Preserving Ordinal Consensus: Towards Feature Selection for Unlabeled Data

Jun Guo,[1] Heng Chang,[1] Wenwu Zhu[1,2,3]
[1]Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen 518055, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3]Beijing National Research Center for Information Science and Technology, Beijing 100084, China
eeguojun@outlook.com, changh17@mails.tsinghua.edu.cn, wwzhu@tsinghua.edu.cn

Abstract

To better pre-process unlabeled data, most existing feature selection methods remove redundant and noisy information by exploring some intrinsic structures embedded in samples. However, these unsupervised studies focus too much on the relations among samples, totally neglecting the feature-level geometric information. This paper proposes an unsupervised triplet-induced graph to explore a new type of potential structure at feature level, and incorporates it into simultaneous feature selection and clustering. In the feature selection part, we design an ordinal consensus preserving term based on a triplet-induced graph. This term enforces the projection vectors to preserve the relative proximity of original features, which contributes to selecting more relevant features. In the clustering part, Self-Paced Learning (SPL) is introduced to gradually learn from 'easy' to 'complex' samples. SPL alleviates the dilemma of falling into the bad local minima incurred by noise and outliers. Specifically, we propose a compelling regularizer for SPL to obtain a robust loss. Finally, an alternating minimization algorithm is developed to efficiently optimize the proposed model. Extensive experiments on different benchmark datasets consistently demonstrate the superiority of our proposed method.

1 Introduction

Real-world data is often redundant or even noisy, which may lead to heavy computational complexity and poor performance (John, Kohavi, and Pfleger 1994; Liu and Motoda 2007). Then, feature selection is proposed to help remove unimportant information (Yang et al. 2011; Witten and Tibshirani 2012), which is beneficial for various applications (Law, Figueiredo, and Jain 2004; Nie et al. 2010; Wang et al. 2016; Cheng, Li, and Liu 2017).

Feature selection can be roughly grouped into three major categories in terms of label availability, *i.e.*, supervised, semi-supervised, and unsupervised (Han and Shen 2016). Supervised feature selection (Jian et al. 2016; Fan et al. 2017a) is utilized to select discriminative features because the class labels of data containing the essential discrimination are provided. However, the acquisition of class label information is very laborious and time-consuming, which makes feature selection based applications more challenging. Semi-supervised feature selection (Xu et al. 2010; Chang et al. 2014) is desired to tackle the predicament of rare labelled samples and abundant unlabelled data. Unsupervised feature selection aims to filter out the unimportant features of unlabeled data. Similar to the class labels in supervised scenarios, cluster structure can be discovered in many ways (Du and Shen 2015; Wei et al. 2017; Du et al. 2018; Guo and Zhu 2018; Zheng et al. 2018).

For unsupervised feature selection, filters (Dash et al. 2002), wrappers (Roth and Lange 2004), and embedding (Hou et al. 2014) are three common branches. Recent literatures (Yang et al. 2011; Li et al. 2012; Qian and Zhai 2013; Wang, Tang, and Liu 2015; Han and Kim 2015; Nie, Zhu, and Li 2016; Zhu et al. 2017; Li et al. 2018) have witnessed fast development of the third branch "embedding", whose goal is to combine feature selection and pseudo label learning into a unified problem. In these works, a commonly-used mechanism is to leverage the manifold structure and sparse learning. Most of them benefit from various geometric information of data. However, they focus too much on the sample-level structures, *i.e.*, relations among samples, failing in fully exploiting the ***feature-level geometric information***. The major reasons and analyses are two-folds:

Above all, unsupervised feature selection aims to select discriminative features meanwhile remove redundant features. These two points are NOT the same. If two features are similar, they usually encourage similar contributions for clustering. If two similar features are both relevant to current tasks, they should be simultaneously selected into the desired feature subset. However, previous works consider this case as feature redundancy, and try to repeal it. We think that feature-level relations can help select more relevant features. Hence, our work prefers to select more discriminative features rather than remove similar features.

Furthermore, in the "embedding" branch of unsupervised feature selection, various proposed graph-based models are utilized to uncover similarities (neighborhood relationships), whereas few of them address the relative proximities (neighborhood rankings). Besides feature-level similarities, relative proximities among features are also important in unsupervised scenarios. For a feature, this type of ordinal

information emphasizes the comparative information among other similar dimensions, *i.e.*, which one is more similar to it. Hence, exploiting this potential structure will help select more relevant features to current clustering tasks.

In order to address the above issues, this paper proposes a joint learning framework of feature selection and clustering. To the best of our knowledge, it is the first attempt for the "embedding" branch of unsupervised feature selection to uncover feature-level ordinal information. We design a triplet-induced graph considering the relative proximity of original features, and incorporate it into the overall self-paced objective function. An alternating minimization algorithm is developed to efficiently optimize the resulting robust model. Experimental analysis on benchmark datasets demonstrates the superiority of our approach.

In summary, our main contributions are as follows.

1) A novel triplet-induced graph is proposed to capture the feature-level relative proximity of data. Based on this pragmatic graph, we design an ordinal consensus preserving function for unsupervised feature selection.

2) Self-Paced Learning (SPL) is incorporated into clustering to decrease the risk of involving in bad local minima. Considering the existence of noise and outliers, we propose a compelling regularizer for SPL to obtain a robust loss.

3) An efficient algorithm is developed for the proposed model of joint feature selection and clustering. Extensive experiments well validate the effectiveness of our work.

2 The Proposed Method

2.1 Problem Definition

Let $\mathbf{X} \in \mathbb{R}^{d_1 \times n}$ denote the original data matrix with n samples and d_1 features. In the "embedding" branch of unsupervised feature selection, the objective function is generally formulated based on the regularized regression, *i.e.,* $\min_{\mathbf{W}} \ell\left(\mathbf{W}^T\mathbf{X}, \mathbf{H}\right) + \beta\|\mathbf{W}\|_{2,1}$, where $\ell(\cdot,\cdot)$ is the loss function, β is a regularization parameter, and $\|\mathbf{W}\|_{2,1}$ stands for the $l_{2,1}$ norm ($\sum_i \sqrt{\sum_j \mathbf{W}_{ij}^2}$).

The typical choices of $\ell(\cdot,\cdot)$ include logistic regression, correntropy, and least square. The target matrix $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n] \in \mathbb{R}^{d_2 \times n}$ is usually the corresponding label matrix in a supervised scenario. The projection matrix $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ $(d_1 > d_2)$ is named feature selection matrix. The $l_{2,1}$ norm guarantees its row-wise sparseness. $\mathbf{W}_{i\cdot}$ will shrink to $\mathbf{0}$ when $\mathbf{X}_{i\cdot}$ is less correlated to the labels.

