2014 IEEE International Conference on Data Mining

Robust Spectral Learning for Unsupervised Feature Selection

Lei Shi*†, Liang Du*, Yi-Dong Shen*
*State Key Laboratory of Computer Science,
Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
†University of Chinese Academy of Sciences, Beijing 100049, China
{shilei,duliang,ydshen}@ios.ac.cn

***Abstract*—In this paper, we consider the problem of unsupervised feature selection. Recently, spectral feature selection algorithms, which leverage both graph Laplacian and spectral regression, have received increasing attention. However, existing spectral feature selection algorithms suffer from two major problems: 1) since the graph Laplacian is constructed from the original feature space, noisy and irrelevant features may have adverse effect on the estimated graph Laplacian and hence degenerate the quality of the induced graph embedding; 2) since the cluster labels are discrete in natural, relaxing and approximating these labels into a continuous embedding can inevitably introduce noise into the estimated cluster labels. Without considering the noise in the cluster labels, the feature selection process may be misguided. In this paper, we propose a Robust Spectral learning framework for unsupervised Feature Selection (RSFS), which jointly improves the robustness of graph embedding and sparse spectral regression. Compared with existing methods which are sensitive to noisy features, our proposed method utilizes a robust local learning method to construct the graph Laplacian and a robust spectral regression method to handle the noise on the learned cluster labels. In order to solve the proposed optimization problem, an efficient iterative algorithm is proposed. We also show the close connection between the proposed robust spectral regression and robust Huber M-estimator. Experimental results on different datasets show the superiority of RSFS.**

I. Introduction

In many tasks of machine learning and data mining, the high dimensionality of data presents challenges to the current learning algorithms. Feature selection, which selects a subset of relevant features, is an effective way to solve these challenges. Recently, a lot of methods have been proposed to address the problem of unsupervised feature selection [1], [2], [3], [4], [5], [6], [7], [8]. These methods usually characterize the structure of data by graph Laplacian, which is defined based on the nearest neighbor graph. In [2], [7], each feature is selected independently according to some specified criterion. Spectral feature selection algorithms, which explore the graph embedding and jointly evaluate features using sparse spectral regression, have received increasing attention in these years. These methods include [1], [3], [4], [5], [6].

Though many spectral feature selection algorithms have been proposed, at least two problems remain not addressed properly. One problem is the construction of graph Laplacian, which can reflect the structure (such as discriminative and geometrical information) of the data. The quality of the constructed graph Laplacian are vitally important to the success of the induced graph embedding (which is also known as pseudo cluster label) and further spectral feature selection algorithms. However, since these methods construct the graph Laplacian from the original uniform feature space, noisy features and outliers will have an adverse effect on the construction of the graph and hence deteriorate the performance of feature selection. Another problem lies on the cluster structure induced by graph embedding. Due to the discrete nature of class labels, these approaches relax and approximate the desired class label to continuous graph embedding, and noises will be inevitably introduced. Since the spectral regression model usually adopts the estimated graph embedding to supervise the evaluation of the importance of features through group sparse induced regularization (i.e. ℓ_{21}-norm), without considering the noise and outliers in the estimated cluster structure, the feature selection process may be misguided.

In this paper, we propose a Robust Spectral learning framework for unsupervised Feature Selection (RSFS), which jointly improves the robustness of graph embedding and sparse spectral regression, and hence more faithful feature selection result could be expected. The basic idea of our method is:

- We utilize the local kernel regression to capture the nonlinear geometrical information, where we adopt the ℓ_1-norm to measure the local learning estimation error. Unlike the traditional ℓ_2-norm used in most existing feature selection approaches, the proposed ℓ_1-norm based local kernel regression is more robust to large reconstruction errors, which are often caused by noisy features and outliers. It has been shown that, by utilizing the structure of scaled partition matrix, the introduced ℓ_1-norm local learning can also be reformulated as a graph embedding problem [9]. In this way, effects of noise and outliers are reduced and hence the structure of the data can be better characterized by the learned graph Laplacian.
- The discrete class label is often relaxed and approximated into continuous values by graph embedding; such continuous relaxation may introduce additional noise, so that the feature selection process may be misled in the sparse spectral regression model. In this paper, we propose a robust sparse spectral regression model by explicitly extracting sparse noise in the continuous approximation. Interestingly, it can be shown that our proposed robust spectral regression model has a dual equivalence with Huber M-estimator in robust statistics [10]. Thus, the robustness of our proposed spectral regression model can be interpreted

1550-4786/14 $31.00 © 2014 IEEE
DOI 10.1109/ICDM.2014.58

as reducing the effects of large regression errors, which are often caused by outliers and noise.

To select the most discriminative features robustly, we perform the robust graph embedding and robust spectral regression simultaneously. We propose an efficient iterative algorithm to solve the proposed optimization problem. Extensive experiments are conducted on data sets with and without explicit noise and outliers. Experimental results show the superiority of RSFS when compared with others.

