Robust Spectral Learning for Unsupervised Feature Selection

Lei Shi, Liang Du and Yi-Dong Shen

Institute of Software, Chinese Academy of Sciences

ICDM 2014, Shenzhen, China

Outline

1. Brief Introduction of Feature Selection
2. Our Proposal
3. Experiments
4. Summary

Brief Introduction of Feature Selection

- select a subset of features from original feature space
  - Reduce computation cost
  - Remove noisy features

Brief Introduction of Feature Selection

- Supervised Feature Selection
  - leverage the supervised information to guide the search of relevant features
- Unsupervised Feature Selection
  - more challenging due to the lack of supervised information (e.g.,class labels)

Main Techniques of Unsupervised Feature Selection

- Keys
  - 1) Structure characterization; 2) Feature Weight Learning
- Main Steps
  - estimate intrinsic structure (e.g, various graph Laplacian)
  - discover the cluster structure (e.g, spectral embedding, Matrix Factorization)
  - Feature weight learning by some criterion or sparsity regularization
- Examples
  - LaplacianScore (NIPS'05), MCFS (KDD'10), NDFS (AAAI'12), RUFS (IJCAI'13)

Limitations of Existing Work

- Graph Laplacian is affected by noisy features
  - Existing method construct graph Laplacian from original feature space, which contains noisy features
  - degenerate the quality of the induces graph embedding
- Estimated cluster structure is with noise
  - relaxing discrete class labels into continuous ones may inevitably introduce noise

Outline

1 Brief Introduction of Feature Selection

2 Our Proposal

3 Experiments

4 Summary

Basic Idea

- jointly improve the robustness of graph embedding and sparse spectral regression
    - employ local kernel regression with ℓ_1 norm to measure the local learning estimation error, which can be reformulated as a graph embedding problem.
    - explicitly extracting sparse noise in the learned graph embedding

Robust Graph Embedding

- Notations
  - Data matrix : $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \in \mathbb{R}^{d \times n}$
  - Partition matrix : $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_c] = [y_{il}] \in \{0, 1\}^{n \times c}$
- Robust Local kernel regression

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

$$s_{ij} = \begin{cases} \frac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{\mathbf{x}_j \in N_i} K(\mathbf{x}_i, \mathbf{x}_j)} & \mathbf{x}_j \in \mathcal{N}_i \\ 0 & \mathbf{x}_j \notin \mathcal{N}_i \end{cases} \quad (2)$$

Robust Graph Embedding

- Equivalent to minimizing the following problem

$$J(\mathbf{F}) = Tr(\mathbf{F}^T \mathbf{M} \mathbf{F}) \tag{3}$$

$\mathbf{M} = (\mathbf{B} - \mathbf{S} - \mathbf{S}^T)$. $\mathbf{B}$ 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 $\mathbf{F}^T\mathbf{F} = \mathbf{I}_c$

Robust Spectral Regression

- assume the noise on the estimated cluster structure is sparse
- introduce a sparse matrix $\mathbf{Z} \in \mathbb{R}^{n \times c}$ to explicitly capture the sparse noise

$$\min_{\mathbf{W},\mathbf{Z}} \quad ||(\mathbf{F} - \mathbf{Z}) - \mathbf{X}^T\mathbf{W}||_F^2, \quad s.t. |\mathbf{Z}|_1 < \eta_1, ||\mathbf{W}||_{2,1} < \eta_2 \qquad (4)$$

Overall Framework

Combining the robust graph embedding and robust sparse spectral regression, we get

$$\begin{aligned}\min_{\mathbf{F},\mathbf{W},\mathbf{Z}} \quad & Tr(\mathbf{F}^T\mathbf{M}\mathbf{F}) + \alpha||(\mathbf{F}-\mathbf{Z}) - \mathbf{X}^T\mathbf{W}||_F^2 \\ & + \beta||\mathbf{W}||_{2,1} + \gamma||\mathbf{Z}||_1 \\ s.t. \quad & \mathbf{F} \in R_+^{n\times c}, \mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\end{aligned} \tag{5}$$