However, in unsupervised cases, the label information is directly unavailable, which makes it challenging for feature selection. In the past decade, researchers determined $\mathbf{H}$ by learning pseudo labels through linear regression (Yang et al. 2011), spectral clustering (Li et al. 2012), K-means clustering (Qian and Zhai 2013), consensus clustering (Liu, Shao, and Fu 2016), and so on. Besides, some works (Han and Kim 2015; Zheng et al. 2019) utilized bi-orthogonal semi Nonnegative Matrix Factorization (NMF) to decompose $\mathbf{H}$ into two matrices: the latent orthogonal bases $\mathbf{U} \in \mathbb{R}^{d_2 \times c}$ and the pseudo-label indicators $\mathbf{V} \in \mathbb{R}^{c \times n}$.

$$\begin{aligned} \min_{\mathbf{W},\mathbf{U},\mathbf{V}} \quad & \ell\left(\mathbf{W}^T\mathbf{X}, \mathbf{U}\mathbf{V}\right) + \beta\|\mathbf{W}\|_{2,1} \\ s.t. \quad & \mathbf{U}^T\mathbf{U} = \mathbf{I},\ \mathbf{V}\mathbf{V}^T = \mathbf{I},\ \mathbf{V} \geq \mathbf{0}, \end{aligned} \tag{1}$$

where c is the number of latent clusters, all elements of $\mathbf{V}$ are non-negative. Hereafter, $\mathbf{I}$ denotes the identity matrix with a compatible size. In this paper, we decompose $\mathbf{H}$ as well, but we have some considerations:

1) **For the constraint on $\mathbf{W}$:** The constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ can not only suppress the feature similarity of arbitrary two selected dimensions, but also avoid arbitrary scaling and the trivial solution of all zeros. For the effectiveness of $\|\mathbf{W}\|_{2,1}$, Theorem 1 guarantees the ideal efficacy of feature selection.

Theorem 1. *(Wang, Nie, and Huang 2014) Given $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ with $d_1 > d_2$, the problem $\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \|\mathbf{W}\|_{2,1}$ is equivalent to $\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \|\mathbf{W}\|_{2,0}$.*

2) **For the constraint on $\mathbf{U}$:** From the perspective of signal processing, $\mathbf{W}$ stands for analytic projection that is compact and precise. Meanwhile, $\mathbf{U}$ represents synthetic decomposition highly relied on residuals. As stated in (Qian and Zhai 2013), when we decompose $\mathbf{H}$ as $\mathbf{H} \simeq \mathbf{U}\mathbf{V}$, the adverse effect induced by noise and outliers is often accumulated in $\mathbf{U}$ but seldom hurts $\mathbf{V}$ severely. Therefore, we put aside the orthogonal constraint on $\mathbf{U}$ for simplicity. In terms of computational complexity, the removal of $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ makes the model more efficient than (Han and Kim 2015; Zheng et al. 2019) in which Singular Value Decomposition (SVD) is used twice for optimization at each iteration.

3) **For the constraints on $\mathbf{V}$:** It is obvious that the constraints $\mathbf{V}\mathbf{V}^T = \mathbf{I}$, $\mathbf{V} \geq \mathbf{0}$ can ensure that each $\mathbf{V}_{\cdot i}$ has only one non-zero element. According to Theorem 2, we can regard $\mathbf{V}$ as a weighted cluster indicator matrix in K-means clustering. Then, we can explicitly utilize the constraints $\mathbf{V}_{\cdot i} \in \{0,1\}^c$, $\|\mathbf{V}_{\cdot i}\|_0 = 1$, $\forall i$ instead of $\mathbf{V}\mathbf{V}^T = \mathbf{I}$, $\mathbf{V} \geq \mathbf{0}$. In such a way, K-means clustering is naturally performed on the selected feature groups $\mathbf{W}^T\mathbf{X}$.

Theorem 2. *(Ding, He, and Simon 2005) The problem*

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & \|\mathbf{Y} - \mathbf{U}\mathbf{V}\|_F^2 \\ s.t. \quad & \mathbf{V}\mathbf{V}^T = \mathbf{I},\ \mathbf{V} \geq \mathbf{0} \end{aligned} \tag{2}$$

is equivalent to relaxed K-means clustering.

Based on the above three considerations, we formulate a basic model of joint feature selection and clustering as

$$\begin{aligned} \min_{\mathbf{W},\mathbf{U},\mathbf{V}} \quad & \ell\left(\mathbf{W}^T\mathbf{X}, \mathbf{U}\mathbf{V}\right) + \beta\|\mathbf{W}\|_{2,1} \\ s.t. \quad & \mathbf{W}^T\mathbf{W} = \mathbf{I},\ \mathbf{V}_{\cdot i} \in \{0,1\}^c,\ \|\mathbf{V}_{\cdot i}\|_0 = 1,\ \forall i. \end{aligned} \tag{3}$$

2.2 Exploiting Feature-Level Ordinal Consensus

In this subsection, we propose a triplet-induced graph to exploit the feature-level relative proximity of data. Based on this pragmatic graph, we further design an ordinal consensus preserving term $\Theta(\mathbf{W})$ for the basic model (3).

Observation 1. *The features and projection vectors share one-to-one correspondences.* □

The projected data $\mathbf{W}^T\mathbf{X}$ can be calculated as $\mathbf{W}_{1\cdot}^T\mathbf{X}_{1\cdot} + \cdots + \mathbf{W}_{i\cdot}^T\mathbf{X}_{i\cdot} + \cdots + \mathbf{W}_{j\cdot}^T\mathbf{X}_{j\cdot} + \cdots + \mathbf{W}_{d_1\cdot}^T\mathbf{X}_{d_1\cdot}$, where $\mathbf{X}_{i\cdot} \in \mathbb{R}^{1\times n}(i = 1, \cdots, d_1)$ is the i^{th} feature of all n samples. $\mathbf{W}_{i\cdot} \in \mathbb{R}^{1\times d_2}(i = 1, \cdots, d_1)$ denotes the corresponding projection vector of $\mathbf{X}_{i\cdot}$. Therefore, we can observe that each $\mathbf{X}_{i\cdot}$ and $\mathbf{W}_{i\cdot}$ do share a one-to-one correspondence.

Readers can regard this one-to-one correspondence as an implicit function. If the projection vector $\mathbf{W}_{i\cdot}$ shrinks to $\mathbf{0}$, the corresponding feature $\mathbf{X}_{i\cdot}$ will be discarded. We can utilize the projection vectors to measure whether the features are related to current tasks. Similar features encourage their corresponding projection vectors to be similar. Meanwhile, if two similar features are relevant to clustering, they should be simultaneously selected. However, previous studies consider this case as feature redundancy, and try to repeal it. We prefer to select more discriminative features rather than remove similar features. Uncovering feature-level relationships contributes to selecting more relevant features.

In the "embedding" branch of unsupervised feature selection, various graphs were proposed to uncover similarities (neighborhood relationships), whereas few of them address the relative proximities (neighborhood rankings). Besides feature-level similarities, relative proximities also play a vital role in unsupervised tasks. To fully exploit this type of ordinal information embedded in the features, we design an *ordinal consensus preserving* term $\Theta(\mathbf{W})$ for Eq.(3). A concrete definition is as follows.

