UNSUPERVISED FEATURE SELECTION WITH ORDINAL LOCALITY

Jun Guo[1], *Yanqing Guo*[2], *Xiangwei Kong*[2], *and Ran He*[3,4,5]

[1]Tsinghua-Berkeley Shenzhen Institute, Tsinghua University
[2]School of Information and Communication Engineering, Dalian University of Technology
[3]National Laboratory of Pattern Recognition, CASIA
[4]Center for Excellence in Brain Science and Intelligence Technology, CAS
[5]University of Chinese Academy of Sciences (UCAS)
eeguojun@outlook.com, {guoyq,kongxw}@dlut.edu.cn, rhe@nlpr.ia.ac.cn

ABSTRACT

Unsupervised feature selection has shown significant potential in distance-based clustering tasks. This paper proposes a novel triplet induced method. Firstly, a triplet-based loss function is introduced to enforce the selected feature groups to preserve ordinal locality of original data, which contributes to distance-based clustering tasks. Secondly, we simplify the orthogonal basis clustering by imposing an orthogonal constraint on the feature projection matrix. Consequently, a general framework for simultaneous feature selection and clustering is discussed. Thirdly, an alternating minimization algorithm is employed to efficiently optimize the proposed model together with rapid convergence. Extensive comparison experiments on several benchmark datasets well validate the encouraging gain in clustering from our proposed method.

Index Terms— Unsupervised feature selection, clustering, triplet, ordinal locality.

1. INTRODUCTION

Feature selection plays an important role in multimedia applications [1]. In practice, high-dimensional features are often redundant, correlated, or even noisy [2, 3],which may lead to adverse effects such as heavy computational complexity and poor classification/clustering performance. Therefore, many feature selection methods are proposed to better explore the properties of high-dimensional data.

Feature selection can be generally grouped into two major categories in terms of label availability, *i.e.*, supervised and unsupervised. Supervised feature selection [4, 5, 6] aims to select discriminative features since the class labels of data containing the essential discrimination are provided. However, the label information is limited, which makes feature-selection-based tasks more challenging. In contrast, unsupervised feature selection is desired to filter out the unimportant features of unlabeled data in unsupervised scenarios.

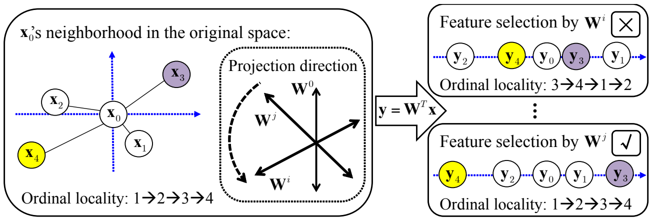
$\mathbf{x}_0$'s neighborhood in the original space:
$\mathbf{x}_3$
$\mathbf{x}_2$
$\mathbf{x}_0$
$\mathbf{x}_1$
$\mathbf{x}_4$
Ordinal locality: 1→2→3→4
Projection direction
$\mathbf{W}^0$
$\mathbf{W}^j$
$\mathbf{W}^i$
$\mathbf{y} = \mathbf{W}^T\mathbf{x}$
Feature selection by $\mathbf{W}^i$ ☒
$\mathbf{y}_2$ $\mathbf{y}_4$ $\mathbf{y}_0$ $\mathbf{y}_3$ $\mathbf{y}_1$
Ordinal locality: 3→4→1→2
⋮
Feature selection by $\mathbf{W}^j$ ✓
$\mathbf{y}_4$ $\mathbf{y}_2$ $\mathbf{y}_0$ $\mathbf{y}_1$ $\mathbf{y}_3$
Ordinal locality: 1→2→3→4

Fig. 1. A basic illustration. $\{\mathbf{x}_i\}_{i=1}^4$ are $\mathbf{x}_0$'s neighbors. Different colors stand for different clusters. Our proposed feature selection method contributes to distance-based clustering by preserving the ranking of $\mathbf{x}_0$'s neighbors as much as possible.

Unsupervised feature selection methods have drawn much attention. Besides filters [7], wrappers [8], and embedding [9], recent investigations leverage the manifold structure and sparse learning mechanism. There are so many works in this topic. Due to page limitation, we only revisit some methods closely related to our work. Cai *et al.* [10] proposed a feature selection approach to preserve the multi-cluster structure of data. Yang *et al.* [11] employed local discriminative score to reflect structure information with an $l_{2,1}$ regularizer. Li *et al.* [12] exploited local discriminative information, manifold structures and feature correlations simultaneously. Qian and Zhai [13] jointly performed robust label learning and robust feature learning. Han and Kim [14] conducted simultaneous orthogonal basis clustering and feature selection by estimating latent cluster centers for the projected data. Wang *et al.* [15] directly applied feature selection into clustering via projection-free sparse learning. Nie *et al.* [16] determined the structure of selected features by adaptively learning a similarity matrix. Liu *et al.* [17] employed consensus clustering for pseudo-labeling and feature selection.

