Reconstruction-based Unsupervised Feature Selection: An Embedded Approach

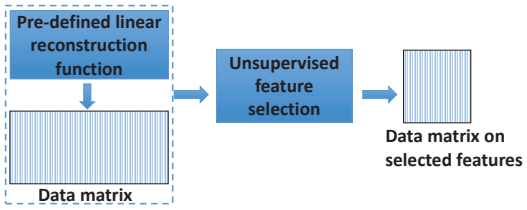
Jundong Li[†], **Jiliang Tang**[‡], **Huan Liu**[†]
[†]Computer Science and Engineering, Arizona State University, USA
[‡]Computer Science and Engineering, Michigan State University, USA
{jundong.li, huan.liu}@asu.edu, tangjili@msu.edu

Abstract

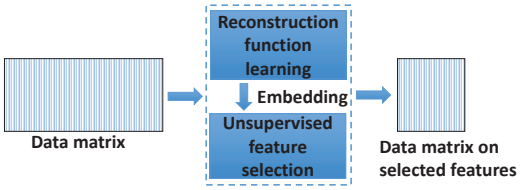
Feature selection has been proven to be effective and efficient in preparing high-dimensional data for data mining and machine learning problems. Since real-world data is usually unlabeled, unsupervised feature selection has received increasing attention in recent years. Without label information, unsupervised feature selection needs alternative criteria to define feature relevance. Recently, data reconstruction error emerged as a new criterion for unsupervised feature selection, which defines feature relevance as the capability of features to approximate original data via a reconstruction function. Most existing algorithms in this family assume predefined, linear reconstruction functions. However, the reconstruction function should be data dependent and may not always be linear especially when the original data is high-dimensional. In this paper, we investigate how to learn the reconstruction function from the data automatically for unsupervised feature selection, and propose a novel reconstruction-based unsupervised feature selection framework REFS, which embeds the reconstruction function learning process into feature selection. Experiments on various types of real-world datasets demonstrate the effectiveness of the proposed framework REFS.

1 Introduction

High-dimensional data arises naturally in many areas, such as machine learning, data mining and computer vision [Jain and Zongker, 1997; Guyon and Elisseeff, 2003]. It poses challenges to many learning tasks due to the *curse of dimensionality* [Duda *et al.*, 2012], i.e., algorithms applicable to low-dimensional data become intractable in high-dimensional space. Besides, data with high dimensionality significantly increases the memory storage requirements and computational costs for data analytics. Moreover, the existence of irrelevant, redundant and noisy features tends to overfit the learning algorithms and results in low efficiency and poor performance. Feature selection [Liu and Motoda, 2007; Li *et al.*, 2016a; Li and Liu, 2017] has been proven to be effective and efficient in handling high-dimensional data.



(a) Existing data reconstruction unsupervised feature selection.



(b) The proposed REFS framework.

Figure 1: Differences between the existing data reconstruction unsupervised feature selection method and the proposed unsupervised feature selection framework.

According to the availability of labels, feature selection methods consist of supervised methods and unsupervised methods. Supervised methods take advantage of the discriminative information encoded in class labels to select the subset of features that are able to distinguish instances from different classes [Tibshirani, 1996; Ding and Peng, 2005; Nie *et al.*, 2008; 2010; Cover and Thomas, 2012; Jian *et al.*, 2016; Li *et al.*, 2017]. Since real-world data is usually unlabeled and collecting labeled data is particular expensive requiring both time and effort, unsupervised feature selection has received increasingly attention in the past few years [He *et al.*, 2005; Cai *et al.*, 2010; Yang *et al.*, 2011; Li *et al.*, 2012; Qian and Zhai, 2013; Li *et al.*, 2015; 2016b; Wei and Philip, 2016]. Due to the lack of label information, unsupervised feature selection algorithms exploit different criteria to define the relevance of features such as data similarity and local discriminative information.

Recently, data reconstruction error [Boutsidis *et al.*, 2008; Masaeli *et al.*, 2010; Farahat *et al.*, 2011] has emerged as another criterion for unsupervised feature selection. It assumes that the original data can be approximated by performing a

reconstruction function on some selected features. The vast majority of existing algorithms in this family assume a predefined and linear reconstruction function, based on which they select features to minimize the reconstruction error as shown in Figure 1(a). However, the reconstruction function should be data dependent and may not always be linear, especially when the original data is high-dimensional. Therefore, we investigate whether we can embed the reconstruction function learning process into feature selection. In essence, we study (1) how can we learn a reconstruction function from the data automatically; and (2) how can we use it for unsupervised feature selection. Our solutions to these two challenges lead to a novel reconstruction-based unsupervised feature selection framework REFS, which embeds the reconstruction function learning into unsupervised feature selection as shown in Figure 1(b). The main contributions of this work are as follows:

- Proposing to learn the reconstruction function from the data for unsupervised feature selection;
- Proposing a novel reconstruction-based unsupervised feature selection framework REFS, which integrates the reconstruction function learning and feature selection into a coherent model; and
- Evaluating the effectiveness of the proposed framework REFS on various types of real-world datasets.

The rest of this paper is organized as follows. In Section 2, we will formulate the proposed framework REFS and introduce an efficient greedy method to solve the optimization problem. Empirical evaluation is presented in Section 3 with discussions. In Section 4, we briefly review related work. The conclusion and future work are presented in Section 5.