Definition 1. Given a triplet of features $\{\mathbf{X}_{i\cdot}, \mathbf{X}_{u\cdot}, \mathbf{X}_{v\cdot}\}$ comprised of $\mathbf{X}_{i\cdot}$ and its two neighbors $\mathbf{X}_{u\cdot}$ and $\mathbf{X}_{v\cdot}$, their corresponding projection vectors form a triplet $\{\mathbf{W}_{i\cdot}, \mathbf{W}_{u\cdot}, \mathbf{W}_{v\cdot}\}$. Denote the function $dis(\cdot,\cdot)$ as a distance metric. Then, the feature selection process is called *ordinal consensus preserving* when the following relative proximities holds: if $dis(\mathbf{X}_{i\cdot}, \mathbf{X}_{u\cdot}) \leq dis(\mathbf{X}_{i\cdot}, \mathbf{X}_{v\cdot})$, then $dis(\mathbf{W}_{i\cdot}, \mathbf{W}_{u\cdot}) \leq dis(\mathbf{W}_{i\cdot}, \mathbf{W}_{v\cdot})$. □

According to the classic Rearrangement Inequality[1], preserving ordinal consensus for feature selection means to determine appropriate $\{\mathbf{W}_{i\cdot}, \mathbf{W}_{u\cdot}, \mathbf{W}_{v\cdot}\}$ that can yield the maximum product of $dis(\mathbf{X}_{i\cdot}, \mathbf{X}_{u\cdot}) - dis(\mathbf{X}_{i\cdot}, \mathbf{X}_{v\cdot})$ and $dis(\mathbf{W}_{i\cdot}, \mathbf{W}_{u\cdot}) - dis(\mathbf{W}_{i\cdot}, \mathbf{W}_{v\cdot})$.

Therefore, determining an appropriate projection matrix $\mathbf{W}$ is identical to optimize the following ordinal consensus preserving objective function over a collection of triplets.

$$\max_{\mathbf{W}} \sum_{i=1}^{d_1} \sum_{u\in\mathcal{N}_i} \sum_{v\in\mathcal{N}_i} \mathbf{D}_{uv}^i \left[dis(\mathbf{W}_{i\cdot}, \mathbf{W}_{u\cdot}) - dis(\mathbf{W}_{i\cdot}, \mathbf{W}_{v\cdot})\right], \tag{4}$$

where $\mathbf{D}^i$ is an antisymmetric matrix whose $(u,v)^{th}$ element is $dis(\mathbf{X}_{i\cdot}, \mathbf{X}_{u\cdot}) - dis(\mathbf{X}_{i\cdot}, \mathbf{X}_{v\cdot})$, and $\mathcal{N}_i$ is a set of indexes for the k nearest neighbors of $\mathbf{X}_{i\cdot}$.

We define $\mathbf{M} \in \mathbb{R}^{d_1\times d_1}$ as a weighting matrix with

$$\mathbf{M}_{ij} \triangleq \begin{cases} \sum\limits_{u\in\mathcal{N}_i} \mathbf{D}_{uj}^i & , j \in \mathcal{N}_i \\ 0 & , j \notin \mathcal{N}_i \end{cases}. \tag{5}$$

According to the proof of Proposition 1 in (Guo et al. 2016), we can easily obtain that the objective function (4) is equivalent to $\min_{\mathbf{W}} \sum_{i=1}^{d_1}\sum_{j=1}^{d_1} \mathbf{M}_{ij} dis(\mathbf{W}_{i\cdot}, \mathbf{W}_{j\cdot})$.

[1] http://en.wikipedia.org/wiki/Rearrangement_inequality.

Therefore, the aforementioned ordinal consensus preserving term $\Theta(\mathbf{W}) = \sum_{i=1}^{d_1}\sum_{j=1}^{d_1} \mathbf{M}_{ij} dis(\mathbf{W}_{i\cdot}, \mathbf{W}_{j\cdot})$. Defined over a set of triplets, Eq.(5) actually defines a novel graph to simultaneously reflect neighborhood relationship as well as ordinal information.

2.3 Self-Paced Joint Learning Model

Non-convex models are often stuck in bad local minima, especially when there exist outliers, heavy noise and missing data. A frequently-used way to alleviate this difficulty is to run the algorithm multiple times with different initializations and then pick the best solution among them (Zhao et al. 2015). Nevertheless, this strategy is time-consuming and inconvenient to implement in unsupervised cases, since there is no explicit criterion for determining a proper solution.

Another heuristic way to handle the dilemma of bad local minima is Self-Paced Learning (SPL) (Kumar, Packer, and Koller 2010). In the past decade, SPL has attracted much attention and been proven to be a powerful technique. It is a learning paradigm mimicking the learning process of human and animal. The samples are not learned randomly but in a significant order which illustrates from 'easy' to gradually more 'complex' ones (Jiang et al. 2015).

A general SPL model is comprised of a weighted loss term on all samples and a self-paced regularizer term imposed on the weights of samples, *i.e.*,

$$\begin{aligned} \min_{\mathbf{w},\mathbf{r}} \quad & \sum_i r_i \ell_i(\mathbf{x}_i, \mathbf{w}) + f(\lambda, \mathbf{r}) \\ s.t. \quad & r_i \in [0,1], \forall i, \end{aligned} \tag{6}$$

where $\mathbf{r} = [r_1, \cdots, r_i, \cdots]^T$ is a vector and r_i is a latent weight for the i^{th} sample. $\ell_i(\mathbf{x}_i, \mathbf{w})$ stands for the loss function for the i^{th} sample. $\mathbf{w}$ is a model parameter. $f(\lambda, \mathbf{r})$ is the self-paced regularizer. λ is a scalar controlling the learning rate. By increasing the penalty on the regularizer step by step during optimization, more samples are automatically chosen for training in a pure self-paced way.

In this paper, we adopt SPL to help the clustering part of our model circumvent the bad local minima. We further develop a self-paced joint learning framework of feature selection and clustering. Finally, the ***overall objective function*** of our proposed method is formulated as follow. For all i,

$$\begin{aligned} \min_{\mathbf{W},\mathbf{U},\mathbf{V},r_i\in[0,1]} \quad & \mathcal{J} = \left\{ \begin{array}{l} \sum\limits_{i=1}^{n} r_i\ell_i + f(\lambda,\mathbf{r}) + \beta\|\mathbf{W}\|_{2,1} \\ +\frac{\alpha}{2}\sum\limits_{i=1}^{d_1}\sum\limits_{j=1}^{d_1} \mathbf{M}_{ij} dis(\mathbf{W}_{i\cdot}, \mathbf{W}_{j\cdot}) \end{array} \right\} \\ s.t. \quad & \mathbf{W}^T\mathbf{W} = \mathbf{I},\ \mathbf{V}_{\cdot i} \in \{0,1\}^c,\ \|\mathbf{V}_{\cdot i}\|_0 = 1, \end{aligned} \tag{7}$$

where ℓ_i stands for $\ell\left(\mathbf{W}^T\mathbf{X}_{\cdot i}, \mathbf{U}\mathbf{V}_{\cdot i}\right)$, β and α are two regularization parameters. On the one hand, Eq.(7) is compatible with a group of regularizers and hard/soft weights, which effectively contribute to bad local minima eradication. On the other hand, it can be incorporated with various graph-based learning methods (Yan et al. 2007).

