Expert Systems With Applications 96 (2018) 64–76

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Robust graph regularized unsupervised feature selection☆

Chang Tang[a,1], Xinzhong Zhu[b,1,*], Jiajia Chen[c,*], Pichao Wang[d], Xinwang Liu[e], Jie Tian[b]

[a] School of Computer Science, China University of Geosciences, Wuhan 430074, China
[b] School of Life Science and Technology, XIDIAN University, Xi'an, Shanxi 710071, China
[c] Department of Pharmacy, Huai'an Second People's Hospital Affiliated to Xuzhou Medical College, Huai'an 223002, China
[d] School of Computing and Information Technology, University of Wollongong, New South Wales 2522, Australia
[e] School of Computer Science, National University of Defense Technology, Changsha 410073, China

ARTICLE INFO

Article history:
Received 11 September 2017
Revised 26 November 2017
Accepted 27 November 2017
Available online 2 December 2017

MSC:
00-01
99-00

Keywords:
Unsupervised feature selection
Local geometric structure
Graph regularization
Similarity preservation

ABSTRACT

Recent research indicates the critical importance of preserving local geometric structure of data in unsupervised feature selection (UFS), and the well studied graph Laplacian is usually deployed to capture this property. By using a squared l_2-norm, we observe that conventional graph Laplacian is sensitive to noisy data, leading to unsatisfying data processing performance. To address this issue, we propose a unified UFS framework via feature self-representation and robust graph regularization, with the aim at reducing the sensitivity to outliers from the following two aspects: i) an $l_{2,1}$-norm is used to characterize the feature representation residual matrix; and ii) an l_1-norm based graph Laplacian regularization term is adopted to preserve the local geometric structure of data. By this way, the proposed framework is able to reduce the effect of noisy data on feature selection. Furthermore, the proposed l_1-norm based graph Laplacian is readily extendible, which can be easily integrated into other UFS methods and machine learning tasks with local geometrical structure of data being preserved. As demonstrated on ten challenging benchmark data sets, our algorithm significantly and consistently outperforms state-of-the-art UFS methods in the literature, suggesting the effectiveness of the proposed UFS framework.

1. Introduction

With the rapid development of data acquisition technology, large amounts of unlabelled and high-dimensional data need to be processed (Chang, Lijuan, Xiao, & Minhui, 2017; Dy, Brodley, Kak, Broderick, & Aisen, 2003; Guyon & Elisseeff, 2003; Javed, Sobral, Bouwmans, & Jung, 2015). These data often contain quite a lot of noisy features, which are detrimental to data processing. As one of the typical method to alleviate this problem, unsupervised feature selection (UFS) attracts more and more attentions. As an essential preprocessing step for improving the performance of data mining tasks such as clustering and classification (Song, HaiYan, & Jing, 2017; Uysal, 2016), UFS aims at selecting a discriminative subset of features from unlabelled high-dimensional data and removing other noisy and unimportant features.

Various methods of UFS have been proposed, and these methods can be generally classified into three categories: filter methods (He, Cai, & Niyogi, 2005; Zhao & Liu, 2007), wrapper methods (Dadaneh, Markid, & Zakerolhosseini, 2016; Kohavi & John, 1997; Maldonado & Weber, 2009; Tabakhi, Moradi, & Akhlaghian, 2014) and embedded methods (Cai, Zhang, & He, 2010; Hou, Nie, Li, & Yi, 2014; Li, Liu, Yang, Zhou, & Lu, 2014; Nie, Wei, & Li, 2016; Wang, Tang, & Liu, 2015; Wang, Liu, Nie, & Huang, 2015; Li & Tang, 2015; Zhao, Wang, Liu et al., 2010; Zhou, Xu, Cheng, Fang, & Pedrycz, 2016; Zhu, Zhu, Hu, Zhang, & Zuo, 2017). Embedded methods are superior to others in many respects, and have received more and more attentions. The recent literature indicates that preserving global pairwise sample similarity and local geometric structure of data is of great importance for feature selection, and furthermore, preserving local geometric data structure becomes clearly more important than preserving global pairwise sample similarity for UFS (Liu, Wang, Zhang, Yin, & Liu, 2014; Nie et al., 2016; Wang, Tang et al., 2015; Zhou et al., 2016; Zhu, Hu, Zhang, & Zuo, 2016). As a well studied model, the graph Laplacian is usually deployed to capture local geometric data structure in UFS and other machine learning tasks.

Though demonstrating promising performance in various applications, we have observed that there are at least two issues for previous graph regularized UFS methods. First, by using a squared l_2-norm, we observe that the conventional graph Laplacian is sensitive to noisy data. Since real world data usually contain lots of

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding authors.
E-mail addresses: tangchang@cug.edu.cn (C. Tang), zxz@zjnu.edu.cn (X. Zhu), jjiachen@outlook.com (J. Chen), pw212@uowmail.edu.au (P. Wang), xinwangliu@nudt.edu.cn (X. Liu), tian@ieee.org (J. Tian).

[1] Chang Tang and Xinzhong Zhu contributed equally as first author to this work.

https://doi.org/10.1016/j.eswa.2017.11.053
0957-4174/

noisy samples and features, the l_2-norm based graph Laplacian would lead to unsatisfying data processing performance. Second, the Frobenius norm has been widely used in many methods to regularize the feature representation term (Liu et al., 2014; Wang, Tang et al., 2015; Zhou et al., 2016), which makes traditional methods sensitive to data outliers, resulting in unsatisfactory feature selection performance.

In order to address the above mentioned two issues, we propose a unified UFS framework via feature self-representation and robust graph regularization, and the proposed framework is robust to noisy data and outliers. Clustering performance with the selected features on ten challenging benchmark data sets demonstrates the effectiveness of the proposed framework. In summary, the main contributions of this paper are highlighted as follows:

- We propose a unified robust graph regularized UFS model in which both global and local structure of data can be well preserved.
- An $l_{2,1}$-norm and l_1-norm are deployed to regularize the feature self-representation term and graph Laplacian term, respectively, which make our model more robust to noisy features and outliers.
- Our proposed l_1-norm based graph Laplacian is readily extendible and we develop an efficient solver for the optimization problem based on the Alternating Direction Method of Multipliers (ADMM) (Boyd, Parikh, Chu, Peleato, & Eckstein, 2011).
- Comprehensive experiments on ten benchmark data sets are conducted to show the effectiveness of the proposed model, and demonstrate its advantage over other state-of-the-art methods.

The rest of this paper is arranged as follows. Some related UFS works are introduced in Section 2. In Section 3, we first propose a robust UFS algorithm based on traditional l_2-norm graph Laplacian regularization, i.e., l_2-norm based graph regularized UFS (l_2-UFS). Considering that the l_2-norm is susceptible to noisy data samples and features, we then introduce an l_1-norm based graph Laplacian regularization term for UFS, i.e., l_1-norm based graph regularized UFS (l_1-UFS), which is more robust against noisy features. Section 4 presents the optimization algorithm for solving the proposed model, and the convergence and computational analysis are also provided in this section. Section 5 explains the connections of our proposed method with previous works. Experimental results and parameters sensitivity analysis are shown in Section 6. Finally, we conclude this paper in Section 7.

2. Related works

Since unlabelled data can be easily captured while obtaining the labels of data instances is typically expensive and time-consuming, it is quite promising and demanding to develop UFS techniques to improve the performance of machine learning and data mining tasks. In UFS, features are selected based on some criterion which evaluates features by their capability of keeping certain properties of original data, such as data distribution, the redundancy of features or local structure. Due to the absence of class label information, structure learning which serves as a pseudo supervised information has been widely used in UFS, and the importance of preserving local geometric structure of data for UFS has been verified in previous research (Liu et al., 2014).

As a commonly used model, the graph Laplacian is usually deployed to capture local geometric data structure in UFS and many other machine learning tasks. Liu et al. (2014) proposed a unified graph regularized UFS framework, their model can be used for supervised, semi-supervised and unsupervised feature selection. Zhou et al. (2016) presented a global and local structure preserving sparse subspace learning model for unsupervised feature selection. In the model, feature selection and subspace learning are realized simultaneously, and the local data structure is also preserved by traditional l_2-norm Laplacian graph. Wang and Wang (2017) combined low-rank approximation and structure learning for UFS. In Shang, Zhang, Jiao, Liu, and Li (2016), by using a dual-graph regularization, the local geometrical information of both data space and feature space are preserved simultaneously. In order to adaptively determine the similarity graph matrix, Nie et al. (2016) proposed an unsupervised feature selection approach which performs feature selection and local structure learning simultaneously. Experimental results of previous UFS works demonstrate the efficacy of the graph regularizer for local geometrical structure preservation.

It should be noted that in previous graph regularized UFS works, traditional l_2-norm are used to measure the similarity between data instances. However, the l_2-norm can be easily dominated by noisy features. Thus the constructed graph based on l_2-norm is also suboptimal which does not reflect necessarily the inherent geometrical structure of the data distribution. This thus has an adverse effect on the feature representation learning. In order to make the graph regularizer more robust to noisy features, we propose an l_1-norm based graph to preserve the local structure of data. In addition, for robustness to outlier data instances, the $l_{2,1}$-norm is used to regularize the feature self-representation term, instead of using the Frobenius norm as previous methods done. We also develop an efficient solver with convergence guarantee for the optimization problem based on the ADMM (Boyd et al., 2011).