2 Reconstruction-based Unsupervised Feature Selection - REFS

We first summarize some notations used in this paper. We use bold uppercase characters to denote matrices (e.g., $\mathbf{A}$), bold lowercase characters to denote vectors (e.g., $\mathbf{b}$). For an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{A}_{ij}$ denotes its (i,j)-th entry, and $\mathbf{a}_i$ denotes its i-th column. The Frobenius norm of the matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ is $||\mathbf{A}||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d}\mathbf{A}_{ij}^2}$. $Tr(\mathbf{A})$ is the trace of matrix $\mathbf{A}$ if it is square. $\mathbf{I}$ is an identity matrix of size n.

2.1 The Objective Function of REFS

Let $\mathcal{F} = \{f_1, f_2, ..., f_d\}$ denote a d-dimensional feature space and $\mathbf{f}_1, \mathbf{f}_2$,...,$\mathbf{f}_d$ are the corresponding d feature vectors. Assume that each feature has been normalized, i.e., for each feature vector $\mathbf{f}_i$, $||\mathbf{f}_i||_2 = 1$. Let $\mathbf{X} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_d] \in \mathbb{R}^{n\times d}$ denotes the original d feature vectors, where each feature vector $\mathbf{f}_i \in \mathbb{R}^{n\times 1}$ contains feature values for all the n data instances. Suppose that the new feature space $\mathcal{S}$ contains k selected features ($k < d$), then $\mathbf{X}_\mathcal{S} = [\mathbf{f}_{\mathcal{S}1}, \mathbf{f}_{\mathcal{S}2}, ..., \mathbf{f}_{\mathcal{S}k}]$ are the k feature vectors on the new feature space $\mathcal{S}$. Our target is to select k features such that they can well represent the original d feature vectors. Specifically, let $\tilde{\mathbf{X}} = [\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, ..., \tilde{\mathbf{f}}_d]$ denote the reconstructed d feature vectors that can be represented from the selected k features, we would like to minimize the reconstruction error between $\mathbf{X}$ and $\tilde{\mathbf{X}}$ as follows:

$$\sum_{j=1}^{d} ||\mathbf{f}_j - \tilde{\mathbf{f}}_j||_2^2 = ||\mathbf{X} - \tilde{\mathbf{X}}||_F^2. \tag{1}$$

It indicates that if these k selected features are the most representative ones, the reconstructed d feature vectors $\tilde{\mathbf{X}} = [\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, ..., \tilde{\mathbf{f}}_d]$ should be able to well approximate the original d feature vectors. Therefore, we intend to find a reconstruction function $\psi(.) : \mathbf{X}_\mathcal{S} \mapsto \tilde{\mathbf{X}}$ that maps $\mathbf{X}_\mathcal{S}$ to $\tilde{\mathbf{X}}$ such that the reconstruction error in Eq. (1) is minimized. Therefore the objective function of the proposed framework REFS is formally defined as:

$$\min_{\mathcal{S},\psi(.)} ||\mathbf{X} - \psi(\mathbf{X}_\mathcal{S})||_F^2. \tag{2}$$

The objective function of the proposed framework REFS in Eq. (2) is an optimization problem with respect to both the selected feature set $\mathcal{S}$ and the reconstruction function $\psi(.)$. In the following subsections, we will give details about how to learn the reconstruction function $\psi(.)$ and how to select features based on $\psi(.)$.

2.2 Learning the Reconstruction Function

In order to accurately measure the representativeness of selected feature space $\mathcal{S}$, the reconstruction function $\psi(.)$ should be able to: (1) make reconstructed features in $\mathcal{S}$ closely approximate their original corresponding features; (2) avoid overfitting such that features in $\mathcal{S}$ can also be generalized to represent features in $\mathcal{F}\backslash\mathcal{S}$; and (3) make reconstructed features preserve the original feature structures.

Let $\tilde{\mathbf{X}}_\mathcal{S} = [\tilde{\mathbf{f}_{\mathcal{S}1}}, \tilde{\mathbf{f}_{\mathcal{S}2}}, ..., \tilde{\mathbf{f}_{\mathcal{S}k}}] \in \mathbb{R}^{n\times k}$ denote the reconstruction of the k selected features. First, the reconstruction function $\psi(.)$ ensures the reconstructed k feature vectors in $\tilde{\mathbf{X}}_\mathcal{S}$ be close to the original k feature vectors in $\mathbf{X}_\mathcal{S}$ by minimizing the following term:

$$\begin{aligned}\min_{\tilde{\mathbf{f}_{\mathcal{S}1}},\tilde{\mathbf{f}_{\mathcal{S}2}},...,\tilde{\mathbf{f}_{\mathcal{S}k}},\mathcal{S}} &\sum_{j=1}^{k} ||\mathbf{f}_{\mathcal{S}j} - \tilde{\mathbf{f}_{\mathcal{S}j}}||_2^2 \\ &= \min_{\tilde{\mathbf{X}},\mathcal{S}} ||\mathbf{X}_\mathcal{S} - \tilde{\mathbf{X}}_\mathcal{S}||_F^2.\end{aligned} \tag{3}$$