The aforementioned unsupervised feature selection methods benefit from various geometrical structures of data. However, the ranking (or topology) information in each sample's neighborhood is often not preserved in many feature selection methods. This kind of topology information, named ***ordinal locality***, is of great importance for distance-based clustering tasks, where samples within the same cluster are close to each

978-1-5090-6067-2/17/$31.00 ©2017 IEEE

other and samples from different clusters are far away. Hence, feature selection poses a critical issue for distance-based clustering tasks, in which ordinal locality is important.

This paper addresses the above issue. A triplet induced method is proposed to select feature groups for distance-based clustering. Our main contributions lie in three-folds:

1) We propose a triplet-based ordinal locality preserving loss function to capture the underlying local structures of data during feature selection. It preserves the relative neighborhood proximities and contributes to distance-based clustering.

2) To facilitate and simplify simultaneous clustering and feature selection, we explicitly impose an orthogonal constraint on the projection matrix and obtain a general framework for simultaneous feature selection and clustering.

3) Based on alternative search strategy, we develop an alternating minimization algorithm to speed up the optimization process. Extensive experiments demonstrate that feature selection can be more conducive to distance-based clustering tasks if ordinal locality is well treated.

2. PRELIMINARIES

2.1. Notation Summary

Bold uppercase letters $(\mathbf{M}, \cdots)$ are matrices. Bold lowercase letters $(\mathbf{m}, \cdots)$ denote vectors. Italic letters $(m, \cdots)$ are scalars. $\mathbf{M}^T$, $\mathbf{M}^{-1}$ and $Tr\left(\mathbf{M}\right)$ are the transpose, inverse and trace of $\mathbf{M}$, respectively. $\mathbf{M}_{i\cdot}$ presents the i^{th} row of $\mathbf{M}$, while $\mathbf{M}_{\cdot j}$ means the j^{th} column of $\mathbf{M}$. $\mathbf{M}_{ij}$ is the j^{th} element in the i^{th} row of $\mathbf{M}$. $\|\mathbf{M}\|_F$ and $\|\mathbf{M}\|_0$ denote the Frobenius norm ($\sqrt{\sum_{i,j}\mathbf{M}_{ij}^2}$) and l_0-norm (number of nonzero entries), respectively. $\|\mathbf{M}\|_{2,0}$ and $\|\mathbf{M}\|_{2,1}$ stand for the $l_{2,0}$-norm ($\sum_i \left\|\sqrt{\sum_j \mathbf{M}_{ij}^2}\right\|_0$) and $l_{2,1}$-norm ($\sum_i \sqrt{\sum_j \mathbf{M}_{ij}^2}$), respectively. $\mathbf{0}$, $\mathbf{1}$ and $\mathbf{I}$ denote the all zeros, all ones and identity matrix with compatible sizes, respectively. $\mathbf{M} \geq \mathbf{0}$ means that all elements of $\mathbf{M}$ are non-negative.

2.2. Regression Based Feature Selection

$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d_1 \times n}$ is the original data matrix with n samples. Based on the regularized regression, feature selection is generally formulated as $\min_{\mathbf{W}} \left\|\mathbf{W}^T\mathbf{X} - \mathbf{H}\right\|_F^2 + \beta\|\mathbf{W}\|_{2,q}$, where $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ $(d_1 > d_2)$ is a projection matrix, namely feature selection matrix. The $l_{2,q}$-norm (q is typically set to 0 or 1) imposed on $\mathbf{W}$ guarantees the sparseness in rows. $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_n] \in \mathbb{R}^{d_2 \times n}$ is a target matrix. When handling multi-class data in a supervised fashion, $\mathbf{H}$ is usually the corresponding label matrix. $\mathbf{W}_{i\cdot}$ will shrink to $\mathbf{0}$ when the i^{th} feature is less correlated to the labels.

In an unsupervised fashion, a frequently used strategy to determine $\mathbf{H}$ is learning accurate pseudo labels by classical machine learning algorithms, such as linear regression [11], spectral clustering [12], and K-means clustering [13]. The recently presented SOCFS [14] used bi-orthogonal semi Nonnegative Matrix Factorization (NMF) to decompose $\mathbf{H}$ into two new matrices: the latent orthogonal bases $\mathbf{U}$ and the pseudo-label indicators $\mathbf{V}$, *i.e.*, $\mathbf{H} \simeq \mathbf{UV}$ with the following constraints $\mathbf{V} \geq \mathbf{0}$, $\mathbf{V}\mathbf{V}^T = \mathbf{I}$, and $\mathbf{U}^T\mathbf{U} = \mathbf{I}$.

3. THE PROPOSED METHOD

Figure 1 gives an intuitive explanation. For distance-based clustering tasks, samples within the same cluster are close to each other and samples from different clusters are far away. In Figure 1, it is $\mathbf{W}^j$ rather than $\mathbf{W}^i$ that has the power to enforce the selected feature groups to preserve the important topology property (ordinal locality) of original data.

