Unsupervised Feature Selection with Adaptive Structure Learning

Reporter: Songling Liu

Data Mining Lab
Big Data Research Center

Overview

- Current feature selection methods
- Motivation
- Method
- Optimization Algorithm
- Experiments
- Conclusion

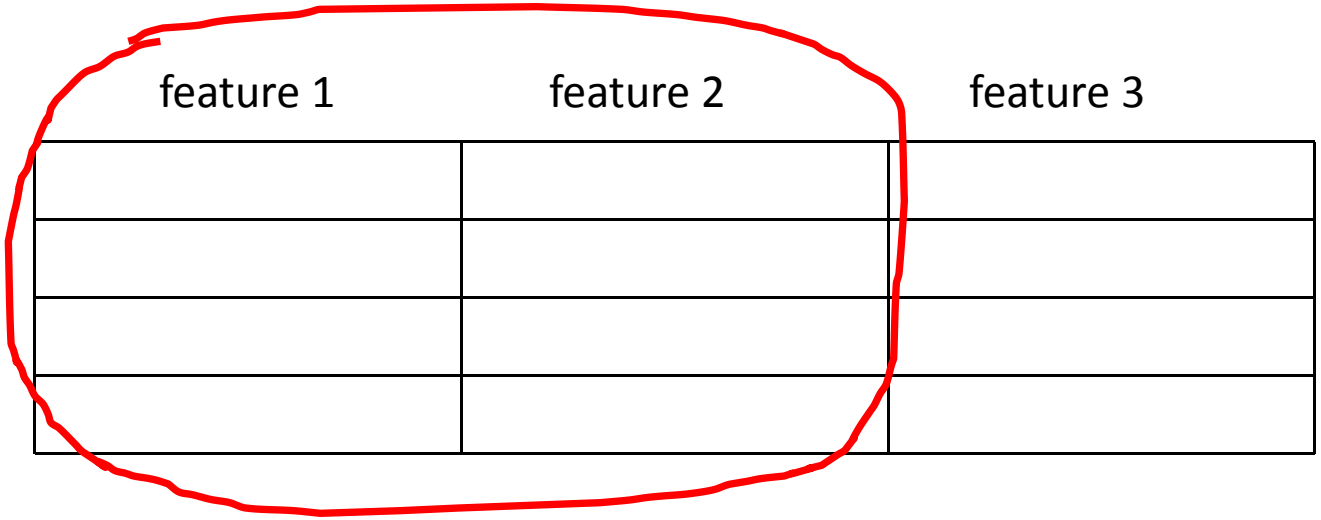
Current feature selection methods

Method type	Structure characterization	Intermediate Analysis	Feature search	Typical algorithm
Filter	Structure learning with All features		Ranking criteria	MaxVar, LapScore, SPEC, EVSC
Embedded Type I	Structure learning with All features		A learning model	TraceRatio, UDFS
Embedded Type II	Structure learning with All features	Clustering	A learning model	MCFS, MRFS, SPFS, FSSL, GLSPFS
Embedded Type III	Structure learning with All features	Clustering	A learning model	JELSR, NDFS, RUFS, CGSSL

Current feature selection methods

Filter:

Addressing this issue by selecting the top ranked features based on some scores computed independently for each feature.



If in variance:1>2>3 and we choose the largest two

Current feature selection methods

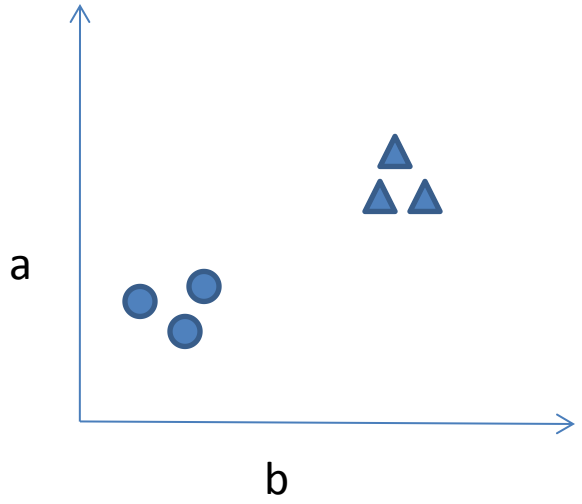
Embedded Method:

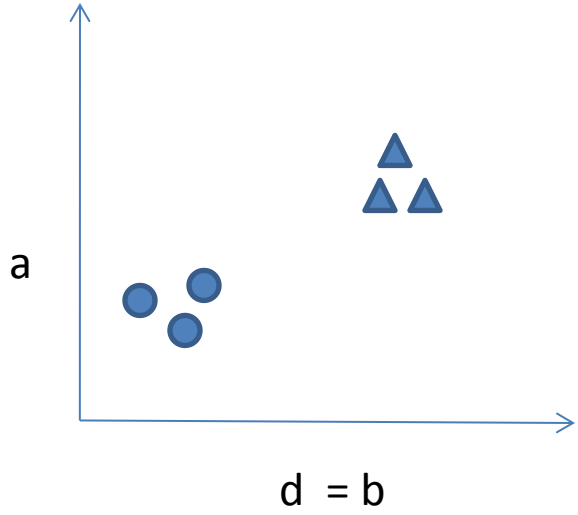
Using all features to estimate the underlying structures of data and select features which can preserve data structures.