The minimization function in Eq. (3) ensures the reconstructed feature vectors in $\mathcal{S}$ well approximate the original corresponding feature vectors. However, the reconstruction function $\psi(.)$ might make $\tilde{\mathbf{X}}_\mathcal{S}$ overfit $\mathbf{X}_\mathcal{S}$ such that the selected features in $\mathcal{S}$ cannot well represent other features in $\mathcal{F}\setminus\mathcal{S}$. Therefore, we propose to minimize the approximation error between $\mathbf{X}_{\mathcal{F}\setminus\mathcal{S}}$ and $\tilde{\mathbf{X}}_{\mathcal{F}\setminus\mathcal{S}}$ with a small weight on the basis of Eq. (3) to avoid overfitting:

$$\min_{\tilde{\mathbf{X}},\mathcal{S}} ||\mathbf{X}_\mathcal{S} - \tilde{\mathbf{X}}_\mathcal{S}||_F^2 + \alpha||\mathbf{X}_{\mathcal{F}\setminus\mathcal{S}} - \tilde{\mathbf{X}}_{\mathcal{F}\setminus\mathcal{S}}||_F^2 \tag{4}$$

where α is introduced to balance the contribution of selected features and unselected features for the feature reconstruction process. Normally, we set it in the range of 0 to 1, indicating that during the feature reconstruction process, the selected

features in $\mathcal{S}$ plays a more important role than unselected features in $\mathcal{F} \setminus \mathcal{S}$.

In addition, the reconstruction function $\psi(.)$ should also preserve feature structures in the original feature space since those structures play important roles in many real-world applications such as natural language processing and bioinformatics [Mitra *et al.*, 2002; Tang *et al.*, 2014]. Feature structures can be represented as a graph where nodes represent features and edges show pairwise feature similarities. Let $\mathcal{G}$ denotes a nearest neighbor graph with d nodes where the i-th node corresponds to the i-th feature vector $\mathbf{f}_i$. In this paper, we consider that there is an edge between feature vector $\mathbf{f}_i$ and feature vector $\mathbf{f}_j$ if $\mathbf{f}_i$ is among the p nearest neighbors of $\mathbf{f}_j$ ($\mathbf{f}_i \in \mathcal{N}_p(\mathbf{f}_j)$) or $\mathbf{f}_j$ is among the p nearest neighbors of $\mathbf{f}_i$ ($\mathbf{f}_j \in \mathcal{N}_p(\mathbf{f}_i)$). Then the adjacency matrix $\mathbf{W}$ for $\mathcal{G}$ is defined as:

$$\mathbf{W}_{ij} = \begin{cases} 1 & \text{if } \mathbf{f}_i \in \mathcal{N}_p(\mathbf{f}_j) \text{ or } \mathbf{f}_j \in \mathcal{N}_p(\mathbf{f}_i) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

To preserve feature structures, the reconstruction function $\psi(.)$ should ensure that two features close to each other in the original data are also close after reconstruction by minimizing the following term:

$$\begin{aligned} &\frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}||\tilde{\mathbf{f}}_i - \tilde{\mathbf{f}}_j||_2^2\mathbf{W}_{ij} \\ =&Tr(\tilde{\mathbf{X}}(\mathbf{D}-\mathbf{W})\tilde{\mathbf{X}}^T) \\ =&Tr(\tilde{\mathbf{X}}\mathbf{L}\tilde{\mathbf{X}}^T), \end{aligned} \tag{6}$$

where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$ and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix.

Integrating components from Eq. (4) and Eq. (6), the reconstruction function $\psi(.)$ can be obtained by solving the following optimization problem:

$$\min_{\tilde{\mathbf{X}},\mathcal{S}} ||\mathbf{X}_{\mathcal{S}} - \tilde{\mathbf{X}}_{\mathcal{S}}||_F^2 + \alpha||\mathbf{X}_{\mathcal{F}\setminus\mathcal{S}} - \tilde{\mathbf{X}}_{\mathcal{F}\setminus\mathcal{S}}||_F^2 + \beta Tr(\tilde{\mathbf{X}}\mathbf{L}\tilde{\mathbf{X}}^T), \tag{7}$$

where the first term makes reconstructed features in $\mathcal{S}$ closely approximate their corresponding original features; the second term penalizes the reconstruction error of unselected features in $\mathcal{F} \setminus \mathcal{S}$ with a small weight α to avoid overfitting; while the third term preserves feature structures after reconstruction, which is controlled by a parameter β. When the selected feature space $\mathcal{S}$ is fixed, the optimization problem in Eq. (7) is a convex optimization problem with respect to $\tilde{\mathbf{X}}$. Therefore, we can automatically learn a reconstruction function $\psi(.)$ from data, which maps $\mathbf{X}_{\mathcal{S}}$ to $\tilde{\mathbf{X}}$.

Let $\mathcal{L}(\tilde{\mathbf{X}}) = ||\mathbf{X}_{\mathcal{S}} - \tilde{\mathbf{X}}_{\mathcal{S}}||_F^2 + \alpha||\mathbf{X}_{\mathcal{F}\setminus\mathcal{S}} - \tilde{\mathbf{X}}_{\mathcal{F}\setminus\mathcal{S}}||_F^2 + \beta Tr(\tilde{\mathbf{X}}\mathbf{L}\tilde{\mathbf{X}}^T)$, the derivative of $\mathcal{L}(\tilde{\mathbf{X}})$ w.r.t. $\tilde{\mathbf{X}}$ is:

$$\frac{\partial\mathcal{L}(\tilde{\mathbf{X}})}{\partial\tilde{\mathbf{X}}} = -(1-\alpha)\pi(\mathbf{X}_{\mathcal{S}} - \tilde{\mathbf{X}}_{\mathcal{S}}) - \alpha(\mathbf{X} - \tilde{\mathbf{X}}) + \beta\tilde{\mathbf{X}}\mathbf{L}. \tag{8}$$

