
5.2. *Experiment Setup*

To validate the effectiveness of our proposed UFSARP, we compared our method with state-of-the-art unsupervised feature selection methods and one baseline (i.e., allfea).

- LapScore [18] evaluates and selects features according to their ability of locality preserving of the manifold structure.
- MCFS [5] selects features by adopting spectral regression with l_1-norm regularization.
- LLCFS [62] incorporates the relevance of each feature into the built-in regularization of the local learning-based clustering algorithm.
- UDFS [61] exploits local discriminative information and feature correlations simultaneously.
- NDFS [31] selects features by using a joint framework of nonnegative spectral analysis and $l_{2,1}$-norm regularized regression.
- RUFS [46] performs robust clustering and robust feature selection simultaneously to select the most critical and discriminative features.

Table 3: Aggregated clustering results measured by NMI (%) of the compared methods

DataSets	UMIST	JAFFE	AR	COIL	LUNG	TOX	Average
AllFea	64.15	81.52	65.48	75.58	60.37	15.87	60.50
LapScore	55.57 ± 2.32	77.28 ± 8.98	63.59 ± 2.36	62.21 ± 4.98	50.14 ± 4.13	10.92 ± 0.68	53.29
MCFS	63.46 ± 4.93	79.04 ± 5.88	66.41 ± 0.85	66.19 ± 6.78	55.68 ± 2.31	16.53 ± 2.68	57.89
LLCFS	63.42 ± 1.42	66.97 ± 3.47	69.01 ± 1.45	64.04 ± 4.34	60.12 ± 4.65	9.68 ± 0.75	55.54
UDFS	65.19 ± 2.96	84.25 ± 1.74	67.49 ± 0.27	44.27 ± 12.61	54.88 ± 4.21	22.16 ± 1.36	56.37
NDFS	71.19 ± 2.77	82.53 ± 3.49	67.89 ± 0.89	56.29 ± 6.91	60.57 ± 1.54	9.07 ± 1.87	57.92
URAFS	62.53 ± 2.23	81.37 ± 3.56	70.42 ± 0.59	69.75 ± 2.17	51.97 ± 4.22	26.16 ± 2.22	60.37
RUFS	68.19 ± 2.61	82.00 ± 3.56	69.54 ± 1.10	70.54 ± 4.48	65.47 ± 1.87	25.79 ± 1.60	63.59
JELSR	71.33 ± 2.06	85.23 ± 3.31	69.02 ± 1.32	71.37 ± 4.97	63.54 ± 2.94	17.46 ± 3.36	62.99
GLSPFS	69.16 ± 0.97	83.20 ± 3.17	69.44 ± 0.84	69.89 ± 4.00	63.50 ± 2.99	23.49 ± 2.77	63.11
FSASL	**72.39 ± 2.39**	**86.42 ± 3.34**	**70.78 ± 0.63**	72.93 ± 4.44	**66.78 ± 1.72**	27.37 ± 1.62	**66.11**
UFSARP	66.30 ± 2.17	**86.42 ± 3.98**	69.92 ± 0.50	**73.81 ± 2.06**	61.34 ± 2.92	**28.30 ± 3.24**	64.35



(a) MEFA200 (b) USPS200

(c) YALE (d) ORL

Figure 3: The clustering NMI versus the number of selected features of the the comparison methods on (a) MEFA200, (b) USPS200, (c) YALE, and (d) ORL.

- JELSR [20][21] combines embedding learning with sparse regression to perform feature selection.

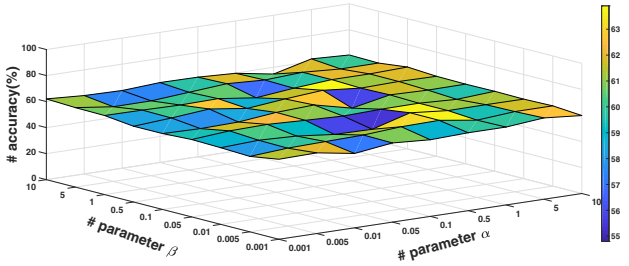
- GLSPFS [34] integrates both global pairwise sample similarity and local geometric data structure to conduct feature selection.
- FSASL [11] unifies the structure construction and feature selection into one framework.
- URAFS [30] performs the feature selection and spectral clustering simultaneously by an uncorrelated regression model.

For the convenience of calculation, we preset several parameters. The size of neighborhood is set as $k = 5$. Note that in the GLSPFS algorithm, the pairwise similarity is determined based on Gaussian Kernel, where the kernel width is searched from grid $\{2^{-3}, 2^{-2}, ..., 2^{3}\}\ \delta_0$, and δ_0 is the mean distance between any two samples. For GLSPFS, we record the best results among three local manifold models, which are locality preserving projection (LPP), locally linear embedding (LLE) and local tangent space alignment (LTSA)[16]. To fairly compare the unsupervised feature selection algorithms, we use grid search strategy to regularize the parameters in all methods. Regarding to the regularization parameters in methods LLCFS, UDFS, NDFS, RUFS, JELSR, GLSPFS, FSASL, and URAFS, we search them from the grid $\{10^{-5}, 10^{-4}, ..., 10^{5}\}$. For UFSARP, α and β are searched from grid $\{10^{-3}, 5 \times 10^{-3}, 10^{-2}, ..., 5, 10\}$, γ is searched in the grid $\{10^{-3}, 10^{-2}, ..., 10^{2}, 10^{3}\}$. Furthermore, to compare the generalized performance of each method, we choose two widely used metrics, i.e., Accuracy (ACC) and Normalized Mutual Information (NMI). Since all the algorithms are affected by different parameters, we would like to repeat the clustering for twenty times with random initialization and record the average result.