II. Related Work

Feature selection is a fundamental problem in machine learning and has been studied extensively in the literature. Based on the availability of class labels, feature selection algorithms can be classified as supervised algorithms and unsupervised algorithms. Based on whether taking the learning algorithm (e.g., a classification algorithm) into consideration when performing feature selection, the feature selection algorithms can be grouped into three categories, including filter, wrapper and embedded methods.

Compared with supervised feature selection, unsupervised feature selection is a much harder problem due to the absence of class labels. Unsupervised filter methods usually assign each feature a score which can indicate the feature's capacity to preserve the structure of data. Top ranked features are selected since they can best preserve the structure of data. The typical methods include Maximum Variance, Laplacian Score [2], SPEC [11] and EVSC [7]. Unsupervised wrapper methods [3] "wrap" the feature selection process into a learning algorithm and leverage the learning results to select features. Unsupervised embedded methods perform feature selection as a part of model training process, e.g., UDFS [4] and NDFS [5].

Among all the unsupervised feature selection methods, spectral feature selection methods have received increasing attention in recent years. The typical spectral feature selection methods include MCFS [3], MSRF [12], FSSL [13], LGDFS [8], UDFS [4], JELSR [14], NDFS [5] and RUFS [6]. Most of these methods involve the following two steps. The first step is to explore the cluster structure of data by spectral analysis of graph Laplacian or by nonnegative matrix factorization, and the second step selects features via sparsity regularization models, i.e. ℓ_1-norm and ℓ_{21}-norm regularized spectral regression, to preserve the estimating cluster structure. MCFS, MRSF and FSSL apply these two steps separately, while UDFS, JELSR, NDFS and RUFS perform them jointly. Most of the above methods pay no special attention to the noise in features and data points when constructing the graph Laplacian, making the learned graph Laplacian unreliable. On the other hand, since the cluster labels are discrete in natural, relaxing and approximating these labels into a continuous embedding will inevitably introduce noise into the estimated cluster labels. The unreliable graph Laplacian and noise in the cluster labels will degenerate the performance of feature selection.

III. The Proposed Method

In this section, we present our proposed Robust Spectral learning framework for unsupervised Feature Selection (RSFS), which selects feature by performing robust graph embedding to effectively learn the cluster structure and robust spectral regression to handle sparse noise on estimated cluster structure simultaneously. After discussing the robust graph embedding and robust sparse spectral regression, we formulate the optimization problem of RSFS. We also present the algorithm to solve the optimization problem of RSFS.

A. Robust Graph Embedding

Discriminative information is very important for feature selection. In supervised scenario, the discriminative information is encoded in the class labels. By exploring the class labels, it's convenient for supervised feature selection algorithms to select discriminative features. However, in unsupervised scenario, there is no label information available. Thus, it is much more difficult to select discriminative features. One way to select discriminative features in unsupervised scenario is to learn pseudo cluster labels (graph embedding), which can guide the feature selection process. [5] and [3] employed spectral analysis to predict cluster labels. However, since the spectral analysis in [5] and [3] depends on the similarity graph constructed from the original feature space, noise and outliers will have an adverse effect on predicting the pesudo labels and hence deteriorate the performance of feature selection. [6] proposed to learn pseudo labels by local learning regularized nonnegative matrix factorization (NMF). Although the loss function of NMF adopts ℓ_{21}-norm, the local learning term employs a square loss, which is sensitive to noise and outliers.

It has also been shown that the local structure of data is very important for exploring the cluster structure of data [15]. By exploring local structure of data, we can get more accurate pseudo cluster labels. Our goal is to design a method which can both utilize the local structure of data and handle noisy features and outliers for robust graph embedding. In the following, we introduce such a robust local learning method.

Let $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \in \mathbb{R}^{d\times n}$ denote the data matrix, whose columns correspond to data instances and rows to features. Suppose these n instances belong to c classes. Denote $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_c] = [y_{il}] \in \{0, 1\}^{n\times c}$ as a *partition matrix* of data matrix $\mathbf{X}$. To utilize the local structure of data, we assume the label of a data point can be predicted by its neighbors. Formally, for each data point $\mathbf{x}_i$, the label predictor $p_{il}(\cdot)$ is constructed based on its neighborhood information $\{(\mathbf{x}_j, y_{jl})|\mathbf{x}_j \in \mathcal{N}_i)\}$, where $\mathcal{N}_i$ is the neighborhood of $\mathbf{x}_i$. Suppose $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_c] \in \mathbb{R}^{n\times c}$, where $\mathbf{p}_l = [p_{1l}(x_1), p_{2l}(x_2), ..., p_{nl}(x_n)]^T \in \mathbb{R}^n$. Then, the objective function can be written as

$$\min_{\mathbf{Y}\in\mathbb{R}^{n\times c}} J(\mathbf{Y}) = \mathcal{L}(\mathbf{Y}, \mathbf{P}), \tag{1}$$

where $\mathcal{L}$ is a loss function which is robust to noise and outliers and $\mathbf{P}$ is the cluster structure estimated by local learning.