We set the derivative in Eq. (8) to be zero, since $\gamma(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\pi(\mathbf{X}_{\mathcal{S}}) + \alpha\mathbf{I} + \beta\mathbf{L}$ is a positive semidefinite matrix, we obtain a closed form solution of $\tilde{\mathbf{X}}$ and a mapping function $\psi(.)$ from $\mathbf{X}_{\mathcal{S}}$ to $\tilde{\mathbf{X}}$ can be derived as follows:

$$\begin{aligned} &\tilde{\mathbf{X}} = \psi(\mathbf{X}_{\mathcal{S}}) \\ =&(\gamma\pi(\mathbf{X}_{\mathcal{S}}) + \alpha\mathbf{X})(\gamma(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\pi(\mathbf{X}_{\mathcal{S}}) + \alpha\mathbf{I} + \beta\mathbf{L})^{-1}, \end{aligned} \tag{9}$$

where $\gamma = 1-\alpha$, $\pi(.)$ is an augment function which augments data matrix in low-dimensional feature space to the data matrix in the original feature space by adding zero column vectors, for example, $\pi(\mathbf{X}_{\mathcal{S}})$ augments the n-by-k matrix to be a n-by-d matrix by adding $d-k$ zero column vectors ($k < d$) to the columns of features in $\mathcal{F} \setminus \mathcal{S}$.

2.3 Selecting Features

Now we embed the learned reconstruction function $\psi(.)$ to the unsupervised feature selection framework REFS. Integrating Eq. (9) into Eq. (2), we can select features by solving the following optimization problem:

$$\begin{aligned} &\min_{\mathcal{S},\psi(.)} ||\mathbf{X} - \psi(\mathbf{X}_{\mathcal{S}})||_F^2 \\ =&\min_{\mathcal{S}} ||\beta\mathbf{X}\mathbf{L}(\gamma(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\pi(\mathbf{X}_{\mathcal{S}}) + \alpha\mathbf{I} + \beta\mathbf{L})^{-1}||_F^2 \end{aligned} \tag{10}$$

The problem in Eq. (10) is an integer programming problem and it is difficult to solve. We propose to use a greedy approach to sequentially minimize the objective function to obtain a local optimal solution and this strategy have been widely used in other unsupervised learning tasks [He, 2010; Chen *et al.*, 2010; Zhang *et al.*, 2011]. Assume current $\mathcal{S}$ contains m selected features ($m < k$), we define:

$$Q_m = \gamma(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\pi(\mathbf{X}_{\mathcal{S}}) + \alpha\mathbf{I} + \beta\mathbf{L} \tag{11}$$

At the very beginning, no features are selected, $\mathbf{X}_{\mathcal{S}}$ is set to be a zero matrix, so:

$$Q_0 = \alpha\mathbf{I} + \beta\mathbf{L} \tag{12}$$

To select the $(m+1)$-th feature, we need to find a feature $\mathcal{C} \in \mathcal{F} \setminus \mathcal{S}$ such that the following term:

$$||\beta\mathbf{X}\mathbf{L}(\gamma(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\pi(\mathbf{X}_{\mathcal{S}}) + \pi(\mathbf{X}_{\mathcal{C}})) + \alpha\mathbf{I} + \beta\mathbf{L})^{-1}||_F^2 \tag{13}$$

is minimized. Here $\pi(\mathbf{X}_{\mathcal{C}})$ augments the n-by-1 feature vector $\mathbf{X}_{\mathcal{C}}$ to a n-by-d matrix by adding zero vectors to features in $\mathcal{F} \setminus \mathcal{C}$. Now we can get:

$$\begin{aligned} Q_{m+1} &= \gamma(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\pi(\mathbf{X}_{\mathcal{S}}) + \pi(\mathbf{X}_{\mathcal{C}})) + \alpha\mathbf{I} + \beta\mathbf{L} \\ &= Q_m + \gamma(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\pi(\mathbf{X}_{\mathcal{C}}) \\ &= Q_m + \gamma\mathbf{I}_{d\times d}(:,\mathcal{C})\mathbf{I}_{d\times d}(\mathcal{C},:) \end{aligned} \tag{14}$$

The computation of the inverse of the matrix Q_{m+1} is very expensive, and we have to compute the inverse of Q_{m+1} $(k-m)$ times to obtain the best $(m+1)$-th feature. Using Sherman-Morrison formula [Golub and Van Loan, 2012], the inverse part can be computed very efficiently:

$$\begin{aligned} Q_{m+1}^{-1} &= (Q_m + \gamma\mathbf{I}_{d\times d}(:,\mathcal{C})\mathbf{I}_{d\times d}(\mathcal{C},:))^{-1} \\ &= Q_m^{-1} - \frac{\gamma Q_m^{-1}\mathbf{I}_{d\times d}(:,\mathcal{C})\mathbf{I}_{d\times d}(\mathcal{C},:)Q_m^{-1}}{1 + \gamma\mathbf{I}_{d\times d}(\mathcal{C},:)Q_m^{-1}\mathbf{I}_{d\times d}(:,\mathcal{C})} \end{aligned} \tag{15}$$

Therefore, we only need to do the matrix inverse operation once, which is the inverse of Q_0, its time complexity is $O(d^3)$. In fact, this cost can be further reduced since Laplacian matrix $\mathbf{L}$ is usually a sparse matrix with $O(d)$ nonzero elements. The proposed framework REFS is summarized in Algorithm 1, we update Eq. (15) sequentially until getting the desired number of selected features.