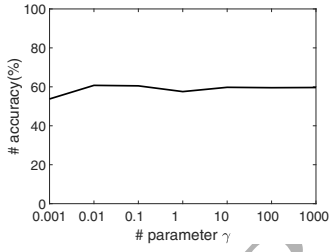
5.3. Clustering with Selected Features

It is difficult to determine the best optimal number of selected features in different algorithms and data sets. So, to better evaluate the performance of unsupervised feature selection algorithms, we demonstrate the average of best results over different number of selected features (the number of selected features for all data sets can be seen in Table 1) with standard deviation. Table 2 and Table 3 show the clustering results measured by Accuracy and NMI respectively. Numbers posted in each cell are the $mean \pm standard\ deviation$ of the performances of algorithms in different data sets, and the last columns in both tables show the average results of different feature selection algorithms over six data sets.

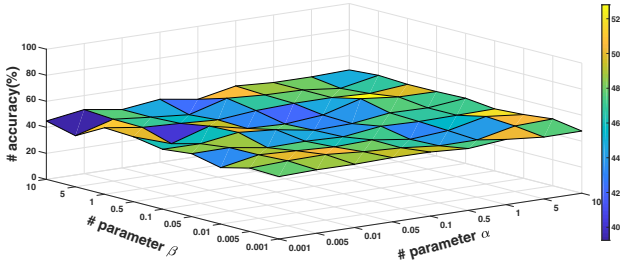
From both Table 2 and Table 3, we find that there is a substantial improvement in clustering result after removing the redundant features from the data sets. Namely, a better performance can be achieved in data clustering by utilizing feature selection methods. It is obvious that UFSARP performs better than LapScore, MCFS, LLCFS, UDFS, RUFS, JELSR, URAFS, and GLSPFS. In Table 2, the Accuracy of UFSARP is slightly inferior to FSASL and URAFS on UMIST data set and AR data set, respectively. However, UFSARP still demonstrates its superiority on the rest data sets. From Table 3, it can be seen that UFSARP does not achieve the best NMI on UMIST and LUNG, and the NMI of UFSARP is slightly inferior to FSASL and URAFS on AR data set. UFSARP gets better result on the rest data sets. Table 2 and Table 3 show that UFSARP achieves good Accuracy, and it does not mean that a good NMI can be obtained simultaneously, such as on AR and LUNG data sets. In particular, our proposed method UFSARP achieves 16.58% and 6.37% improvement regarding to Accuracy and NMI respectively with less than 10% features.



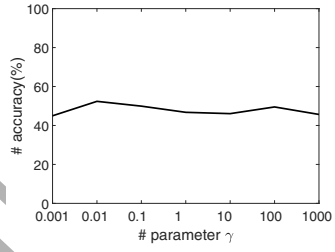
(a) Variations of clustering Accuracy(%) versus parameters α and β on MEFA200



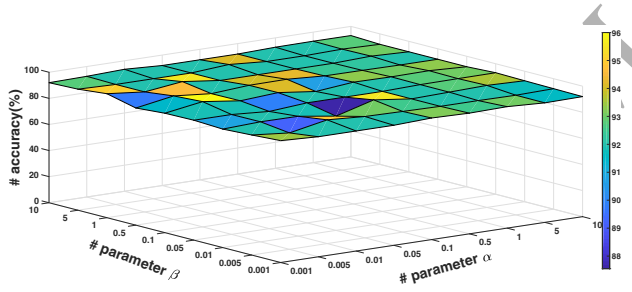
(b) Variations of clustering Accuracy(%) versus parameters γ on MEFA200



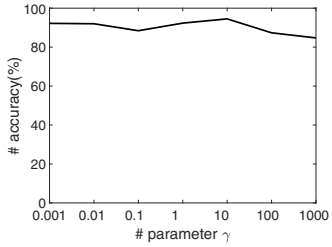
(c) Variations of clustering Accuracy(%) versus parameters α and β on ORL



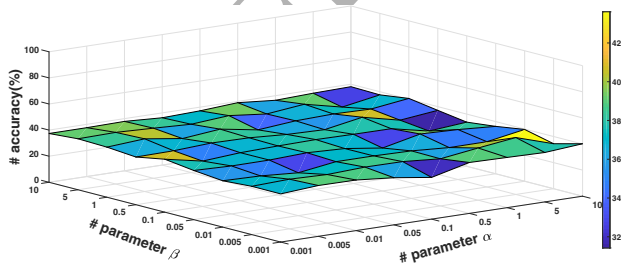
(d) Variations of clustering Accuracy(%) versus parameters γ on ORL



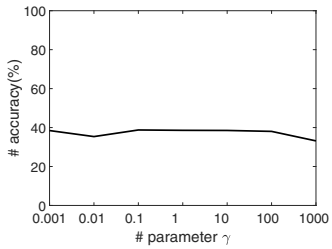
(e) Variations of clustering Accuracy(%) versus parameters α and β on USPS200