3 Optimization

3.1 Optimization Procedure

For the convenience of calculation, we adopt the common least square as a metric to establish each loss function $\ell(\cdot,\cdot)$

and pairwise distance function $dis(\cdot,\cdot)$. The overall model (7) can be optimized by alternative search strategy.

1) **U-step**: To update $\mathbf{U}$ with other variables fixed, we solve $\min_{\mathbf{U}} ||\left(\mathbf{W}^T\mathbf{X}-\mathbf{U}\mathbf{V}\right)diag\left(\sqrt{\mathbf{r}}\right)||_F^2$. The function $\sqrt{\mathbf{r}}$ denotes the element-wise square root of $\mathbf{r}$, and $diag\left(\sqrt{\mathbf{r}}\right)$ returns a diagonal matrix with the elements of vector $\sqrt{\mathbf{r}}$ on the main diagonal. We set the first-order partial derivative w.r.t. $\mathbf{U}$ to zero and obtain

$$\mathbf{U}=\mathbf{W}^T\mathbf{X}diag\left(\mathbf{r}\right)\mathbf{V}^T\left[\mathbf{V}diag\left(\mathbf{r}\right)\mathbf{V}^T\right]^{-1}. \tag{8}$$

2) **V-step**: To update $\mathbf{V}$ when other variables are fixed, we decouple the problem and assign the nearest cluster centroid for each sample. We can adopt an exhaustive search to solve each sub-problem

$$\begin{array}{ll}\min\limits_{\mathbf{V}_{\cdot i}} & \left\|\mathbf{W}^T\mathbf{x}_i-\mathbf{U}\mathbf{V}_{\cdot i}\right\|_F^2 \\ s.t. & \mathbf{V}_{\cdot i}\in\{0,1\}^c,\ \|\mathbf{V}_{\cdot i}\|_0=1.\end{array} \tag{9}$$

3) **W-step**: The $l_{2,1}$ norm of $\mathbf{W}$ equals $\sum_{i=1}^{d_1}\sqrt{||\mathbf{W}_{i\cdot}||_2^2}$.

Lemma 1. *(He et al. 2014) For a fixed u, there exists a conjugate function $\psi(\cdot)$, such that $\sqrt{u^2+\varepsilon}=\inf_{p\in\mathbb{R}}\{\frac{1}{2}pu^2+\psi(p)\}$, where p is determined by $\delta(u)=1/\sqrt{u^2+\varepsilon}$.*

According to Lemma 1, we solve $\mathbf{W}$ by alternately minimizing the following augmented function

$$\min_{\mathbf{P},\mathbf{W}:\mathbf{W}^T\mathbf{W}=\mathbf{I}}\left\{\begin{array}{c}\left\|\left(\mathbf{W}^T\mathbf{X}-\mathbf{U}\mathbf{V}\right)diag\left(\sqrt{\mathbf{r}}\right)\right\|_F^2 \\ +\beta\sum\limits_{i=1}^{d_1}\left\{\frac{\mathbf{P}_{ii}}{2}\left\|\mathbf{W}_{i\cdot}\right\|_2^2+\psi_i\left(\mathbf{P}_{ii}\right)\right\} \\ +\alpha Tr\left(\mathbf{W}^T\mathbf{L}\mathbf{W}\right)\end{array}\right\}, \tag{10}$$

where $\mathbf{L}\triangleq\mathbf{G}-\frac{\mathbf{M}+\mathbf{M}^T}{2}$ is the Laplacian matrix and $\mathbf{G}$ is a diagonal matrix whose $(i,i)^{th}$ element equals $\sum_{j=1}^{n}\frac{\mathbf{M}_{ij}+\mathbf{M}_{ji}}{2}$. $\mathbf{P}$ is a $d_1\times d_1$ diagonal matrix storing the auxiliary variables. $\{\psi_i\}_{i=1}^{d_1}$ are conjugate functions.

Based on Lemma 1, Eq.(10) is alternately minimized as[2]

$$\mathbf{P}_{ii}^{t+1}=1/\sqrt{\|\mathbf{W}_{i\cdot}^t\|_2^2+\varepsilon}\,, \tag{11}$$

$$\mathbf{W}^{t+1}=\underset{\mathbf{W}:\mathbf{W}^T\mathbf{W}=\mathbf{I}}{\arg\min}\ Tr\left(\mathbf{W}^T\mathbf{Q}\mathbf{W}\right). \tag{12}$$

where t means the t^{th} iteration. $\mathbf{Q}=\alpha\mathbf{L}+\mathbf{X}diag\left(\mathbf{r}\right)\mathbf{X}^T-\mathbf{X}diag\left(\mathbf{r}\right)\mathbf{V}^T\left[\mathbf{V}diag\left(\mathbf{r}\right)\mathbf{V}^T\right]^{-1}\mathbf{V}diag\left(\mathbf{r}\right)\mathbf{X}^T+\frac{\beta}{2}\mathbf{P}^{t+1}$. The solution of Eq.(12) is computed by performing eigen decomposition of $\mathbf{Q}$. The optimal $\mathbf{W}$ is made up of d_2 eigenvectors corresponding to the d_2 smallest eigenvalues.

4) **r-step**: For $f(\lambda,\mathbf{r})$, (Xu, Tao, and Xu 2015) designed $\sum_{i=1}^{n}\left(1+e^{-\lambda}-r_i\right)\ln\left(1+e^{-\lambda}-r_i\right)+r_i\ln r_i-\lambda r_i$ to obtain a soft weighting manner. When λ and other variables are fixed in each iteration, r_i is calculated as $r_i=\frac{1+e^{-\lambda}}{1+e^{\ell_i-\lambda}}$.

Considering the existence of noise and outliers, we propose a new regularizer to obtain a robust loss: $f(\lambda,\mathbf{r},\mathbf{z})=\sum_{i=1}^{n}z_i\left(1+e^{-\lambda}-r_i\right)\ln\left(1+e^{-\lambda}-r_i\right)+z_ir_i\ln r_i-\lambda r_i$,

[2] $\varepsilon=10^{-6}$ when the denominator is zero, and 0 otherwise.

Algorithm 1: The algorithm to solve Eq.(7)

Input: Data matrix $\mathbf{X}\in\mathbb{R}^{d_1\times n}$; number of nearest neighbors k, latent clusters c, and projected dimension d_2; regularization parameters β and α.

Output: m selected features.