3. Proposed framework

3.1. Notations

Throughout this paper, matrices are written as boldface capital letters and vectors are denoted as boldface lower case letters. For an arbitrary matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, $\mathbf{M}_{ij}$ denotes its (i, j)th entry, $\mathbf{m}_i$ and $\mathbf{m}^j$ denotes the ith row and jth column of $\mathbf{M}$, respectively. $\mathrm{Tr}(\mathbf{M})$ is the trace of $\mathbf{M}$ if $\mathbf{M}$ is square and $\mathbf{M}^T$ is the transpose of $\mathbf{M}$. $\langle \mathbf{A}, \mathbf{B}\rangle$ is the standard inner product between two matrices. $\mathbf{I}_m$ is the identity matrix with size $m \times m$ (denoted by $\mathbf{I}$ if the size is obviously known). The $l_{2,1}$-norm of matrix $\mathbf{M}$ is defined as $||\mathbf{M}||_{2,1} = \sum_{i=1}^{m} ||\mathbf{m}_i||_2 = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} \mathbf{M}_{ij}{}^2}$. $||\mathbf{M}||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{M}_{ij}{}^2}$ is the well-known Frobenius norm of $\mathbf{M}$.

3.2. Self-representation based UFS

Self-similarity widely exists in real world, which can be understood as a part of an object is similar to other parts of itself, e.g., trees and leaves (Eloy, 2011), coastlines (Mandelbrot, 1967), and images (Buades, Coll, & Morel, 2005). Self-similarity has also been successfully used in image data processing, such as image restoration and denoising (Buades et al., 2005), super- resolution(Freedman & Fattal, 2011; Yang, Huang, & Yang, 2010). Self-representation, as a specific form of self-similarity used for data processing, has been extensively used in machine learning and data mining fields. e.g., the low rank representation model (LRR) can be regarded as a self-representation model where the data matrix is used as the bases matrix. In LRR, each sample can be represented as a linear combination of other samples.

Self-representation has also been used for UFS. Zhu et al. proposed a regularized self-representation (RSR) model for unsupervised feature selection (Zhu, Zuo, Zhang, Hu, & Shiu, 2015). In RSR, $\mathbf{X} \in \mathcal{R}^{n\times d}$ represents the data matrix, where n and d are the numbers of samples and features, respectively. Each row of $\mathbf{X}$ represents a sample, and each column of $\mathbf{X}$ represents one feature dimension. The feature selection problem is formulated as a multi-

output regression problem:

$$\min_{\mathbf{W}} ||\mathbf{X} - \mathbf{XW}||_{2,1} + \lambda||\mathbf{W}||_{2,1}, \tag{1}$$

where $\mathbf{W}$ is the feature weight matrix. The $l_{2,1}$-norm imposes row-sparsity on the reconstruction residual for robustness to outliers (i.e., the first term of Eq. (1)), the second term is used to guide the selection of feature subset. λ is a positive constant which balances the two terms. Although the RSR model can effectively identify the most representative features, it ignores the local geometric structure of data. The importance of preserving local geometric data structure has been well recognized in the recent literature on dimensionality reduction (Chen, Ma, & Liu, 2013; Gu, Li, & Han, 2011; Liu et al., 2014; Saul & Roweis, 2003). This motivates us to propose a robust graph regularized UFS framework in which the local geometric structure of data can be well preserved.

3.3. Robust graph regularized UFS

In this subsection, we propose a unified robust graph regularized UFS framework based on feature self-representation. Traditional l_2-norm based graph is first introduced for local geometric structure preservation. Considering that l_2-norm based graph is sensitive to noisy data, we propose a robust l_1-norm based graph regularization term.

The pairwise similarity graph is often used to capture the local geometric structure of data (He & Niyogi, 2005; Roweis & Saul, 2000; Zhang, Yang, Zhao, & Ge, 2007), and it also has been embedded into some UFS works (Liu et al., 2014; Wang, Tang et al., 2015). Inspired by its promising results, we propose a unified robust graph regularized UFS framework based on feature self-representation. The objective of our framework is as follows:

$$\min_{\mathbf{W}} ||\mathbf{X} - \mathbf{XW}||_{2,1} + \lambda||\mathbf{W}||_{2,1} + \beta\mathcal{G}(\mathbf{W}), \tag{2}$$

where the first term is used to regularize the feature reconstruction error and capture the global structure of data, and the second term is used to regularize the feature reconstruction coefficient matrix, and the third term $\mathcal{G}(\mathbf{W})$ is the graph regularization term which is used to capture the local geometric structure of data; $\lambda > 0$ and $\beta > 0$ are two parameters for balancing the $l_{2,1}$ norm and graph regularization terms.

3.3.1. l_2-UFS

Traditional l_2-norm based graph such as local linear embedding (LLE) (Roweis & Saul, 2000), linear preserve projection (LPP) (He & Niyogi, 2005), and local tangent space alignment (LTSA) (Zhang et al., 2007) can be employed to modelling $\mathcal{G}(\mathbf{W})$. Without loss of generality, we use the LPP to formulate our l_2-norm based graph regularization term.

In LPP, a similarity matrix $\mathbf{S}$ of data points is computed according to the following equation:

$$\mathbf{S}_{ij} = \begin{cases} \exp\left(\dfrac{||\mathbf{x}_i - \mathbf{x}_j||^2}{-2\sigma^2}\right), & \mathbf{x}_i \in \mathcal{N}_n(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_n(\mathbf{x}_i); \\ 0, & otherwise, \end{cases} \tag{3}$$

where $\mathcal{N}_n(\mathbf{x}_i)$ denotes the set of n nearest neighbours of $\mathbf{x}_i$ and σ is a width parameter. LPP optimizes a linear transformation $\mathbf{W}$ by

$$\min_{\mathbf{W}} \sum_{i,j=1}^{n} ||\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j||^2\mathbf{S}_{ij}. \tag{4}$$

Eq. (4) enforces that two similar sample points in original space should be also similar in the transformation space. Our l_2-norm based graph regularized UFS model (l_2-UFS) can be formulated as:

$$\min_{\mathbf{W}} ||\mathbf{X} - \mathbf{XW}||_{2,1} + \lambda||\mathbf{W}||_{2,1} + \beta \sum_{i,j=1}^{n} ||\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j||^2\mathbf{S}_{ij}. \tag{5}$$

By some simple algebra, Eq. (5) can be rewritten as the following compact form

$$\min_{\mathbf{W}} ||\mathbf{X} - \mathbf{XW}||_{2,1} + \lambda||\mathbf{W}||_{2,1} + \beta Tr\left(\mathbf{W}^T\mathbf{X}^T\mathbf{LXW}\right), \tag{6}$$

where $\mathbf{L}$ is the Laplacian matrix and $\mathbf{L} = \mathbf{D} - \mathbf{S}$. $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_{j=1}^{n}\mathbf{S}_{ij}$, and $\mathbf{S}$ is computed by Eq. (3). In next section, we will derive an iterative reweighted least-squares (IRLS) algorithm to solve Eq. (6) and its convergence analysis is also discussed.

3.3.2. l_1-UFS

In Eq. (5), the last graph regularization term is used to preserve the local geometric structure of data. However, it is well known that the least squares based regularisation function can be easily dominated by noisy data samples. To handle this problem, we propose an l_1-norm based graph regularization, which makes the regularisation term more robust against noisy data samples and outliers.

Due to the fact that the Laplacian matrix $\mathbf{L}$ in Eq. (6) is real symmetric, it can be decomposed into the following form using the eigen decomposition technique:

$$\mathbf{L} = \mathbf{UVU}^T, \tag{7}$$

then the graph regularization term in Eq. (6) can be rewritten as

$$\begin{aligned} Tr\left(\mathbf{W}^T\mathbf{X}^T\mathbf{LXW}\right) &= Tr\left(\mathbf{W}^T\mathbf{X}^T\mathbf{UVU}^T\mathbf{XW}\right) \\ &= ||\mathbf{V}^{\frac{1}{2}}\mathbf{U}^T\mathbf{XW}||_F^2 = ||\mathbf{AW}||_F^2, \end{aligned} \tag{8}$$

where $\mathbf{A} = \mathbf{V}^{\frac{1}{2}}\mathbf{U}^T\mathbf{X}$. As can be seen, Eq. (8) is quadratic. To promote robustness to noisy data, we propose to use l_1-norm instead of Frobenius norm. That induces the proposed l_1-norm based graph regularisation term:

$$\mathcal{G}(\mathbf{W}) = ||\mathbf{AW}||_1. \tag{9}$$

Finally, our l_1-norm based graph regularized UFS model (l_1-UFS) can be formulated as:

$$\min_{\mathbf{W}} ||\mathbf{X} - \mathbf{XW}||_{2,1} + \lambda||\mathbf{W}||_{2,1} + \beta||\mathbf{AW}||_1. \tag{10}$$

The key advantage of using the l_1-norm based graph regularisation for UFS instead of the conventional l_2-norm based graph regularisation lies in its sparsity. It is well-known that l_1-norm has a shrinkage property thus promotes sparsity. Intuitively, there often exists noise and outliers in the reconstructed data matrix $\mathbf{W}^T\mathbf{X}^T$, the magnitude of $||\mathbf{AW}||_F^2$ becomes very large for those outlying data points, and leads to the whole objective function being dominated by the noise and outliers. In contrast, $||\mathbf{AW}||_1$ becomes sparse due to the use of l_1-norm, consequently suppressing the impact of outliers and noises (Kodirov, Xiang, Fu, & Gong, 2016).

The problem in Eq. (10) is convex but non-smooth. Due to the l_1-norm in the third term, solving it is thus more difficult than Eq. (6). In next section, we develop an efficient solver for Eq. (10) based on the Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011).

4. Optimization and algorithms

In this section, we will give the optimization procedure of Eqs. (6) and (10) in detail. Then the optimized $\mathbf{W}$ can be use for selecting representative features.