(f) Variations of clustering Accuracy(%) versus parameters γ on USPS200

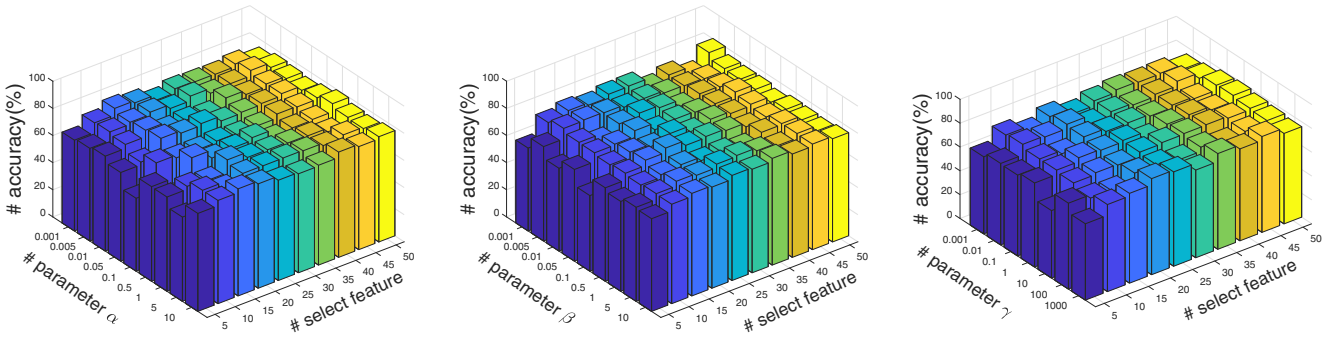


(g) Variations of clustering Accuracy(%) versus parameters α and β on YALE



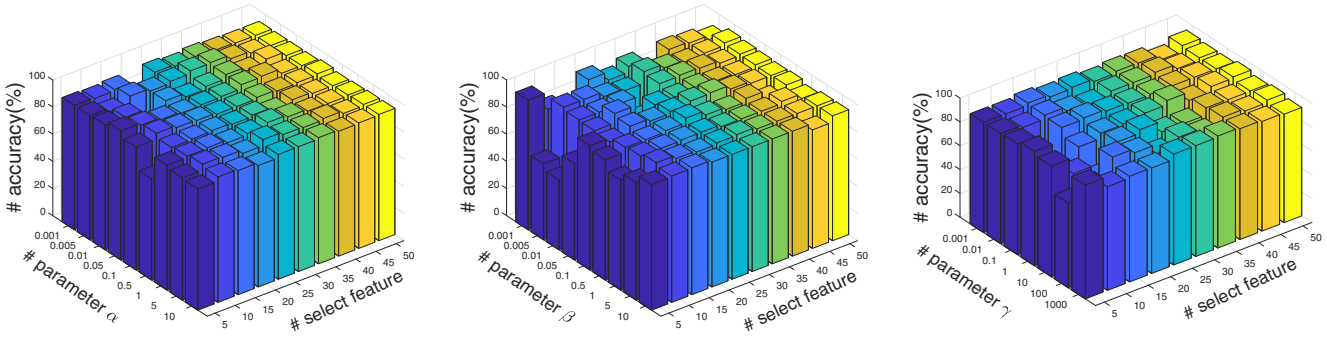
(h) Variations of clustering Accuracy(%) versus parameters γ on YALE

Figure 4: Variations of clustering Accuracy(%) versus parameters α, β and γ on (a)(b)MEFA200, (c)(d)ORL, (e)(f)USPS200 and (g)(h)YALE data sets.



(a) α (b) β (c) γ

Figure 5: The Clustering Accuracy of the proposed method versus parameters α and β with γ fixed on the JAFFE face image data set.



(a) α (b) β (c) γ

Figure 6: The Clustering Accuracy of the proposed method versus parameters α and β with γ fixed on the USPS200 biomedical data set.

Also, we compare the Accuracy of different algorithms by selecting a various number of features. These experiments are conducted on data sets of ORL, YALE, USPS200, and MEFA200. The results are shown in Fig. 2 and Fig. 3, which indicate that our method results in much higher Accuracy than those conventional feature selection methods. In different dimensions, UFSARP performs better than other approaches in most cases. Especially in lower dimensions, UFSARP still achieves good results. The reasons may be that UFSARP performs feature selection and graph embedding simultaneously, which encourages the reconstructed data to maintain local manifold structure after feature selection. Fig. 2 and Fig. 3 show that as the number of selected features increases, the Accuracy increases in general, but the best results may not appear in most selected features. Intuitively, if the number of selected features is too small, it is not enough to represent the key information of data. Conversely, a large number of selected features may increase the redundant information.

In summary, comparing with other algorithms, our proposed method is more stable and has an average higher accuracy. As seen from Fig. 2, Fig. 3, Table 2, and Table 3, we have the following conclusions.

(1) All the feature selection methods generally get better performance than using all features, which indicates that feature selection can reduce redundancy and enhance performance.

(2) Both clustering Accuracy and NMI indicate that UFSARP performs better than other approaches in most cases. The main reason is that UFSARP not only unifies the manifold learning and feature selection into one process, but also retains the value of local residual struture in the data reconstruction process.