3.1. Triplet-Induced Ordinal Locality

Ordinal locality describes the local structures of data. Besides the neighborhood relationships between data points, it emphasizes the ranking information of each data point's neighbors. Different from similarity/distance preserving loss functions based on doublets, we novelly propose an ordinal locality preserving loss function with triplets in this section.

Suppose the selected feature group for arbitrary original sample $\mathbf{x}_i$ is denoted by $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$, thus $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$. We first provide a definition of ***ordinal locality preserving***.

Definition 1. Given a triplet $(\mathbf{x}_i, \mathbf{x}_u, \mathbf{x}_v)$ comprised of $\mathbf{x}_i$ and its neighbors $\mathbf{x}_u$ and $\mathbf{x}_v$, their corresponding selected feature groups also form a triplet $(\mathbf{y}_i, \mathbf{y}_u, \mathbf{y}_v)$. Let $dist\left(\cdot,\cdot\right)$ be a distance metric. Then, we say that the feature selection process is ***ordinal locality preserving*** when the following condition holds: if $dist\left(\mathbf{x}_i, \mathbf{x}_u\right) \leq dist\left(\mathbf{x}_i, \mathbf{x}_v\right)$, then $dist\left(\mathbf{y}_i, \mathbf{y}_u\right) \leq dist\left(\mathbf{y}_i, \mathbf{y}_v\right)$. □

Based on Definition 1 and the rearrangement inequality[1], determining appropriate feature groups for each data point is identical to optimize the following ordinal locality preserving loss function (1) over a collection of triplets.

$$\max_{\mathbf{Y}} \sum_{i=1}^{n} \sum_{u \in \mathcal{N}_i} \sum_{v \in \mathcal{N}_i} \mathbf{S}_{uv}^i \left[dist\left(\mathbf{y}_i, \mathbf{y}_u\right) - dist\left(\mathbf{y}_i, \mathbf{y}_v\right)\right], \qquad (1)$$

where $\mathcal{N}_i$ denotes a set of sequence numbers indicating the k nearest neighbors of $\mathbf{x}_i$. $\mathbf{S}^i$ is an antisymmetric matrix with $(u, v)^{th}$ element equals $dist\left(\mathbf{x}_i, \mathbf{x}_u\right) - dist\left(\mathbf{x}_i, \mathbf{x}_v\right)$. Inspired by [18], we denote $\mathbf{C} \in \mathbb{R}^{n \times n}$ as a weighting matrix with

$$\mathbf{C}_{ij} = \begin{cases} \sum\limits_{u \in \mathcal{N}_i} \mathbf{S}_{uj}^i & , j \in \mathcal{N}_i \\ 0 & , j \notin \mathcal{N}_i \end{cases}. \qquad (2)$$

Proposition 1. *The ordinal locality preserving loss function (1) is equivalent to* $\min_{\mathbf{Y}} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{C}_{ij} dist\left(\mathbf{y}_i, \mathbf{y}_j\right)$.

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

Proof. Recall that $\mathbf{S}^i$ is an antisymmetric matrix, so $\mathbf{S}^i_{uv} = -\mathbf{S}^i_{vu}$. Then, (1) is equivalent to

$$\max_{\mathbf{Y}} \left\{ \begin{array}{l} -\sum\limits_{i=1}^{n} \sum\limits_{u \in \mathcal{N}_i} \sum\limits_{v \in \mathcal{N}_i} \mathbf{S}^i_{vu} dist\left(\mathbf{y}_i, \mathbf{y}_u\right) \\ -\sum\limits_{i=1}^{n} \sum\limits_{v \in \mathcal{N}_i} \sum\limits_{u \in \mathcal{N}_i} \mathbf{S}^i_{uv} dist\left(\mathbf{y}_i, \mathbf{y}_v\right) \end{array} \right\}, \tag{3}$$

which can be reformulated as (4) according to (2).

$$\max_{\mathbf{Y}} \left\{ \begin{array}{l} -\sum\limits_{i=1}^{n} \sum\limits_{u=1}^{n} \mathbf{C}_{iu} dist\left(\mathbf{y}_i, \mathbf{y}_u\right) \\ -\sum\limits_{i=1}^{n} \sum\limits_{v=1}^{n} \mathbf{C}_{iv} dist\left(\mathbf{y}_i, \mathbf{y}_v\right) \end{array} \right\} \tag{4}$$

The ultimate form $\min_{\mathbf{Y}} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{C}_{ij} dist\left(\mathbf{y}_i, \mathbf{y}_j\right)$ can be easily derived from (4). □

For simplicity, we use squared Euclidean distance to establish each pairwise distance. Then, the aforementioned ordinal locality preserving loss function can be consequently written as $\min_{\mathbf{Y}} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{C}_{ij} \left\|\mathbf{y}_i - \mathbf{y}_j\right\|_2^2$, which has an equivalent compact matrix form: $\min_{\mathbf{Y}} Tr\left(\mathbf{YLY}^T\right)$ as well as $\min_{\mathbf{W}} Tr\left(\mathbf{W}^T\mathbf{XLX}^T\mathbf{W}\right)$. $\mathbf{L} \triangleq \mathbf{D} - \frac{\mathbf{C}+\mathbf{C}^T}{2}$ is the Laplacian matrix and $\mathbf{D}$ is a diagonal matrix whose $(i,i)^{th}$ element equals $\sum_{j=1}^{n} \frac{\mathbf{C}_{ij}+\mathbf{C}_{ji}}{2}$. Note that our proposed loss function has a similar form to doublet-based loss functions with Laplacian matrix. However, it is essentially different from them due to a specific calculation (2) for the weighting matrix.