Relaxing the elements in **F** into continuous ones,

$$\begin{aligned}\min_{\mathbf{F},\mathbf{W},\mathbf{Z}} \quad & Tr(\mathbf{F}^T\mathbf{M}\mathbf{F}) + \alpha||(\mathbf{F}-\mathbf{Z}) - \mathbf{X}^T\mathbf{W}||_F^2 \\ & + \beta||\mathbf{W}||_{2,1} + \gamma||\mathbf{Z}||_1 \\ s.t. \quad & \mathbf{F} \in R_+^{n\times c}, \mathbf{F}^T\mathbf{F} = \mathbf{I}_c\end{aligned} \tag{6}$$

Algorithm to Sovle **RSFS**

We develop an coordinate descent algorithm to alternatively minimizing the objective function with respect to **W**, **Z**, and **F**. This procedure is repeated until convergence.
Please refer to our paper for more details.

Outline

1. Brief Introduction of Feature Selection
2. Our Proposal
3. Experiments
4. Summary

Experimental setup

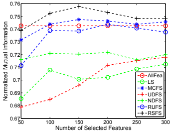
- Comparing methods
  - Random, LS (NIPS2005), MCFS (KDD2010), UDFS (IJCAI2011), NDFS (AAAI2012), RUFS(IJCAI2013)
- Data sets

Table: Summary of data sets

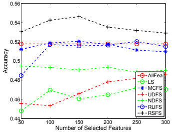
Dataset	Size	Dimensions	Classes
BBCSport	737	1000	5
WebKB4	4199	1000	4
ORL	400	1024	40
COIL20	1440	1024	20
MNIST	4000	784	10
Jaffe	213	676	10

- Perform k-means clustering on the selected features. ACC and NMI are reported

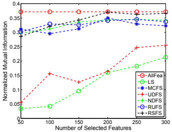
Clustering on Data Sets without Explicit Noise



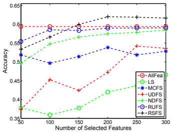
(a) ORL



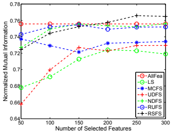
(b) ORL



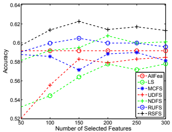
(c) WebKB4



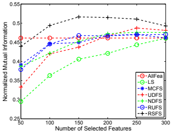
(d) WebKB4



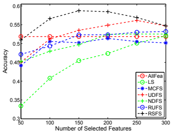
(e) COIL20



(f) COIL20



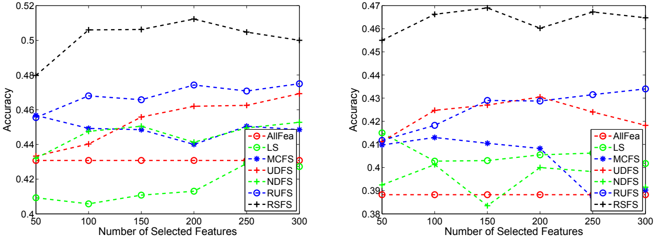
(g) MNIST



(h) MNIST

Figure: Clustering accuracy and normalized mutual information versus the number of selected features on all the data sets

Clustering on Data Sets with Malicious Occlusion



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

Figure: Clustering Accuracy on ORL with different ratio of noisy images

Outline

1. Brief Introduction of Feature Selection
2. Our Proposal
3. Experiments
4. Summary

Summary

- we proposed a unified robust spectral framework for unsupervised feature selection
  - robust graph embedding
  - robust sparse spectral regression
- experimental results verified the effectiveness of our proposed method
  - in datasets with and without explicit noise

Thanks for your attention!
Questions?

Code available at : http://kingsleyshi.com/codes/RSFS.rar