Embedded Unsupervised Feature Selection

Suhang Wang, Jiliang Tang, Huan Liu
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University, USA
{suhang.wang, jiliang.tang, huan.liu}@asu.edu

Abstract

Sparse learning has been proven to be a powerful technique in supervised feature selection, which allows to embed feature selection into the classification (or regression) problem. In recent years, increasing attention has been on applying spare learning in unsupervised feature selection. Due to the lack of label information, the vast majority of these algorithms usually generate cluster labels via clustering algorithms and then formulate unsupervised feature selection as sparse learning based supervised feature selection with these generated cluster labels. In this paper, we propose a novel unsupervised feature selection algorithm EUFS, which directly embeds feature selection into a clustering algorithm via sparse learning without the transformation. The Alternating Direction Method of Multipliers is used to address the optimization problem of EUFS. Experimental results on various benchmark datasets demonstrate the effectiveness of the proposed framework EUFS.

Introduction

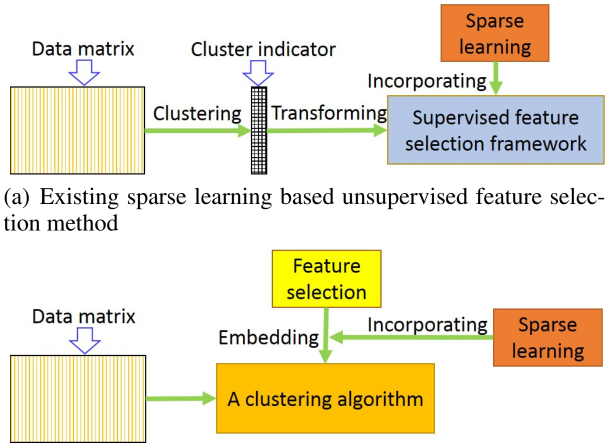
In many real-world applications such as data mining and machine learning, one is often faced with high-dimensional data (Jain and Zongker 1997; Guyon and Elisseeff 2003). Data with high dimensionality not only significantly increases the time and memory requirements of the algorithms, but also degenerates many algorithms' performance due to the curse of dimensionality and the existence of irrelevant, redundant and noisy dimensions(Liu and Motoda 2007). Feature selection, which reduces the dimensionality by selecting a subset of most relevant features, has been proven to be an effective and efficient way to handle high-dimensional data (John *et al.* 1994; Liu and Motoda 2007).

In terms of the label availability, feature selection methods can be broadly classified into supervised methods and unsupervised methods. The availability of the class label allows supervised feature selection algorithms (Duda *et al.* 2001; Nie *et al.* 2008; Zhao *et al.* 2010; Tang *et al.* 2014) to effectively select discriminative features to distinguish samples from different classes. Sparse learning has been proven to be a powerful technique in supervised feature selection (Nie *et al.* 2010; Gu and Han 2011; Tang and Liu 2012a), which enables feature selection to be embedded in the classification (or regression) problem. As most data is unlabeled and it is very expensive to label the data, unsupervised feature selection attracts more and more attentions in recent years (Wolf and Shashua 2005; He *et al.* 2005; Boutsidis *et al.* 2009; Yang *et al.* 2011; Qian and Zhai 2013; Alelyani *et al.* 2013).

Without label information to define feature relevance, a number of alternative criteria have been proposed for unsupervised feature selection. One commonly used criterion is to select features that can preserve the data similarity or manifold structure constructed from the whole feature space (He *et al.* 2005; Zhao and Liu 2007). In recent years, applying sparse learning in unsupervised feature selection has attracted increasing attention. These methods usually generate cluster labels via clustering algorithms and then transform unsupervised feature selection into sparse learning based supervised feature selection with these generated cluster labels such as Multi-cluster feature selection (MCFS) (Cai *et al.* 2010), Nonnegative Discriminative Feature Selection (NDFS) (Li *et al.* 2012), and Robust Unsupervised Feature Selection (RUFS) (Qian and Zhai 2013).

In this paper, we propose a novel unsupervised feature selection algorithm, i.e., Embedded Unsupervised Feature Selection (EUFS). Unlike existing unsupervised feature selection methods such as MCFS, NDFS or RUFS, which transform unsupervised feature selection into sparse learning based supervised feature selection with cluster labels generated by clustering algorithms, we directly embed feature selection into a clustering algorithm via sparse learning without the transformation (see Figure 1). This work theoretically extends the current state-of-the-art unsupervised feature selection, algorithmically expands the capability of unsupervised feature selection, and empirically demonstrates the efficacy of the new algorithm. The major contributions of this paper are summarized next.

- Providing a way to directly embed unsupervised feature selection algorithm into a clustering algorithm via sparse learning instead of transforming it into sparse learning based supervised feature selection with cluster labels;
- Proposing an embedded feature selection framework EUFS, which selects features in unsupervised scenarios with sparse learning; and



(a) Existing sparse learning based unsupervised feature selection method

(b) Embedded unsupervised feature selection

Figure 1: Differences between the existing sparse learning based unsupervesd feature selection methods and the proposed embedded unsupervised feature selection

- Conducting experiments on various datasets to demonstrate the effectiveness of the proposed framework EUFS.

The rest of this paper is organized as follows. In Section 2, we give details about the embedded unsupervised feature selection framework EUFS. In Section 3, we introduce a method to solve the optimization problem of the proposed framework. In Section 4, we show empirical evaluation with discussion. In section 5, we present the conclusion with future work.