3.2. Overall Objective Function

We carry forward the idea of decomposing $\mathbf{H}$ but adopt a different strategy from [14]. In [14], the orthogonal and nonnegative constraints on the pseudo-label indicators $\mathbf{V}$ can ensure that each column of $\mathbf{V}$ has only one non-zero element so that the learned $\mathbf{V}$ can be much closer to the true scaled-label matrix. As indicated in [13], when employing NMF or its variants to decompose $\mathbf{H}$, the adverse effect induced by outliers and noise will be primarily accumulated in the learned cluster centers but will not hurt the indicators severely. Therefore, we focus on fully exploiting the pseudo indicators $\mathbf{V}$, regardless of whether the latent cluster centers $\mathbf{U}$ is orthogonal or not. Based on this consideration, the orthogonal constraint on the latent cluster centers $\mathbf{U}$ is discarded.

What's more, we impose an orthogonal constraint on $\mathbf{W}$ to suppress the feature similarity of arbitrary two selected dimensions. The orthogonal constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ can also avoid arbitrary scaling and the trivial solution of all zeros. Consequently, we obtain the overall formulation:

$$\begin{array}{ll} \min\limits_{\mathbf{W},\mathbf{U},\mathbf{V}} & F = \left\{ \begin{array}{l} \left\|\mathbf{W}^T\mathbf{X} - \mathbf{UV}\right\|_F^2 + \beta\|\mathbf{W}\|_{2,1} \\ +\alpha Tr\left(\mathbf{W}^T\mathbf{XLX}^T\mathbf{W}\right) \end{array} \right\} \\ s.t. & \mathbf{W}^T\mathbf{W} = \mathbf{I},\ \mathbf{V} \geq \mathbf{0},\ \mathbf{VV}^T = \mathbf{I}, \end{array} \tag{5}$$

where β and α are scalar constants which control the relative importance of corresponding terms.

3.3. Discussion

A general framework for simultaneous feature selection and clustering can be derived from our proposed model (5). On the one hand, it can be highly related to K-means clustering on the projected data $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$. On the other hand, it can be incorporated with various graph-based learning methods. Specifically, we employ the similarity-based and maximum-margin-based graphs. Further discussion is as follow:

1) When $\alpha = 0$, model (5) is equivalent to simultaneous feature selection and K-means clustering. This can be well verified by Theorem 1 in the next section.

2) When $\mathbf{C}_{ij} = \left\{ \begin{array}{ll} \exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2/\sigma\right) & , j \in \mathcal{N}_i \\ 0 & , j \notin \mathcal{N}_i \end{array} \right.$ (heat kernel [10]) or $\mathbf{C}_{ij}$ is calculated with other weighting methods in each sample's neighborhood, a similarity-based graph is constructed. With doublet relationship on each pair of samples, data's similarity can be preserved.

3) When $\mathbf{C}_{ij} = -\frac{1}{n}$, a maximum-margin-based graph with $\mathbf{L} = \frac{1}{n}\mathbf{1} - \mathbf{I}$ is constructed. By doublets, the global relationships can be preserved. This can maximize data's total separability meanwhile minimize the within-class scatter, which is consistent with [19].

4) When $\mathbf{C}_{ij}$ is calculated by (2), a triplet-induced graph is constructed. This is our proposed model, which is important for distance-based clustering tasks. With triplet relations in each neighborhood, ordinal locality can be preserved.

4. OPTIMIZATION

4.1. Half-quadratic Technique

According to half-quadratic (HQ) theory [20], we have:

Lemma 1. *For a fixed m, there exists a conjugate function $\psi(\cdot)$, such that $\sqrt{m^2+\varepsilon} = \inf_{r\in\mathbb{R}}\left\{\frac{r}{2}m^2 + \psi(r)\right\}$. The infimum could be reached at $r = 1/\sqrt{m^2+\varepsilon}$.*

The HQ technique can be employed to optimize the problem (5) by alternately minimizing its augmented function $\hat{F}$.

$$\begin{array}{ll} \min\limits_{\mathbf{W},\mathbf{U},\mathbf{V},\mathbf{R}} & \hat{F} = \left\{ \begin{array}{c} \left\|\mathbf{W}^T\mathbf{X} - \mathbf{UV}\right\|_F^2 \\ +\beta \sum\limits_{i=1}^{d_1} \left\{ \frac{\mathbf{R}_{ii}}{2} \left\|\mathbf{W}_{i\cdot}\right\|_2^2 + \psi_i\left(\mathbf{R}_{ii}\right) \right\} \\ +\alpha Tr\left(\mathbf{W}^T\mathbf{XLX}^T\mathbf{W}\right) \end{array} \right\} \\ s.t. & \mathbf{W}^T\mathbf{W} = \mathbf{I},\ \mathbf{V} \geq \mathbf{0},\ \mathbf{VV}^T = \mathbf{I}, \end{array} \tag{6}$$

where $\mathbf{R}$ is a $d_1 \times d_1$ diagonal matrix storing the auxiliary variables. $\{\psi_i\}_{i=1}^{d_1}$ are conjugate functions.