4.1. Solver of l_2-UFS

Eq. (6) is convex but not smooth, we derive an iterative reweighted least-squares (IRLS) algorithm to solve it. For the IRLS algorithm, supposing the current estimation of $\mathbf{W}$ is $\mathbf{W}^t$, we define two diagonal weighting matrices $\mathbf{G}_1^t$ and $\mathbf{G}_2^t$ by $g_{1,i}^t = 1/2||\mathbf{x}_i - \mathbf{x}_i\mathbf{W}^t||_2$ and $g_{2,i}^t = 1/2||\mathbf{w}_i^t||_2$, and then $\mathbf{W}^{t+1}$ is updated by solving the following weighted least squares problem:

$$\begin{aligned}\mathbf{W}^{t+1} &= \arg\min_{\mathbf{W}} Q\left(\mathbf{W}|\mathbf{W}^t\right) \\ &= \arg\min_{\mathbf{W}} \begin{Bmatrix} Tr\left((\mathbf{X}-\mathbf{X}\mathbf{W})^T\mathbf{G}_1^t(\mathbf{X}-\mathbf{X}\mathbf{W})\right) \\ +\lambda Tr\left(\mathbf{W}^T\mathbf{G}_2^t\mathbf{W}\right)+\beta Tr\left(\mathbf{W}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{W}\right)\end{Bmatrix}.\end{aligned} \tag{11}$$

Taking the derivative of Eq. (11) with respect to $\mathbf{W}$ and setting it to zero, we have the closed form solution of $\mathbf{W}^{t+1}$ as follows:

$$\mathbf{W}^{t+1} = \begin{Bmatrix} \left((\mathbf{G}_2^t)^{-1}\mathbf{X}^T\mathbf{G}_1^t\mathbf{X}+\lambda\mathbf{I}+\beta(\mathbf{G}_2^t)^{-1}\mathbf{X}^T\mathbf{L}\mathbf{X}\right)^{-1} \\ \cdot(\mathbf{G}_2^t)^{-1}\mathbf{X}^T\mathbf{G}_1^t\mathbf{X}\end{Bmatrix}. \tag{12}$$

In order to avoid the overflow error during the iteration process, a sufficiently small value ε is introduced by defining $g_{1,i}^t = 1/\max\left(2||\mathbf{x}_i - \mathbf{x}_i\mathbf{W}^t||_2, \varepsilon\right)$ and $g_{2,i}^t = 1/\max\left(2||\mathbf{w}_i^t||_2, \varepsilon\right)$. The IRLS algorithm for solving l_2-UFS is summarized in Algorithm 1. As for the stopping criterion for Algorithm 1, we utilize the loss function in Eq. (6). When the loss variation ratio is below 10^{-6}, we stop the iteration. Empirically, we set the maximum iteration number as 100.

Algorithm 1 IRLS algorithm for solving l_2-UFS.

Input: Data matrix $\mathbf{X} \in R^{n\times d}$, Laplacian matrix $\mathbf{L}$, parameters λ and β.
Initialization: $t = 0$, $\mathbf{G}_1^t = I$, $\mathbf{G}_2^t = \mathbf{I}$;
while not converged do
1. Update $\mathbf{W}$ using Eq. (12);
2. Update $\mathbf{G}_1^t$ by $g_{1,i}^t = 1/2||\mathbf{x}_i - \mathbf{x}_i\mathbf{W}^t||_2$,
Update $\mathbf{G}_2^t$ by $g_{2,i}^t = 1/2||\mathbf{w}_i^t||_2$;
3. Update t by: $t = t + 1$;
end while
Output: W.
Feature selection: Sort each feature of $\mathbf{X}$ according to $||\mathbf{w}_i||_2, (i = 1, 2, \cdots, d)$ in descending order and select the top-n ranked ones.

4.2. Solver of l_1-UFS

Due to the three non-smooth terms in Eq. (10), we resort to the Alternating Direction Method of Multipliers (ADMM) for solving it.

First, we introduce an auxiliary variable $\mathbf{Y} = \mathbf{A}\mathbf{W}$ to make the objective function separable, then the Augmented Lagrangian function of Eq. (10) is:

$$\begin{aligned}\mathcal{L}(\mathbf{W}, \mathbf{Y}, \mathbf{F}, \mu) &= ||\mathbf{X} - \mathbf{X}\mathbf{W}||_{2,1} + \lambda||\mathbf{W}||_{2,1} + \beta||\mathbf{Y}||_1 \\ &\quad + \langle\mathbf{F}, \mathbf{Y} - \mathbf{A}\mathbf{W}\rangle + \frac{\mu}{2}||\mathbf{Y} - \mathbf{A}\mathbf{W}||_F^2 \\ &= ||\mathbf{X} - \mathbf{X}\mathbf{W}||_{2,1} + \lambda||\mathbf{W}||_{2,1} + \beta||\mathbf{Y}||_1 \\ &\quad + \frac{\mu}{2}||\mathbf{Y} - \mathbf{A}\mathbf{W} + \frac{1}{\mu}\mathbf{F}||_F^2 - \frac{1}{2\mu}||\mathbf{F}||_F^2,\end{aligned} \tag{13}$$

where $\mathbf{F}$ is the Lagrange multiplier, and $\mu > 0$ controls the penalty for violating the linear constraints. Now, we can update $\mathbf{W}$ and $\mathbf{Y}$ in an alternating manner.

W-subproblem: Fix $\mathbf{Y}$, $\mathbf{F}$ and update $\mathbf{W}$. We need to solve the following problem:

$$\arg\min_{\mathbf{W}} ||\mathbf{X} - \mathbf{X}\mathbf{W}||_{2,1} + \lambda||\mathbf{W}||_{2,1} + \frac{\mu}{2}||\mathbf{Y} - \mathbf{A}\mathbf{W} + \frac{1}{\mu}\mathbf{F}||_F^2. \tag{14}$$

Similar to the solver of l_2-UFS in Section 4.1, the optimization of $\mathbf{W}$ can be obtained by the IRLS algorithm. We define two similar diagonal weighting matrices as Eq. (11) and rewrite problem in Eq. (14) in the trace form, then $\mathbf{W}^{t+1}$ is updated by solving the following weighted least squares problem:

$$\mathbf{W}^{t+1} = \arg\min_{\mathbf{W}} \begin{Bmatrix} Tr\left((\mathbf{X}-\mathbf{X}\mathbf{W})^T\mathbf{G}_1^t(\mathbf{X}-\mathbf{X}\mathbf{W})\right)+\lambda Tr\left(\mathbf{W}^T\mathbf{G}_2^t\mathbf{W}\right) \\ +\frac{\mu}{2}Tr\left((\mathbf{Y}-\mathbf{A}\mathbf{W}+\frac{1}{\mu}\mathbf{F})^T(\mathbf{Y}-\mathbf{A}\mathbf{W}+\frac{1}{\mu}\mathbf{F})\right)\end{Bmatrix}. \tag{15}$$

Taking the derivative of Eq. (15) with respect to $\mathbf{W}$ and setting it to zero, we have the closed form solution of $\mathbf{W}^{t+1}$ as follows:

$$\begin{aligned}\mathbf{W}^{t+1} &= \left(2\left(\mathbf{G}_2^t\right)^{-1}\mathbf{X}^T\mathbf{G}_1^t\mathbf{X} + 2\lambda\mathbf{I} + \mu\left(\mathbf{G}_2^t\right)^{-1}\mathbf{A}^T\mathbf{A}\right)^{-1} \\ &\quad \cdot\left(\mathbf{G}_2^t\right)^{-1}\left(2\mathbf{X}^T\mathbf{G}_1^t\mathbf{X} + \mu\mathbf{A}^T\mathbf{Y} + \mathbf{A}^T\mathbf{F}\right).\end{aligned} \tag{16}$$

Y-subproblem: Fix $\mathbf{W}$, $\mathbf{F}$ and update $\mathbf{Y}$. We solve the following objective function:

$$\arg\min_{\mathbf{Y}} \beta||\mathbf{Y}||_1 + \frac{\mu}{2}||\mathbf{Y} - \mathbf{A}\mathbf{W} + \frac{1}{\mu}\mathbf{F}||_F^2. \tag{17}$$

$\mathbf{Y}$ can be solved by using the soft-thresholding operator:

$$\mathbf{Y} = sign\left(\mathbf{A}\mathbf{W} - \frac{1}{\mu}\mathbf{F}\right)\max\left(|\mathbf{A}\mathbf{W} - \frac{1}{\mu}\mathbf{F}| - \frac{\beta}{\mu}\right). \tag{18}$$

The optimization process of solving l_1-UFS by using ADMM is summarized in Algorithm 2. Similar to Algorithm 1, we utilize the loss function in Eq. (10). When the loss variation ratio is below 10^{-6}, we break the iteration. We also empirically set the maximum iteration number as 100.

Algorithm 2 Optimization process of solving l_1-UFS by using ADMM.

Input: Data matrix $\mathbf{X} \in R^{n\times d}$, Laplacian matrix $\mathbf{L}$, parameters λ and β.
Initialization: $k = 0$, $t = 0$, $\mathbf{G}_1^t = I$, $\mathbf{G}_2^t = \mathbf{I}$, $\mathbf{F}_0 = \mathbf{0}$, $\mathbf{Y}_0 = \mathbf{0}$, $\mu_0 = 0.1$, $\mu_{\max} = 10^{10}$, $\rho = 1.1$;
while not converged do
  while not converged do
    1.1. Update $\mathbf{W}_{k+1}$ using Eq. (16);
    1.2. Update $\mathbf{G}_1^t$ and $\mathbf{G}_2^t$;
    1.3. Update t by: $t = t + 1$;
  end while
2. Update $\mathbf{Y}_{k+1}$ using Eq. (18);
3. Update $\mathbf{F}_{k+1}$: $\mathbf{F}_{k+1} = \mathbf{F}_{k+1} + \mu_{k+1}\left(\mathbf{A}\mathbf{W}_{k+1} - \mathbf{Y}_{k+1}\right)$;
4. Update μ_{k+1}: $\mu_{k+1} = \min\left(\rho\mu_k, \mu_{\max}\right)$;
5. Update k by: $k = k + 1$;
end while
Output: W and **Y**.
Feature selection: Sort each feature of $\mathbf{X}$ according to $||\mathbf{w}_i||_2, (i = 1, 2, \cdots, d)$ in descending order and select the top-n ranked ones.