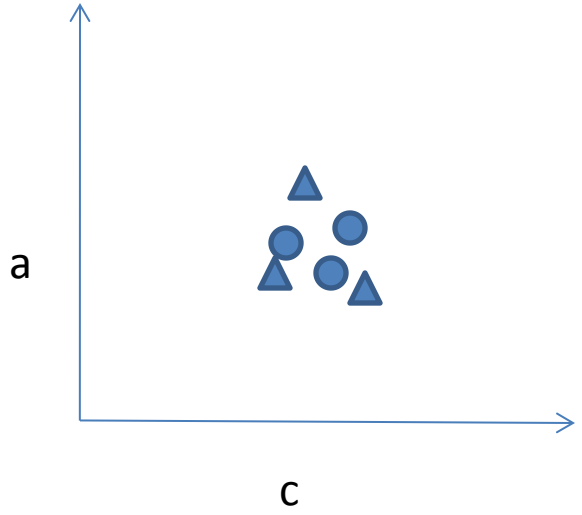
Common drawback--Using all features which could be redundant and include noisy features.

Current feature selection methods

Suppose a dataset with
4 features a,b,c and d.



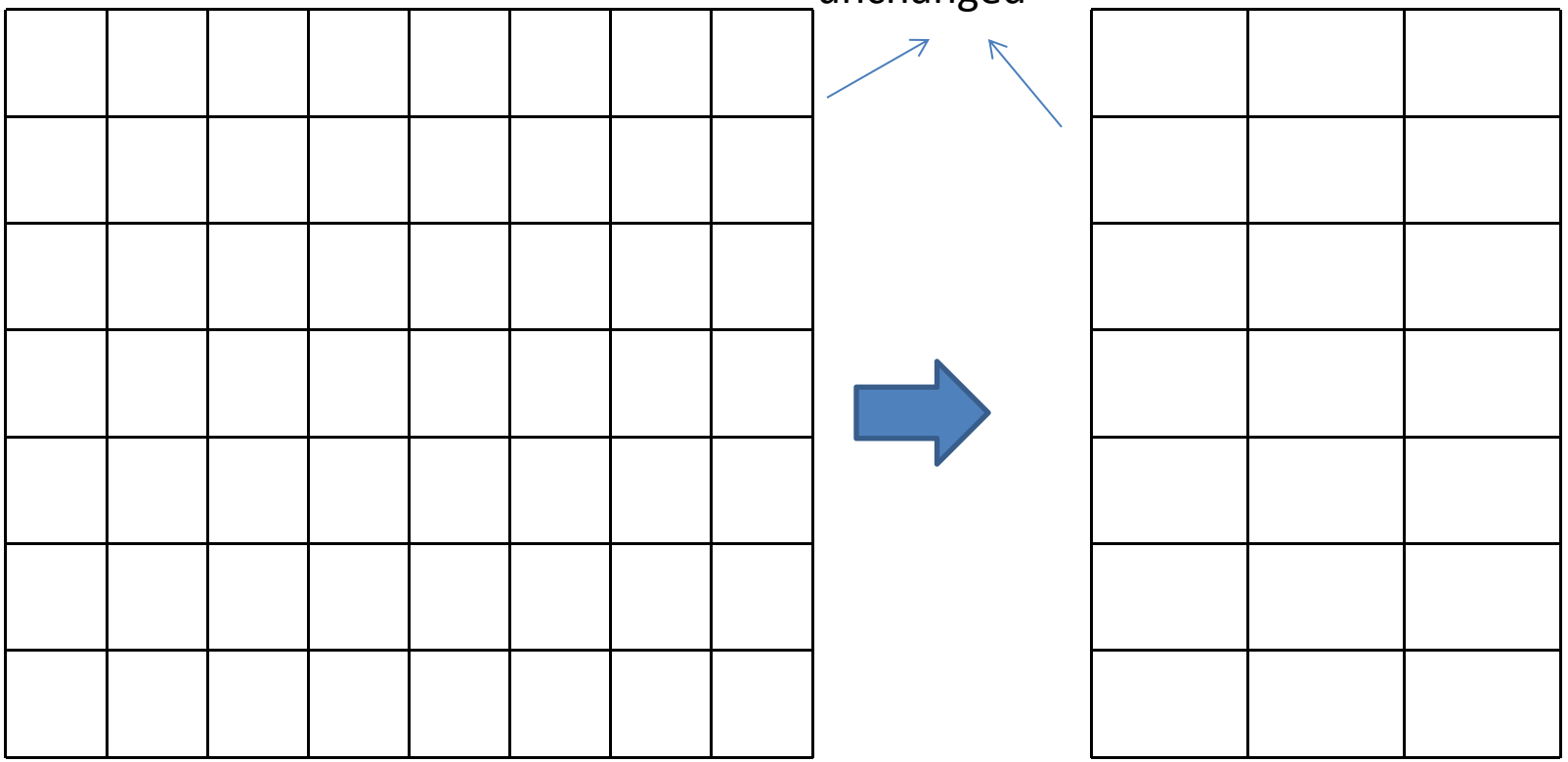




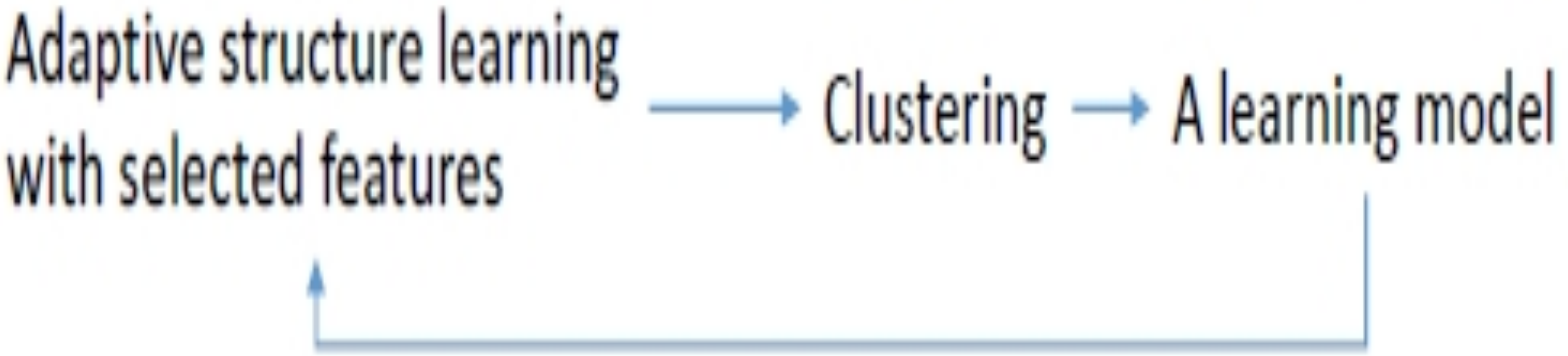
Data structure estimating VS Feature selection

We want to preserve the **similarity** after feature selection

Pairwise distance prefered to be unchanged



Motivation



- A unified learning framework which performs structure learning and feature selection **simultaneously**.
- Select good features from well-estimated data structure.
- Estimate data structure with good features.

Problem I

How to preserve pairwise global structure?

PCA? MaxVar?

However, such dense similarity becomes less discriminative for high dimension data, especially when there are many unfavorable features in the original high dimensional space.

Method

Adaptive Global Structure Learning

using the sparse reconstruction coeffcients to extract the global structure of data.

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

Method

the selected features should preserve such global and sparse reconstruction structure

$$\min_{\mathbf{S},\mathbf{W}} \sum_{i=1}^{n} ||\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{X}\mathbf{s}_i||^2 + \alpha||\mathbf{S}||_1 + \gamma||\mathbf{W}||_{21}$$
$$\text{s.t.} \quad \mathbf{S}_{ii} = 0, \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}$$

Compared with the last equation:

1) The global structure captured by S can be used to **guide the search of relevant features**;

2) the global structure can also be better estimated.

Problem II

How to preserve pairwise local structure?

LLE? Graph Laplacian?

They would be inevitably affected by the redundant and noisy features. Moreover, the iterative updating of discrete neighborhood relationship using the result of feature selection still sffuers from the lack of theoretical guarantee of its convergence.