4.2. Optimization Procedure

$\hat{F}\left(\mathbf{W}, \mathbf{U}, \mathbf{V}, \mathbf{R}\right)$ can be alternately minimized as follows:

1) Update all the diagonal elements of $\mathbf{R}$ in parallel:

$$\mathbf{R}_{ii} = 1/\sqrt{\left\|\mathbf{W}_{i\cdot}\right\|_2^2 + \varepsilon}. \tag{7}$$

Algorithm 1 The algorithm to solve Eq.(6)

```
Input:
    Data matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d_1 \times n}$;
    Number of each sample's nearest neighbors $k$;
    Parameters $d_2$, $c$, $\beta$ and $\alpha$.
Output:
    Feature selection matrix $\mathbf{W}$ and cluster indicator matrix $\mathbf{V}$.
Compute $\mathbf{C}$ via (2) and its corresponding Laplacian matrix $\mathbf{L}$;
Initialize $\mathbf{W}^{(0)}$ with $d_2$ different columns randomly selected from a $d_1 \times d_1$ identity matrix, $t = 0$;
while not convergence do
    $t \leftarrow t + 1$;
    Update $\mathbf{R}^{(t)}$ via (7);
    Update $\mathbf{U}^{(t)}$ and $\mathbf{V}^{(t)}$ by K-means;
    Update $\mathbf{W}^{(t)}$ by eigen decomposition;
end while
```

Remark 1. *The $l_{2,1}$-norm of $\mathbf{W}$ equals $\sum_{i=1}^{d_1} \sqrt{\|\mathbf{W}_{i\cdot}\|_2^2}$, whose corresponding minimizer function is infinite near the origin. According to [3, 11], we add a small perturbation ε to each $\|\mathbf{W}_{i\cdot}\|_2^2$. It is obvious that this kind of regularized $l_{2,1}$-norm approximates the $l_{2,1}$-norm of $\mathbf{W}$ when $\varepsilon \rightarrow 0$.*

2) To update $(\mathbf{U}, \mathbf{V})$ with fixed $\mathbf{W}$, we need to conduct orthogonal semi-NMF [21] on projected data $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$.

Theorem 1. *The orthogonal semi-NMF problem: $\min_{\mathbf{U},\mathbf{V}} \|\mathbf{Y} - \mathbf{U}\mathbf{V}\|_F^2$, s.t. $\mathbf{V} \geq \mathbf{0}, \mathbf{V}\mathbf{V}^T = \mathbf{I}$ is equivalent to relaxed K-means clustering.*

Proof. Please refer to [22, 23]. □

Based on Theorem 1, we update $(\mathbf{U}, \mathbf{V})$ via K-means clustering following the settings in the proof. We set the first-order partial derivative of $f(\mathbf{W}, \mathbf{U}, \mathbf{V})$ w.r.t. $\mathbf{U}$ to zero and obtain the zero gradient condition $\mathbf{U} = \mathbf{W}^T\mathbf{X}\mathbf{V}^T$.

3) To update $\mathbf{W}$ with fixed $(\mathbf{U}, \mathbf{V})$, we substitute $\mathbf{U} = \mathbf{W}^T\mathbf{X}\mathbf{V}^T$ into $f(\mathbf{W}, \mathbf{U}, \mathbf{V})$ and solve the objective function $\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} Tr\left(\mathbf{W}^T\mathbf{G}\mathbf{W}\right)$ by performing eigen decomposition on $\mathbf{G} = \frac{\beta}{2}\mathbf{R} + \mathbf{X}\left(\alpha\mathbf{L} + \mathbf{I} - \mathbf{V}^T\mathbf{V}\right)\mathbf{X}^T$. The optimal $\mathbf{W}$ is comprised of d_2 eigenvectors corresponding to the d_2 smallest eigenvalues.

As summarized in Algorithm 1, the aforementioned update steps are alternatively performed until convergence.