4.3. Convergence and computational complexity analysis

4.3.1. Convergence analysis

In order to analyse the convergence of Algorithm 1, we need to use the concept of surrogate function and give its definition as follows (Hunter & Lange, 2000).

Definition 1. Supposing there is a loss function $L(\theta)$ need to be minimized, we transfer its minimization of the observed data to another function $P(\theta|\theta^t)$ depending on the current iterate θ^t through the complete data. The key ingredient in making this

transfer successful is the fact that $L(\theta) - P(\theta|\theta^t)$ attains its maximum at $\theta = \theta^t$ (t denotes the tth iteration), then $P(\theta|\theta^t)$ is called the surrogate function of $L(\theta)$.

Then, we let

$$H(\mathbf{W}) = ||\mathbf{X} - \mathbf{XW}||_{2,1} + \lambda||\mathbf{W}||_{2,1} + \beta Tr\left(\mathbf{W}^T\mathbf{X}^T\mathbf{LXW}\right) \quad (19)$$

and

$$J(\mathbf{W}) = H(\mathbf{W}) - Q\left(\mathbf{W}|\mathbf{W}^t\right). \quad (20)$$

Then we can get the following theorem:

Theorem 1. *$Q\left(\mathbf{W}|\mathbf{W}^t\right)$ is a surrogate function, i.e., $J(\mathbf{W})$ attains its maximum when $\mathbf{W} = \mathbf{W}^t$.*

Proof. From Eq. (20), we have

$$\begin{aligned} J\left(\mathbf{W}^t\right) &= H\left(\mathbf{W}^t\right) - Q\left(\mathbf{W}^t|\mathbf{W}^t\right) \\ &= \sum_i ||\mathbf{x}_i - \mathbf{x}_i\mathbf{W}^t||_2 + \lambda\sum_j ||\mathbf{W}_j^t||_2 \\ &\quad - \left(\sum_i \frac{||\mathbf{x}_i - \mathbf{x}_i\mathbf{W}^t||_2^2}{2||\mathbf{x}_i - \mathbf{x}_i\mathbf{W}^t||_2} + \lambda\sum_i \frac{||\mathbf{W}_j^t||_2^2}{2||\mathbf{W}_j^t||_2}\right) \\ &= \frac{1}{2}\left(\sum_i ||\mathbf{x}_i - \mathbf{x}_i\mathbf{W}^t||_2 + \lambda\sum_j ||\mathbf{W}_j^t||_2\right). \end{aligned}$$

Then, one can easily obtains that $\mathbf{W}$, $J(\mathbf{W}^t) - J(\mathbf{W}) \geq 0$ and $Q(\mathbf{W}|\mathbf{W}^t)$ is a surrogate function. □

The loss function $H(\mathbf{W})$ can be minimized by iteratively minimizing the surrogate function $Q(\mathbf{W}|\mathbf{W}^t)$, and we can obtain the following theorem:

Theorem 2. *Let $\mathbf{W}^{t+1} = \arg\min_{\mathbf{W}} Q(\mathbf{W}|\mathbf{W}^t)$, we have $H(\mathbf{W}^{t+1}) \leq H(\mathbf{W}^t)$.*

Proof. $H(\mathbf{W}^{t+1}) = H(\mathbf{W}^{t+1}) - Q(\mathbf{W}^{t+1}|\mathbf{W}^t) + Q(\mathbf{W}^{t+1}|\mathbf{W}^t) \leq H(\mathbf{W}^t) - Q(\mathbf{W}^t|\mathbf{W}^t) + Q(\mathbf{W}^{t+1}|\mathbf{W}^t) \leq H(\mathbf{W}^t) - Q(\mathbf{W}^t|\mathbf{W}^t) + Q(\mathbf{W}^t|\mathbf{W}^t) = H(\mathbf{W}^t)$. □

For the convergence analysis of Algorithm 2, although the theoretical convergence proof of ADMM for global optimum does not exist, we can prove that Algorithm 2 can converge to a local stable point in a similar way.

4.3.2. Computational complexity analysis

In l_2-UFS, we mainly need to update $\mathbf{W}$ in each iteration, in which the computational complexity is basically $\mathcal{O}(d^3 + d^2n)$, where d and n are the number of features and samples, respectively. Hence, the time complexity of l_2-UFS is $\mathcal{O}(T(d^3 + d^2n))$, where T is the total number of iterations. By some simple algebra, we have

$$\left(\mathbf{X}^T\left(\mathbf{G}_1^t + \beta\mathbf{L}\right)\mathbf{X} + \lambda\mathbf{G}_2^t\right)^{-1}\mathbf{X}^T = \Lambda\mathbf{X}^T\left(\left(\mathbf{G}_1^t + \beta\mathbf{L}\right)\mathbf{X}\Lambda\mathbf{X}^T + \mathbf{I}\right)^{-1} \quad (21)$$

where $\Lambda = \frac{1}{\lambda}(\mathbf{G}_2^t)^{-1}$, and $\mathbf{I}$ is a $n \times n$ identity matrix. Therefore, we can convert a $d \times d$ matrix inverse problem to an $n \times n$ one. Then, the time complexity for solving $\mathbf{W}$ at each iteration becomes $\mathcal{O}(\min\{n, d\}^3)$.

In l_1-UFS, the optimization process is divided into two subproblems. We first need to perform eigen-decomposition of the Laplacian matrix $\mathbf{L}$, which has a complexity of $\mathcal{O}(n^3)$. Fortunately, since the Laplacian matrix is sparse, the complexity can be reduced to $\mathcal{O}(rn^2)$, where r is the ratio of nonzero entries in $\mathbf{L}$. At the step of estimation of $\mathbf{W}$, the time complexity is the same as l_2-UFS (i.e., $\mathcal{O}(T_1(\min\{n, d\}^3))$), where T_1 is the total number of iterations for solving $\mathbf{W}$. The remaining parts have linear complexity $\mathcal{O}(n)$. In the experiments section, we will give the running time comparison of different methods on different datasets.

Table 1
Statistics of the Datasets.

Datasets	Instance	Feature	Class	Type
Yale	165	1024	15	Face images
ORL	400	1024	40	Face images
orlraws10P	100	10304	10	Face images
warpPIE10P	210	2420	10	Face images
warpAR10P	130	2400	10	Face images
Isolet	1560	617	26	Speech Signal
Prostate_GE	102	5966	2	Biomedical MicroArray data
CLL_SUB_111	111	11340	3	Biomedical MicroArray data
USPS	9298	256	10	Digit images
COIL20	1440	1024	20	Digit images

5. Connections with previous methods

The most related previous works are global and local structure preservation for feature selection (GLSPFS) (Liu et al., 2014), the regularized self-representation feature selection (RSR) (Zhu et al., 2015), and the global and local structure preserving sparse subspace learning model for unsupervised feature selection (GLoss) (Zhou et al., 2016). Following we give a detailed analysis about the connections between our proposed methods and these works.

5.1. Connection with GLSPFS

Liu et al. (2014) proposed a global and local structure preservation framework for feature selection which integrates both global pairwise sample similarity and local geometric data structure to conduct feature selection. Their method can be used for both unsupervised, semi-supervised and supervised feature selection. However, in their model, they use the Frobenius norm and traditional l_2-norm based graph to regularize the feature reconstruction term and local geometrical structure preservation term, respectively. It is well known that the Frobenius norm and traditional l_2-norm are sensitive to noisy features and outlier samples. In our models, we use $l_{2,1}$ norm and l_1 norm based graph instead to enhance the robustness of feature selection, as described by Eqs. (6) and (10).

5.2. Connection with RSR

In RSR, the $l_{2,1}$ norm is used to regularize both the feature reconstruction term and feature representation coefficient matrix, but this model ignores the local structure preservation which is critical for UFS. Our models use the Laplacian graph regularization term for preserving the local structure of data. Therefore, the RSR model can be regarded as a special case of our framework (when $\beta = 0$ in Eq. (2)).

5.3. Connection with GLoss

In GLoss, the features are reconstructed from the learned subspace spanned by the selected features. The model also uses the Frobenius norm and traditional l_2-norm based graph to regularize the feature reconstruction term and local geometrical structure preservation term, respectively. The experimental results in next section will show that our proposed l_1-UFS preforms better than GLoss, which verify that l_1-UFS is more robust than GLoss.

6. Experiments

In this section, the evaluation of our l_2-UFS and l_1-UFS models on several public data sets is presented. Meanwhile, the proposed

Table 2
Clustering results (ACC, NMI and ARI) of different feature selection algorithms on different datasets.