There are many choices for the local predictor p. In order to effectively capture the structure of data, we choose kernel regression as our local predictor. The basic idea of kernel regression is that the prediction of a data point takes the weighted average of the target values of the training data points. The weight is defined by the kernel function. Formally, for each data point $\mathbf{x}_i$, a local kernel regression $p_{il}(\cdot)$ is constructed to estimate the cluster label of $\mathbf{x}_i$, i.e.,

$$p_{il}(\mathbf{x}_i) = \frac{\sum_{\mathbf{x}_j\in\mathcal{N}_i} K(\mathbf{x}_i, \mathbf{x}_j) y_{jl}}{\sum_{\mathbf{x}_j\in\mathcal{N}_i} K(\mathbf{x}_i, \mathbf{x}_j)} \tag{2}$$

where $\mathcal{N}_i$ is the neighborhood of $\mathbf{x}_i$. Define a matrix $\mathbf{S} = [s_{ij}] \in R^{n\times n}$ as follows

$$s_{ij} = \begin{cases} \frac{K(\mathbf{x}_i,\mathbf{x}_j)}{\sum_{\mathbf{x}_j\in N_i} K(\mathbf{x}_i,\mathbf{x}_j)} & \mathbf{x}_j \in \mathcal{N}_i \\ 0 & \mathbf{x}_j \notin \mathcal{N}_i \end{cases} \tag{3}$$

Thus, we have $\mathbf{p}_l = \mathbf{S}\mathbf{y}_l \quad and \quad \mathbf{P} = \mathbf{S}\mathbf{Y}$.

In order to alleviate the side effect of irrelevant and noisy features, here we employ ℓ_1-norm, which reduces the effect of large fitting error. Thus, we have,

$$\min_{\mathbf{Y}\in\mathbb{R}^{n\times c}} J(\mathbf{Y}) = \sum_{l=1}^{c} ||\mathbf{y}_l - \mathbf{S}\mathbf{y}_l||_1. \tag{4}$$

Although the above objective function with respect to the partition matrix $\mathbf{Y}$ is attractive, it's difficult to derive a quadratic form. Following [9], we employ the scaled partition matrix $\mathbf{G} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}$. Balanced clusters, which can lead to better performance in practice, is obtained by the scaling procedure. It can be proved that

$$\min_{\mathbf{G}} \quad ||\mathbf{G} - \mathbf{S}\mathbf{G}||_1$$

is equivalent to minimizing the following problem [9],

$$J(\mathbf{F}) = Tr(\mathbf{F}^T\mathbf{M}\mathbf{F}) \tag{5}$$

where $\mathbf{M} = (\mathbf{B} - \mathbf{S} - \mathbf{S}^T)$. $\mathbf{B}$ is the degree matrix of $(\mathbf{S} + \mathbf{S}^T)$. $\mathbf{F} = [\mathbf{f}_1, ..., \mathbf{f}_c]$ is defined as $\mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}$, and $\mathbf{F}^T\mathbf{F} = \mathbf{I}_c$.

B. Robust Sparse Spectral Regression

The graph embedding is discrete in natural. By relaxing and approximating it in continuous values, noise is inevitably introduced. Without considering the noise on the estimated cluster structure, the feature selection process may be misguided. Motivated by the recent development in robust principle component analysis [16], we propose a robust spectral regression model, which assumes the learned cluster structure may be arbitrarily corrupted, but the corruption is sparse. We introduce a sparse matrix $\mathbf{Z} \in \mathbb{R}^{n\times c}$ to explicitly capture the sparse noise. Thus, the goal of robust spectral regression is to approximate $\mathbf{F}$ as

$$\min_{\mathbf{W},\mathbf{Z}} \quad ||(\mathbf{F} - \mathbf{Z}) - \mathbf{X}^T\mathbf{W}||_F^2, \quad s.t. |\mathbf{Z}|_1 < \eta_1, ||\mathbf{W}||_{2,1} < \eta_2, \tag{6}$$

where η_1 and η_2 are very small positive numbers. $\mathbf{W}$ is the spectral regression coefficients where ℓ_{21}-norm is imposed to pursue row-wise sparsity; Such property makes it suitable for the task of feature selection [17]. Specifically, $\mathbf{w}_i$ shrinks to zeros if the i-th feature is less relevant to the estimated cluster structure. Interestingly, it can be shown later in Eq. (12), Eq. (13) and Eq. (14) that the above problem is equivalent to minimizing the regression error with Huber M-estimator, which actually reduces the large regression error caused by noise and outliers.

