Unsupervised Feature Selection with Adaptive Structure Learning

Liang Du and Yi-Dong Shen

Institute of Software, Chinese Academy of Sciences

10-13 August 2015, SIGKDD, Sydney Australian

Outline

Feature Selection

- Data with high dimensionality are often encountered in many real applications.
  - Features are often correlated or redundant, or sometimes noisy.
- Such high dimensionality presents great challenges to learning methods.
  - The curse of dimensionality, computation and storage cost.
- Feature selection techniques can be used to effectively keep few informative features.
  - Supervised methods select those features to respect the the label information.
  - Unsupervised methods select those features to preserve the underlying structure of data.

Unsupervised Filter Feature Selection Methods

The basic procedure

1. It first employs *all the input features* to characterize the underlying structure of data,
   - e.g., the pairwise similarity, the graph Laplacian.
2. It then selects features to preserve such structures based on certain evaluation criteria.

Typical methods

- Laplacian Score [NIPS, 2006], SPEC [ICML, 2007], EVSC [ICML, 2011]

Unsupervised Embedded Feature Selection Methods (I)

The basic procedure

1. It first employs *all the input features* to characterize the underlying structure of data.
2. It then *jointly* selects features to preserve such structure.

Typical methods

- TraceRatio [IJCAI, 2008], UDFS [IJCAI, 2011]

Unsupervised Embedded Feature Selection Methods (II)

The basic procedure

1. It first employs *all the input features* to characterize the underlying structure of data.
2. It then flats the cluster structure via graph embedding or other clustering module.
   - an *intermediate cluster analysis sub-step* is involved.
3. It selects those features that are best aligned to the embedding via sparse spectral regression.

Typical methods

- MCFS [KDD, 2010], MRSF [IJCAI, 2011], FSSL [IJCAI, 2011], SPFS [TKDE, 2013], GLSPFS [TNNLS, 2014]

Unsupervised Embedded Feature Selection Methods (III)

The basic procedure

1. It first employs *all the input features* to characterize the underlying structure of data.
2. It then flats the cluster structure via graph embedding or other clustering module.
3. It selects those features that are best aligned to the embedding via sparse spectral regression.
4. The selected features are used to iteratively improve the intermediate cluster analysis sub-step.
   - the *intermediate cluster analysis* is updated with selected features.

Typical methods

- JELSR [IJCAI, 2011; TC, 2014], NDFS[AAAI, 2012], RUFS [IJCAI, 2013], CGSSL [TKDE, 2014], RSFS [ICDM, 2014]

Unsupervised Embedded Feature Selection Methods (IV)

The basic procedure

1. It first employs *all the input features* to characterize the underlying structure of data.
2. It then flats the cluster structure via graph embedding or other clustering module.
3. It selects those features that are best aligned to the embedding via sparse spectral regression.
4. The selected features are used to iteratively re-capture the underlying structure of data.
   - the *underlying structure of data* is re-captured with selected features.

Typical method

- LLCFS [PAMI, 2011]
  - It actually optimizes two different objectives for structure learning and feature selection. Its theoretic convergence can not be guaranteed and emprically performance is poor.

Limitation of existing unsupervised FS algorithms

- For the problem of unsupervised feature selection, we have to face the *chicken-and-egg dilemma* between **structure characterization** and **feature learning**.
  - on the one hand, one need the true structures of data to identify the informative features.
  - on the other hand, one need the informative features to accurately estimate the true structures of data.

- Most existing unsupervised feature selection methods failed to accurately estimate the structue of data **only with the informative features**.

Outline

Basic Idea

- **Structure characterization**, we extract both the global and local strcuture of data.
  - extract the global structure via sparse representation.
  - extract the local structure with probabilistic neighborhood.
- **Unsupervised Feature Learning**, we use these structures to guide the search of relevant features.
  - flat the structures of data via graph embedding.
  - estimate the informative features via sparse spectral regression.
- Perform **Structure characterization** and **Unsupervised Feature Learning** in a unified framework.
  - The structures are adaptively refined according to the results of feature selection.
  - Better structure characterization often leads to select better features.