(a) ACC% ± std%

Datasets	Baseline	LS	MCFS	RSR	GLoSS	SGOFS	l_2-UFS	$l_{1,F}$-UFS	l_1-UFS
Yale	41.12±3.65	40.54±2.31	44.02±3.31	39.51±1.71	37.90±1.51	46.21±3.71	39.87±1.69	40.67±2.14	48.19±2.92
ORL	56.09±2.28	48.94±2.65	50.34±2.51	54.18±2.78	54.48±1.38	57.68±2.28	56.40±2.28	56.85±2.34	58.24±2.38
orlraws10P	72.86±6.37	61.84±4.32	75.94±7.64	75.37±6.48	76.84±5.36	75.75±5.84	75.89±6.23	76.65±5.35	77.84±5.34
warpPIE10P	27.34±1.67	42.59±2.16	38.83±3.22	34.68±2.34	41.83±2.25	41.22±3.73	36.73±2.14	41.42±2.75	46.64±3.44
warpAR10P	22.23±2.43	46.76±4.38	25.45±2.74	33.25±1.74	35.26±3.33	39.25±2.74	34.67±1.37	39.76±2.25	41.77±2.74
Isolet	58.43±3.76	53.78±2.44	56.65±3.73	53.92±2.73	61.45±3.67	64.36±3.68	58.94±3.37	58.57±2.68	60.84±2.37
Prostate_GE	58.66±2.46	55.94±1.67	59.93±2.84	60.53±1.88	60.83±2.17	63.10±2.75	60.91±1.38	61.27± 1.54	63.82±1.69
CLL_SUB_111	53.34±3.49	45.63±6.62	46.21±6.71	45.89±6.42	47.06±2.36	65.45±1.49	47.65±6.38	60.34± 1.76	65.77±1.81
USPS	66.47±2.14	62.67±2.03	72.54±2.79	72.39±4.61	67.67±3.91	65.62±2.87	70.39±4.81	72.91± 1.51	73.92±1.58
COIL20	58.45±5.61	49.34±3.67	59.28±3.39	60.98±3.68	58.17±1.84	61.84±2.37	60.86±5.53	61.75±5.29	63.55±5.47

(b) NMI% ± std%

Datasets	Baseline	LS	MCFS	RSR	GLoSS	SGOFS	l_2-UFS	$l_{1,F}$-UFS	l_1-UFS
Yale	48.81±2.25	48.52±1.87	49.79±2.86	45.48±1.82	47.28±1.43	53.32±3.27	47.76±1.35	46.93±1.68	57.30±2.31
ORL	72.55±1.76	71.67±1.22	72.83±1.56	72.43±1.49	73.08±1.15	76.41±1.69	71.06±1.22	75.26±1.62	77.87±1.71
orlraws10P	79.94±3.78	70.52±3.31	85.83±4.72	82.88±3.58	83.63±4.48	80.95±3.54	83.51±4.53	84.16±4.76	86.42±4.34
warpPIE10P	27.94±3.73	47.56±2.32	39.78±2.45	37.61±2.89	43.88±1.48	44.84±2.63	38.62±3.17	44.64±2.64	53.21±3.37
warpAR10P	18.22±3.83	48.42±3.42	22.31±2.79	28.56±1.61	32.67±3.35	43.61±2.52	37.43±2.73	39.81±2.10	44.54±2.92
Isolet	74.67±1.37	68.63±1.08	70.26±1.37	69.67±1.76	73.27±1.91	74.76±1.59	72.83±1.67	73.53±1.34	74.71±1.00
Prostate_GE	2.17±0.05	2.76±0.98	2.27±0.78	3.57±0.84	3.86±1.06	6.71±1.97	4.66±0.16	4.46±1.13	8.78±0.73
CLL_SUB_111	18.27±4.56	11.87±9.87	10.19±5.17	9.80±5.58	10.73±5.60	31.89±6.24	11.18±5.29	14.62±6.07	20.63±5.62
USPS	61.63±1.62	59.47±1.60	66.56±0.44	66.16±1.77	61.36±1.82	60.32±1.63	67.53±2.02	67.77± 1.91	68.37±1.01
COIL20	74.31±2.97	64.78±1.69	70.78±1.35	72.73±1.39	66.27±1.83	70.76±1.85	69.11±1.62	73.72±1.71	75.84±1.39

(c) ARI% ± std%

Datasets	Baseline	LS	MCFS	RSR	GLoSS	SGOFS	l_2-UFS	$l_{1,F}$-UFS	l_1-UFS
Yale	22.63±3.41	20.66±3.51	22.21±3.39	19.16±2.92	18.01±2.87	24.34±3.27	20.18±2.94	21.34±2.82	26.04±2.67
ORL	43.02±2.57	32.67±2.27	34.72±2.71	38.06±2.80	39.14±2.64	46.07±2.91	41.17±2.91	42.46±2.71	47.84±2.63
orlraws10P	61.82±8.21	47.91±6.19	67.48±10.41	62.19±7.04	64.43±7.16	62.21±6.82	63.81±7.64	65.91±6.89	68.01±6.57
warpPIE10P	6.10±1.81	24.24±3.20	16.24±1.76	13.85±1.90	17.67±2.64	17.27±1.63	15.73±1.74	16.79±1.68	25.34±1.61
warpAR10P	1.86±2.27	26.68±3.35	4.68±2.16	10.49±2.47	13.25±2.38	18.46±2.91	14.52±2.37	16.32±2.67	17.16±2.59
Isolet	54.38±2.78	43.85±1.94	47.85±1.74	43.10±1.92	49.50±1.34	56.05±1.07	47.63±1.26	49.80±1.18	49.96±1.04
Prostate_GE	2.24±0.02	0.29±0.65	2.98±0.01	3.09±0.46	5.62±1.01	7.61±0.73	4.51±0.68	6.31±0.91	8.25±0.84
CLL_SUB_111	8.73±0.00	2.81±6.71	4.23±1.73	5.37±2.17	6.75±5.16	10.42±2.64	6.88±2.75	9.00±2.57	11.20±2.63
USPS	52.55±2.64	48.42±2.49	55.14±3.17	56.28±2.76	54.09±3.11	50.60±2.49	56.97±3.26	57.11±3.36	59.37±3.22
COIL20	54.49±5.12	41.93±2.34	53.14±3.51	55.91±2.17	56.50±2.64	58.27±1.78	56.69±2.15	57.29±2.07	60.03±2.77

models are compared with several state-of-the-art UFS methods, and the effect of the model parameters on the performance is reported. The implementation source code of the two models will be publicly released.[1]

6.1. Datasets

Ten different public available datasets were used for evaluation, including one speech signal dataset Isolet[2] (Dietterich & Bakiri, 1991), two biomedical MicroArray datasets Prostate_GE (Rd et al., 2002) and CLL_SUB_111, two digit image datasets USPS and COIL20, and five face image datasets. i.e. Yale, ORL, orlraws10P, warpPIE10P and warpAR10P[3] (Cai et al., 2010). The statistics of these datasets are summarized in Table 1.

6.2. Comparison methods and parameters setting

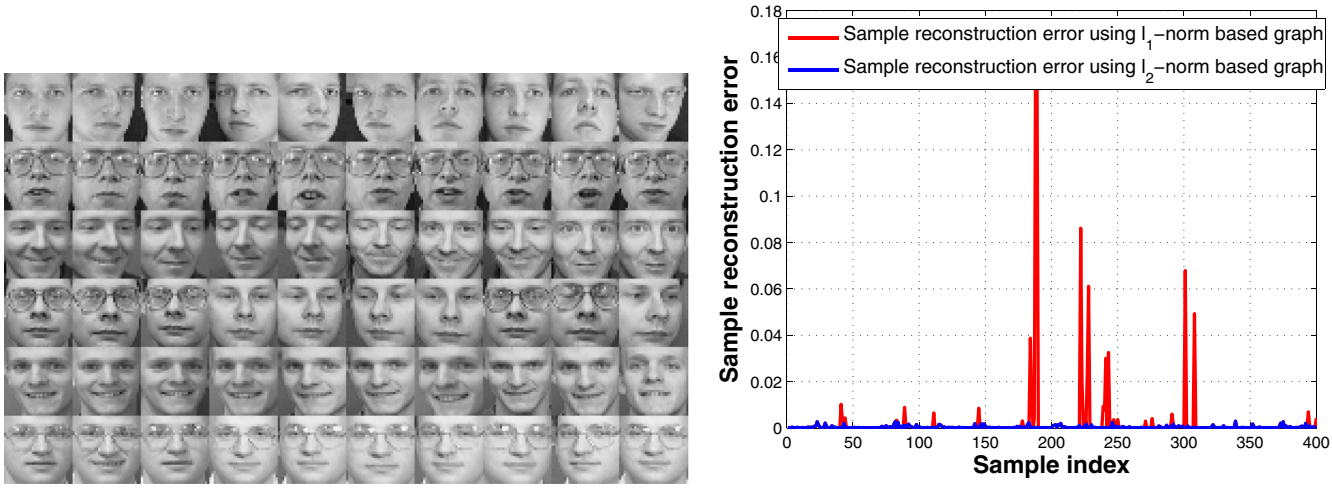
To validate the effectiveness of our l_2-UFS and l_1-UFS, we follow the common experiment setting of UFS, i.e., the clustering performances with selected features are evaluated. We compare our models with the following representative methods:

- **Baseline**: All of the original features are adopted;
- **LS**: Laplacian Score (He et al., 2005), in which features are selected with the most consistency with Gaussian Laplacian matrix;

[1] http://www.tangchang.net.

[2] Isolet dataset is publicly available from http://www.archive.ics.uci.edu/ml/datasets/ISOLET (for computation efficiency, only the first group (with 1560 instances) of data are used for experimental test.)

[3] All of the datasets are publicly available from http://www.featureselection.asu.edu/datasets.php.



(a) Original images (b) Sampe reconstruction errors

Fig. 1. Reconstruction errors of sample points that contain noise.

Table 3
Running time (s) comparison of different methods on different datasets.