We briefly review Algorithm 1. In line 1, we build the feature adjacency matrix $\mathbf{W}$ on the original data $\mathbf{X}$ and calculate the Laplacian matrix $\mathbf{L}$. We compute Q_0 according to Eq. (12) (line 3) and then get its inverse in line 4. In line 8, for each candidate feature j, we update the inverse of Q_j using Eq. (15). The best feature j is obtained in line 13 by Eq. (13). The procedure repeats (line 5-17) until the desired number of features is obtained, which is k.

Algorithm 1 Reconstruction-based Unsupervised Feature Selection (REFS)

Input: $\mathbf{X} \in \mathbb{R}^{n\times d}$, parameters α and β, number of nearest neighbors p and the number of features to select k
Output: k most representative features

1: Construct feature adjacency matrix $\mathbf{W}$ and calculate the Laplacian matrix $\mathbf{L}$
2: Initialize $\mathcal{S} \leftarrow \emptyset$, $i \leftarrow 0$, $j \leftarrow 1$, $\gamma \leftarrow 1-\alpha$
3: $Q_0 \leftarrow \alpha\mathbf{I} + \beta\mathbf{L}$
4: Compute Q_0^{-1}
5: **while** $i < k$ **do**
6: **while** $j \leq d$ **do**
7: **if** $j \in (\mathcal{F} \setminus \mathcal{S})$ **then**
8: $Q_j^{-1} \leftarrow Q_i^{-1} - \frac{\gamma Q_i^{-1}\mathbf{I}_{d\times d}(:,j)\mathbf{I}_{d\times d}(j,:)Q_i^{-1}}{1+\gamma\mathbf{I}_{d\times d}(j,:)Q_i^{-1}\mathbf{I}_{d\times d}(:,j)}$
9: $v_j \leftarrow ||\beta\mathbf{X}\mathbf{L}Q_j^{-1}||_F^2$
10: **end if**
11: $j \leftarrow j+1$
12: **end while**
13: $j \leftarrow \arg\min_j v_j$
14: $\mathcal{S} \leftarrow \mathcal{S} \cup f_j$
15: $i \leftarrow i+1$
16: $Q_i^{-1} \leftarrow Q_j^{-1}$
17: **end while**
18: return $\mathcal{S}$

3 Experiments

In this section, we conduct experiments to evaluate the performance of the proposed framework REFS. After introducing the experimental settings, we compare the proposed framework REFS with the state-of-the-art unsupervised feature selection methods. Further experiments are designed to investigate the effects of parameters α and β on REFS.

3.1 Experimental Settings

We choose 8 benchmark datasets of various types for evaluation, including three image datasets, i.e., object image dataset COIL20[1] [Nene *et al.*, 1996], face image dataset ORL[2], handwritten digit datasets USPS [Hull, 1994]; two text datasets, i.e., RELATHE and BASEHOCK[3]; two microarray datasets, i.e., Lung [Bhattacharjee et al., 2001] and GLIOMA [Nutt et al., 2003] and one spoken letter recognition dataset Isolet[4]. Detailed datasets description are summarized in Table 1.

Table 1: Dataset description

Dataset	size	# of features	# of classes
COIL20	1440	1024	20
ORL	400	1024	40
USPS	9298	256	10
RELATHE	1427	4322	2
BASEHOCK	1993	4862	2
Lung	203	3312	5
GLIOMA	50	4433	4
Isolet	1560	617	26

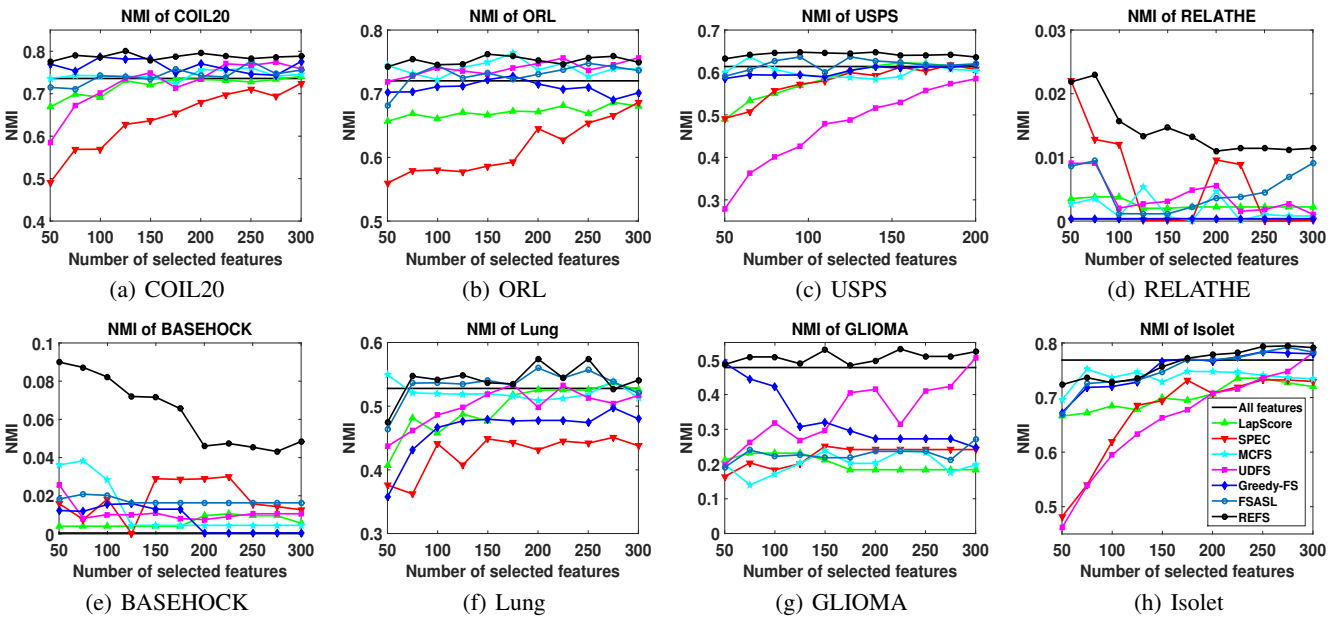
Following a common way to assess unsupervised feature selection [Cai *et al.*, 2010; Yang *et al.*, 2011; Li *et al.*, 2012], we use the clustering performance to evaluate the quality of selected features. Two commonly used clustering performance evaluation metrics, i.e., Accuracy (ACC) and Normalized Mutual Information (NMI) [Cai *et al.*, 2010] are used in this paper. With a specific clustering algorithm, the clustering performance can reflect the quality of selected features. The higher the ACC and NMI values are, the better feature selection performance is.