C. The Objective Function of RSFS

In the previous subsections, we present a robust method to explore the graph embedding and a robust spectral regression to handle sparse noise on the estimated cluster structure. By combining the robust graph embedding (Eq. (5)) and robust sparse spectral regression (Eq. (6)) into a unified framework, we obtain the following objective function,

$$\begin{aligned} \min_{\mathbf{F},\mathbf{W},\mathbf{Z}} \quad & Tr(\mathbf{F}^T\mathbf{M}\mathbf{F}) + \alpha||(\mathbf{F} - \mathbf{Z}) - \mathbf{X}^T\mathbf{W}||_F^2 \\ & + \beta||\mathbf{W}||_{2,1} + \gamma||\mathbf{Z}||_1 \\ s.t. \quad & \mathbf{F} \in R_+^{n\times c}, \mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}} \end{aligned} \tag{7}$$

where $\alpha, \beta, \gamma \in \mathbb{R}_+$ are parameters. Since the elements in $\mathbf{F}$ are discrete values in nature, the optimization problem in Eq. (7) is an NP-hard problem [15]. By relaxing these discreate values to continuous ones [15], [18], the objective function in Eq. (7) can be relaxed to

$$\begin{aligned} \min_{\mathbf{F},\mathbf{W},\mathbf{Z}} \quad & Tr(\mathbf{F}^T\mathbf{M}\mathbf{F}) + \alpha||(\mathbf{F} - \mathbf{Z}) - \mathbf{X}^T\mathbf{W}||_F^2 \\ & + \beta||\mathbf{W}||_{2,1} + \gamma||\mathbf{Z}||_1 \\ s.t. \quad & \mathbf{F} \in R_+^{n\times c}, \mathbf{F}^T\mathbf{F} = \mathbf{I}_c \end{aligned} \tag{8}$$

D. Algorithm to Solve RSFS

In this subsection, we present an efficient algorithm to solve the optimization problem in Eq. (8). Let

$$\begin{aligned} \mathcal{L}(\mathbf{W},\mathbf{Z},\mathbf{F}) = & Tr(\mathbf{F}^T\mathbf{M}\mathbf{F}) + \alpha||(\mathbf{F} - \mathbf{Z}) - \mathbf{X}^T\mathbf{W}||_F^2 \\ & + \beta||\mathbf{W}||_{2,1} + \gamma||\mathbf{Z}||_1, \end{aligned} \tag{9}$$

where three variables $\mathbf{W}$, $\mathbf{Z}$ and $\mathbf{F}$ are involved. Due to the non-smoothness of the row-sparsity induced ℓ_{21}-norm, we develop an coordinate descent algorithm to alternatively minimizing the above objective function with respect to $\mathbf{W}$, $\mathbf{Z}$, and $\mathbf{F}$, separately. This procedure is repeated until convergence.

1) Optimize $\mathbf{W}$ for fixed $\mathbf{F}$ and $\mathbf{Z}$: The optimization problem for updating $\mathbf{W}$ is equivalent to minimizing

$$\mathcal{L}_1 = ||(\mathbf{F} - \mathbf{Z}) - \mathbf{X}^T\mathbf{W}||_F^2 + \frac{\beta}{\alpha}||\mathbf{W}||_{2,1} \tag{10}$$

Let $\frac{\partial\mathcal{L}_1}{\partial\mathbf{W}} = 2\mathbf{X}(\mathbf{X}^T\mathbf{W} - (\mathbf{F} - \mathbf{Z})) + 2\frac{\beta}{\alpha}\mathbf{D}\mathbf{W} = 0$, thus we get the close-form solution to update $\mathbf{W}$,

$$\mathbf{W} = (\mathbf{X}\mathbf{X}^T + \frac{\beta}{\alpha}\mathbf{D})^{-1}\mathbf{X}(\mathbf{F} - \mathbf{Z}) \tag{11}$$

where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \frac{1}{2||\mathbf{w}_i||}$[1].

2) Optimize $\mathbf{Z}$ for fixed $\mathbf{W}$ and $\mathbf{F}$: The optimization problem of updating $\mathbf{Z}$ is equivalent to minimizing

$$\mathcal{L}_2 = ||\mathbf{E} - \mathbf{Z}||_F^2 + \frac{\gamma}{\alpha}||\mathbf{Z}||_1, \quad \mathbf{E} = \mathbf{F} - \mathbf{X}^T\mathbf{W}. \tag{12}$$

The optimization problem in Eq. (12) can be solved efficiently via the *soft-thresholding operator* [19] and the closed form solution is as follows,

$$\mathbf{Z}_{ij} = \begin{cases} 0, & if \quad |\mathbf{E}_{ij}| \le \frac{\gamma}{2\alpha} \\ (1 - \frac{\gamma}{2\alpha|\mathbf{E}_{ij}|})\mathbf{E}_{ij}, & otherwise \end{cases} \tag{13}$$

[1]To avoid zero values, We use a very small constant σ to regularize $\mathbf{D}_{ii} = \frac{1}{2\sqrt{\mathbf{w}_i\mathbf{w}_i^T + \sigma}}$ [17].