Adaptive Global Structure Learning

- Global structure learning via sparse representation

$$\min_{\mathbf{S}} \quad \sum_{i=1}^{n} \left(||\mathbf{x}_i - \mathbf{X}\mathbf{s}_i||^2 + \alpha ||\mathbf{s}_i||_1 \right) \quad \text{s.t.} \quad \mathbf{S}_{ii} = 0 \qquad (1)$$

  We use the sparse reconstruction coefficients $\mathbf{S}$ to extract the global structure of data.

- Adaptive global structure Learning with selected features

$$\min_{\mathbf{S},\mathbf{W}} \quad \sum_{i=1}^{n} ||\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{X}\mathbf{s}_i||^2 + \alpha ||\mathbf{S}||_1 + \gamma ||\mathbf{W}||_{21} \qquad (2)$$

$$\text{s.t.} \quad \mathbf{S}_{ii} = 0, \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}$$

  - The selected features $\mathbf{W}$ should preserve the global structure captured by $\mathbf{S}$.
  - The global structure $\mathbf{S}$ can be refined with the selected informative features $\mathbf{W}$.

Adaptive Local Structure Learning

- Local structure learning via probabilistic neighborhood

$$\min_{\mathbf{P}} \sum_{i,j} (||\mathbf{x}_i - \mathbf{x}_j||_2^2 \mathbf{P}_{ij} + \mu \mathbf{P}_{ij}^2), \text{ s.t. } \mathbf{P}\mathbf{1}_n = \mathbf{1}_n, \mathbf{P} \geq 0 \quad (3)$$

  We use the sparse reconstruction coefficients $\mathbf{P}$ to extract the local structure of data.

- Adaptive local structure learning with selected features

$$\min_{\mathbf{P},\mathbf{W}} \quad \sum_{i,j}^{n} (||\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j||_2^2 \mathbf{P}_{ij} + \mu \mathbf{P}_{ij}^2) + \gamma ||\mathbf{W}||_{21} \quad (4)$$

$$\text{s.t.} \quad \mathbf{P}\mathbf{1}_n = \mathbf{1}_n, \mathbf{P} \geq \mathbf{0}, \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}$$

  - The selected features $\mathbf{W}$ should preserve the local structure captured by $\mathbf{P}$.
  - The local structure $\mathbf{P}$ can be refined with the selected informative features $\mathbf{W}$.

Unsupervised Feature Selection with Adaptive Structure Learning (FSASL)

- The formulation of FSASL

$$\min_{\mathbf{W},\mathbf{S},\mathbf{P}} \quad \left(||\mathbf{W}^T\mathbf{X} - \mathbf{W}^T\mathbf{X}\mathbf{S}||^2 + \alpha||\mathbf{S}||_1\right)$$
$$+ \quad \beta\sum_{i,j}^{n}\left(||\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j||^2\mathbf{P}_{ij} + \mu\mathbf{P}_{ij}^2\right) + \gamma||\mathbf{W}||_{21}$$
$$\text{s.t.} \quad \mathbf{S}_{ii} = 0, \mathbf{P}\mathbf{1}_n = \mathbf{1}_n, \mathbf{P} \geq \mathbf{0}, \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}$$

- The benefits of FSASL
  - Given $\mathbf{S}$ and $\mathbf{P}$, FSASL selects those features to well respect both the global and local structure of data.
  - Given $\mathbf{W}$, FSASL estimates the global and local structure of data in a transformed space, i.e. $\mathbf{W}^T\mathbf{X}$, where the adverse effect of noisy features is largely alleviated by sparse regularization.

The Optimization Algorithm

We derive an alternative iterative algorithm to solve the problem.

1. Given $\mathbf{W}$ and $\mathbf{P}$, optimize it w.r.t $\mathbf{S}$.
2. Given $\mathbf{W}$ and $\mathbf{S}$, optimize it w.r.t $\mathbf{P}$.
3. Given $\mathbf{S}$ and $\mathbf{P}$, optimize it w.r.t $\mathbf{W}$.

We repeat the following steps until it converges.