Find k nearest neighbors for each feature $\mathbf{X}_{i\cdot}$. Compute $\mathbf{M}$ via Eq.(5) and its Laplacian matrix $\mathbf{L}$;

Initialize $\mathbf{W}^{(0)}$ with d_2 columns randomly selected from a $d_1\times d_1$ identity matrix. Initialize $\mathbf{U}^{(0)}$ and $\mathbf{V}^{(0)}$ by K-means clustering on $\mathbf{W}^{(0)}\mathbf{X}$, $\mathbf{r}^{(0)}=\mathbf{1}^n$, $t=0$, $\mu=1.1$, and $\lambda=10^{-6}$;

repeat

  $t\leftarrow t+1$, $\lambda\leftarrow\lambda\mu$;

  Update $\mathbf{U}^{(t)}$ via Eq.(8);

  Update $\mathbf{V}^{(t)}$ by exhaustive search;

  Update $\mathbf{P}^{(t)}$ via Eq.(11);

  Update $\mathbf{W}^{(t)}$ by eigen decomposition;

  Update $\mathbf{r}^{(t)}$ via $r_i=\frac{1+e^{-\lambda}}{1+e^{\frac{\ell_i-\lambda}{z_i}}}$ for each i;

until *convergence*;

Sort all features according to $\|\mathbf{W}_{i\cdot}\|_2\,(i=1,\cdots,d_1)$ in descending order and select the top-m ranked ones.

where $\mathbf{z}$ is a vector storing n independent Bernoulli random variables $\{z_i\}_{i=1}^n$ with the probability s of being 1. Then, the optimal $\mathbf{r}$ has a closed-form solution $r_i=\frac{1+e^{-\lambda}}{1+e^{\frac{\ell_i-\lambda}{z_i}}}$.

The robust effect can be demonstrated in each r_i's closed-form solution[3]: with the probability $1-s$, r_i approaches 0; with the probability s, r_i approaches the solution of the plain regularizer in (Xu, Tao, and Xu 2015). Our proposed regularizer helps improve the compatibility for noise and outliers.

3.2 Algorithmic Analysis

Variables $\mathbf{U}$, $\mathbf{V}$, $\mathbf{P}$, $\mathbf{W}$, and $\mathbf{r}$ are alternately optimized for several iterations in Algorithm 1. Since the augmented function $\hat{\mathcal{J}}$ of Eq.(7) is bounded below and minimized in each iteration, the sequences generated by Algorithm 1 will be converging, *i.e.*, $\hat{\mathcal{J}}\left(\mathbf{U}^{t+1},\mathbf{V}^{t+1},\mathbf{P}^{t+1},\mathbf{W}^{t+1},\mathbf{r}^{t+1}\right)\leq\hat{\mathcal{J}}\left(\mathbf{U}^t,\mathbf{V}^t,\mathbf{P}^t,\mathbf{W}^t,\mathbf{r}^t\right)$. Concrete analysis for each variable can be yielded by referring to: unsupervised feature selection with K-means clustering/exhaustive search (Wang et al. 2015), half-quadratic theory for $l_{2,1}$-norm (He et al. 2014), eigen decomposition with an $l_{2,1}$-norm regularizer (Yang et al. 2011), and self-paced learning (Jiang et al. 2015).

Generally, $c\ll\min(d_1,n)$. $\mathbf{V}$ is sparse, *i.e.*, each column $\mathbf{V}_{\cdot i}$ has only one non-zero element. The computational cost for $\mathbf{U}$ is $\mathcal{O}(ncd_2)$, which is the same as $\mathbf{V}$. The computational cost for $\mathbf{P}$ is $\mathcal{O}(d_1d_2)$, which is highly related to the $l_{2,1}$ norm of $\mathbf{W}\in\mathbb{R}^{d_1\times d_2}$. The computational cost for $\mathbf{W}$ is $\mathcal{O}(d_1^3)$, which involves the eigen decomposition. The

[3] The experimental results are insensitive to $s\in[0.7,0.9]$ since there are not so many outliers or noise in our datasets.

Table 1: Description of Benchmark Datasets.

Dataset	# of Samples	# of Features	# of Classes	Type
LUNG	203	3312	5	cancer
COIL20	1440	1024	20	object
Isolet1	1560	617	26	spoken letter
USPS	9298	256	10	written digit
AT&T	400	644	40	human face
UMIST	575	644	20	human face

computational cost for $\mathbf{r}$ is $\mathcal{O}(nd_2)$. Therefore, the total time complexity of our algorithm is $\mathcal{O}(d_1^3)$ for each iteration. The overall time cost tends to be small since our algorithm converges in a few iterations in the experiments.

Moreover, we can list the time complexities of our competitors: ① MCFS (Cai, Zhang, and He 2010): $\mathcal{O}(d_1 n^2 + cd_2^3 + cnd_2^2 + d_1 log\, d_1)$. ② UDFS (Yang et al. 2011): $\mathcal{O}(d_1^3)$. ③ NDFS (Li et al. 2012): $\mathcal{O}(cn + d_1^3)$. ④ RUFS (Qian and Zhai 2013): $\mathcal{O}(cn^2 + d_1^3)$. ⑤ SOCFS (Han and Kim 2015): $\mathcal{O}(c^2 n + n^3)$. Hence, our proposed method has an acceptable time complexity.

4 Experiments

4.1 Experimental Setup

Datasets. We utilize six datasets: LUNG (Bhattacharjee et al. 2001), COIL20 (Nene et al. 1996), Isolet1 (Fanty and Cole 1990), USPS (Hull 1994), AT&T (Samaria and Harter 1994), and UMIST[4]. Detailed information is in Table 1.

Comparing algorithms. Our method is compared to the following unsupervised feature selection algorithms:

① All Features: All original features are used as baseline.

② Laplacian Score (He, Cai, and Niyogi 2005): Features corresponding to the largest Laplacian scores are selected to preserve the local manifold structure well.

③ MCFS (Cai, Zhang, and He 2010): Features are selected by sparse regression and spectral analysis.

④ UDFS (Yang et al. 2011): Features are selected by joint $l_{2,1}$-norm minimization and discriminative analysis.

⑤ NDFS (Li et al. 2012): Features are selected by joint $l_{2,1}$-norm regression and nonnegative spectral analysis.

⑥ RUFS (Qian and Zhai 2013): Features are selected by joint $l_{2,1}$-norm regression and NMF with local learning.

⑦ SOCFS (Han and Kim 2015): Features are selected by bi-orthogonal semi-NMF.

We also evaluate UDFS, NDFS, RUFS, and SOCFS with feature-level graph regularizations. Note that (Zheng et al. 2019) is like a method combining SOCFS and doublet-induced graph. Our proposed framework (7) has two variants: ① $\alpha = 0$; ② doublet-induced (similarity) graph. We also evaluate them on benchmark datasets.

Settings. Some parameters need to be pre-determined. Consistent with (Han and Kim 2015), d_2 is set as the number of latent clusters c. The number of neighboring parameter k is set to 5 for all related methods on all datasets to specify the size of neighborhood. Due to dimension limitation, the number of selected features is set as $\{50, 80, \cdots, 200\}$ for the USPS dataset. For other datasets, we set the numbers of selected features as $\{50, 100, \cdots, 300\}$. For NDFS, we fix $\gamma = 10^8$ to guarantee the orthogonality. For similarity graph, we employ Gaussian kernel with $\sigma = 1$ to calculate the similarity $\mathbf{M}_{ij}$. Other parameters may be different for each method on these datasets. We report the best results from the optimal parameters for all algorithms. All the results in the tables and figures are produced by their published works. Our experiments adopt K-means algorithm whose performance depends on initialization. Following (Cai, Zhang, and He 2010), we repeat all experiments 20 times with random initialization. Normalized Mutual Information (NMI) (Cai, Zhang, and He 2010) is employed to measure the performance in clustering. The larger NMI is, the better performance is.