By substituting Eq. (13) into Eq. (12), we get

$$\min \mathcal{L}_2 = \sum_{ij} \overline{\mathbf{E}}_{ij}, \tag{14}$$

where

$$\overline{\mathbf{E}}_{ij} = \begin{cases} \mathbf{E}_{ij}^2, & if \quad |\mathbf{E}_{ij}| \leq \frac{\gamma}{2\alpha} \\ \frac{\gamma}{\alpha}|\mathbf{E}_{ij}| - (\frac{\gamma}{2\alpha})^2, & otherwise \end{cases}$$

The right part of Eq. (14) is denoted as Huber-estimator in robust statistics [10]. Based on the duality between Eq. (12) and Eq. (14), we can find that Eq. (12) imposes an ℓ_2-norm on small errors ($|\mathbf{E}_{ij}| \leq \frac{\gamma}{2\alpha}$) and imposes an ℓ_1-norm on large errors ($|\mathbf{E}_{ij}| > \frac{\gamma}{2\alpha}$). Different from other feature selection methods [3], [5], which use spectral regression with squared loss, our method can adaptively impose the ℓ_1-norm on large errors, which are often caused by noise and outliers. In this way, our method can have better performance even when the data are noisy or corrupted.

*3) Optimize **F** for fixed **W** and **Z**:* By incorporating the orthogonal constraint of $\mathbf{F}$ into the objective function via Langrange multiplier, it is equivalent to optimizing

$$\mathcal{L}_3 = Tr(\mathbf{F}^T\mathbf{M}\mathbf{F}) + \alpha||\mathbf{F}-\mathbf{A}||_F^2 + \frac{\nu}{2}||\mathbf{F}^T\mathbf{F}-\mathbf{I}_c||_F^2 s.t. \quad \mathbf{F} \geq 0 \tag{15}$$

where $\mathbf{A} = \mathbf{X}^T\mathbf{W}+\mathbf{Z}$. In practice, ν is set to be large to ensure the orthogonal condition. Inspired by recent development in non-negative matrix factorization community [20] , we present an iterative multiplicative updating rule to solve Eq. (15). Let $\Phi \in \mathbb{R}^{n\times c}$ be the Lagrangian multiplier, then we have

$$\begin{aligned} \mathcal{L}(\mathbf{F}) = & Tr(\mathbf{F}^T\mathbf{M}\mathbf{F}) + \alpha||\mathbf{F}-\mathbf{A}||_F^2 \\ & + \frac{\nu}{2}||\mathbf{F}^T\mathbf{F}-\mathbf{I}_c||_F^2 + Tr(\Phi\mathbf{F}^T). \end{aligned} \tag{16}$$

Setting $\frac{\partial L(\mathbf{F})}{\partial \mathbf{F}} = 0$, we get $\Phi = -2(\mathbf{M}\mathbf{F}+\alpha\mathbf{F}+\nu\mathbf{F}\mathbf{F}^T\mathbf{F}-\nu\mathbf{F}-\alpha\mathbf{A})$. By employing the KKT condition $\Phi_{ij}\mathbf{F}_{ij} = 0$, we get

$$[\mathbf{M}\mathbf{F} + \alpha\mathbf{F} + \nu\mathbf{F}\mathbf{F}^T\mathbf{F} - \nu\mathbf{F} - \alpha\mathbf{A}]_{ij}\mathbf{F}_{ij} = 0. \tag{17}$$

Algorithm 1 The Optimization Algorithm of RSFS

Input: The data matrix $\mathbf{X} \in R^{d\times n}$, the parameters α, β, γ, ν and k.

Output: Sort all the d features according to $||\mathbf{W}_i||_2 \quad (i = 1, ..., d)$ in descending order and select the top q ranked features.

1: Construct the k-nearest neighbor graph and calculate $\mathbf{S}$ by Eq. (3)
2: Calculate $\mathbf{B}$ as the degree matrix of $(\mathbf{S}+\mathbf{S}^T)$ and $\mathbf{M} = \mathbf{B}-\mathbf{S}-\mathbf{S}^T$
3: The iteration step $t = 1$; Initialize $\mathbf{F}^t \in \mathbb{R}^{n\times c}$ and $\mathbf{Z}^t \in \mathbb{R}^{n\times c}$; set $\mathbf{D}^t \in \mathbb{R}^{d\times d}$ as an identity matrix
4: **repeat**
5: $\quad \mathbf{W}^{t+1} = (\mathbf{X}\mathbf{X}^T + \frac{\beta}{\alpha}\mathbf{D}^t)^{-1}\mathbf{X}(\mathbf{F}^t - \mathbf{Z}^t)$
6: $\quad$ update $\mathbf{Z}$ by Eq. (13)
7: $\quad \mathbf{F}_{ij}^{t+1} \leftarrow \mathbf{F}_{ij}^t\sqrt{\frac{[\mathbf{M}^-\mathbf{F}^t+\nu\mathbf{F}^t+\alpha\mathbf{A}^+]_{ij}}{[\mathbf{M}^+\mathbf{F}^t+\alpha\mathbf{F}+\nu\mathbf{F}^t(\mathbf{F}^t)^T\mathbf{F}^t+\alpha\mathbf{A}^-]_{ij}}}$
8: $\quad$ update the diagonal matrix $\mathbf{D}$ as $\mathbf{D}_{ii}^{t+1} = \frac{1}{2||\mathbf{w}_i^{t+1}||}$;
9: $\quad$ t = t+1;
10: **until** Convergence criterion satisfied
11: Sort each feature according to $||\mathbf{w}_i||$ and select the top ranked ones.