(3) With the increasing number of selected features, the accuracy increase generally. However, the result will no longer be promoted when the number of selected features increases to a certain extent. Besides, UFSARP can still achieve good results at lower dimensions, since the local similarity matrix P is adaptively learnt.

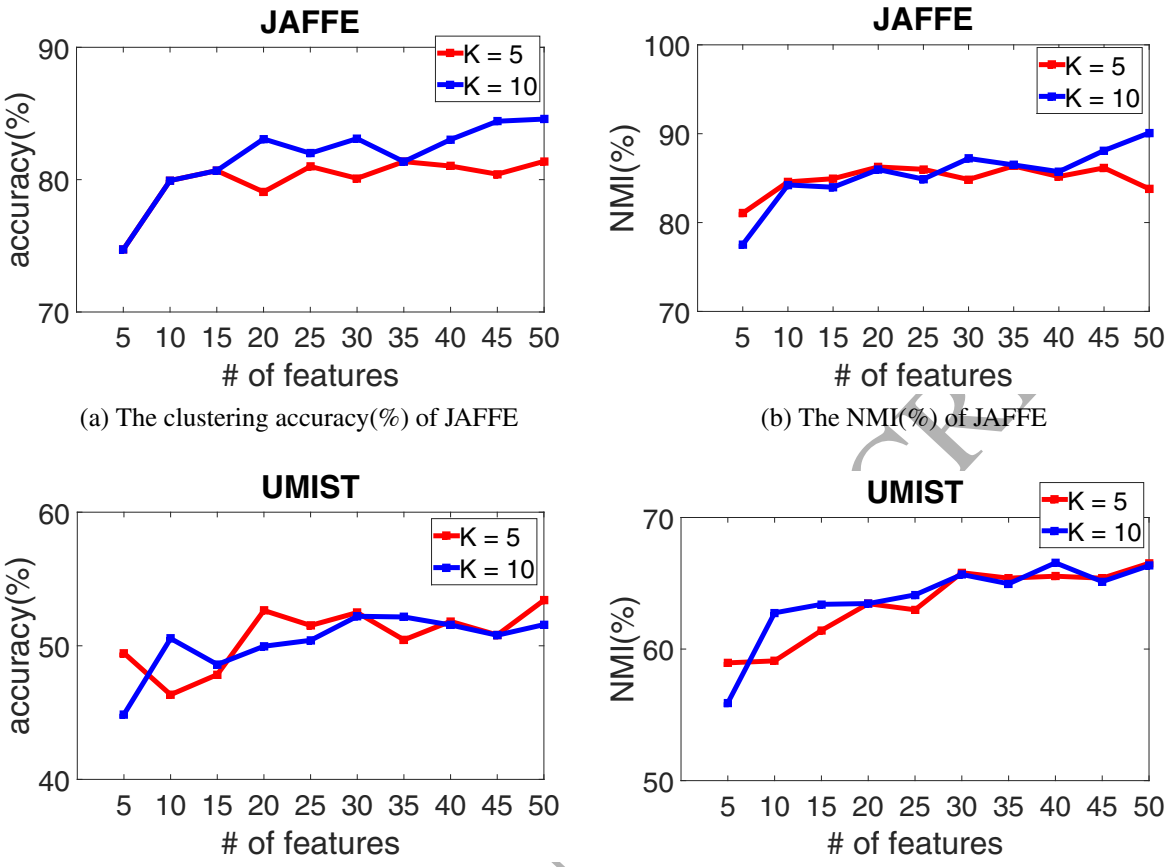
5.4. Parameter Sensitivity

We study the sensitivity of the regularization parameters α, β and γ. When we fix the value of some parameters, we keep other parameters changed at the optimal value. First, we focus on the influence of α and β by fixing γ. Parameters α and β are used to control the trade off between the global structure and the local structure. Fig. 4(a), (c), (e), (g) show the clustering Accuracy variation of our method with respect to different values of parameters α and β. The graphs are generally flat which indicates that the performance of our method is not sensitive to the values of α and β. Then we focus on the influence of parameter γ by fixing others. As we change the value of γ, the variances of performance are demonstrated in Fig. 4(b), (d), (f), (h). Our method is robust to different values of γ on these data sets when it is in the range of $[10^{-2}, 10^{-1}]$ or $[10, 10^{2}]$, which indicates that a suitable value of γ can guarantee a better row sparsity of W. The results show that our method is robust in terms of γ.

We are also interested in the sensitivity of the number of selected features. The results are shown in Fig. 5 and Fig. 6. Generally speaking, UFSARP obtains the best performance when the range of the number of selected features is large, i.e., the range is [20, 50]. Besides, it is easy to find that parameters α, β and γ do not affect the Accuracy too much when there are enough selected features. Specifically, in Fig. 5, the Accuracy is depressing in the case of few selected features. The reason is intuitive to understand that, for face images, too few features can not distinguish between different samples. Fig. 6 shows that UFSARP is very robust to the number of selected features on UFSARP. Even in scenario of 5 selected features, the algorithm still achieves good results. Because USPS200 is handwritten data set, it is easy to find significant features between samples.

In short, the experimental results show that our method is not very sensitive to α, β, γ, and even the number of selected features, i.e., our method is pretty robust to different parameters.