[4]http://www.sheffield.ac.uk/eee/research/iel/research/face

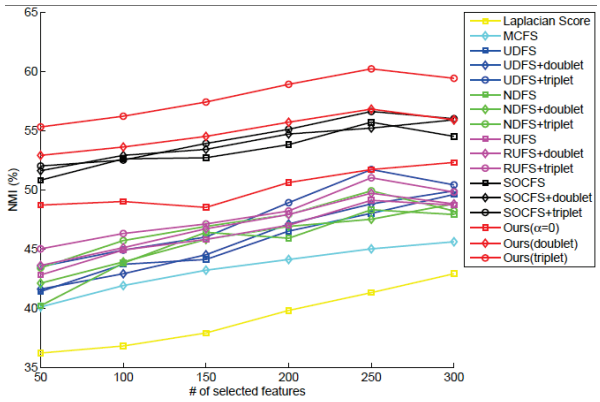
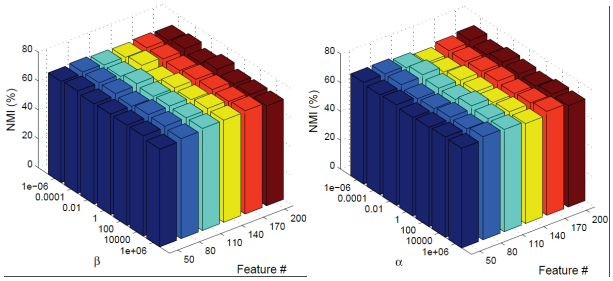


Figure 1: NMI of different methods with different selected feature numbers over LUNG dataset.



(a) $\alpha = 1$ (b) $\beta = 100$

Figure 2: NMI over USPS dataset with different β, α, and selected feature numbers.

4.2 Clustering with Selected Features

The comparison results are reported in Table 2. The number in the parentheses denotes the number of selected features when the performance is achieved. Table 2 significantly demonstrates the superiority of our approach on all datasets. We have three additional **observations** based on comparison in Table 2. First of all, simultaneous feature selection and clustering achieves competitive results than select-

Table 2: Clustering results (NMI%±STD). The best results are in boldface.

	LUNG	COIL20	Isolet1	USPS	AT&T	UMIST
All Features	51.7±5.4	76.3±1.8	75.9±1.6	60.9±0.8	80.5±1.8	42.1±2.3
Laplacian Score	42.9±5.0 (300)	71.8±2.0 (300)	73.1±1.5 (300)	59.5±2.1 (200)	80.4±1.8 (300)	45.1±3.4 (200)
MCFS	45.6±4.5 (300)	74.9±2.2 (150)	74.4±1.9 (200)	61.2±1.8 (200)	80.2±1.9 (200)	45.1±3.2 (150)
UDFS	49.6±5.1 (300)	74.7±1.6 (300)	73.6±1.6 (300)	56.8±1.4 (200)	80.6±1.8 (150)	44.9±2.7 (300)
UDFS + doublet	49.9±5.0 (300)	75.0±1.8 (300)	73.9±1.9 (250)	57.0±1.5 (200)	81.2±1.9 (200)	45.1±2.9 (250)
UDFS + triplet	51.7±5.1 (250)	75.4±1.7 (250)	74.4±1.7 (250)	57.5±1.5 (170)	82.5±1.8 (150)	45.6±2.7 (300)
NDFS	48.3±5.2 (250)	76.0±1.6 (300)	78.4±1.8 (250)	60.7±1.3 (140)	80.3±1.8 (300)	47.8±3.1 (150)
NDFS + doublet	48.8±5.0 (300)	76.2±1.7 (250)	78.8±1.8 (300)	62.7±1.5 (170)	80.9±2.0 (300)	48.0±2.9 (150)
NDFS + triplet	49.9±5.0 (250)	76.9±1.9 (250)	79.1±1.7 (250)	63.5±1.3 (140)	82.2±1.9 (300)	48.5±2.8 (200)
RUFS	49.1±5.1 (250)	77.0±2.2 (150)	78.9±1.1 (300)	61.5±1.4 (170)	80.9±1.7 (300)	46.4±3.0 (150)
RUFS + doublet	49.7±5.2 (250)	77.3±2.4 (200)	79.2±1.3 (250)	61.9±1.7 (200)	81.1±1.7 (300)	46.9±3.1 (200)
RUFS + triplet	51.0±5.0 (250)	77.8±2.1 (150)	79.7±1.2 (250)	62.5±1.6 (170)	82.3±1.7 (300)	47.2±3.0 (200)
SOCFS	55.7±6.2 (250)	74.8±2.3 (300)	78.3±1.9 (300)	61.6±1.4 (110)	81.1±1.6 (100)	49.4±3.2 (50)
SOCFS + doublet	55.9±6.0 (300)	75.0±2.2 (250)	79.2±2.0 (300)	61.9±1.1 (110)	81.4±1.3 (200)	50.0±3.0 (100)
SOCFS + triplet	56.6±5.9 (250)	75.3±2.1 (250)	80.0±2.0 (250)	62.2±1.0 (110)	82.3±1.2 (100)	50.3±3.0 (100)
Ours ($\alpha = 0$)	52.3±6.3 (300)	74.7±2.6 (250)	77.3±2.1 (250)	62.1±1.7 (200)	79.8±1.9 (150)	48.3±3.5 (50)
Ours (doublet)	56.8±6.1 (250)	77.5±2.3 (250)	78.9±2.0 (300)	62.9±1.5 (200)	83.6±1.6 (200)	51.5±3.3 (100)
Ours (triplet)	**60.2±5.8** (250)	**80.1±2.2** (200)	**82.2±1.6** (200)	**64.5±1.0** (200)	**86.2±1.6** (200)	**52.6±3.1** (100)

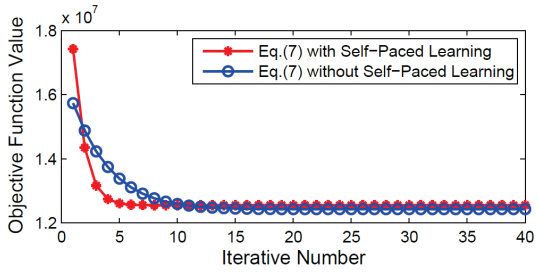


Figure 3: Convergence curves over COIL20 dataset.

ing features one by one or using two-step strategies. Then, uncovering feature-level geometric information of data improves the performance of state-of-the-art methods. Last but not least, triplet-induced graph outperforms doublet-induced graph for unsupervised feature selection.

Figure 1 contains the clustering results of NMI under each selected feature number setting from various unsupervised feature selection methods over the LUNG dataset. The concrete **explanation and analysis** is as follows. Firstly, our method learns the feature selection matrix and the pseudo-labels simultaneously, which can select discriminative features in unsupervised cases. Secondly, our method substitutes the orthogonal constraint on latent cluster centers by directly projecting data into an orthogonal subspace. Thus, orthogonal basis learning and feature selection are naturally combined together. Thirdly, our method enforces the projection vectors to preserve feature-level ordinal information of original data, which contributes to selecting more relevant features. Our work is the continuation and distillation of the previous structure-based feature selection methods.