Embedded Unsupervised Feature Selection

Throughout this paper, matrices are written as boldface capital letters and vectors are denoted as boldface lowercase letters. For an arbitrary matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, $\mathbf{M}_{ij}$ denotes the (i, j)-th entry of $\mathbf{M}$ while $\mathbf{m}_i$ and $\mathbf{m}^j$ mean the i-th row and j-th column of $\mathbf{M}$ respectively. $||\mathbf{M}||_F$ is the Frobenius norm of $\mathbf{M}$ and Tr($\mathbf{M}$) is the trace of $\mathbf{M}$ if $\mathbf{M}$ is square. $\langle\mathbf{A},\mathbf{B}\rangle$ equals Tr($\mathbf{A}^T\mathbf{B}$), which is the standard inner product between two matrices. $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is a vector whose elements are all 1. The $l_{2,1}$-norm is defined as $||\mathbf{M}||_{2,1} = \sum_{i=1}^{m}||\mathbf{m}_i|| = \sum_{i=1}^{m}\sqrt{\sum_{j=1}^{n}\mathbf{M}_{ij}^2}$).

Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ be the data matrix with each row $\mathbf{x}_i \in \mathbb{R}^{1\times d}$ being a data instance. We use $\mathcal{F} = \{f_1, \dots, f_d\}$ to denote the d features and $\mathbf{f}_1, \dots, \mathbf{f}_d$ are the corresponding feature vectors. Assume that each feature has been normalized, i.e., $||\mathbf{f}_j||_2 = 1$ for $j = 1, \dots, d$. Suppose that we want to cluster $\mathbf{X}$ into k clusters $(C_1, C_2, \dots, C_k)$ under the matrix factorization framework as:

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \ ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_F^2 \\ &s.t.\ \mathbf{U} \in \{0,1\}^{N\times k}, \mathbf{U}^T\mathbf{1} = \mathbf{1} \end{aligned} \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{N\times k}$ is the cluster indicator and $\mathbf{V} \in \mathbb{R}^{d\times k}$ is the latent feature matrix. The problem in Eq.(1) is difficult to solve due to the constraint on $\mathbf{U}$. Following the common relaxation for label indicator matrix (Von Luxburg 2007; Tang and Liu 2012b), the constraint on $\mathbf{U}$ is relaxed to orthogonality, i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{U} \geq \mathbf{0}$. After the relaxation, Eq.(1) can be rewritten as:

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

Another significance of the orthogonality constraint on $\mathbf{U}$ is to allow us to perform feature selection via $\mathbf{V}$, which can be stated by the follow theorem:

Theorem 1. *Let* $\mathbf{X} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_d]$*, and* $||\mathbf{f}_i|| = 1$ *for* $i = 1, \dots, d$*. We use* $\mathbf{U}\mathbf{V}^T$ *to reconstruct* $\mathbf{X}$*, i.e.,* $\hat{\mathbf{X}} = \mathbf{U}\mathbf{V}^T$*. If* $\mathbf{U}$ *is orthogonal, then we can perform feature selection via* $\mathbf{V}$*.*

Proof. Since $\hat{\mathbf{X}} = \mathbf{U}\mathbf{V}^T$, we have $\hat{\mathbf{f}}_i = \mathbf{U}\mathbf{v}_i^T$. Then

$$||\hat{\mathbf{f}}_i||_2 = ||\mathbf{U}\mathbf{v}_i^T||_2 = \left(\mathbf{v}_i\mathbf{U}^T\mathbf{U}\mathbf{v}_i\right)^{1/2} = ||\mathbf{v}_i||_2 \tag{3}$$

Consider the case that $||\mathbf{v}_i||_2$ is close to 0, which indicates that the reconstructed feature representation $||\hat{\mathbf{f}}_i||_2$ is close to 0. $||\mathbf{f}_i|| = 1$ means $\mathbf{f}_i$ is not well reconstructed via $\hat{\mathbf{f}}_i$, which suggests that this corresponding feature could be not representative and we should exclude such features to have a better reconstruction. One way to do this is to add a selection matrix diag($\mathbf{p}$) to $\mathbf{X}$ and $\mathbf{V}$ as,

$$||\mathbf{X}\text{diag}(\mathbf{p}) - \mathbf{U}(\text{diag}(\mathbf{p})\mathbf{V})^T||_F^2 \tag{4}$$

where $\mathbf{p} = \{0, 1\}^d$ with $p_i = 1$ if the i-th feature is selected and otherwise $p_i = 0$, which completes the proof. □

With Theorem 1, if we want to select m features for the clustering algorithm in Eq.(2), we can rewrite it as:

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \ ||\mathbf{X}\text{diag}(\mathbf{p}) - \mathbf{U}(\text{diag}(\mathbf{p})\mathbf{V})^T||_F^2 \\ &s.t.\ \mathbf{U}^T\mathbf{U} = \mathbf{I}, \mathbf{U} \geq \mathbf{0} \\ &\quad \mathbf{p} \in \{0,1\}^d, \mathbf{p}^T\mathbf{1} = m \end{aligned} \tag{5}$$

The constraint on $\mathbf{p}$ makes Eq.(5) mixed integer programming (Boyd and Vandenberghe 2004), which is difficult to solve. We relax the problem in the following way. First, the following theorem suggests that we can ignore the selection matrix on $\mathbf{X}$ as