- Given $\mathbf{W}$ and $\mathbf{P}$, the optimal value of $\mathbf{S}$ can be obtained by solving the following LASSO problem.

$$\min_{\mathbf{s}_i} \quad ||\mathbf{x}_i^{'} - \mathbf{X}^{'}\mathbf{s}_i||^2 + \alpha|\mathbf{s}_i|, \quad \text{s.t.} \quad \mathbf{S}_{ii} = 0 \tag{5}$$

where $\mathbf{X}^{'} = \mathbf{W}^T\mathbf{X}$.

The Optimization Algorithm

- Given $\mathbf{W}$ and $\mathbf{S}$, we have to solve

$$\begin{aligned} \min_{\mathbf{p}_i^T} \quad & \sum_{j=1}^{n} ||\mathbf{x}_i^{'} - \mathbf{x}_j^{'}||^2 \mathbf{P}_{ij} + \mu ||\mathbf{P}_{ij}||^2, \\ \text{s.t.} \quad & \mathbf{1}_n^T \mathbf{p}_i = 1, \mathbf{P}_{ij} \geq 0 \end{aligned} \tag{6}$$

  which can be reformulated as the euclidean projection of a vector onto the probability simplex, where the optimal value of $\mathbf{P}$ can be efficiently obtained without iterations.

- It should be pointed out that the regularization term μ can be emprically determined by the neighborhood size k, which is more intuitive and easy to tune.

The Optimization Algorithm

- Given $\mathbf{S}$ and $\mathbf{P}$, we have

$$\min_{\mathbf{W}} \quad Tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^T) + \gamma||\mathbf{W}||_{21} \qquad (7)$$
$$\text{s.t.}\mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}$$

  where $\mathbf{L_S} = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^T$, $\mathbf{L_P} = \mathbf{D_P} - (\mathbf{P} + \mathbf{P}^T)/2$ and let $\mathbf{L} = \mathbf{L_S} + \beta\mathbf{L_P}$.
- Instead of solving a generalized eigen-problem, we solve it with the following two-steps:
  1. Solve the eigen-problem $\mathbf{L}\mathbf{Y} = \Lambda\mathbf{Y}$ to get $\mathbf{Y}$ corresponding to the c smallest eigenvalues;
  2. Find $\mathbf{W}$ which satisfies $\mathbf{X}^T\mathbf{W} = \mathbf{Y}$ by solving the following optimization problem:

$$\min_{\mathbf{W}} \quad ||\mathbf{Y} - \mathbf{X}^T\mathbf{W}||^2 + \gamma||\mathbf{W}||_{21} \qquad (8)$$

The Optimization Algorithm

Input: The data matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$, the regularization parameters α, β, γ, μ, the dimension of the transformed data c.

repeat

For each i, update the i-th column of $\mathbf{S}$ by solving the problem in Eq. (5);

For each i, update the i-th row of $\mathbf{P}$ using the algorithm for the euclidean projection of a vector onto the probability simplex;

Compute the overall graph Laplacian $\mathbf{L} = \mathbf{L_S} + \beta \mathbf{L_P}$;

Compute $\mathbf{W}$ according to Eq. (8);

until Converges

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

Outline

Datasets

Table: Summary of the benchmark data sets and the number of selected features

Data Sets	sample	feature	class	selected features
MFEA	2000	240	10	$[5, 10, \ldots, 50]$
USPS49	1673	256	2	$[5, 10, \ldots, 50]$
UMIST	575	644	20	$[5, 10, \ldots, 50]$
JAFFE	213	676	10	$[5, 10, \ldots, 50]$
AR	840	768	120	$[5, 10, \ldots, 50]$
COIL	1440	1024	20	$[5, 10, \ldots, 50]$
LUNG	203	3312	5	$[10, 20, \ldots, 150]$
TOX	171	5748	4	$[10, 20, \ldots, 150]$

Compared Algorithms

- LapScore [NIPS, 2006]
- MCFS [KDD, 2010]
- LLCFS [PAMI, 2011]
- UDFS [IJCAI, 2011]
- NDFS [AAAI, 2012]
- SPFS [TKDE, 2013]
- RUFS [IJCAI, 2013]
- JELSR [TC, 2014]
- GLSPFS [TNNLS, 2014]
- FSASL

code: `https://github.com/csliangdu/FSASL`

Experiment Protocol

- With the selected features, we evaluate the performance in terms of k-means clustering measured by Accuracy (ACC) and Normalized Mutual Information (NMI).
- For the compared methods, we tune the parameters in relative-large ranges and record the best resutl according to the grid-search strategy.
- We report the clustering results aggregated from different number of selected features with significance test.