5.5. Effect of Neighborhood Size and Running Time



(a) The clustering accuracy(%) of JAFFE

(b) The NMI(%) of JAFFE

(c) The clustering accuracy(%) of UMIST

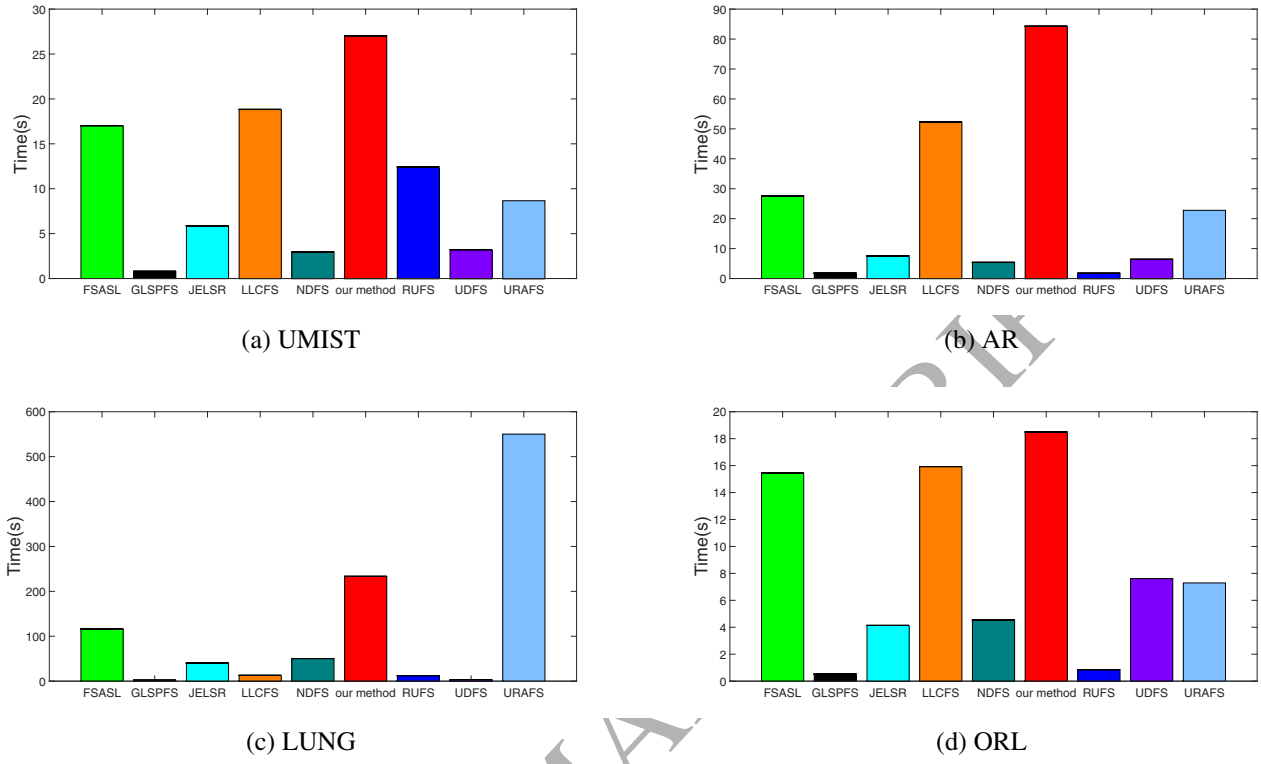
(d) The NMI(%) of UMIST

Figure 7: Comparison on different size of neighborhood. (a)(b) JAFFE, (c)(d) UMIST.

In the previous experiment, the size of neighborhood K is set as 5. We investigate the effect of neighborhood size and has conducted another experiment for $K = 10$ on data sets of JAFFE and UMIST. The results are shown in Fig.7.

From Fig.7 we can see that the proposed method obtains better performance when $K = 10$ on JAFFE data set. While, on data set UMIST, there is a better accuracy when $K = 5$, and the NMI is higher when $K = 10$. Therefore, the overall results of accuracy and NMI are similar on both data sets when $K = 5$ and $K = 10$.

In addition, we conduct another experiment to show the running time of our method that runs on the real data set. All of the experiments are implemented on MATLAB R2014b, and the codes are run on a Windows 10 machine with 2.80-GHz i7-7700HQ CPU, 16GB main memory. Fig. 8 shows the running time of the methods in our experiments. We compared the proposed UFSARP with seven related methods FSASL, GLSPFS, JELSR, LLCFS, NDFS, URAFS, and UDFS. From the results in Fig. 8, the running time of UFSARP is slightly inferior to other methods except URAFS on LUNG data set. However, the longer running time of UFSARP is within a reasonable range that is acceptable since it has comparably promising experimental results.



(a) UMIST (b) AR

(c) LUNG (d) ORL

Figure 8: Comparison on running time. (a)UMIST, (b)AR, (c)LUNG, and (d)ORL. The X-axis represents the feature selection methods, while y-axis records the running time.

6. Conclusion

In this paper, we proposed a novel unsupervised feature selection method to improve the credibility of reconstruction structure. In our method, the local structure and the feature selection are integrated within a unified framework. Namely, the framework processes the data reconstruction and feature selection simultaneously within one single function. In addition, by considering the data local residuals, local manifold structure can be better preserved in the data reconstruction process. The extensive experiments have been conducted on the real-world benchmark data sets, and experimental results demonstrate that our method performs superior to all others.