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \ ||\mathbf{X} - \mathbf{U}(\text{diag}(\mathbf{p})\mathbf{V})^T||_F^2 \\ &s.t.\ \mathbf{U}^T\mathbf{U} = \mathbf{I}, \mathbf{U} \geq \mathbf{0} \\ &\quad \mathbf{p} \in \{0,1\}^d, \mathbf{p}^T\mathbf{1} = m \end{aligned} \tag{6}$$

Theorem 2. *The optimization problems in Eq.(5) and Eq.(6) are equivalent.*

Proof. One way to prove Theorem 2 is to show that the objective functions in Eq.(5) and Eq.(6) are equivalent. For Eq.(5), we have

$$\begin{aligned} &||\mathbf{X}\text{diag}(\mathbf{p}) - \mathbf{U}(\text{diag}(\mathbf{p})\mathbf{V})^T||_F^2 \\ &= \sum_{i=1}^{d} ||p_i\mathbf{f}_i - p_i\mathbf{U}\mathbf{v}_i^T||_F^2 \\ &= \sum_{i:p_i=1} ||\mathbf{f}_i - \mathbf{U}\mathbf{v}_i^T||_F^2 \end{aligned} \tag{7}$$

And for Eq.(6), we have

$$\begin{aligned} & ||\mathbf{X} - \mathbf{U}(\text{diag}(\mathbf{p})\mathbf{V})^T||_F^2 \\ = & \sum_{i=1}^{d} ||\mathbf{f}_i - p_i\mathbf{U}\mathbf{v}_i^T||_F^2 \\ = & \sum_{i:p_i=1} ||\mathbf{f}_i - \mathbf{U}\mathbf{v}_i^T||_F^2 + (N - m) \end{aligned} \tag{8}$$

which complete the proof. □

We observe that diag($\mathbf{p}$) and $\mathbf{V}$ is as the form of diag($\mathbf{p}$)$\mathbf{V}$ in Eq.(6). Since $\mathbf{p}$ is a binary vector and $N - m$ rows of the diag($\mathbf{p}$) are all zeros, diag($\mathbf{p}$)$\mathbf{V}$ is a matrix where elements of many rows are all zeros. This motivates us to absorb the diag($\mathbf{p}$) into $\mathbf{V}$, i.e., $\mathbf{V}$ = diag($\mathbf{p}$)$\mathbf{V}$, and add $l_{2,1}$ norm on $\mathbf{V}$ to achieve feature selection as

$$\begin{aligned} & \arg\min_{\mathbf{U},\mathbf{V}} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_F^2 + \alpha||\mathbf{V}||_{2,1} \\ & s.t.\ \mathbf{U}^T\mathbf{U} = \mathbf{I}, \mathbf{U} \geq \mathbf{0} \end{aligned} \tag{9}$$

Since we forces some rows of $\mathbf{V}$ close to 0, $\mathbf{U}$ and $\mathbf{V}$ may poorly reconstruct some data instances. Reconstructing errors from these instances may easily dominate the objective function because of the squared errors. To make the model robust to these instances, we adopt robust analysis, i.e., replace the loss function by $l_{2,1}$-norm, as follows

$$\begin{aligned} & \arg\min_{\mathbf{U},\mathbf{V}} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_{2,1} + \alpha||\mathbf{V}||_{2,1} \\ & s.t.\ \mathbf{U}^T\mathbf{U} = \mathbf{I}, \mathbf{U} \geq \mathbf{0} \end{aligned} \tag{10}$$

To take advantage of information from attribute-value part, i.e, $\mathbf{X}$, similar data instances should have similar labels, according to the spectral analysis (Von Luxburg 2007), we further add the following term to force similar instances with similar labels as:

$$\min \text{Tr}(\mathbf{U}^T\mathbf{L}\mathbf{U}) \tag{11}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is the Laplacian matrix and $\mathbf{D}$ is a diagonal matrix with its elements defined as $D_{ii} = \sum_{j=1}^{n} S_{ij}$. $\mathbf{S} \in \mathbb{R}^{N \times N}$ denotes the similarity matrix based on $\mathbf{X}$, which is obtained through RBF kernel as

$$S_{ij} = e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\sigma^2}} \tag{12}$$

Putting Eq.(10) and Eq.(11) together, the proposed framework EUFS is to solve the following optimization problem:

$$\begin{aligned} & \arg\min_{\mathbf{U},\mathbf{V}} ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_{2,1} + \alpha||\mathbf{V}||_{2,1} + \beta\text{Tr}(\mathbf{U}^T\mathbf{L}\mathbf{U}) \\ & s.t.\ \mathbf{U}^T\mathbf{U} = \mathbf{I}, \mathbf{U} \geq \mathbf{0} \end{aligned} \tag{13}$$

Optimization Algorithm

The objective function in Eq.(13) is not convex in both $\mathbf{U}$ and $\mathbf{V}$ but is convex if we update the two variables alternatively. Following (Huang *et al.* 2014), we use Alternating Direction Method of Multiplier (ADMM) (Boyd *et al.* 2011) to optimize the objective function. By introducing two auxiliary variables $\mathbf{E} = \mathbf{X} - \mathbf{U}\mathbf{V}^T$ and $\mathbf{Z} = \mathbf{U}$, we can convert Eq.(13) into the following equivalent problem,

$$\begin{aligned} & \arg\min_{\mathbf{U},\mathbf{V},\mathbf{E},\mathbf{Z}} ||\mathbf{E}||_{2,1} + \alpha||\mathbf{V}||_{2,1} + \beta\text{Tr}(\mathbf{Z}^T\mathbf{L}\mathbf{U}) \\ & s.t. \quad \mathbf{E} = \mathbf{X} - \mathbf{U}\mathbf{V}^T, \mathbf{Z} = \mathbf{U}, \mathbf{U}^T\mathbf{U} = \mathbf{I}, \mathbf{Z} \geq \mathbf{0} \end{aligned} \tag{14}$$