4.3 Parameter Sensitivity and Convergence Study

We find that the numbers of hyper-parameters for MCFS, UDFS, NDFS, RUFS, and SOCFS are 4, 4, 6, 5, and 5, respectively. However, not all of them are major parameters to be fine-tuned. Hence, these papers do not list all hyper-parameters in their algorithms. Parameters for minor cases, such as determining convergence tolerance and avoiding singularity or zero-denominator, can be set to small values, *e.g.*, 10^{-7}. The number of selected features m should be determined by users. In our experiments, all methods can have the same range of m for fairness, *e.g.*, $\{50, 80, \cdots, 200\}$ for USPS dataset and $\{50, 100, \cdots, 300\}$ for other datasets. The number of latent clusters c is given prior, which is common in existing works. Consistent with the settings in RUFS and SOCFS, the dimension of projected space d_2 is set as the number of latent clusters c. The number of neighboring parameter k is set to 5 for all used datasets to specify the size of neighborhood. This setting is also consistent to previous works, *e.g.*, MCFS, NDFS, and RUFS. The remaining parameters are major parameters that should be fine-tuned for each method. If there is no specific constraint, all methods will tune them in a range of $\{10^{-6}, 10^{-4}, \cdots, 10^{6}\}$.

Then, we study the sensitivity of parameters for our proposed method. As aforementioned, we follow SOCFS (Han and Kim 2015) to set the dimension of projected space d_2 as the number of latent clusters c, meanwhile, k is set to 5 for all methods over all datasets for fair comparison. Parameter $\mu \in [1.1, 1.5]$ is a step-size to monotone increase λ from an initial value 10^{-6}. We just follow previous work (Fan et al. 2017b) to set μ and initialize λ. They are common parameters in self-paced learning, which do not need to be tuned.

Hence, there are two major parameters to be fine-tuned in our algorithm, *i.e.*, β and α. Consistent with previous works UDFS, NDFS, RUFS, and SOCFS, we tune them by grid-search strategy in the range of $\{10^{-6}, 10^{-4}, \cdots, 10^{6}\}$. We

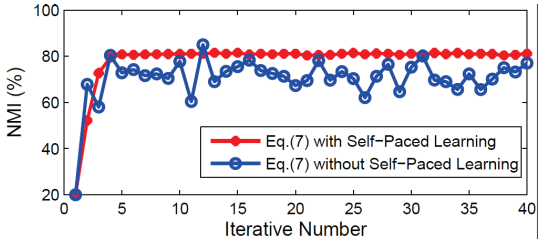


Figure 4: Tendency of NMI over COIL20 dataset.

report the NMI results of USPS dataset and similar trends can be observed on other datasets as well. The experiment results are shown in Figure 2. We can see that our method is not sensitive to β and α with relatively wide ranges.

We also find that our proposed algorithm converges rapidly. It converges in less than 40 iterations in most of our experiments. Without loss of generality, we report the convergence curves over COIL20 dataset in Figure 3. To better understand the behavior of Self-Paced Learning (SPL), we plot the tendency curves of NMI for our model (7) in Figure 4. The clustering performance improves rapidly in the first few iterations because more and more 'easy' samples are selected in these phases. As the iteration number increases, more and more samples are considered for optimizing. Due to some 'complex' samples or outliers, the improvements gradually become steady and inconspicuous.

5 Related Work

Recent years have witnessed many efforts devoted to the "embedding" branch of unsupervised feature selection. Due to 8-page limitation, we only choose some methods that are closely related to our proposed approach. These investigations have emerged to leverage the manifold structure and sparse learning mechanism.

Cai *et al.* (2010) proposed an unsupervised feature selection method to preserve the multi-cluster structure of data. In (Yang et al. 2011), the local discriminative score was introduced to reflect structure information with an $l_{2,1}$ regularizer. In (Li et al. 2012), the local discriminative information, manifold structures, and features' correlations were simultaneously exploited. Qian and Zhai (2013) jointly performed robust label learning and robust feature learning. Wang *et al.* (2015) directly embedded feature selection into clustering via sparse learning without projection. Han and Kim (2015) conducted simultaneous orthogonal basis clustering and feature selection by estimating the latent cluster centers for the projected data. In (Du and Shen 2015; Nie, Zhu, and Li 2016), the structure of selected features was determined by the adaptively learned similarity matrix. Liu *et al.* (2016) employed consensus clustering for pseudo-labeling and feature selection. Guo *et al.* (2017) defined a new type of graph to preserve the ordinal locality of samples for unsupervised feature selection. In (Zhu et al. 2017), a hypergraph was adaptively learned to better uncover the structure of unlabeled multimedia data when selecting features. Li *et al.* (2018) exploited the shared features by all instances and instance-specific features tailored to each instance. However, most of these previous works focus too much on various sample-level structures, totally neglecting the feature-level geometric information.

6 Conclusion

In this paper, we proposed a novel self-paced unsupervised feature selection method to facilitate and simplify clustering tasks. A triplet-induced graph has been proposed to enforce the projection vectors to preserve the feature-level ordinal consensus of original data, which contributes to selecting more relevant features. Meanwhile, we have designed a compelling self-paced regularizer, resulting in a robust framework of simultaneous feature selection and clustering. Based on alternative search strategy, an iterative minimization algorithm has been developed for efficient optimization. Extensive experiments demonstrate the effectiveness of uncovering feature-level geometric structures for unsupervised feature selection. Comparison results also validate that our method outperforms the state-of-the-art alternatives.

Acknowledgments

This work is funded by National Natural Science Foundation of China Major Project No. U1611461 and National Program on Key Basic Research Project No. 2015CB352300. Jun Guo is partially supported by the 2018 Tencent Rhino-Bird Elite Training Program. We would like to thank the anonymous reviewers for their helpful comments. We also thank Prof. Yi Ma from UC Berkeley, Prof. Shengyu Zhang and Dr. Guangyong Chen from Tencent for the valuable discussions. We sincerely thank Prof. Ran He from Institute of Automation, Chinese Academy of Sciences (CASIA), Prof. Yanqing Guo from Dalian University of Technology (DUT), and Prof. Xiangwei Kong from Zhejiang University for reviewing an earlier version of this paper.

References

Bhattacharjee, A.; Richards, W. G.; Staunton, J.; Li, C.; Monti, S.; Vasa, P.; Ladd, C.; Beheshti, J.; Bueno, R.; Gillette, M.; et al. 2001. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *National Academy of Sciences (NAS)* 98(24):13790–13795.

Cai, D.; Zhang, C.; and He, X. 2010. Unsupervised feature selection for multi-cluster data. In *SIGKDD*, 333–342.

Chang, X.; Nie, F.; Yang, Y.; and Huang, H. 2014. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, 1171–1177.

Cheng, K.; Li, J.; and Liu, H. 2017. Unsupervised feature selection in signed social networks. In *SIGKDD*, 777–786.