Experimental Results

Table: Aggregated clustering results measured by Accuracy (%) of the compared methods.

Data Sets	AllFea	LapScore	MCFS	LLCFS	UDFS	NDFS	SPFS	RUFS	JELSR	GLSPFS	FSASL
MFEA	68.73	51.78 ± 5.51 0.00	51.04 ± 8.13 0.00	60.38 ± 8.58 0.00	48.94 ± 3.32 0.00	67.13 ± 7.53 0.01	**68.20 ± 9.43 0.22**	64.58 ± 7.99 0.00	67.01 ± 8.37 0.01	61.00 ± 8.70 0.00	**69.94 ± 7.19 1.00**
USPS49	77.70	69.21 ± 8.95 0.00	53.74 ± 3.50 0.00	94.96 ± 1.44 0.03	94.05 ± 1.13 0.00	68.12 ± 8.18 0.00	83.43 ± 6.66 0.00	85.86 ± 2.58 0.00	95.16 ± 0.55 0.00	94.75 ± 0.61 0.00	**95.95 ± 0.48 1.00**
UMIST	42.40	36.73 ± 1.18 0.00	44.46 ± 3.26 0.00	47.31 ± 0.83 0.00	48.04 ± 1.92 0.00	52.80 ± 2.26 0.00	46.72 ± 1.70 0.00	50.87 ± 1.95 0.00	53.52 ± 1.54 0.01	50.53 ± 0.59 0.00	**54.92 ± 1.89 1.00**
JAFFE	71.57	67.62 ± 8.49 0.00	73.56 ± 4.83 0.00	64.79 ± 4.08 0.00	75.48 ± 1.63 0.00	74.98 ± 2.15 0.00	73.93 ± 2.85 0.00	75.75 ± 2.53 0.00	77.77 ± 1.87 0.00	75.46 ± 1.61 0.00	**79.29 ± 2.24 1.00**
AR	30.26	25.29 ± 2.89 0.00	29.05 ± 1.19 0.00	**34.22 ± 2.70 0.05**	30.87 ± 0.35 0.00	32.34 ± 1.52 0.00	31.06 ± 2.14 0.00	34.84 ± 1.90 0.04	34.19 ± 2.52 0.02	34.12 ± 1.60 0.00	**36.11 ± 0.75 1.00**
COIL	59.17	45.60 ± 6.16 0.00	51.50 ± 5.38 0.00	50.84 ± 3.76 0.00	31.40 ± 16.89 0.00	44.22 ± 6.33 0.00	56.94 ± 3.43 0.00	59.20 ± 3.28 0.00	59.53 ± 4.01 0.03	57.96 ± 2.27 0.00	**60.93 ± 2.50 1.00**
LUNG	72.46	58.97 ± 5.24 0.00	70.42 ± 3.41 0.00	71.58 ± 5.85 0.00	65.46 ± 3.88 0.00	75.52 ± 1.57 0.00	73.49 ± 3.43 0.00	77.35 ± 2.62 0.00	77.86 ± 3.12 0.00	77.83 ± 2.70 0.00	**81.93 ± 1.63 1.00**
TOX	43.65	40.25 ± 0.65 0.00	43.10 ± 1.86 0.00	39.28 ± 0.49 0.00	47.14 ± 0.75 0.00	38.28 ± 1.64 0.00	39.93 ± 1.13 0.00	47.67 ± 0.83 0.00	43.96 ± 1.56 0.00	47.38 ± 1.93 0.00	**49.17 ± 0.67 1.00**
Average	58.24	49.43	52.11	57.92	55.17	56.67	59.21	62.02	63.63	62.38	**66.03**

Experimental Results

Table: Aggregated clustering results measured by Normalized Mutual Information (%) of the compared methods.