which can be solved by the following ADMM problem

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V},\mathbf{E},\mathbf{Z},\mathbf{Y}_1,\mathbf{Y_2},\mu} & ||\mathbf{E}||_{2,1} + \alpha||\mathbf{V}||_{2,1} + \beta\text{Tr}(\mathbf{Z}^T\mathbf{L}\mathbf{U}) \\ & + \langle\mathbf{Y}_1, \mathbf{Z} - \mathbf{U}\rangle + \Big\langle\mathbf{Y}_2, \mathbf{X} - \mathbf{U}\mathbf{V}^T - \mathbf{E}\Big\rangle \\ & + \frac{\mu}{2}(||\mathbf{Z} - \mathbf{U}||_F^2 + ||\mathbf{X} - \mathbf{U}\mathbf{V}^T - \mathbf{E}||_F^2) \\ s.t. \quad & \mathbf{U}^T\mathbf{U} = \mathbf{I}, \mathbf{Z} \geq \mathbf{0} \end{aligned} \tag{15}$$

where $\mathbf{Y}_1, \mathbf{Y}_2$ are two Lagrangian multipliers and μ is a scalar to control the penalty for the violation of equality constraints $\mathbf{E} = \mathbf{X} - \mathbf{U}\mathbf{V}^T$ and $\mathbf{Z} = \mathbf{U}$.

Update E

To update $\mathbf{E}$, we fix the other variables except $\mathbf{E}$ and remove terms that are irrelevant to $\mathbf{E}$. Then Eq.(15) becomes

$$\min_{\mathbf{E}} \frac{1}{2}||\mathbf{E} - (\mathbf{X} - \mathbf{U}\mathbf{V}^T + \frac{1}{\mu}\mathbf{Y}_2)||_F^2 + \frac{1}{\mu}||\mathbf{E}||_{2,1} \tag{16}$$

The equation has a closed form solution by the following Lemma (Liu *et al.* 2009)

Lemma 3. *Let* $\mathbf{Q} = [\mathbf{q}_1; \mathbf{q}_2; ...; \mathbf{q}_m]$ *be a given matrix and* λ *a positive scalar. If the the optimal solution of*

$$\min_{\mathbf{W}} \frac{1}{2}||\mathbf{W} - \mathbf{Q}||_F^2 + \lambda||\mathbf{W}||_{2,1} \tag{17}$$

is $\mathbf{W}^*$*, then the* i*-th row of* $\mathbf{W}^*$ *is*

$$\mathbf{w}_i^* = \begin{cases} (1 - \frac{\lambda}{||\mathbf{q}_i||})\mathbf{q}_i, & \textit{if } ||\mathbf{q}_i|| > \lambda \\ \mathbf{0}, & \textit{otherwise} \end{cases} \tag{18}$$

Apparently, if we let $\mathbf{Q} = \mathbf{X} - \mathbf{U}\mathbf{V}^T + \frac{1}{\mu}\mathbf{Y}_2$, then using Lemma 3, $\mathbf{E}$ can be updated as

$$\mathbf{e}_i = \begin{cases} (1 - \frac{1}{\mu||\mathbf{q}_i||})\mathbf{q}_i, & \text{if } ||\mathbf{q}_i|| > \frac{1}{\mu} \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{19}$$

Update V

To update $\mathbf{V}$, we fix the other variables except $\mathbf{V}$ and remove terms that are irrelevant to $\mathbf{V}$, then Eq.(15) becomes

$$\min_{\mathbf{V},\mathbf{U}^T\mathbf{U}=\mathbf{I}} \frac{\mu}{2}||\mathbf{X} - \mathbf{U}\mathbf{V}^T - \mathbf{E} + \frac{1}{\mu}\mathbf{Y}_2||_F^2 + \alpha||\mathbf{V}||_{2,1} \tag{20}$$

Using the fact that $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, we can reformulate Eq.(20) as

$$\min_{\mathbf{V}} \frac{1}{2}||\mathbf{V} - (\mathbf{X} - \mathbf{E} + \frac{\mathbf{1}}{\mu}\mathbf{Y}_2)^T\mathbf{U}||_F^2 + \frac{\alpha}{\mu}||\mathbf{V}||_{2,1} \tag{21}$$

Again, the above equation has a closed form solution according to Lemma 3. Let $\mathbf{K} = (\mathbf{X} - \mathbf{E} + \frac{1}{\mu}\mathbf{Y}_2)^T\mathbf{U}$, then

$$\mathbf{v}_i = \begin{cases} (1 - \frac{\alpha}{\mu||\mathbf{k}_i||})\mathbf{k}_i, & \text{if } ||\mathbf{k}_i|| > \frac{\alpha}{\mu} \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{22}$$

Update $\mathbf{Z}$

Similarly, to update $\mathbf{Z}$, we fix $\mathbf{U}, \mathbf{V}, \mathbf{E}, \mathbf{Y}_1, \mathbf{Y}_2, \mu$ and remove terms irrelevant to $\mathbf{Z}$, then Eq.(15) becomes

$$\min_{\mathbf{Z}\geq\mathbf{0}} \frac{\mu}{2}||\mathbf{Z}-\mathbf{U}||_F^2 + \beta\text{Tr}(\mathbf{Z}^T\mathbf{L}\mathbf{U}) + \langle\mathbf{Y}_1, \mathbf{Z}-\mathbf{U}\rangle \quad (23)$$