In UFSARP, the complexity level is relatively higher than other frameworks. In the future, we plan to further investigate the local residual, and alleviate the complexity.

7. Acknowledgement

This work is supported in part by the National Natural Science Foundation of China under Grants 61402118, 61772141, 61702110, 61772141, and 61673123, in part by the Guangdong Provincial Natural Science Foundation, under Grant 17ZK0422, and in part by the Guangdong Provincial Science & Technology Project under Grants 2016B010108007, and in part by

the Guangzhou Science & Technology Project under Grants 2016201604030034, 201802030011, 201802010042 ,201804010347, and 201802010026

8. Reference

References

[1] G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon, Merlinłrapid analysis of dense genetic maps using sparse gene flow trees, Nature genetics., vol. 30, no. 1, 2002, pp. 97.
[2] M. Belkin, and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural computation., vol. 15, no. 6, pp. 1373-1396, 2003.
[3] C. M. Bishop, Neural networks for pattern recognition, Oxford university press., 1995.
[4] S.Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. vol. 3, no. 1, 2011
[5] D. Cai, C. Zhang, and X. He, Unsupervised feature selection for multi-cluster data, In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining., July. 2010, pp. 333-342.
[6] E. J. Cands, J. K. Romberg, and T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Commun. Pure Appl. Math.,vol. 59, no. 8, 2006, pp. 1207-1223.
[7] E. J. Cands and T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies, IEEE Trans. Inf. Theory., vol. 52, no. 12, Dec. 2006, pp. 5406-5425.
[8] X. Chen, J. Yang, and Z. Jin, An Improved Linear Discriminant Analysis with L_1-Norm for Robust Feature Extraction, International Conference on Pattern Recognition., IEEE Computer Society, 2014, pp. 1585-1590.
[9] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, Learning with ℓ^1-graph for image analysis, IEEE transactions on image processing., vol. 19 no. 4, 2010, pp. 858-866.
[10] D. L. Donoho, For most large underdetermined systems of linear equations the minimal ‘1-norm solution is also the sparsest solution, Commun. Pure Appl. Math., vol. 59, no. 6, 2006, pp. 797-829.
[11] L. Du, and Y. D. Shen, Unsupervised feature selection with adaptive structure learning, In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining., ACM, Aug. 2015, pp. 209-218.
[12] J. G. Dy, and C. E. Brodley, Feature selection for unsupervised learning, Journal of machine learning research., vol. 5, Aug. 2004, pp. 845-889.
[13] X. Fang, Y. Xu, X. Li, Z. Fan, H. Liu, and Y. Chen, Locality and similarity preserving embedding for feature selection, Neurocomputing., vol. 128, no. 5, 2014, pp. 304-315.
[14] X. Fang, Y. Xu, X. Li, Z. Lai, S. Teng, and L. Fei, Orthogonal self-guided similarity preserving projection for classification and clustering, Neural Networks., vol. 88, 2017, pp. 1-8.
[15] Y. A. Ghassabeh, F. Rudzicz, and H. A. Moghaddam, Fast incremental lda feature extraction, Pattern Recognition., vol. 48, no. 6, 2015, pp. 1999-2012.
[16] I. Guyon, and A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research., vol. 3, no. 6, 2003, pp. 1157-1182.
[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Gene selection for cancer classification using support vector machines, Machine learning., vol. 46, no. 1-3, 2002, pp. 389-422.
[18] X. He, D. Cai, and P. Niyogi, Laplacian score for feature selection, International Conference on Neural Information Processing Systems., vol. 18, 2005, pp. 507-514.
[19] X. He, M. Ji, C. Zhang, and H. Bao, A variance minimization criterion to feature selection using laplacian regularization, IEEE Transactions on Pattern Analysis and Machine Intelligence., vol. 33, no. 10, 2011, pp. 2013-2025.
[20] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, Joint embedding learning and sparse regression: a framework for unsupervised feature selection, IEEE Transactions on Cybernetics., vol. 44, no. 6, 2014, pp. 793-804.
[21] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, Feature selection via joint embedding learning and sparse regression, International Joint Conference on Ijcai, 2011, pp. 1324-1329.
[22] S. H. Huang, Supervised feature selection: a tutorial, Artificial Intelligence Research., vol. 4, no, 2, 2015.