Though the non-negative constraint has been adopted in NDFS [5], the optimization schema developed in NDFS can not be used directly. The problem is due to the fact that $\mathbf{M}$ and $\mathbf{A}$ in Eq. (17) are mix signed. To tackle this problem, we introduce $\mathbf{M} = \mathbf{M}^+ - \mathbf{M}^-$ and $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$ [20], where $\mathbf{M}_{ij}^+ = (|\mathbf{M}_{ij}| + \mathbf{M}_{ij})/2$ and $\mathbf{M}_{ij}^- = (|\mathbf{M}_{ij}| - \mathbf{M}_{ij})/2$. We get

$$[(\mathbf{M}^+ - \mathbf{M}^-)\mathbf{F} + \alpha\mathbf{F} + \nu\mathbf{F}\mathbf{F}^T\mathbf{F} - \nu\mathbf{F} - \alpha(\mathbf{A}^+ - \mathbf{A}^-)]_{ij}\mathbf{F}_{ij} = 0.$$

Then, we obtain the following updating rule

$$\mathbf{F}_{ij} \leftarrow \mathbf{F}_{ij}\sqrt{\frac{[\mathbf{M}^-\mathbf{F} + \nu\mathbf{F} + \alpha\mathbf{A}^+]_{ij}}{[\mathbf{M}^+\mathbf{F} + \alpha\mathbf{F} + \nu\mathbf{F}\mathbf{F}^T\mathbf{F} + \alpha\mathbf{A}^-]_{ij}}}. \tag{18}$$

We summarize the optimization algorithm in Algorithm 1.

IV. Experiments

A. Data Sets

Six data sets are used in our experiments, including two face data sets, i.e., ORL [3] and Jaffe [21], one object image data set, i.e., COIL20 [3], two text data sets, i.e., BBCSport [22] and WebKB4 [23], and one handwritten data set, i.e., MNIST [24]. Table I gives a summary of these data sets.

TABLE I. Summary of data sets

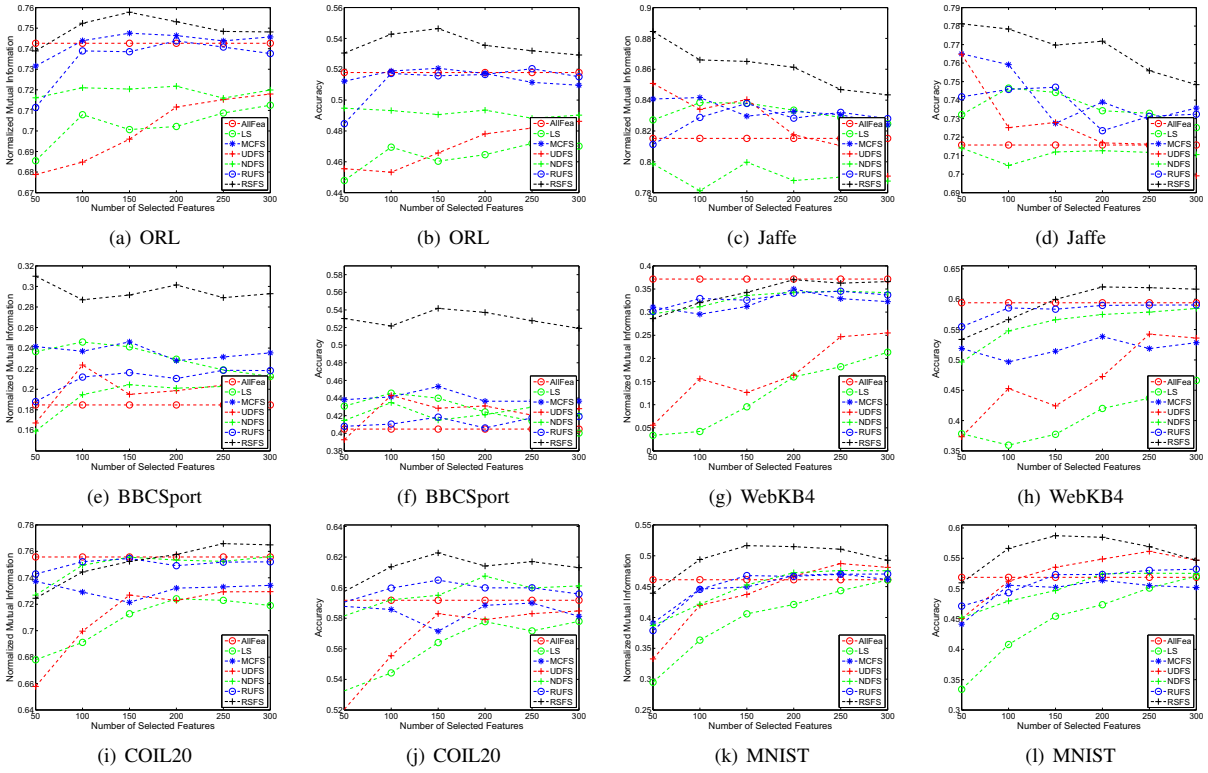
Dataset	Size	Dimensions	Classes
BBCSport	737	1000	5
WebKB4	4199	1000	4
ORL	400	1024	40
COIL20	1440	1024	20
MNIST	4000	784	10
Jaffe	213	676	10

B. Methods to Compare

We systematically compare 3 weak baselines (AllFea, Laplacian Score [2], MCFS [3]) and 3 strong baselines (UDFS [4], NDFS [5] and RUFS [6]) in unsupervised feature selection literatures.