We can rewrite Eq.(23) by putting the second and third terms to the quadratic term and get a compact form

$$\min_{\mathbf{Z}\geq\mathbf{0}} ||\mathbf{Z}-\mathbf{T}||_F^2 \quad (24)$$

where $\mathbf{T}$ is defined as

$$\mathbf{T} = (\mathbf{U} - \frac{1}{\mu}\mathbf{Y}_1 - \frac{\beta}{\mu}\mathbf{L}\mathbf{U}) \quad (25)$$

Eq.(24) can be further decomposed to element-wise optimization problems as

$$\min_{Z_{ij}\geq 0} (Z_{ij} - T_{ij})^2 \quad (26)$$

Clearly, the optimal solution of the above problem is

$$Z_{ij} = \max(T_{ij}, 0) \quad (27)$$

Update $\mathbf{U}$

Optimizing Eq.(15) with respect to $\mathbf{U}$ yields the equation

$$\min_{\mathbf{U}^T\mathbf{U}=\mathbf{I}} \langle\mathbf{Y}_1, \mathbf{Z}-\mathbf{U}\rangle + \left\langle\mathbf{Y}_2, \mathbf{X}-\mathbf{U}\mathbf{V}^T-\mathbf{E}\right\rangle + \frac{\mu}{2}(||\mathbf{Z}-\mathbf{U}||_F^2 + ||\mathbf{X}-\mathbf{U}\mathbf{V}^T-\mathbf{E}||_F^2) + \beta\text{Tr}(\mathbf{Z}^T\mathbf{L}\mathbf{U}) \quad (28)$$

By expanding Eq.(28) and dropping terms that are independent of $\mathbf{U}$, we arrive at

$$\min_{\mathbf{U}^T\mathbf{U}=\mathbf{I}} \frac{\mu}{2}||\mathbf{U}||_F^2 - \mu\langle\mathbf{N}, \mathbf{U}\rangle \quad (29)$$

where $\mathbf{N}$ is defined as

$$\mathbf{N} = \frac{1}{\mu}\mathbf{Y}_1 + \mathbf{Z} - \beta\mathbf{L}\mathbf{Z} + (\mathbf{X} - \mathbf{E} + \frac{1}{\mu}\mathbf{Y}_2)\mathbf{V} \quad (30)$$

We can further write the above equation into a more compact form as

$$\min_{\mathbf{U}^T\mathbf{U}=\mathbf{I}} ||\mathbf{U}-\mathbf{N}||_F^2 \quad (31)$$

And now we have converted the objective function of updating $\mathbf{U}$ to the classical Orthogonal Procrutes problem (Schönemann 1966) and can be solved using the following lemma (Huang *et al.* 2014)

Lemma 4. *Given the objective in Eq.(31), the optimal* $\mathbf{U}$ *is defined as*

$$\mathbf{U} = \mathbf{P}\mathbf{Q}^T \quad (32)$$

where $\mathbf{P}$ *and* $\mathbf{Q}$ *are the left and right singular vectors of the economic singular value decomposition (SVD) of* $\mathbf{N}$.

Update $\mathbf{Y}_1$, $\mathbf{Y}_2$ and μ

After updating the variables, we now need to update the ADMM parameters. According to (Boyd *et al.* 2011), they are updated as follows

$$\mathbf{Y}_1 = \mathbf{Y}_1 + \mu(\mathbf{Z}-\mathbf{U}) \quad (33)$$

$$\mathbf{Y}_2 = \mathbf{Y}_2 + \mu(\mathbf{X}-\mathbf{U}\mathbf{V}^T-\mathbf{E}) \quad (34)$$

$$\mu = \max(\rho\mu, \mu_{max}) \quad (35)$$

Here, $\rho > 1$ is a parameter to control the convergence speed and μ_{max} is a larger number to prevent μ becomes too large.

With these updating rules, EUFS algorithm is summarized in Algorithm 1.

Algorithm 1 Embedded Unsupervised Feature Selection

```
Input: $\mathbf{X} \in \mathbf{R}^{N\times d}, \alpha, \beta, n$, latent dimensional $k$
Output: n features for the dataset
Initialize $\mu = 10^{-3}, \rho = 1.1, \mu_{max} = 10^{10}$, $\mathbf{U} = \mathbf{0}, \mathbf{V} = \mathbf{0}$ (or initialized using K-means)
repeat
    Calculate $\mathbf{Q} = \mathbf{X} - \mathbf{U}\mathbf{V}^T + \frac{1}{\mu}\mathbf{Y}_2$
    Update $\mathbf{E}$
        $\mathbf{e}_i = \begin{cases} (1 - \frac{1}{\mu||\mathbf{q}_i||})\mathbf{q}_i, & \text{if } ||\mathbf{q}_i|| > \frac{1}{\mu} \\ \mathbf{0}, & \text{otherwise} \end{cases}$   (36)
    Calculate $\mathbf{K} = (\mathbf{X} - \mathbf{E} + \frac{\mathbf{1}}{\mu}\mathbf{Y}_2)^T\mathbf{U}$
    Update $\mathbf{V}$
        $\mathbf{v}_i = \begin{cases} (1 - \frac{\alpha}{\mu||\mathbf{k}_i||})\mathbf{k}_i, & \text{if } ||\mathbf{k}_i|| > \frac{\alpha}{\mu} \\ \mathbf{0}, & \text{otherwise} \end{cases}$   (37)
    Calculate $\mathbf{T}$ using Eq.(25)
    Update $\mathbf{Z}$ using Eq.(27)
    Calculate $\mathbf{N}$ according to Eq.(30)
    Update $\mathbf{U}$ by Lemma 4
    Update $\mathbf{Y}_1, \mathbf{Y}_2, \mu$
until convergence
Sort each feature of $\mathbf{X}$ according to $||\mathbf{v}_i||_2$ in descending order and select the top-n ranked ones
```