[23] Z. Kang, C. Peng, and Q. Cheng, Kernel-driven Similarity Learning, Neurocomputing., vol. 267, 2017, pp. 210-219.
[24] Z. Kang, C. Peng, and Q. Cheng, Twin Learning for Similarity and Clustering: A Unified Kernel Approach, AAAI., 2017, pp. 2080-2086.
[25] Z. Kang, L. Wen, W. Chen, and Z. Xu, Low-rank kernel learning for graph-based clustering, Knowledge-Based Systems., 2018.
[26] Z. Kang, C. Peng, Q. Cheng, and Z. Xu, Unified Spectral Clustering with Optimal Graph, AAAI., 2018.
[27] R. Kohavi, and G. H. John, Wrappers for feature subset selection, Artificial intelligence., vol. 97, no. 1-2, 1997, pp. 273-324.
[28] X. Kong, P. S. Yu, Semi-supervised feature selection for graph classification, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining., 2010, pp. 793-802.
[29] W. J. Krzanowski, Selection of variables to preserve multivariate data structure, using principal components, Journal of the Royal Statistical Society., vol. 36, no. 1, 1987, pp. 22-33.
[30] X. Li, H. Zhang, R. Zhang, Y. Liu, and F. Nie, Generalized Uncorrelated Regression with Adaptive Graph for Unsupervised Feature Selection, IEEE Transactions on Neural Networks and Learning Systems., 2018, Doi. 10.1109.
[31] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, Unsupervised featrue selection using nonnegative spectral analysis, AAAI., 2012.
[32] J. Liu, S. Ji, and J. Ye, Multi-task feature learning via efficient $l_{2,1}$-norm minimization, In Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence., AUAI Press, Jun. 2009, pp. 339-348.
[33] H. Liu, X. Wu, and S. Zhang, A new supervised feature selection method for pattern classification, Computational Intelligence., vol. 30, no. 2, 2014, pp. 342-361.
[34] X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu, Global and local structure preservation for feature selection, IEEE Transactions on NNLS, Vol. 25, no. 6, 2014, pp. 1083-1095.
[35] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, and C. Yuan, Low-rank preserving projections, IEEE transactions on cybernetics., vol. 46 no. 8, 2016, pp. 1900-1913.
[36] L. Ma, M. Li, Y. Gao, T. Chen, X. Ma, and L. Qu, A novel wrapper approach for feature selection in object-based image classification using polygon-based cross-validation, IEEE Geoscience and Remote Sensing Letters., vol. 14, no. 3, 2017, pp. 409-413.
[37] S. Maldonado, and R. Weber, A wrapper method for feature selection using support vector machines, Information Sciences., vol. 179, no. 13, 2009, pp. 2208-2217.
[38] A. M. Martłnez, and A. C. Kak, PCA versus LDA, IEEE transactions on pattern analysis and machine intelligence., vol. 23, no. 2, 2001, pp. 228-233.
[39] P. Mitra, C. A. Murthy, and S. K. Pal, Unsupervised feature selection using feature similarity, Pattern Analysis and Machine Intelligence IEEE Transactions on., vol. 24, no. 3, 2002, pp. 301-312.
[40] C. A. Murthy, Bridging Feature Selection and Extraction: Compound Feature Generation, IEEE Educational Activities Department., vol. 29, no. 4, 2017.
[41] M. N. Murty, and V. S. Devi, Feature Extraction and Feature Selection, Introduction to Pattern Recognition and Machine Learning, IISc Lecture Notes Series., vol. 5, 2015, pp. 75-110.
[42] F. Nie, H. Huang, X. Cai, and C. H. Ding, Efficient and robust feature selection via joint $l_{2,1}$-norms minimization, In Advances in neural information processing systems., 2010, pp. 1813-1821.
[43] F. Nie, X. Wang, and H. Huang, Clustering and projected clustering with adaptive neighbors, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining., ACM, 2014, pp. 977-986.
[44] M. Peng, Q. Xie, H Wang, Y Zhang, and G. Tian, Bayesian sparse topical coding, IEEE Transactions on Knowledge and Data Engineering, 2018
[45] M. Peng, Q. Xie, Y. Zhang, H. Wang, X. Zhang, J. Huang, and G. Tian, Neural sparse topical coding, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018
[46] M. Qian and C. Zhai, Robust unsupervised feature selection, IJCAI, 2013, pp. 1621-1627.
[47] L. Qiao, S. Chen, and X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognition., vol. 43, no. 1, 2010, pp. 331-341.
[48] J. Ren, Z. Qiu, W. Fan, H. Cheng, and P. S. Yu, Forward Semi-Supervised Feature Selection, Advances in

Knowledge Discovery and Data Mining, Pacific-Asia Conference, PAKDD., Osaka, Japan, Vol. 5012, May. 2008, pp. 970-976.

[49] V. Roth, and T. Lange, Feature selection in clustering problems, In Advances in neural information processing systems., 2004, pp. 473-480.

[50] S. T. Roweis, and L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, science., vol. 290, no. 5500, 2000, pp. 2323-2326.

[51] R. J. Schalkoff, Digital image processing and computer vision, New York: Wiley., vol. 286, 1989.

[52] D. L. Swets, and J. J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Transactions on pattern analysis and machine intelligence., vol. 18, no. 8, 1996, pp. 831-836.

[53] S. Tabakhi, P. Moradi, and F. Akhlaghian, An unsupervised feature selection algorithm based on ant colony optimization, Engineering Applications of Artificial Intelligence., vol. 32, 2014, pp. 112-123.

[54] J. B. Tenenbaum, V. De Silva, and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, science., vol. 290, no. 5500, 2000, pp. 2319-2323.

[55] X. D. Wang, C. H. E. N. Rung-Ching, C. Q. Hong, and Z. Q. Zeng, Unsupervised feature analysis with sparse adaptive learning, Pattern Recognition Letters., vol. 102, 2018, pp. 89-94.

[56] S. Wang, J. Tang, and H. Liu, Embedded unsupervised feature selection, AAAI., 2015, pp. 470-476.

[57] J. Weston, A. Elisseeff, B. Scholkopf, and M. Tipping, Use of the zero-norm with linear models and kernel methods, Journal of machine learning research., vol. 3, Mar. 2003, pp. 1439-1461.