Dash, M.; Choi, K.; Scheuermann, P.; and Liu, H. 2002. Feature selection for clustering − a filter solution. In *ICDM*, 115–122.

Ding, C.; He, X.; and Simon, H. D. 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SIAM Conf. Data Mining*, volume 5, 606–610.

Du, L., and Shen, Y. D. 2015. Unsupervised feature selection with adaptive structure learning. In *SIGKDD*, 209–218.

Du, X.; Nie, F.; Wang, W.; Yang, Y.; and Zhou, X. 2018. Exploiting combination effect for unsupervised feature selection by $l_{2,0}$ norm. *TNNLS*.

Fan, M.; Chang, X.; Zhang, X.; Wang, D.; and Du, L. 2017a. Top-k supervise feature selection via ADMM for integer programming. In *IJCAI*, 1646–1653.

Fan, Y.; He, R.; Liang, J.; and Hu, B. 2017b. Self-paced learning: an implicit regularization perspective. *AAAI* 1877–1883.

Fanty, M., and Cole, R. 1990. Spoken letter recognition. In *NIPS*, 220–226.

Guo, J., and Zhu, W. 2018. Dependence guided unsupervised feature selection. In *AAAI*, 2232–2239.

Guo, J.; Guo, Y.; Kong, X.; Zhang, M.; and He, R. 2016. Discriminative analysis dictionary learning. In *AAAI*, 1617–1623.

Guo, J.; Guo, Y.; Kong, X.; and He, R. 2017. Unsupervised feature selection with ordinal locality. In *ICME*, 1213–1218.

Han, D., and Kim, J. 2015. Unsupervised simultaneous orthogonal basis clustering feature selection. In *CVPR*, 5016–5023.

Han, Y., and Shen, Y. 2016. Partially supervised graph embedding for positive unlabelled feature selection. In *IJCAI*, 1548–1554.

He, R.; Zheng, W.; Tan, T.; and Sun, Z. 2014. Half-quadratic-based iterative minimization for robust sparse representation. *TPAMI* 36(2):261–275.

He, X.; Cai, D.; and Niyogi, P. 2005. Laplacian score for feature selection. In *NIPS*, 507–514.

Hou, C.; Nie, F.; Li, X.; Yi, D.; and Wu, Y. 2014. Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *TCYB* 44(6):793–804.

Hull, J. J. 1994. A database for handwritten text recognition research. *TPAMI* 16(5):550–554.

Jian, L.; Li, J.; Shu, K.; and Liu, H. 2016. Multi-label informed feature selection. In *IJCAI*, 1627–1633.

Jiang, L.; Meng, D.; Zhao, Q.; Shan, S.; and Hauptmann, A. G. 2015. Self-paced curriculum learning. In *AAAI*, 2694–2700.

John, G. H.; Kohavi, R.; and Pfleger, K. 1994. Irrelevant features and the subset selection problem. In *ICML*, 121–129.

Kumar, M. P.; Packer, B.; and Koller, D. 2010. Self-paced learning for latent variable models. In *NIPS*, 1189–1197.

Law, M. H. C.; Figueiredo, M. A. T.; and Jain, A. K. 2004. Simultaneous feature selection and clustering using mixture models. *TPAMI* 26(9):1154–1166.

Li, Z.; Yang, Y.; Liu, J.; Zhou, X.; and Lu, H. 2012. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*, 1026–1032.

Li, J.; Wu, L.; Dani, H.; and Liu, H. 2018. Unsupervised personalized feature selection. In *AAAI*, 3514–3521.

Liu, H., and Motoda, H. 2007. Computational methods of feature selection. Technical report, CRC Press.

Liu, H.; Shao, M.; and Fu, Y. 2016. Consensus guided unsupervised feature selection. In *AAAI*, 1874–1880.

Nene, S. A.; Nayar, S. K.; Murase, H.; et al. 1996. Columbia object image library (COIL-20). Technical report, CUCS-005-96.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. 2010. Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. In *NIPS*, 1813–1821.

Nie, F.; Zhu, W.; and Li, X. 2016. Unsupervised feature selection with structured graph optimization. In *AAAI*, 1302–1308.

Qian, M., and Zhai, C. 2013. Robust unsupervised feature selection. In *IJCAI*, 1621–1627.

Roth, V., and Lange, T. 2004. Feature selection in clustering problems. In *NIPS*, 473–480.

Samaria, F. S., and Harter, A. C. 1994. Parameterisation of a stochastic model for human face identification. In *Workshop on Applicat. Comput. Vision*, 138–142.

Wang, S.; Nie, F.; Chang, X.; Yao, L.; Li, X.; and Sheng, Q. Z. 2015. Unsupervised feature analysis with class margin optimization. In *ECML/PKDD*, 383–398.

Wang, K.; He, R.; Wang, L.; Wang, W.; and Tan, T. 2016. Joint feature selection and subspace learning for cross-modal retrieval. *TPAMI* 38(10):2010–2023.

Wang, D.; Nie, F.; and Huang, H. 2014. Unsupervised feature selection via unified trace ratio formulation and k-means clustering (TRACK). In *ECML/PKDD*, 306–321.

Wang, S.; Tang, J.; and Liu, H. 2015. Embedded unsupervised feature selection. In *AAAI*, 470–476.

Wei, X.; Xie, S.; Cao, B.; and Yu, P. S. 2017. Rethinking unsupervised feature selection: From pseudo labels to pseudo must-links. In *ECML/PKDD*, 272–287.

Witten, D. M., and Tibshirani, R. 2012. A framework for feature selection in clustering. *Journal of the American Statistical Association* 105(490):713–726.

Xu, Z.; King, I.; Lyu, M. R. T.; and Jin, R. 2010. Discriminative semi-supervised feature selection via manifold regularization. *TNN* 21(7):1033–1047.

Xu, C.; Tao, D.; and Xu, C. 2015. Multi-view self-paced learning for clustering. In *IJCAI*, 3974–3980.

Yan, S.; Xu, D.; Zhang, B.; Zhang, H. J.; Yang, Q.; and Lin, S. 2007. Graph embedding and extensions: A general framework for dimensionality reduction. *TPAMI* 29(1):40–51.

Yang, Y.; Shen, H.; Ma, Z.; Huang, Z.; and Zhou, X. 2011. $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, volume 22, 1589–1594.

Zhao, Q.; Meng, D.; Jiang, L.; Xie, Q.; Xu, Z.; and Hauptmann, A. G. 2015. Self-paced learning for matrix factorization. In *AAAI*, 3196–3202.

Zheng, W.; Zhu, X.; Zhu, Y.; and Zhang, S. 2018. Robust feature selection on incomplete data. In *IJCAI*, 3191–3197.

Zheng, X.; Guo, Y.; Guo, J.; and Kong, X. 2019. $U^2F^2S^2$: Uncovering feature-level similarities for unsupervised feature selection. *Neural Process. Lett.* 49(3):1071–1091.

Zhu, X.; Zhu, Y.; Zhang, S.; Hu, R.; and He, W. 2017. Adaptive hypergraph learning for unsupervised feature selection. In *IJCAI*, 3581–3587.