Parameter Initialization

One way to initialize $\mathbf{U}$ and $\mathbf{V}$ is to simply set them to be $\mathbf{0}$. As the algorithm runs, the objective function will gradually converge to the optimal value. To accelerate the convergence speed, following the common way of initializing NMF, we can use k-means to initialize $\mathbf{U}$ and $\mathbf{V}$. To be specific, we apply k-means to cluster rows of $\mathbf{X}$ and get the soft cluster indicator $\mathbf{U}$. $\mathbf{V}$ is simply set as $\mathbf{X}^T\mathbf{U}$. μ is typically set in the range of 10^{-6} to 10^{-3} initially depending on the datasets and is updated in each iteration. μ_{max} is set to be a large value such as 10^{10} to give μ freedom to increase but prevent it from being too large. ρ is empirically set to 1.1 in our algorithm. The larger ρ is , the faster μ becomes larger and the more we penalize the deviation of the equality constraint, which makes it converges faster. However, we may sacrifice some precision of the final objective function with large ρ.

Convergence Analysis

The convergence of our algorithm depends on the convergence of the ADMM. The detailed convergence proof of ADMM can be found in (Goldstein *et al.* 2012; Boyd *et al.* 2011). The convergence criteria can be set as $\frac{|\mathbf{J}_{t+1}-\mathbf{J}_t|}{\mathbf{J_t}} < \epsilon$, where $\mathbf{J}_t$ is the objective function value in Eq.(14) and ϵ is some tolerance value. In practice, we can control the number of iterations by setting a maximum iteration value. Our experiments find that our algorithm converges within 110 iterations for all the datasets we used.

Time Complexity Analysis

The computation cost for $\mathbf{E}$ depends on the computation of $\mathbf{Q} = \mathbf{X} - \mathbf{U}\mathbf{V}^T + \frac{1}{\mu}\mathbf{Y}_2$ and update of $\mathbf{E}$. Since $\mathbf{U}$ is sparse, i.e., each row of $\mathbf{U}$ only has one nonzero element, then the computation cost is O(Nd) and O(Nd), respectively.

Similarly, the computation cost for $\mathbf{V}$ involves the computation of $\mathbf{K} = (\mathbf{X} - \mathbf{E} + \frac{1}{\mu}\mathbf{Y}_2)^T\mathbf{U}$ and update of $\mathbf{V}$, which is O(Nd) again due to the sparsity of $\mathbf{U}$.

The main computation cost for $\mathbf{Z}$ is the computation of $\mathbf{T} = (\mathbf{U} - \frac{1}{\mu}\mathbf{Y}_1^T - \frac{\beta}{\mu}\mathbf{L}\mathbf{U})$, which is O($k^2$) due to the sparsity of both $\mathbf{U}$ and $\mathbf{L}$.

The main computation cost of $\mathbf{U}$ involves the computation of $\mathbf{N}$ and its SVD decomposition, which is O(Ndk) and O(Nk^2). The computational cost for $\mathbf{Y_1}$ and $\mathbf{Y_2}$ are both O(Nd). Therefore, the overall time complexity is O($Ndk + Nk^2$). Since $d \gg k$, the final computation cost if O(Ndk) for each iteration.

Experimental Analysis

In this section, we conduct experiments to evaluate the effectiveness of EUFS [1]. After introducing datasets and experimental settings, we compare EUFS with the state-of-the-art unsupervised feature selection methods. Further experiments are conducted to investigate the effects of important parameters on EUFS.

Datasets

The experiments are conducted on 6 publicly available benchmark datasets, including one Mass Spectrometry (MS) dataset ALLAML (Fodor 1997), two microarray datasets, i.e., Prostate Cancer gene expression (Prostate-GE) [2] (Singh *et al.* 2002) and TOX-171, two face image datasets, i.e., PIX10P and PIE10P[3] and one object image dataset COIL20[4] (Nene *et al.* 1996). The statistics of the datasets used in the experiments are summarized in Table 1.

Table 1: Statistics of the Dataset

Dataset	# of Samples	# of Features	# of Classes
ALLAML	72	7192	2
COIL20	1440	1024	20
PIE10P	210	1024	10
TOX-171	171	5748	4
PIX10P	100	10000	10
Prostate-GE	102	5996	2

Experimental Settings

Following the common way to evaluate unsupervised feature selection algorithms, we assess EUFS in terms of clustering performance (Zhao and Liu 2007; Li *et al.* 2012). We compare EUFS with the following representative unsupervised feature selection algorithms:

- **All Features**: All original features are adopted
- **LS**: Laplacian Score (He *et al.* 2005) which evaluates the importance of a feature through its power of locality preservation
- **MCFS**: Multi-Cluster Feature Selection (Cai *et al.* 2010) which selects features using spectral regression with l_1-norm regularization
- **NDFS**: Nonnegative Discriminative Feature Selection (Li *et al.* 2012) which selects features by a joint feamewrok of nonnegative spectral analysis and $l_{2,1}$ regularized regression
- **RUFS**: Robust Unsupervised Feature Selection (Qian and Zhai 2013) which jointly performs robust label learning via local learning regularized robust orthogonal nonnegative matrix factorization and robust feature learning via joint $l_{2,1}$-norms minimization.