3.2 Performance Comparison

We compare REFS with the following state-of-the-art unsupervised feature selection algorithms.

- **All Features**: All original features are adopted.
- **LapScore**: Laplacian score [He *et al.*, 2005] which selects features that best preserve the local manifold structure of data.
- **SPEC**: Features are selected by spectral analysis [Zhao and Liu, 2007] and it is an extension of Laplacian score.
- **MCFS**: Multi-Cluster unsupervised feature selection [Cai *et al.*, 2010] which selects features by two steps - first performing spectral regression and then applying Lasso.
- **UDFS**: Unsupervised discriminative feature selection [Yang *et al.*, 2011] which selects features by exploiting local discriminative information and feature correlation simultaneously.
- **GreedyFS**: GreedyFS is an unsupervised feature selection algorithm based on data reconstruction which predefines a reconstruction function [Farahat *et al.*, 2011].
- **FSASL**: FSASL is an unsupervised method which performs data manifold learning and feature selection simultaneously [Du and Shen, 2015].

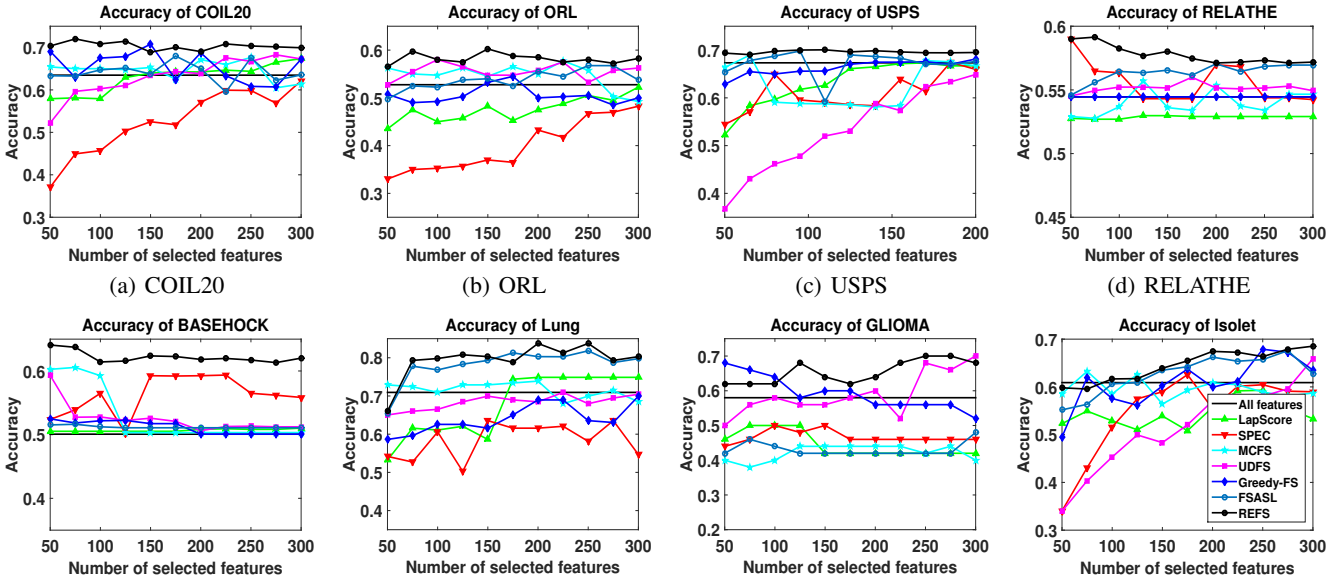
[1] http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html
[2] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
[3] http://featureselection.asu.edu/datasets.php
[4] http://archive.ics.uci.edu/ml/datasets/ISOLET



(a) COIL20 (b) ORL (c) USPS (d) RELATHE

(e) BASEHOCK (f) Lung (g) GLIOMA (h) Isolet

Figure 2: Comparisons of clustering results (NMI) from different unsupervised feature selection algorithms.