Datasets	Baseline	LS	MCFS	RSR	GloSS	SGOFS	l_2-UFS	$l_{1,F}$-UFS	l_1-UFS
Yale	–	0.063	0.562	0.178	1.513	22.801	0.381	0.394	0.405
ORL	–	0.047	0.644	0.22	1.651	27.326	0.553	0.589	0.599
orlraws10P	–	0.062	6.96	70.03	950.324	6244.347	113.584	125.038	134.347
warpPIE10P	–	0.031	1.331	1.517	19.971	188.607	2.824	3.514	4.064
warpAR10P	–	0.016	1.051	1.252	19.012	147.672	2.473	2.997	3.596
Isolet	–	0.187	2.961	0.19	20.505	10.658	2.421	6.679	8.735
Prostate_GE	–	0.031	2.318	15.538	280.633	1157.323	31.88	47.297	57.954
CLL_SUB_111	–	0.078	5.714	96.806	1064.729	7359.825	167.893	180.641	198.067
USPS	–	35.628	55.173	2.424	984.851	6839.482	148.024	153.904	167.016
COIL20	–	0.234	3.332	0.42	17.227	20.528	2.871	7.894	11.043

- **MCFS**: Multi-cluster feature selection (Cai et al., 2010), it uses the l_1-norm to regularize the feature selection process as a spectral information regression problem;
- **RSR**: Regularized self-representation feature selection method (Zhu et al., 2015), which uses the $l_{2,1}$-norm to measure the fitting error and also promote sparsity;
- **GLoSS**: Global and local structure preserving sparse subspace learning model for UFS (Zhou et al., 2016), which can simultaneously realize feature selection and subspace learning.
- **SGOFS**: UFS with structured graph optimization (Nie et al., 2016), which performs feature selection and local structure learning simultaneously.

For LS, MCFS, RSR and SGOFS, we use the authors' released Matlab implementation. For GLoSS, since the authors have not released their codes, we implement it ourself by using Matlab. For all the methods, we use MATLAB R2013b to run the codes on a machine with an Intel i5 2.50GHz CPU and 8GB RAM.

Like previous methods, we also use the selected features to perform K-means clustering. Two widely used evaluation metrics, i.e., accuracy (ACC) and normalized mutual information (NMI), are employed to evaluate the performance of clusters. The larger ACC and NMI represent better performance. Denote by q_i the clustering results and by p_i the true label of x_i. ACC is defined as follows:

$$ACC = \frac{\sum_{i=1}^{n} \delta(p_i, map(q_i))}{n}, \tag{22}$$

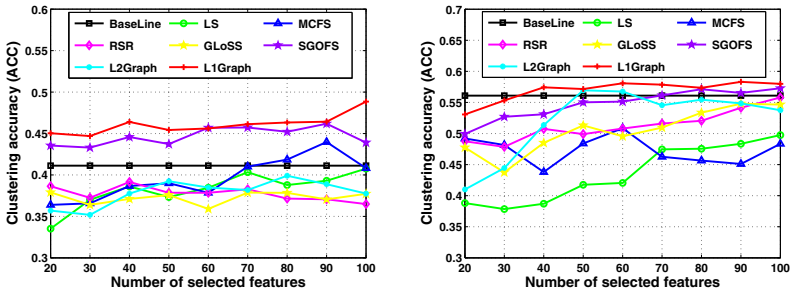
where $\delta(x, y) = 1$ if $x = y$, otherwise $\delta(x, y) = 0$. $map(q_i)$ is the best mapping function that permutes clustering labels to match the true labels using the KuhnMunkres algorithm. Given two variables P and Q, NMI is defined as

$$NMI(P, Q) = \frac{I(P, Q)}{\sqrt{H(P)H(Q)}} \tag{23}$$

where $H(P)$ and $H(Q)$ are the entropies of P and Q, respectively, and $I(P, Q)$ is the mutual information between P and Q. For clustering, P and Q are the clustering results and the true labels, respectively. NMI reflects the consistency between clustering results and ground truth labels.

In addition, since the Rand Index (Hubert & Arabie, 1985) is also used for measuring data clustering algorithms, it is a measure of the similarity between two data clusterings. A corrected-for-chance version of the Rand index is the Adjusted Rand Index (ARI), it can correct the classification rates for chance and remove the bias for chance correct classifications. Compared to traditional classification rates, its superiority for assessing the quality of a clustering solution has been demonstrated in Steinley and Douglas (2004). Thus, in this work, we also use ARI as one of the evaluation metrics.

Several parameters need to be set in l_2-UFS, l_1-UFS and other previous methods. For LS, GLoSS, MCFS, SOGFS, l_2-UFS and l_1-UFS, we fixed the neighbourhood size to 5 for all the datasets since we focus on the ability to select important features but not the neighbourhood size for building the similarity graph, this setting was also used in other previous literatures. For SOGFS, the projection dimension m is tuned from $\{\frac{d}{3}, \frac{2d}{3}\}$ where d is the original feature dimension. In order to make fair comparison of different UFS methods, we tuned the rest parameters for all methods by a "grid-search" strategy from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. Because the optimal number of selected features is unknown, we set different number of selected features for all datasets, the selected



(a) Yale (b) ORL

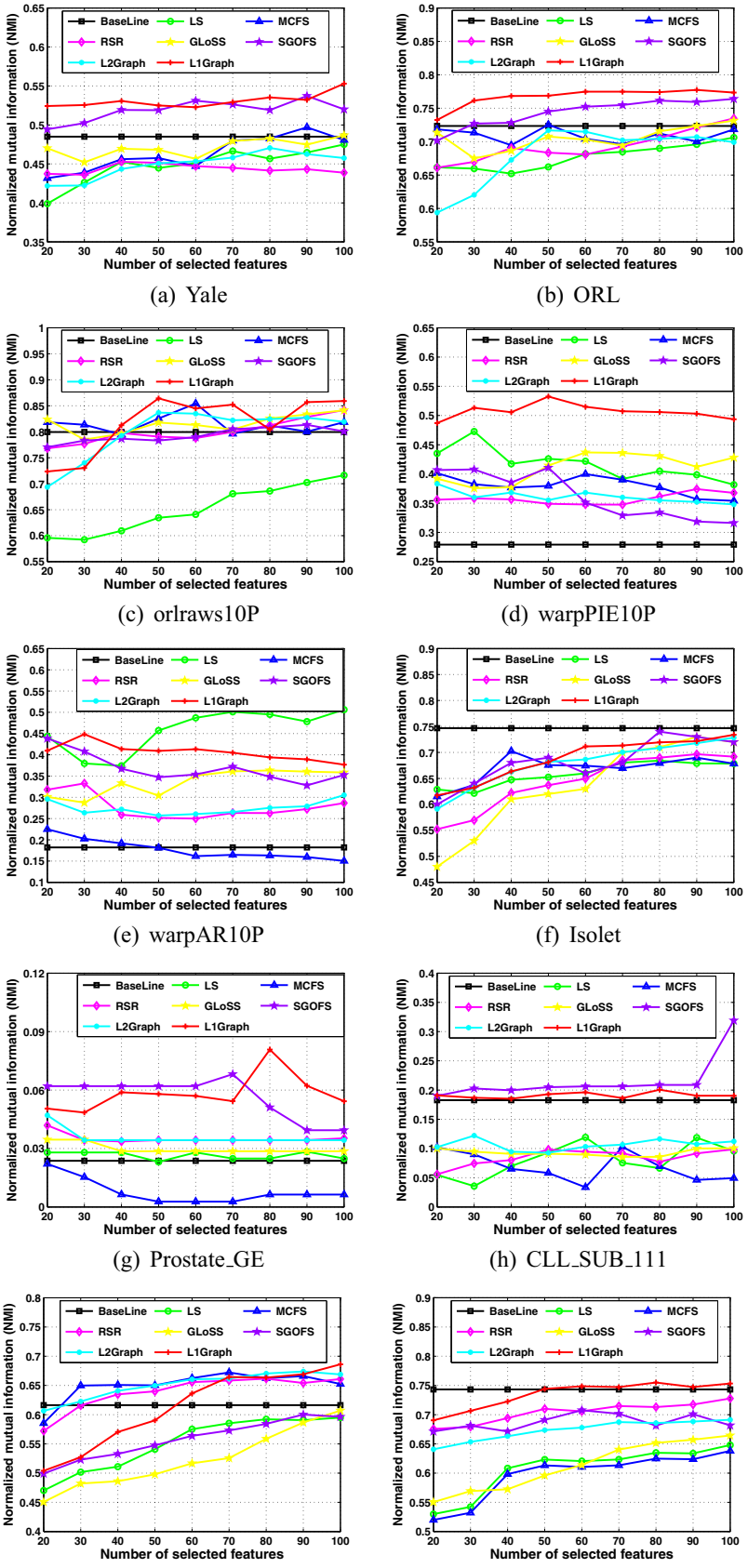
(c) orlraws10P (d) warpPIE10P

(e) warpAR10P (f) Isolet

(g) Prostate_GE (h) CLL_SUB_111

(i) USPS (j) COIL20

Fig. 2. ACC of using different selected numbers of features by different methods.



(a) Yale (b) ORL

(c) orlraws10P (d) warpPIE10P

(e) warpAR10P (f) Isolet

(g) Prostate_GE (h) CLL_SUB_111

(i) USPS (j) COIL20

Fig. 3. NMI of using different selected numbers of features by different methods.