Two widely used evaluation metrics, *accuracy* (ACC) and *normalized mutual information* (NMI), are employed to evaluate the quality of clusters. The larger ACC and NMI are, the better performance is.

There are some parameters to be set. Following (Qian and Zhai 2013), for LS, MCFS, NDFS, RUFS and EUFS, we fix the neighborhood size to be 5 for all the datasets. To fairly compare different unsupervised feature selection methods, we tune the parameters for all methods by a "grid-search" strategy from $\{10^{-6}, 10^{-4}, \ldots, 10^4, 10^6\}$. For EUFS, we set the latent dimension as the number of clusters. How to determine the optimal number of selected features is still an open problem (Tang and Liu 2012a), we set the number of selected features as {50, 100, 150, …, 300} for all datasets. Best clustering results from the optimal parameters are reported for all the algorithms. In the evaluation , we use K-means to cluster samples based on the selected features. Since K-means depends on initialization, following previous work, we repeat the experiments 20 times and the average results with standard deviation are reported.

Experimental Results

The experimental results of different methods on the datasets are summarized in Table 2 and Table 3. We make the following observations:

[1] The implementation of EUFS can be found from http://www.public.asu.edu/ swang187/

[2] ALLAML and Prostate-GE are publicly available from https://sites.google.com/site/feipingnie/file

[3] TOX-171, PIX10P and PIE10P are publicly available from http://featureselection.asu.edu/datasets.php

[4] COIL20 is publicly available from http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html

Table 2: Clustering results(ACC%±std) of different feature selection algorithms on different datasets. The best results are highlighted in bold. The number in parentheses is the number of features when the performance is achieved

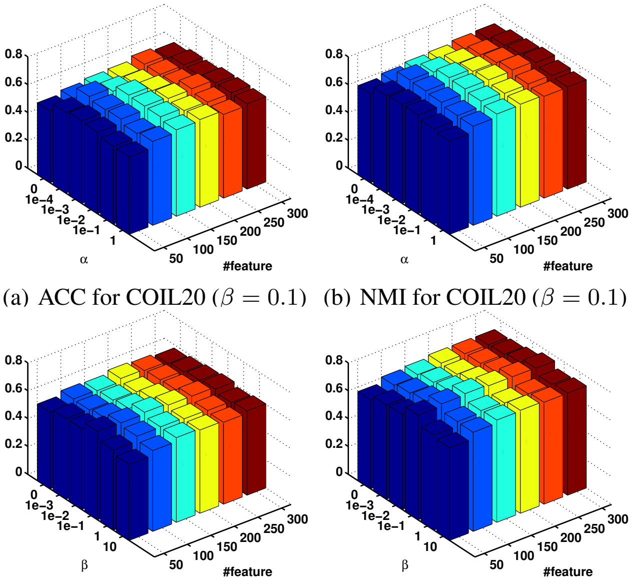
Dataset	ALL Features	Laplacian Score	MCFS	NDFS	RUFS	EUFS
ALLAML	67.3±6.72	73.2±5.52(150)	68.4±10.4(100)	69.4±0.00(100)	72.2±0.00(150)	**73.6±0.00**(100)
COIL20	53.6±3.83	55.2±2.84(250)	59.7±4.03(250)	60.1±4.26(300)	62.7±3.51(150)	**63.4±5.47**(100)
PIE10P	30.8±2.29	36.0±2.95(100)	44.3±3.20(50)	40.5±4.51(100)	42.6±4.61(50)	**46.4±2.69**(50)
TOX-171	41.5±3.88	47.5±3.33(200)	42.5±5.15(100)	46.1±2.55(100)	47.8±3.78(300)	**49.5±2.57**(100)
PIX10P	74.3±12.1	76.6±8.10(150)	75.9±8.59(200)	76.7±8.52(200)	73.2±9.40(300)	**76.8±5.88**(150)
Prostate-GE	58.1±0.44	57.5±0.49(300)	57.3±0.50(300)	58.3±0.50(100)	59.8±0.00(50)	**60.4±0.80**(100)

Table 3: Clustering results(NMI%±std) of different feature selection algorithms on different datasets. The best results are highlighted in bold. The number in parentheses is the number of features when the performance is achieved

Dataset	ALL Features	Laplacian Score	MCFS	NDFS	RUFS	EUFS
ALLAML	8.55±5.62	15.0±1.34(100)	11.7±12.2(50)	7.20±0.30(300)	12.0±0.00(150)	**15.1±0.00**(100)
COIL20	70.6±1.95	70.3±1.73(300)	72.4±1.90(150)	72.1±1.75(300)	73.1±1.69(150)	**77.2±2.75**(100)
PIE10P	32.2±3.47	38.5±1.44(50)	**54.3±3.39**(50)	46.0±3.14(100)	49.6±5.15(50)	49.8±3.10(150)
TOX-171	17.8±5.20	**30.5±2.70**(150)	17.7±6.88(100)	22.3±2.41(300)	28.8±2.71(300)	26.0±2.41(100)
PIX10P	82.8±6.48	84.3±4.63(150)	85.0±4.95(200)	84.8±4.76(200)	81.1±6.23(300)	**85.1±4.30**(50)
Prostate-GE	1.95±0.27	1.59±0.21(300)	1.53±0.21(300)	2.02±0.25(100)	2.86±0.00(50)	**3.36±0.48**(100)