- AllFea, which selects all the features.
- LS[2] [2], which selects those features that can best preserve the local manifold structure of data.
- MCFS[3] [3], which selects the features by adopting spectral regression with ℓ_1-norm regularization. The neighborhood size is set to 5.
- UDFS[4] [4], which exploits local discriminative information and feature correlations simultaneously and considers the manifold structure as well. The parameters are searched from the grid $\{10^{-9}, 10^{-6}, 10^{-3}, 1, 10^3, 10^6, 10^9\}$ and the neighborhood size is 5 as used in [4].
- NDFS[5] [5], which selects features by a joint framework of nonnegative spectral analysis and $\ell_{2,1}$-norm regularized regression. The parameters are searched from the grid $\{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2, 10^4, 10^6\}$ and the neighborhood size is 5 as used in [5].

[2]http://www.cad.zju.edu.cn/home/dengcai/Data/code/LaplacianScore.m
[3]http://www.cad.zju.edu.cn/home/dengcai/Data/code/MCFS_p.m
[4]http://www.cs.cmu.edu/~yiyang/UDFS.rar
[5]https://sites.google.com/site/zcliustc/home/publication/AAAI2012.m



(a) ORL (b) ORL (c) Jaffe (d) Jaffe

(e) BBCSport (f) BBCSport (g) WebKB4 (h) WebKB4

(i) COIL20 (j) COIL20 (k) MNIST (l) MNIST

Fig. 1. Clustering accuracy and normalized mutual information versus the number of selected features on all the data sets

- RUFS[6] [6], which selects feature by robust nonnegative matrix factorization and robust regression. The parameters are searched from the grid $\{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2, 10^4, 10^6\}$ and the neighborhood size is 5 as used in [6].
- RSFS[7], which is proposed in this paper. We tune the parameters from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3\}$. The neighborhood size is set to be 5.

For each method, the parameters are searched in the grid as described. For the selected features, the K-means algorithm is applied 20 times with random initiation and the best average result is reported. Clustering Accuracy and Normalized Mutual Information are used to evaluate the performance of different algorithms.

C. Clustering on Data Sets without Explicit Noise

We first evaluate the performance of the seven methods on data sets without explicit noise. The clustering results, in terms of NMI and ACC, are reported in Figure 1. We have the following observations. Firstly, by selecting features jointly and utilizing discriminative information, MCFS, UDFS, NDFS, RUFS and RSFS have a better performance than LS. By learning graph embedding and performing feature selection simultaneously, RSFS, RUFS and NDFS have a better performance over most of the datasets. By considering outliers and noise, RSFS and RUFS achieve better performance than other methods. At last, out proposed RSFS achieves the best performance. This can be explained by the following main reasons. First, the robust graph embedding learning method can learn better cluster structure. It also explores the local structure of data, which has been shown to be important for data analysis. Second, by assuming sparse noise in the learned pseudo label matrix, we propose a robust regression model to handle the noise. Third, the robust graph embedding method and the robust spectral regression are performed jointly to handle noise and outliers in data.

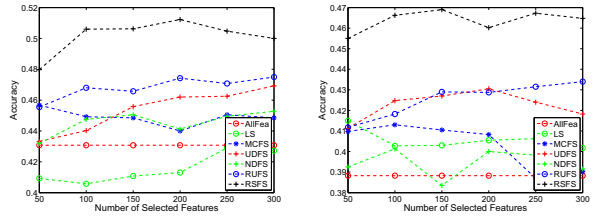
[6]http://web.engr.illinois.edu/~mqian2/upload/research/RUFS/RUFS.m