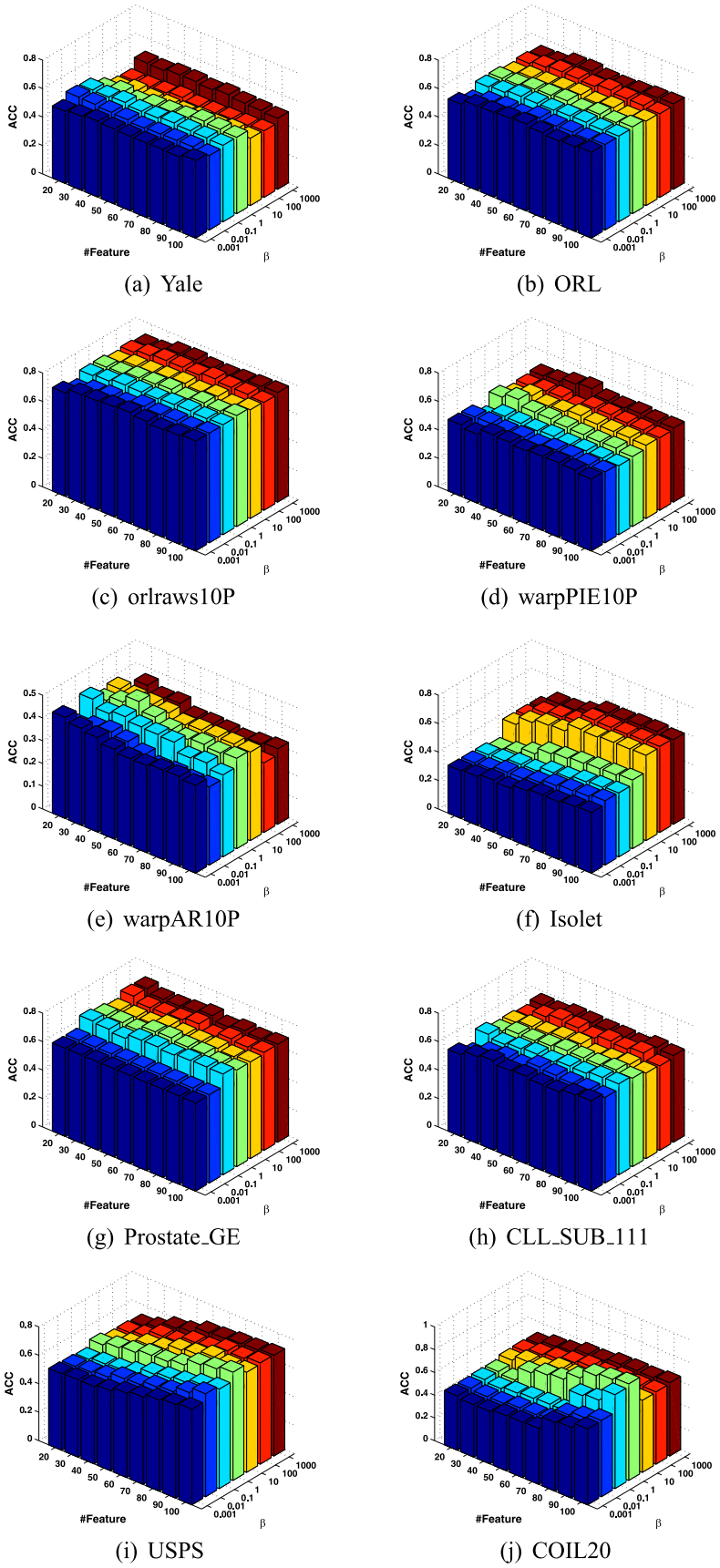
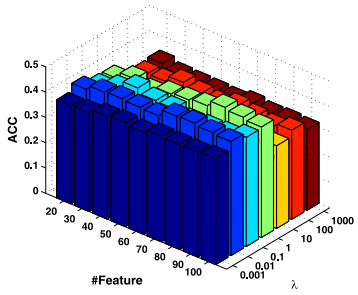
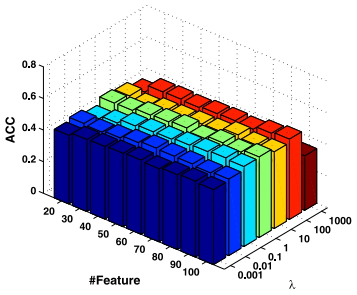


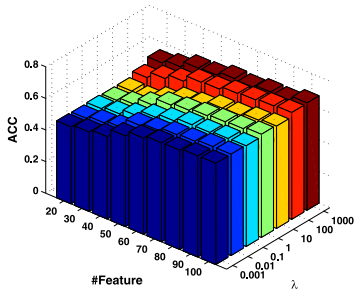
Fig. 4. The clustering accuracy of l_1-UFS w.r.t. parameter β while fixing parameters $\lambda = 1$. Here, β is reported with a grid search strategy, varying in $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$.



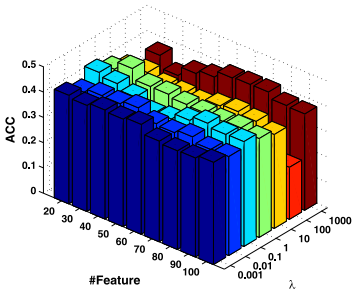
(a) Yale



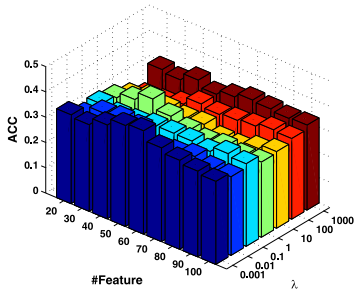
(b) ORL



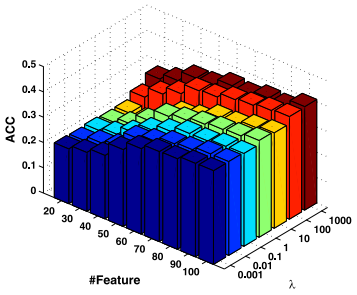
(c) orlraws10P



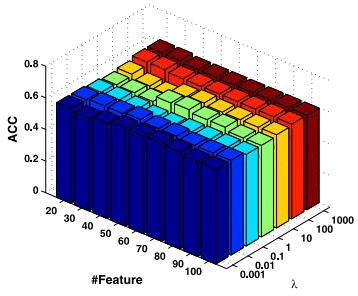
(d) warpPIE10P



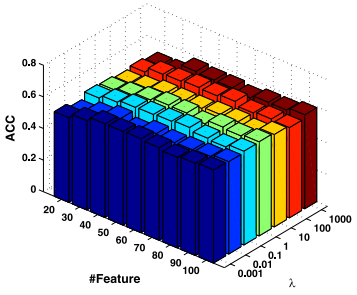
(e) warpAR10P



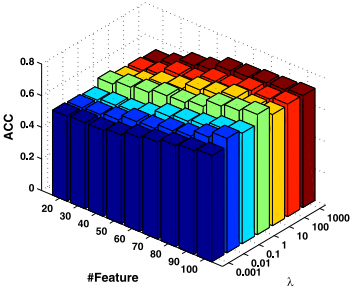
(f) Isolet



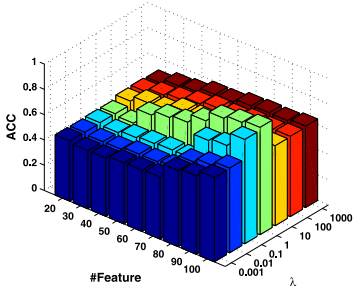
(g) Prostate_GE



(h) CLL_SUB_111



(i) USPS



(j) COIL20

Fig. 5. The clustering accuracy of l_1-UFS w.r.t. parameter λ while fixing parameters $\beta = 1$. Here, λ is reported with a grid search strategy, varying in $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$.

feature number was tuned from $\{20, 30, \ldots, 90, 100\}$, and the best clustering results from the optimal parameters are reported for all the algorithms. After completing the feature selection process, we use the K-means algorithm to cluster the samples using the selected features. As K-means algorithm depends on initialization, we run it 100 times with random starting points and report the average value.

6.3. Experimental results

The experimental results of different methods in terms of ACC, NMI and ARI on different datasets are summarized in Table 2(a)–(c). The best three results are highlighted in descending order with red, green and blue fonts, respectively. The results have shown that the proposed methods can perform better on most of the datasets than other state-of-the-art methods. There are two major reasons. First, the use of the $l_{2,1}$ to constrain the row sparsity on both the coefficient matrix and reconstruction error matrix can make feature selection naturally and enforce robustness to outlier samples, respectively. Second, the graph regularization terms can well preserve the local geometrical structure of data, which has been verified as an important property in UFS. Moreover, the l_1-UFS performs better than l_2-UFS, which demonstrates the robustness to noisy data by using the l_1-norm based graph in l_1-UFS.

In order to give an intuitive comparison, we show the sample reconstruction errors of each data sample in ORL dataset. Note that with the error matrix $\mathbf{E} = \mathbf{X} - \mathbf{XW} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]$, $||\mathbf{e}_i||_2$ can be used to recognize the noisy samples, i.e., the noisy samples will have larger reconstruction errors. Fig. 1 gives an example. Some original face images from ORL dataset are shown in Fig. 1(a). As can be seen, some face images are with eye glasses, i.e., the eye glasses can be regarded as noise. Hence, the reconstruction errors of those face images with eye glasses should be larger than other images. Fig. 1(b) plots the sample reconstruction of each sample, it verifies that l_1-norm based graph is more robust to noise than traditional l_2-norm based graph, i.e., by using l_1-norm based graph, more noisy samples can be steadily recognized.

For demonstrating the robustness to outliers by using the $l_{2,1}$-norm in our framework, we also give the results by replacing the $l_{2,1}$-norm with Frobenius norm in Eq. (10) (represented as $l_{1,F}$-UFS). As can be seen, the l_1-UFS can outperform $l_{1,F}$-UFS, which verifies the effectiveness of the $l_{2,1}$-norm based regularization. We can also observe that our l_2-UFS performs better than RSR, which demonstrates the importance of local geometrical structure preservation in UFS again.