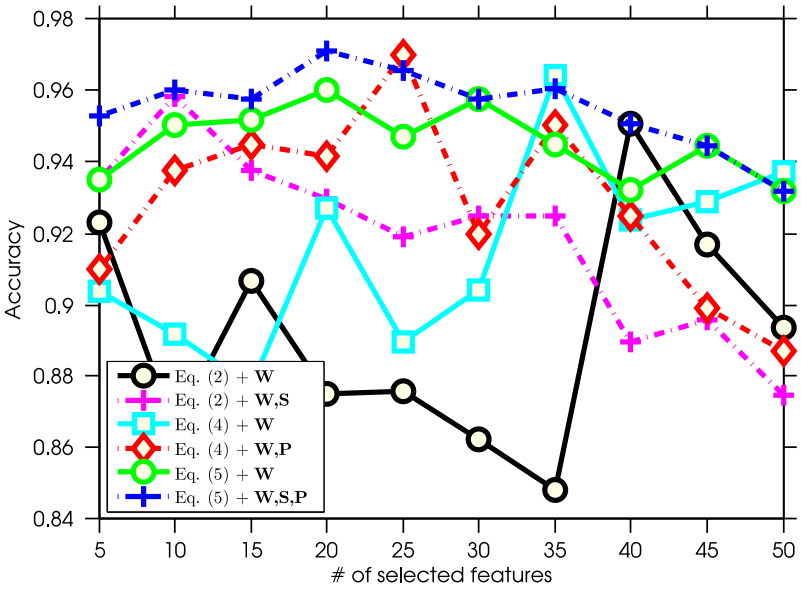
Data Sets	AllFea	LapScore	MCFS	LLCFS	UDFS	NDFS	SPFS	RUFS	JELSR	GLSPFS	FSASL
MFEA	70.33	53.74 ± 4.77 0.00	54.72 ± 9.14 0.00	52.77 ± 9.76 0.00	49.19 ± 3.83 0.00	64.97 ± 7.54 0.03	**64.92 ± 8.27 0.11**	63.98 ± 7.22 0.00	**64.51 ± 9.07 0.06**	59.26 ± 7.59 0.00	**66.70 ± 6.71 1.00**
USPS49	23.51	15.88 ± 17.98 0.00	4.60 ± 2.57 0.00	72.03 ± 5.56 0.03	68.12 ± 4.46 0.00	12.27 ± 9.62 0.00	38.10 ± 16.66 0.00	41.73 ± 7.23 0.00	72.28 ± 2.24 0.00	70.43 ± 2.57 0.00	**75.88 ± 2.28 1.00**
UMIST	64.15	55.57 ± 2.32 0.00	63.46 ± 4.93 0.00	63.42 ± 1.42 0.00	65.19 ± 2.96 0.00	71.19 ± 2.77 0.01	64.90 ± 3.06 0.00	68.19 ± 2.61 0.00	71.33 ± 2.06 0.00	69.16 ± 0.97 0.00	**72.39 ± 2.39 1.00**
JAFFE	81.52	77.28 ± 8.98 0.00	79.04 ± 5.88 0.00	66.97 ± 3.47 0.00	84.25 ± 1.74 0.00	82.53 ± 3.49 0.00	80.01 ± 3.06 0.00	82.00 ± 3.56 0.00	85.23 ± 3.31 0.00	83.20 ± 3.17 0.00	**86.42 ± 3.34 1.00**
AR	65.48	63.59 ± 2.36 0.00	66.41 ± 0.85 0.00	69.01 ± 1.45 0.01	67.49 ± 0.27 0.00	67.89 ± 0.89 0.00	66.94 ± 1.11 0.00	69.54 ± 1.10 0.01	69.02 ± 1.32 0.00	69.44 ± 0.84 0.00	**70.78 ± 0.63 1.00**
COIL	75.58	62.21 ± 4.98 0.00	66.19 ± 6.78 0.00	64.04 ± 4.34 0.00	44.27 ± 12.61 0.00	56.29 ± 6.91 0.00	69.91 ± 4.38 0.00	70.54 ± 4.48 0.00	71.37 ± 4.97 0.00	69.89 ± 4.00 0.00	**72.93 ± 4.44 1.00**
LUNG	60.37	50.14 ± 4.13 0.00	55.68 ± 2.31 0.00	60.12 ± 4.65 0.00	54.88 ± 4.21 0.00	60.57 ± 1.54 0.00	61.75 ± 3.32 0.00	65.47 ± 1.87 0.00	63.54 ± 2.94 0.00	63.50 ± 2.99 0.00	**66.78 ± 1.72 1.00**
TOX	15.87	10.92 ± 0.68 0.00	16.53 ± 2.68 0.00	9.68 ± 0.75 0.00	22.16 ± 1.36 0.00	9.07 ± 1.87 0.00	10.13 ± 1.03 0.00	23.58 ± 1.60 0.00	17.46 ± 3.36 0.00	23.49 ± 2.77 0.00	**25.79 ± 1.62 1.00**
Average	57.10	48.67	50.83	57.26	56.94	53.10	57.08	60.63	64.34	63.55	**67.21**