4.3. Algorithmic Analysis

According to Lemma 1, with fixed $(\mathbf{W}, \mathbf{U}, \mathbf{V})$, the equation $F(\mathbf{W}, \mathbf{U}, \mathbf{V}) = \inf_{\mathbf{P}} \hat{F}(\mathbf{W}, \mathbf{U}, \mathbf{V}, \mathbf{R})$ holds. It follows that

$$\min_{\mathbf{W},\mathbf{U},\mathbf{V}} F(\mathbf{W}, \mathbf{U}, \mathbf{V}) = \min_{\mathbf{W},\mathbf{U},\mathbf{V},\mathbf{R}} \hat{F}(\mathbf{W}, \mathbf{U}, \mathbf{V}, \mathbf{R}). \quad (8)$$

Therefore, minimizing $F(\mathbf{W}, \mathbf{U}, \mathbf{V})$ is equivalent to minimizing $\hat{F}(\mathbf{W}, \mathbf{U}, \mathbf{V}, \mathbf{R})$ on the enlarged domain. According to the convergence proof in [19] and the properties of HQ [4, 20], we obtain $\hat{F}\left(\mathbf{W}^{(t+1)}, \mathbf{U}^{(t+1)}, \mathbf{V}^{(t+1)}, \mathbf{R}^{(t+1)}\right) \leq \hat{F}\left(\mathbf{W}^{(t)}, \mathbf{U}^{(t)}, \mathbf{V}^{(t)}, \mathbf{R}^{(t+1)}\right) \leq \hat{F}\left(\mathbf{W}^{(t)}, \mathbf{U}^{(t)}, \mathbf{V}^{(t)}, \mathbf{R}^{(t)}\right)$. At each alternative minimization step, the objective function is non-increasing. In practice, we can normalize each nonzero $\mathbf{C}_{ij}$ to $[0, 1]$ by row. Thus, the Laplacian matrix is positive semi-definite and $F(\mathbf{W}, \mathbf{U}, \mathbf{V})$ is bounded below. Based on (8), $\hat{F}(\mathbf{W}, \mathbf{U}, \mathbf{V}, \mathbf{R})$ is also bounded. Algorithm 1 will converge to the local optimum solution and may converge to the global optimum solution.

Table 1. Dataset Description.

Dataset	# of Classes	# of Samples	# of Features
COIL20	20	1440	1024
Isolet1	26	1560	617
LUNG	5	203	3312
USPS	10	9298	256
AT&T	40	400	644
UMIST	20	575	644

The time complexity for $\mathbf{R} \in \mathbb{R}^{d_1 \times d_1}$ is $\mathcal{O}(d_1 d_2)$, which is highly related to the $l_{2,1}$-norm of $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$. We employ K-means clustering to simultaneously update $\mathbf{U} \in \mathbb{R}^{d_2 \times c}$ and $\mathbf{V} \in \mathbb{R}^{c \times n}$, where c is the number of clusters. The corresponding computational cost is $\mathcal{O}(ncd_2T)$, where T is the number of iterations in K-means. The time complexity for $\mathbf{W}$ is $\mathcal{O}(d_1^3)$, which involves the eigen decomposition. Therefore, the overall computational complexity of our algorithm is $\mathcal{O}(d_1^3)$, where d_1 is the dimensionality of original data. Note that the eigen decomposition is the most computational operation. If d_1 is very large, dimensionality reduction for preprocessing is desirable. Fortunately, our proposed triplet-based idea can be easily extended to an effective technique for unsupervised dimensionality reduction.

5. EXPERIMENTS

5.1. Datasets and Comparing Algorithms

In this section, we evaluate the performance of our proposed method on several benchmark datasets: one object image dataset (COIL20[2]), one spoken letter recognition dataset (Isolet1[2]), one cancer dataset (LUNG[2]), one handwritten digit dataset (USPS[2]), and two face image datasets (AT&T[3] and UMIST[4]). Detailed information is listed in Table 1.

Our method is compared to the following unsupervised feature selection algorithms: ① All Features: All original features are adopted as the baseline in the experiments. ② Laplacian Score (LS): Features corresponding to the largest Laplacian scores are selected to preserve the local manifold structure well [24]. ③ Multi-Cluster Feature Selection (MCFS) [10]. ④ Unsupervised Discriminative Feature Selection

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

[3] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

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

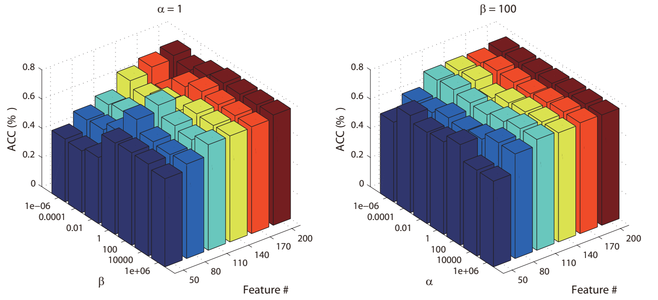

Fig. 2. Clustering accuracy (ACC) over USPS dataset with different β, α and selected feature numbers.

(UDFS) [11]. ⑤ Nonnegative Discriminative Feature Selection (NDFS) [12]. ⑥ Robust Unsupervised Feature Selection (RUFS) [13]. ⑦ Simultaneous Orthogonal Basis Clustering Feature Selection (SOCFS) [14]. Please notice that ③∼⑦ have already been introduced in Section 1. We also evaluate SOCFS with a triplet-induced regularization. As aforementioned, our proposed general framework (5) for simultaneous feature selection and clustering has three variants: ① $\alpha = 0$; ② similarity-based graph; ③ maximum-margin-based graph. We also evaluate them on benchmark datasets.