[7]http://kingsleyshi.com/codes/RSFS.rar

D. Clustering on Data Sets with Malicious Occlusion

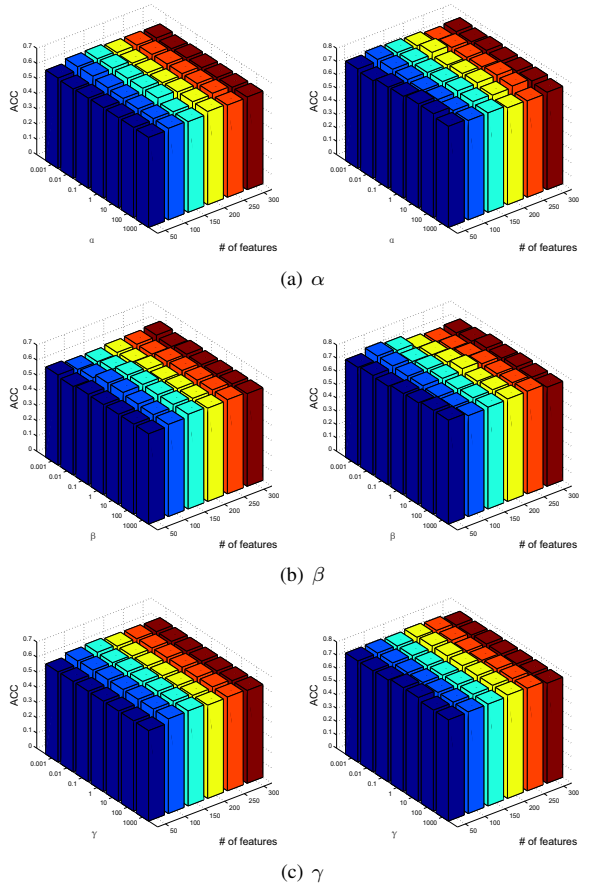
In this subsection, we describe the experimental results on data sets with explicit noise and outliers. In this experiment, we use the ORL data set, which contains 400 gray scale images of 40 individuals. In order to impose some noise to the original ORL data set, different ratio (0.2, 0.3) of images are randomly selected and partially occluded with random blocks. 10 tests were conducted on different randomly chosen percentage of outliers, and the average performance over the 10 data sets is reported.

Figure 2 shows the clustering results in term of ACC for the methods over datasets with different ratio of noise. We have two observations. First, our proposed method can achieve the best performance over all the corrupted data sets. Second, the improvement between our method and other methods increases when the ratio of corruption varies from 0.2 to 0.3.



(a) Corrupted ratio = 0.2 (b) Corrupted ratio = 0.3

Fig. 2. Clustering Accuracy on ORL with different ratio of noisy images



(a) α

(b) β

(c) γ

Fig. 3. Clustering Accuracy with different parameters

E. Effects of the Parameters

In this subsection, we study the sensitiveness of parameters. Due to space limit, we only report the results on COIL20 and Jaffe in Figure 3. Figure 3 shows the best clustering accuracy with respect to each of the parameters over the selected features. The results show that our method is not very sensitive to the parameters α, β and γ.

V. Conclusion

We have proposed a novel robust unsupervised feature selection framework, called RSFS, which is a unified framework that jointly performs robust graph embedding learning and robust sparse spectral regression. To solve optimization problem of RSFS, an efficient iterative algorithm was proposed. The extensive experimental results show that our proposed method outperforms other state-of-the-art methods.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments and suggestions. This work is supported in part by China National 973 program 2014CB340301 and NSFC grant 61379043.

References

[1] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy." in *AAAI*, 2010.

[2] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *NIPS*, vol. 186, 2005, p. 189.

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

[4] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "l 2, 1-norm regularized discriminative feature selection for unsupervised learning," in *IJCAI*. AAAI Press, 2011, pp. 1589–1594.

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

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

[7] Y. Jiang and J. Ren, "Eigenvalue sensitive feature selection," in *ICML*, 2011, pp. 89–96.

[8] L. Du, Z. Shen, X. Li, P. Zhou, and Y.-D. Shen, "Local and global discriminative learning for unsupervised feature selection," in *ICDM*, 2013, pp. 131–140.

[9] J. Sun, Z. Shen, H. Li, and Y. Shen, "Clustering via local regression," in *ECML/PKDD*. Springer, 2008, pp. 456–471.

[10] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011, vol. 114.

[11] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *ICML*, 2007, pp. 1151–1157.

[12] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *AAAI*, 2010, pp. 673–678.

[13] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *IJCAI*, 2011, pp. 1294–1299.

[14] C. Hou, F. Nie, D. Yi, and Y. Wu, "Feature selection via joint embedding learning and sparse regression," in *IJCAI*, 2011, pp. 1324–1329.

[15] J. Shi and J. Malik, "Normalized cuts and image segmentation," *PAMI*, vol. 22, no. 8, pp. 888–905, 2000.

[16] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[17] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint l2, 1-norms minimization," *NIPS*, vol. 23, pp. 1813–1821, 2010.

[18] Y. Yang, H. T. Shen, F. Nie, R. Ji, and X. Zhou, "Nonnegative spectral clustering with discriminative regularization." in *AAAI*, 2011.

[19] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106.

[20] Q. Gu and J. Zhou, "Co-clustering on manifolds," in *KDD*. ACM, 2009, pp. 359–368.

[21] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *PAMI*, vol. 21, no. 12, pp. 1357–1362, 1999.

[22] D. Greene and P. Cunningham, "Producing accurate interpretable clusters from high-dimensional data," in *PKDD*, 2005, pp. 486–494.

[23] C. Ding and T. Li, "Adaptive dimension reduction using discriminant analysis and k-means clustering," in *ICML*. ACM, 2007, pp. 521–528.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.