The Effect of Adaptive Structure Learning

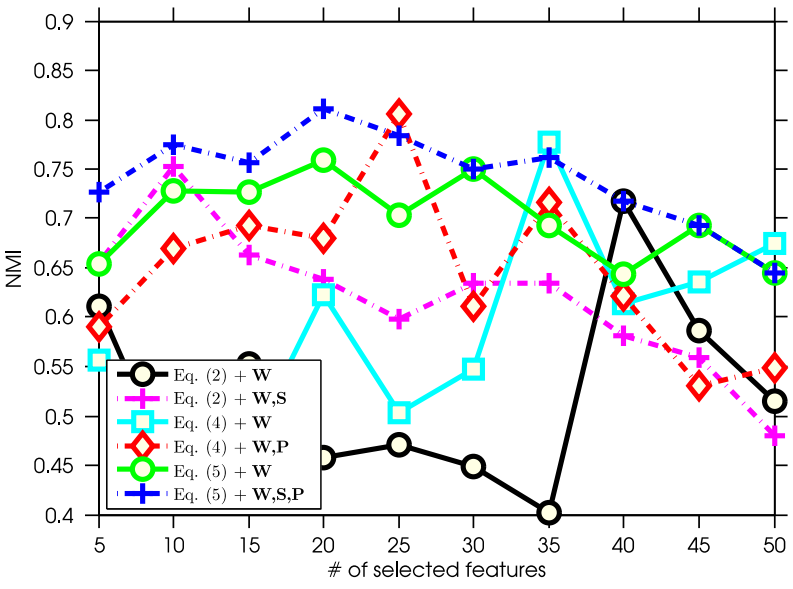
Question: Does the adaptive structure learning lead to select more informative features?
We design six different settings to empirically investigate the effect of adaptive structure learning.

- Global structure guided FS, with/without adaptive structure learning
- Local structure guided FS, with/without adaptive structure learning
- Global and local structures guided FS, with/without adaptive structure learning

The Effect of Adaptive Structure Learning



(a)



(b)