5.2. Experimental Setup

Following the experimental settings in previous works, accuracy (ACC) is employed to measure the performance in clustering. The larger ACC is, the better performance is. There are some parameters to be set in advance. The dimension of projected space d_2 is set as the number of clusters c. The number of neighboring parameter k is set to 5 for all datasets to specify the size of neighborhood. For similarity-based graph, we calculate $\mathbf{C}_{ij}$ via heat kernel weighting [10] with $\sigma = 1$. For NDFS, we fix $\gamma = 10^8$ to guarantee the orthogonality.

For fair comparison, we tune the parameters for all unsupervised feature selection algorithms by grid-search strategy from $\{10^{-6}, 10^{-4}, 10^{-2}, \cdots, 10^{6}\}$. For USPS dataset, the number of features is set as $\{50, 80, \cdots, 200\}$ due to dimension limitation. We set the numbers of selected features as $\{50, 100, \cdots, 300\}$ for other datasets. Different parameters may be utilized for different datasets. We report the best results from the optimal parameters for all the algorithms. In our experiments, we adopt K-means algorithm whose performance depends on initialization. Following [10], all experiments are repeated 20 times with random initialization. The mean and standard deviation (STD) of clustering accuracy (ACC) for all algorithms are reported. All results in the tables are produced by their published works.

5.3. Results and Analysis

We report the comparison results of clustering in Table 2. We have the following observations. Firstly, preserving the ordinal locality of data in feature selection achieves competitive performance in contrast to state-of-the-art methods. Secondly, triplet-induced feature selection is more effective than other doublet-based methods. Thirdly, simultaneous feature selection and clustering can achieve better results than selecting features one by one or using two-step strategies. Finally, feature selection is necessary, which can not only significantly reduce the number of features, but also improve the performance in distance-based clustering tasks.

As shown in Table 2, our method outperforms its competitors on all datasets. The main reasons are: 1) Our method enforces the selected feature groups to preserve the relative neighborhood proximity of original data. This is effective for distance-based clustering tasks. 2) Our method learns the feature selection matrix and the pseudo-label indicators simultaneously, which can select discriminative features in unsupervised scenarios. 3) Our method substitutes the orthogonal constraint on latent cluster centers by directly projecting data into an orthogonal subspace. Orthogonal basis learning and feature selection are naturally combined together.

Next, we study the sensitiveness of parameters. Due to the space limit, we only report the results in terms of ACC over USPS dataset. The experimental results are shown in Figure 2. From Figure 2, we can see that our method is not sensitive to β and α with relatively wide ranges.

From Figure 3, we also find that our proposed algorithm converges rapidly. It converges in less than 40 iterations in most of our experiments. Due to 6-page limitation, we only report the convergence curve of UMIST dataset.

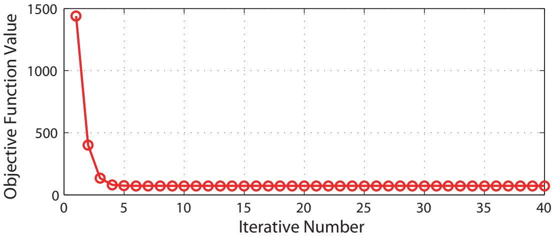

Fig. 3. The convergence curve of UMIST dataset.

6. CONCLUSION AND FUTURE WORK

This paper has proposed a triplet-induced unsupervised feature selection method. A triplet-based loss function has been defined to enforce the selected feature groups to preserve the ordinal locality of original data. Meanwhile, we explicitly introduce a general framework for simultaneous clustering and feature selection. Based on half-quadratic minimization, an alternating optimization algorithm has been developed for efficient optimization. Extensive experiments show that triplet relations contribute to distance-based clustering tasks and our proposed method outperforms state-of-the-art feature-selection-based clustering methods. In the future, it will be interesting to regard each dimension of feature as a data point, and exploit feature-wise topological structures.

Table 2. Clustering results (ACC%±STD) of several unsupervised feature selection algorithms on benchmark datasets. The best results are in boldface. The number in the parentheses denotes the number of selected features for the reported performance.