(a) COIL20 (b) ORL (c) USPS (d) RELATHE

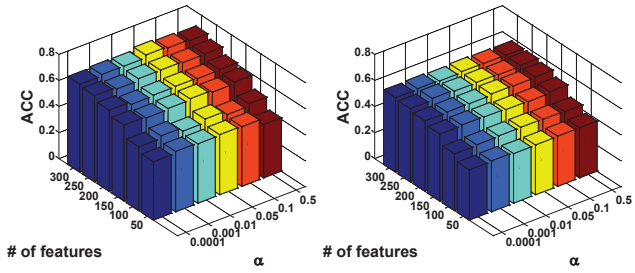
(e) BASEHOCK (f) Lung (g) GLIOMA (h) Isolet

Figure 3: Comparisons of clustering results (ACC) from different unsupervised feature selection algorithms.

For LapScore, MCFS, UDFS and FSASL, we specify the number of neighborhood size to be 5 to construct the Laplacian matrix on the data instances following previous work. In REFS, we also set the number nearest neighborhood size p to be 5, but the Laplacian matrix is built on the feature vectors instead of on the data instances. MCFS, UDFS, FSASL and REFS all have different regularization parameters. For a fair comparison, we tune these regularization parameters for all methods by grid search and report the best performance. How to determine the optimal number of selected features is still a challenging problem, thus we vary the number of selected features as $\{50, 75, 100, ..., 275, 300\}$ for all datasets except USPS. Since USPS has 256 features, we set the number of selected features as $\{50, 65, 80, ..., 185, 200\}$. In the evaluation phase, we use K-means to cluster samples based on the selected features.

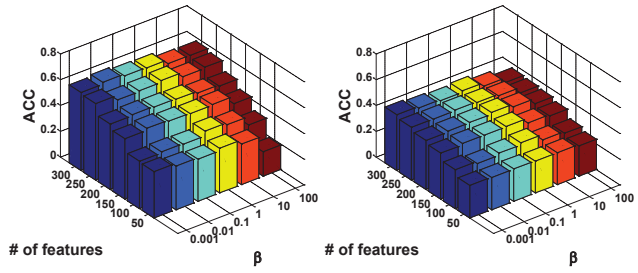
Figure 2 and Figure 3 show the clustering performance in terms of NMI and ACC by different unsupervised feature selection algorithms. We make the following observations:

- In most situations, feature selection algorithms are nec-



(a) COIL20 (b) Isolet

Figure 4: ACC of REFS on COIL20 and Isolet datasets with respect to different α and feature numbers ($\beta = 0.1$).



(a) COIL20 (b) Isolet

Figure 5: ACC of REFS on COIL20 and Isolet datasets with respect to different β and feature numbers ($\alpha = 0.1$).

essary and effective, which can improve the clustering performance.

- The proposed framework REFS can achieve better clustering performance even with a very small number of features, such as 50 features. These results suggest that REFS can select very representative features.

- In text data, i.e., RELATHE and BASEHOCK, and gene data, i.e., Lung and GLIOMA, REFS greatly improves the clustering performance in terms of both NMI and ACC. It may be because of the strong feature dependencies in these datasets, such as synonyms or antonyms words in text data and genes with similar functions.

- REFS outperforms the representative data reconstruction algorithm GreedyFS. We perform pairwise Wilcoxon signed-rank test between them and the test results show that REFS is significantly better. There are two reasons for the improvement - (1) REFS learns the reconstruction function from the data automatically while GreedyFS predefines a reconstruction function; and (2) REFS considers both the feature reconstruction ability and feature structures preserving ability; while GreedyFS only considers the reconstruction ability.

3.3 Effects of Parameters α and β

REFS has two important parameters α and β. The parameter α prevents the overfiting of the reconstruction function. We first investigate how it affects the performance of REFS by varying its value as $\{10^{-4}, 10^{-3}, 0.01, 0.05, 0.1, 0.2, 0.5\}$ when β is 0.1. The other parameter β controls how strongly the reconstruction function preserves the original feature structures. We vary its value as $\{10^{-5}, 10^{-4}, 10^{-3}, 0.01, 0.05, 0.1, 0.2, 0.5, 1, 5, 10, 100\}$ when α is 0.1. Performance variance results are presented in Figure 4 and Figure 5. Due to space limit, we only report the results in terms of ACC on COIL20 and Isolet datasets. We can observe that the performance is not sensitive to both α and β, the clustering performance does not vary much when α and β are in a wide range. However, the clustering results are relatively more sensitive to the number of selected features, which is still an open problem in feature selection.

4 Related Work

We briefly review some related work on reconstruction-based and preserving-based feature selection, which is most related to our framework REFS. Convex Principle Feature Selection (CPFS) [Masaeli *et al.*, 2010] reformulates the feature selection problem as a convex continuous optimization problem that minimizes a mean-squared-reconstruction error with linear and sparsity constraint. GreedyFS [Farahat *et al.*, 2011] uses a projection matrix to project the original data onto the span of some representative feature vectors. Preserving-based feature selection methods select features that best preserve similarities between data instances. Laplacian Score (LapScore) [He *et al.*, 2005] evaluates the importance of a feature through its power of locality preservation, i.e., preserving neighborhood structures of data. Spectral Feature Selection (SPEC) [Zhao and Liu, 2007] extends the idea of LapScore and proposes a unified framework to rank features based on general similarity matrix. These works are substantially different from our proposed framework REFS - (1) REFS does not rely on any predefined reconstruction function while directly learns the reconstruction function from data; and (2) REFS is from the feature preserving perspective which measures feature correlation instead of data similarity.