- Feature selection is necessary and effective. The selected subset of the features can not only reduce the computation cost, but also improve the clustering performance;
- Robust analysis is also important for unsupervised feature selection, which helps us select more relevant features and improve the performance;
- EUFS tends to achieve better performance with usually fewer selected features such as 50 or 100; and
- Most of the time, the proposed framework EUFS outperforms baseline methods, which demonstrates the effectiveness of the proposed algorithm. There are two major reasons. First, we directly embed feature selection in the process of clustering using sparse learning and the norm of the latent feature reflects the quality of the reconstruction and thus the importance of the original feature. Second, the graph regularize helps to learn better cluster indicators that fits the existing manifold structure, which leads to a better latent feature matrix. Finally, we introduce robust analysis to ensure that these poorly reconstructed instances have less effect on feature selection.

We also perform parameter analysis for some important parameters of EUFS. Due to space limit, we only report the results on COIL20 in Figure2. The experimental results show that our method is not very sensitive to α and β. However, the performance is relatively sensitive to the number of selected features, which is a common problem for many unsupervised feature selection methods.



(a) ACC for COIL20 ($\beta = 0.1$) (b) NMI for COIL20 ($\beta = 0.1$)

(c) ACC for COIL20 ($\alpha = 0.1$) (d) NMI for COIL20 ($\alpha = 0.1$)

Figure 2: ACC and NMI of EUFS with different α, β and feature numbers on datasets COIL20

Conclusion

We propose a new unsupervised feature selection approach, EUFS, which directly embeds feature selection into a clustering algorithm via sparse learning. It eliminates the need for transforming unsupervised feature selection into the sparse learning based supervised feature selection with pseudo labels. Nonnegative orthogonality is applied on the cluster indicator to make the problem tractable and ensure that feature selection on latent features has similar effects as on original features. $l_{2,1}$-norm is applied on the cost function to reduce the effects of the noise introduced by the reconstruction of $\mathbf{X}$ and feature selection on $\mathbf{V}$. Experimental results on 6 different real world datasets validate the unique contributions of EUFS. Future work is to investigate if EUFS can be extended to dimensionality reduction algorithms.

Acknowledgments

This material is based upon work supported by, or in part by, the National Science Foundation (NSF) under grant number IIS-1217466.

References

Salem Alelyani, Jiliang Tang, and Huan Liu. Feature selection for clustering: A review. In *Data Clustering: Algorithms and Applications*, pages 29–60. CRC Press, 2013.

Christos Boutsidis, Petros Drineas, and Michael W Mahoney. Unsupervised feature selection for the k-means clustering problem. In *Advances in Neural Information Processing Systems*, pages 153–161, 2009.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 333–342. ACM, 2010.

Richard O Duda, Peter E Hart, and David G Stork. Pattern classification. 2nd. *Edition. New York*, 2001.

Stephen P. A. Fodor. Dna sequencing: Massively parallel genomics. 277(5324):393–395, 1997.

Tom Goldstein, Brendan ODonoghue, and Simon Setzer. Fast alternating direction optimization methods. *CAM report*, pages 12–35, 2012.

Quanquan Gu and Jiawei Han. Towards feature selection in network. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1175–1184. ACM, 2011.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514, 2005.

Jin Huang, Feiping Nie, Heng Huang, and Chris Ding. Robust manifold nonnegative matrix factorization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(3):11, 2014.

Anil Jain and Douglas Zongker. Feature selection: Evaluation, application, and small sample performance, 1997.

George H John, Ron Kohavi, Karl Pfleger, et al. Irrelevant features and the subset selection problem. In *ICML*, volume 94, pages 121–129, 1994.

Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, and Hanqing Lu. Unsupervised feature selection using nonnegative spectral analysis. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada.*, 2012.

Huan Liu and Hiroshi Motoda. *Computational methods of feature selection*. CRC Press, 2007.

Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 339–348. AUAI Press, 2009.

Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). Technical report, Technical Report CUCS-005-96, 1996.

Feiping Nie, Shiming Xiang, Yangqing Jia, Changshui Zhang, and Shuicheng Yan. Trace ratio criterion for feature selection. In *AAAI*, volume 2, pages 671–676, 2008.

Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint ?2, 1-norms minimization. In *Advances in Neural Information Processing Systems*, pages 1813–1821, 2010.

Mingjie Qian and Chengxiang Zhai. Robust unsupervised feature selection. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1621–1627. AAAI Press, 2013.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Dinesh Singh, Phillip G Febbo, Kenneth Ross, Donald G Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A Renshaw, Anthony V D'Amico, Jerome P Richie, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.

Jiliang Tang and Huan Liu. Feature selection with linked data in social media. In *Proceedings of the Twelfth SIAM International Conference on Data Mining, Anaheim, California, USA, April 26-28, 2012.*, pages 118–128, 2012.

Jiliang Tang and Huan Liu. Unsupervised feature selection for linked social media data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 904–912. ACM, 2012.

Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. In *Data Classification: Algorithms and Applications*. 2014.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Lior Wolf and Amnon Shashua. Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. *The Journal of Machine Learning Research*, 6:1855–1887, 2005.

Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. l2, 1-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1589, 2011.

Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1151–1157. ACM, 2007.

Zheng Zhao, Lei Wang, and Huan Liu. Efficient spectral feature selection with minimum redundancy. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, 2010.