Figure: Clustering results w.r.t. 6 different settings of FSASL on USPS200

Parameter Sensitivity

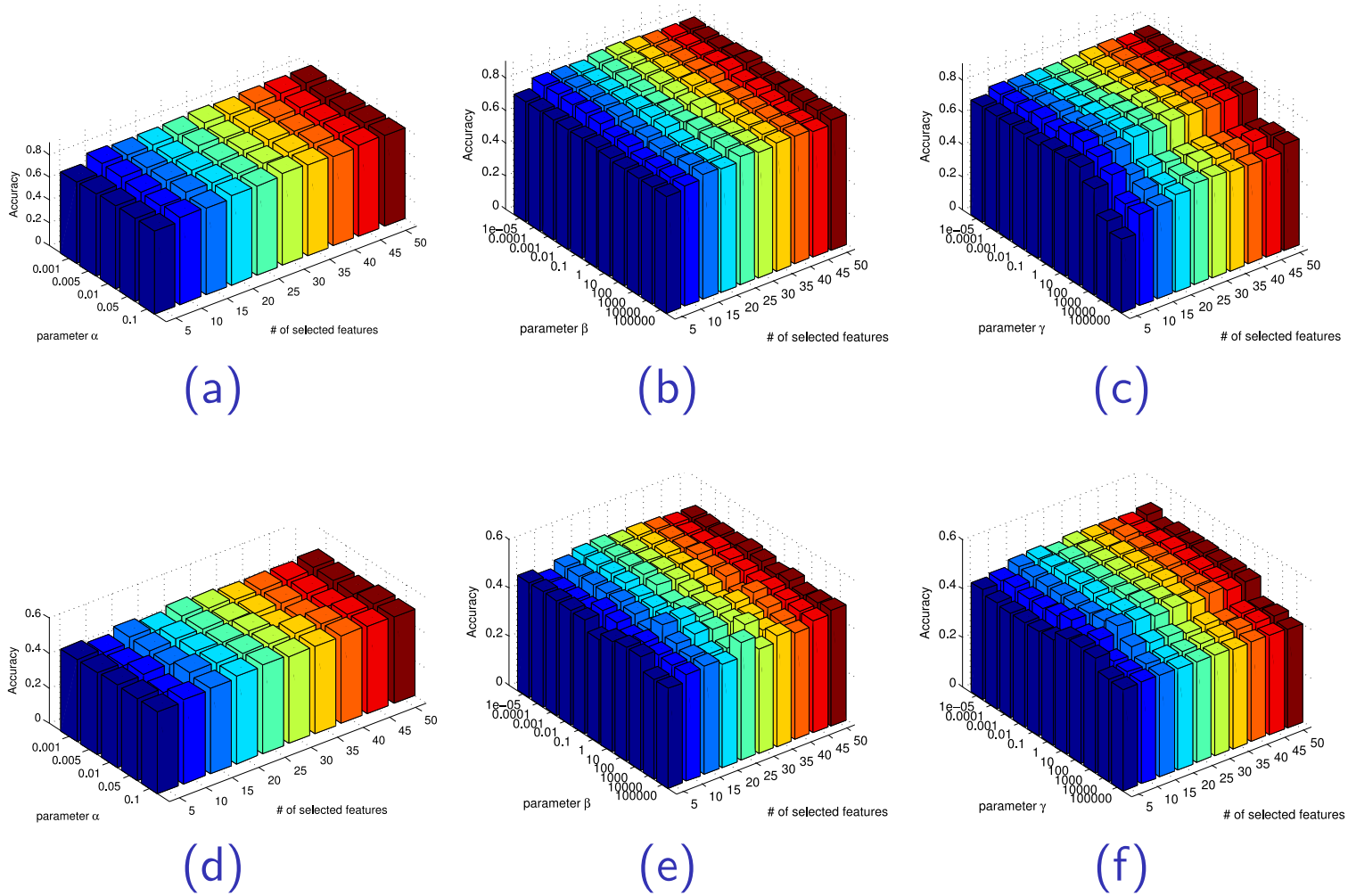


Figure: Accuracy of different parameters on JAFFE (a-c) and TOX (d-f).

Summary

- We investigate most of existing unsupervised embedded methods and further classify them into four closely related but different types. These analyses provide more insight into what should be further emphasized on the development of more essential unsupervised feature selection algorithm.
- We propose a novel unified learning framework, called unsupervised Feature Selection with Adaptive Structure Learning (FSASL in short), to fulfil the gap between two essential sub tasks, i.e. structure learning and feature learning. In this way, these two tasks can be mutually improved.
- Comprehensive experiments on benchmark data sets show that our method achieves statistically significant improvement over state-of-the-art feature selection methods.

Acknowledgement

We would like to thank Prof. Feiping Nie and Prof. Mingyu Fan for their helpful suggestions to improve this paper.
This work is supported in part by the China National 973 program 2014CB340301, the Natural Science Foundation of China (NSFC) grant 61379043, 61322211.

Thanks!

Q&A