[58] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, Robust face recognition via sparse representation, IEEE transactions on pattern analysis and machine intelligence., vol. 31, no. 2, 2009, pp. 210-227.

[59] S. Yan, and H. Wang, Semi-supervised Learning by Sparse Representation, Siam International Conference on Data Mining., DBLP, SDM, Sparks, Nevada, USA, Apr. 2009, pp. 792-801.

[60] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin, Graph embedding and extensions: A general framework for dimensionality reduction, IEEE transactions on pattern analysis and machine intelligence., vol. 29, no. 1, 2007, pp. 40-51.

[61] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning, In IJCAI proceedings-international joint conference on artificial intelligence., vol. 22, no. 1, July. 2011, pp. 1589.

[62] H. Zeng, and Y. M. Cheung, Feature selection and kernel learning for local learning-based clustering, IEEE transactions on pattern analysis and machine intelligence., vol. 33, no. 8, 2011, pp. 1532-1547.

[63] Z. Zhao, L. Wang, H. Liu, and J. Ye, On similarity preserving feature selection, IEEE Transactions on Knowledge and Data Engineering., vol. 25, no. 3, 2013, pp. 619-632.

[64] Z. Zhao, and H. Liu, Spectral feature selection for supervised and unsupervised learning, In Proceedings of the 24th international conference on Machine learning., ACM, June. 2007, pp. 1151-1157.

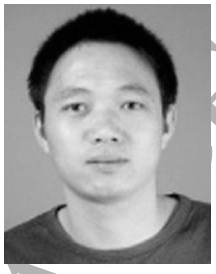
[65] W. Zheng, Z. Lin, and H. Wang, $L1$-norm kernel discriminant analysis via bayes error bound optimization for robust feature extraction, IEEE Transactions on Neural Networks and Learning Systems., vol. 25, no. 4, 2014, pp. 793.

[66] X. Zhu, X. Li, S. Zhang, C. Ju, and X. Wu, Robust joint graph sparse coding for unsupervised spectral feature selection, IEEE transactions on neural networks and learning systems., vol. 28, no. 6, 2017, pp. 1263-1275.

[67] X. Zhu, X. Wu, W. Ding, and S. Zhang, Feature selection by joint graph sparse coding, In Proceedings of the 2013 SIAM International Conference on Data Mining., Society for Industrial and Applied Mathematics, May. 2013, pp. 803-811.

Luyao Teng received Bachelor degree from Monash University, Melbourne, Australia, in 2012, and Master degree in University of Melbourne, Melbourne, Australia, in 2014. She currently is a PhD Candidate in Victoria University, Melbourne, Australia.
Her research interests include pattern recognition and machine learning.

Zhenye Feng received the B.S. degree in computer science and technology from Guangdong University of Technology, Guangzhou, China, in 2016. He is currently pursuing the M.S. degree in computer science and technology from Guangdong University of Technology, Guangzhou, China.
His current research interest includes pattern recognition and machine learning.



Xiaozhao Fang received the M.S. degree and the Ph.D. degree in computer science and technology from the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China, in 2008 and 2016, respectively.
He is currently with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China. His current research interests include pattern recognition, data mining, and machine learning.

Shaohua Teng received his Ph. D. Degrees in Industry Engineering at Guangdong University of Technology, Guangdong, China, in 2008. From 1982 to 1998, he was a teacher at the Jiangxi Normal University. Since 2005, he has been a professor with Guangdong University of Technology.
He is a senior member of Chinese Association of Automation and China Computer Federation. He is also a member of the IEEE. His research interests include big data, data mining, network security, cooperative work, and Petri net theory and applications. He has applied for 7 patents on his invention. He has published 30 papers on computer magazines and international conferences and 2 books. He earned the Provincial Science and Technology Award, and Guangdong outstanding teacher. He is famous teacher.

Hua Wang received his Ph.D. degree from the University of Southern Queensland, Australia. He is now a full time Professor at Victoria University. He was a professor at the University of Southern Queensland before he joined Victoria University. Hua has more than ten years teaching and working experience in Applied Informatics at both enterprise and university. He has expertise in electronic commerce, business process modeling and enterprise architecture. As an Chief Investigator, three Australian Research Council (ARC) Discovery grants have been awarded since 2006, and 200 peer reviewed scholar papers have been published. Six Ph.D. students have already graduated under his principal supervision.

Peipei Kang received the M.S. degree in School of Computer Science and Technology from Guangdong University of Technology, Guangzhou, China. She is currently pursuing the Ph.D. degree in School of

Computer Science and Technology, Guangdong University of Technology, Guangzhou, China.
Her current research interests include pattern recognition and machine learning.

Yanchun Zhang is a full Professor and Director of Centre for Applied Informatics at Victoria University since 2004. Dr Zhang obtained a PhD degree in Computer Science from The University of Queensland in 1991. His research interests include databases, data mining, web services and e-health. He has published over 300 research papers in international journals and conference proceedings including VLDBJ, ACM Transactions on Computer and Human Interaction (TOCHI), IEEE Transactions on Knowledge and Data Engineering (TKDE), SIGMOD and ICDE conferences as well as medical/health journals. Dr. Zhang is a founding editor and editor-in-chief of World Wide Web Journal (Springer) and Health Information Science and Systems Journal (Springer). He is Chairman of International Web information Systems Engineering Society (WISE).