	COIL20	Isolet1	LUNG	USPS	AT&T	UMIST
All Features	59.4±4.9	57.9±3.6	70.0±8.9	65.7±2.4	60.9±3.4	42.1±2.3
LS	56.3±4.8 (300)	55.6±3.3 (300)	60.1±9.5 (300)	62.5±4.4 (200)	61.3±3.5 (300)	45.1±3.4 (200)
MCFS	58.7±5.3 (300)	60.7±4.0 (150)	64.3±7.9 (300)	65.2±4.2 (200)	61.2±3.7 (200)	45.1±3.2 (150)
UDFS	58.9±5.1 (250)	57.9±3.0 (250)	66.2±7.8 (300)	62.4±3.1 (200)	61.7±3.8 (150)	44.9±2.7 (300)
NDFS	59.2±5.0 (300)	64.6±4.4 (250)	66.9±9.1 (300)	64.9±3.1 (200)	61.4±3.5 (300)	47.8±3.1 (150)
RUFS	59.9±4.9 (250)	62.8±3.8 (300)	68.4±8.3 (250)	65.8±3.1 (170)	61.6±3.2 (300)	46.4±3.0 (150)
SOCFS	60.4±4.7 (300)	64.9±4.4 (300)	74.0±8.9 (300)	66.1±2.0 (170)	62.7±3.1 (100)	49.4±3.2 (50)
SOCFS + triplet	62.4±4.6 (200)	65.1±4.2 (250)	75.1±8.6 (150)	67.6±2.5 (110)	63.8±3.7 (200)	50.2±3.3 (50)
Ours ($\alpha = 0$)	61.2±5.5 (250)	62.1±4.6 (250)	66.7±9.3 (300)	66.4±4.0 (200)	59.7±3.5 (300)	46.3±3.4 (250)
Ours (similarity)	62.0±4.7 (200)	66.0±4.2 (300)	73.0±9.1 (300)	67.7±2.6 (200)	61.7±3.1 (250)	50.6±3.2 (200)
Ours (max-margin)	62.5±4.6 (200)	65.4±3.9 (300)	75.3±9.3 (250)	67.8±2.5 (200)	62.8±3.0 (250)	50.1±3.3 (250)
Ours (triplet)	**64.6±4.5** (250)	**67.5±3.8** (300)	**78.8±8.7** (150)	**68.7±2.2** (170)	**64.5±3.4** (250)	**51.1±3.6** (200)

7. ACKNOWLEDGMENTS

This work is funded by the NSFC (Grant No. 61622310, 61402079, 61473289, 61175003, 61172109), the Foundation for Innovative Research Groups of the NSFC (Grant No. 71421001), the State Key Development Program (Grant No. 2016YFB1001001), the Open Project Program of the National Laboratory of Pattern Recognition.

8. REFERENCES

[1] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *TPAMI*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.

[2] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *TPAMI*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.

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

[4] R. He, T. Tan, L. Wang, and W. Zheng, "$l_{2,1}$ regularized correntropy for robust feature selection," in *CVPR*, 2012, pp. 2504–2511.

[5] S. Wang, X. Chang, X. Li, Q. Z. Sheng, and W. Chen, "Multi-task support vector machines for feature selection with shared knowledge discovery," *Signal Processing*, vol. 120, pp. 746–753, Mar. 2016.

[6] L. Jian, J. Li, K. Shu, and H. Liu, "Multi-label informed feature selection," in *IJCAI*, 2016, pp. 1627–1633.

[7] M. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature selection for clustering – a filter solution," in *ICDM*, 2002, pp. 115–122.

[8] V. Roth and T. Lange, "Feature selection in clustering problems," in *NIPS*, 2004, pp. 473–480.

[9] S. Yang, S. Yan, C. Zhang, and X. Tang, "Bilinear analysis for kernel selection and nonlinear feature extraction," *TNN*, vol. 18, no. 5, pp. 1442–1452, Sep. 2007.

[10] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *SIGKDD*, 2010, pp. 333–342.

[11] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou, "$l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," in *IJCAI*, 2011, vol. 22, pp. 1589–1594.

[12] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *AAAI*, 2012, pp. 1026–1032.

[13] M. Qian and C. Zhai, "Robust unsupervised feature selection," in *IJCAI*, 2013, pp. 1621–1627.

[14] D. Han and J. Kim, "Unsupervised simultaneous orthogonal basis clustering feature selection," in *CVPR*, 2015, pp. 5016–5023.

[15] S. Wang, J. Tang, and H. Liu, "Embedded unsupervised feature selection," in *AAAI*, 2015, pp. 470–476.

[16] F. Nie, W. Zhu, and X. Li, "Unsupervised feature selection with structured graph optimization," in *AAAI*, 2016, pp. 1302–1308.

[17] H. Liu, M. Shao, and Y. Fu, "Consensus guided unsupervised feature selection," in *AAAI*, 2016, pp. 1874–1880.

[18] J. Guo, Y. Guo, X. Kong, M. Zhang, and R. He, "Discriminative analysis dictionary learning," in *AAAI*, 2016, pp. 1617–1623.

[19] S. Wang, F. Nie, X. Chang, L. Yao, X. Li, and Q. Z. Sheng, "Unsupervised feature analysis with class margin optimization," in *ECML/PKDD*, 2015, pp. 383–398.

[20] R. He, W. Zheng, T. Tan, and Z. Sun, "Half-quadratic-based iterative minimization for robust sparse representation," *TPAMI*, vol. 36, no. 2, pp. 261–275, Feb. 2014.

[21] C. Ding, T. Li, and M. Jordan, "Convex and semi-nonnegative matrix factorizations," *TPAMI*, vol. 32, no. 1, pp. 45–55, Jan. 2010.

[22] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *SDM*, 2005, vol. 5, pp. 606–610.

[23] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix tri-factorizations for clustering," in *SIGKDD*, 2006, pp. 126–135.

[24] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *NIPS*, 2005, pp. 507–514.