5 Conclusion

We propose a novel reconstruction-based unsupervised feature selection framework REFS, which embeds the reconstruction function learning process to feature selection. During the reconstruction function learning phase, we take into account both the reconstruction ability and preserving ability of reconstructed features. The optimization problem of REFS is an integer programming problem and it is difficult to solve; hence we propose to use an efficient method to select features sequentially to obtain a local optimal solution. Experimental results on various types of datasets show that REFS outperforms the state-of-the-art unsupervised feature selection algorithms in clustering performance. Further directions can be focused on finding a global optimal solution to optimize the reconstruction problem in this paper.

Acknowledgements

This material is, in part, supported by National Science Foundation (NSF) under grant number 1614576.

References

[Bhattacharjee et al., 2001] Arindam Bhattacharjee et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, 2001.

[Boutsidis *et al.*, 2008] Christos Boutsidis, Michael W Mahoney, and Petros Drineas. Unsupervised feature selection for principal components analysis. In *KDD*, 2008.

[Cai *et al.*, 2010] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *KDD*, 2010.

[Chen *et al.*, 2010] Chun Chen, Zhengguang Chen, Jiajun Bu, Can Wang, Lijun Zhang, and Cheng Zhang. G-optimal design with laplacian regularization. In *AAAI*, 2010.

[Cover and Thomas, 2012] Thomas M Cover and Joy A Thomas. *Elements of information theory*. 2012.

[Ding and Peng, 2005] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *JBCB*, 2005.

[Du and Shen, 2015] Liang Du and Yi-Dong Shen. Unsupervised feature selection with adaptive structure learning. In *KDD*, 2015.

[Duda *et al.*, 2012] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. 2012.

[Farahat *et al.*, 2011] Ahmed K Farahat, Ali Ghodsi, and Mohamed S Kamel. An efficient greedy method for unsupervised feature selection. In *ICDM*, 2011.

[Golub and Van Loan, 2012] Gene H Golub and Charles F Van Loan. *Matrix computations*. 2012.

[Guyon and Elisseeff, 2003] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *JMLR*, 2003.

[He *et al.*, 2005] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *NIPS*, 2005.

[He, 2010] Xiaofei He. Laplacian regularized d-optimal design for active learning and its application to image retrieval. *TIP*, 2010.

[Hull, 1994] Jonathan J. Hull. A database for handwritten text recognition research. *TPAMI*, 1994.

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

[Jian *et al.*, 2016] Ling Jian, Jundong Li, Kai Shu, and Huan Liu. Multi-label informed feature selection. In *IJCAI*, 2016.

[Li and Liu, 2017] Jundong Li and Huan Liu. Challenges of feature selection for big data analytics. *IEEE Intelligent Systems*, 2017.

[Li *et al.*, 2012] Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, and Hanqing Lu. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*, 2012.

[Li *et al.*, 2015] Jundong Li, Xia Hu, Jiliang Tang, and Huan Liu. Unsupervised streaming feature selection in social media. In *CIKM*, 2015.

[Li *et al.*, 2016a] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *arXiv preprint arXiv:1601.07996*, 2016.

[Li *et al.*, 2016b] Jundong Li, Xia Hu, Ling Jian, and Huan Liu. Toward time-evolving feature selection on dynamic networks. In *ICDM*, 2016.

[Li *et al.*, 2017] Jundong Li, Liang Wu, Osmar R Zaïane, and Huan Liu. Toward personalized relational learning. In *SDM*, 2017.

[Liu and Motoda, 2007] Huan Liu and Hiroshi Motoda. *Computational methods of feature selection*. 2007.

[Masaeli *et al.*, 2010] Mahdokht Masaeli, Yan Yan, Ying Cui, Glenn Fung, and Jennifer G Dy. Convex principal feature selection. In *ICDM*, 2010.

[Mitra *et al.*, 2002] Pabitra Mitra, CA Murthy, and Sankar K. Pal. Unsupervised feature selection using feature similarity. *TPAMI*, 2002.

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

[Nie *et al.*, 2008] Feiping Nie, Shiming Xiang, Yangqing Jia, Changshui Zhang, and Shuicheng Yan. Trace ratio criterion for feature selection. In *AAAI*, 2008.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint 2, 1-norms minimization. In *NIPS*, 2010.

[Nutt et al., 2003] Catherine L Nutt et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. 2003.

[Qian and Zhai, 2013] Mingjie Qian and Chengxiang Zhai. Robust unsupervised feature selection. In *IJCAI*, 2013.

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

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *JRSS(B)*, 1996.

[Wei and Philip, 2016] Xiaokai Wei and S Yu Philip. Unsupervised feature selection by preserving stochastic neighbors. In *AISTATS*, 2016.

[Yang *et al.*, 2011] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. l2, 1-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, 2011.

[Zhang *et al.*, 2011] Lijun Zhang, Chun Chen, Jiajun Bu, Deng Cai, Xiaofei He, and Thomas S Huang. Active learning based on locally linear reconstruction. *TPAMI*, 2011.

[Zhao and Liu, 2007] Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In *ICML*, 2007.