In order to illustrate the effect of feature selection to clustering, we compare the clustering results using all features and selected features given by different methods. Figs. 2 and 3 plot the ACC value and NMI value respectively with respect to different number of selected features. From the figures, we can observe that, in most cases, the proposed l_1-UFS gives the best results. It is worth noting that when using fewer features, our method can obtain higher clustering accuracy than the baseline in most cases, which validates that our method can save the clustering time and improve clustering accuracy. As to the l_2-UFS, it also shows better results than RSR, this once again verifies the importance of local geometrical structure preservation for UFS.

6.4. Parameter sensitivity

There are two parameters in our proposed models (i.e., λ and β). To study the sensitivity of the models with regard to the parameters in Eqs. (6) and (10), experiments were conducted by fixing $\lambda = 1$ and varying β or fixing $\beta = 1$ and varying λ respectively. Since l_1-UFS performs far better than l_2-UFS, following we plot the ACC results of l_1-UFS. Figs. 4 and 5 show the ACC values of l_1-UFS on different datasets for different λ, β and selected number of features, respectively. Results have shown that our l_1-UFS performs stably well and is to some extent robust to parameters λ and β, but they are relatively sensitive to the number of selected features, which is a commonality of almost all the UFS methods.

6.5. Running time comparison

In order to give an intuitive efficiency comparison of different methods, we show the running time of different methods on different datasets in Table 3. Since the convergence conditions and iteration times of different methods vary according to the size of different datasets, here we show the average time cost of each method on each dataset with 10 iteration running times. As can be seen, although our proposed methods are not the fastest, they perform faster than GloSS and SGOFS in most cases, which demonstrate the good efficiency of the proposed l_2-UFS and l_1-UFS.

7. Conclusion

In this paper, we propose a unified robust unsupervised feature selection framework via feature self-representation and graph regularization. An $l_{2,1}$-norm and an l_1-norm are used to regularize the feature representation residual matrix and the graph regularization term. By this way, the proposed framework is able to reduce the effect of noisy data on feature selection. Furthermore, the proposed l_1-norm based graph Laplacian is readily extendible, which can be easily integrated into other UFS methods and machine learning tasks with local geometrical structure of data being preserved. Experimental results on ten challenging benchmark data sets demonstrate that our model consistently outperforms state-of-the-art UFS methods in the literature, suggesting the effectiveness of the proposed UFS framework.

Acknowledgments

This work is partly supported by the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) under Grant No. CUG170654 and the National Natural Science Foundation of China under Grant No. 61701451 and 61601261.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.eswa.2017.11.053.

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning, 3*(1), 1–122.

Buades, A., Coll, B., & Morel, J. M. (2005). A non-local algorithm for image denoising. In *IEEE conference on computer vision and pattern recognition* (pp. 60–65). IEEE.

Cai, D., Zhang, C., & He, X. (2010). Unsupervised feature selection for multi-cluster data. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 333–342). ACM.

Chang, T., Lijuan, C., Xiao, Z., & Minhui, W. (2017). Gene selection for microarray data classification via subspace learning and manifold regularization. *Medical & Biological Engineering & Computing*. doi:10.1007/s11517-017-1751-6.

Chen, J., Ma, Z., & Liu, Y. (2013). Local coordinates alignment with global preservation for dimensionality reduction.. *IEEE Transactions on Neural Networks and Learning Systems, 24*(1), 106–117.

Dadaneh, B. Z., Markid, H. Y., & Zakerolhosseini, A. (2016). Unsupervised probabilistic feature selection using ant colony optimization. *Expert Systems with Applications, 53*, 27–42.

Dietterich, T. G., & Bakiri, G. (1991). A general method for improving multi–class inductive learning programs. In *AAAI conference on artificial intelligence* (pp. 572–577). AAAI.

Dy, J. G., Brodley, C. E., Kak, A., Broderick, L. S., & Aisen, A. M. (2003). Unsupervised feature selection applied to content-based retrieval of lung images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(3), 373–378.

Eloy, C. (2011). Leonardo's rule, self-similarity and wind-induced stresses in trees. *Physical Review Letters, 107*(25). 258101–258101.
Freedman, G., & Fattal, R. (2011). Image and video upscaling from local self-examples. *ACM Transactions on Graphics, 30*(2), 474–484.
Gu, Q., Li, Z., & Han, J. (2011). Joint feature selection and subspace learning. In *International joint conference on artificial intelligence* (pp. 1294–1299). AAAI.
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learnign Research, 3*, 1157–1182.
He, X., Cai, D., & Niyogi, P. (2005). Laplacian score for feature selection.. In *Advances in neural information processing systems* (pp. 507–514).
He, X., & Niyogi, P. (2005). Locality preserving projections. In *Advances in neural information processing systems* (pp. 186–197).
Hou, C., Nie, F., Li, X., & Yi, D. (2014). Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *IEEE Transactions Cybernetics, 44*(6), 793–804.
Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.
Hunter, D. R., & Lange, K. (2000). [optimization transfer using surrogate objective functions]: Rejoinder. *Journal of Computational and Graphical Statistics, 9*(1), 52–59.
Javed, S., Sobral, A., Bouwmans, T., & Jung, S. K. (2015). Or-pca with dynamic feature selection for robust background subtraction. In *ACM/SIGAPP symposium on applied computing* (pp. 86–91). ACM.
Kodirov, E., Xiang, T., Fu, Z., & Gong, S. (2016). Learning robust graph regularisation for subspace clustering. In *British machine vision conference* (pp. 138.1–138.12).
Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1–2), 273–324.
Li, Z., Liu, J., Yang, Y., Zhou, X., & Lu, H. (2014). Clustering-guided sparse structural learning for unsupervised feature selection. *IEEE Transactions on Knowledge and Data Engineering, 26*(9), 2138–2150.
Liu, X., Wang, L., Zhang, J., Yin, J., & Liu, H. (2014). Global and local structure preservation for feature selection. *IEEE Transactions on Neural Networks and Learning Systems, 25*(6), 1083–1095.
Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using support vector machines. *Information Sciences, 179*(13), 2208–2217.
Mandelbrot, B. (1967). How long is the coast of britain.. *Science, 156*(3775), 636–638.
Nie, F., Wei, Z., & Li, X. (2016). Unsupervised feature selection with structured graph optimization. In *AAAI conference on artificial intelligence* (pp. 1302–1308). AAAI.
Rd, P. E., Ornstein, D. K., Paweletz, C. P., Ardekani, A., Hackett, P. S., Hitt, B. A., et al. (2002). Serum proteomic patterns for detection of prostate cancer. *Journal of Urology, 169*(4), 1576–1578.
Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding.. *Science, 290*(5500), 2323–2326.
Saul, L. K., & Roweis, S. T. (2003). Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research, 4*(2), 119–155.
Shang, R., Zhang, Z., Jiao, L., Liu, C., & Li, Y. (2016). Self-representation based dual-graph regularized feature selection clustering. *Neurocomputing, 171*(1), 1242–1253.
Song, Q., HaiYan, J., & Jing, L. (2017). Feature selection based on fda and f-score for multi-class classification. *Expert Systems with Applications, 81*, 22–27.
Steinley, & Douglas (2004). Properties of the hubert-arabie adjusted rand index.. *Psychological Methods, 9*(3), 386–396.
Tabakhi, S., Moradi, P., & Akhlaghian, F. (2014). An unsupervised feature selection algorithm based on ant colony optimization. *Engineering Applications of Artificial Intelligence, 32*(6), 112123.
Uysal, A. K. (2016). An improved global feature selection scheme for text classification. *Expert Systems with Applications, 43*, 82–92.
Wang, S., Tang, J., & Liu, H. (2015). Embedded unsupervised feature selection.. In *AAAI conference on artificial intelligence* (pp. 470–476). AAAI.
Wang, S., & Wang, H. (2017). Unsupervised feature selection via low-rank approximation and structure learning. *Knowledge-Based Systems, 124*, 70–79.
Wang, X., Liu, Y., Nie, F., & Huang, H. (2015). Discriminative unsupervised dimensionality reduction. In *International conference on artificial intelligence* (pp. 3925–3931). AAAI.
Yang, C. Y., Huang, J. B., & Yang, M. H. (2010). Exploiting self-similarities for single frame super-resolution. In *Asian conference on computer vision* (pp. 497–510). Springer.
Li, Z., & Tang, J. (2015). Unsupervised feature selection via nonnegative spectral analysis and redundancy control. *IEEE Transactions on Image Processing, 24*(12), 5343.
Zhang, T., Yang, J., Zhao, D., & Ge, X. (2007). Linear local tangent space alignment and application to face recognition. *Neurocomputing, 70*(79), 1547–1553.
Zhao, Z., & Liu, H. (2007). Spectral feature selection for supervised and unsupervised learning. In *International conference on machine learning* (pp. 1151–1157). IEEE.
Zhao, Z., Wang, L., & Liu, H. (2010). Efficient spectral feature selection with minimum redundancy. In *AAAI conference on artificial intelligence* (pp. 673–678). AAAI.
Zhou, N., Xu, Y., Cheng, H., Fang, J., & Pedrycz, W. (2016). Global and local structure preserving sparse subspace learning: An iterative approach to unsupervised feature selection. *Pattern Recognition, 53*, 87–101.
Zhu, P., Hu, Q., Zhang, C., & Zuo, W. (2016). Coupled dictionary learning for unsupervised feature selection. In *AAAI conference on artificial intelligence* (pp. 2422–2428). AAAI.
Zhu, P., Zhu, W., Hu, Q., Zhang, C., & Zuo, W. (2017). Subspace clustering guided unsupervised feature selection. *Pattern Recognition, 66*, 364–374.
Zhu, P., Zuo, W., Zhang, L., Hu, Q., & Shiu, S. C. K. (2015). Unsupervised feature selection by regularized self-representation. *Pattern Recognition, 48*(2), 438–446.