Method

Adaptive Local Structure Learning

Learn a euclidean distance induced probabilistic neighborhood matrix

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

Using matrix P,graph laplacian can be characterized as

$$\mathbf{L}_\mathbf{P} = \mathbf{D}_\mathbf{P} - (\mathbf{P} + \mathbf{P}^T)/2$$

Where Dp's element is $\sum_j (\mathbf{P}_{ij} + \mathbf{P}_{ji})/2$

Method

Using the sparse matrix to select **infomative features**

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

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

we can learn a better probabilistic neighborhood graph for local structure characterization

Method

Unsupervised Feature Selection with Adaptive Structure Learning

$$\min_{\mathbf{W},\mathbf{S},\mathbf{P}} \left(||\mathbf{W}^T\mathbf{X} - \mathbf{W}^T\mathbf{X}\mathbf{S}||^2 + \alpha||\mathbf{S}||_1 \right)$$

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

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

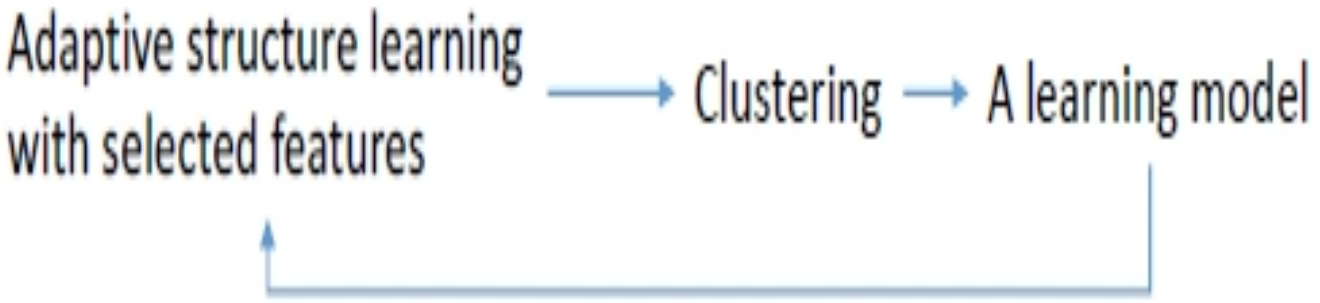
Method

When S and P fixed

It selects those **features** to well respect both the global and local structure of data;

When W fixed

It learns the **global and local structure** of data in a transformed space



Optimization Algorithm

First, when W and P are fixed,

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

Where $\mathbf{X}^{'} = \mathbf{W}^T\mathbf{X}$

We get sparsity representation of instances,which means global structure

Optimization Algorithm

Next, when W and S are fixed,

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

With transformation

$$\min_{\mathbf{P}_i^T} \quad \frac{1}{2}||\mathbf{p}_i^T - \mathbf{a}_i^T||^2, \quad \text{s.t.} \quad \mathbf{p}_i^T \mathbf{1}_n = 1, 0 \leq \mathbf{p}_{ij}^T \leq 1$$

Where $\mathbf{A}_{ij} = -\frac{1}{2\mu}||\mathbf{x}_i^{'} - \mathbf{x}_j^{'}||^2,$

Optimization Algorithm

Next, when S and P are fixed,

$$\min_{\mathbf{W}} \quad ||\mathbf{W}^T\mathbf{X} - \mathbf{W}^T\mathbf{X}\mathbf{S}||^2 + \beta \sum_{i,j}^{n} ||\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j||^2 \mathbf{P}_{ij} + \gamma ||\mathbf{W}||_{21}$$

$$\text{s.t.} \quad \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I} \tag{9}$$

Using $\mathbf{L_S} = (\mathbf{I} - \mathbf{S})(\mathbf{I} - \mathbf{S})^T$, $\mathbf{L_P} = \mathbf{D_P} - (\mathbf{P} + \mathbf{P}^T)/2$ and let $\mathbf{L} = \mathbf{L_S} + \beta\mathbf{L_P}$, the above problem can be rewritten as

$$\min_{\mathbf{W}} \quad Tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^T) + \gamma ||\mathbf{W}||_{21} \tag{10}$$

$$\text{s.t.} \quad \mathbf{W}^T\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{I}$$

Optimization Algorithm

Algorithm 2 The optimization algorithm of FSASL

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. (6);

For each i, update the i-th row of $\mathbf{P}$ using Algorithm 1;

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

Compute $\mathbf{W}$ by Eq. (12) or Eq. (14);

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.

Experiments

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	64.94 ± 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	84.77 ± 1.59 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	48.40 ± 16.89 0.00	52.22 ± 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	49.17 ± 0.83 0.00	43.96 ± 1.56 0.00	47.38 ± 1.93 0.00	**50.12** **± 0.67** **1.00**
Average	58.24	49.43	55.98	57.92	59.29	56.67	59.21	62.2	63.62	62.38	**66.15**

Experiments

Table 2: 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	54.19 ± 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	63.14 ± 1.05 0.00	72.03 ± 5.56 0.03	68.12 ± 4.46 0.00	62.27 ± 9.62 0.00	68.10 ± 16.66 0.00	71.73 ± 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	25.79 ± 1.60 0.00	17.46 ± 3.36 0.00	23.49 ± 2.77 0.00	**27.37 ± 1.62 1.00**
Average	57.10	48.67	58.14	57.26	57.56	59.35	60.83	64.65	64.34	63.55	67.41

Experiments

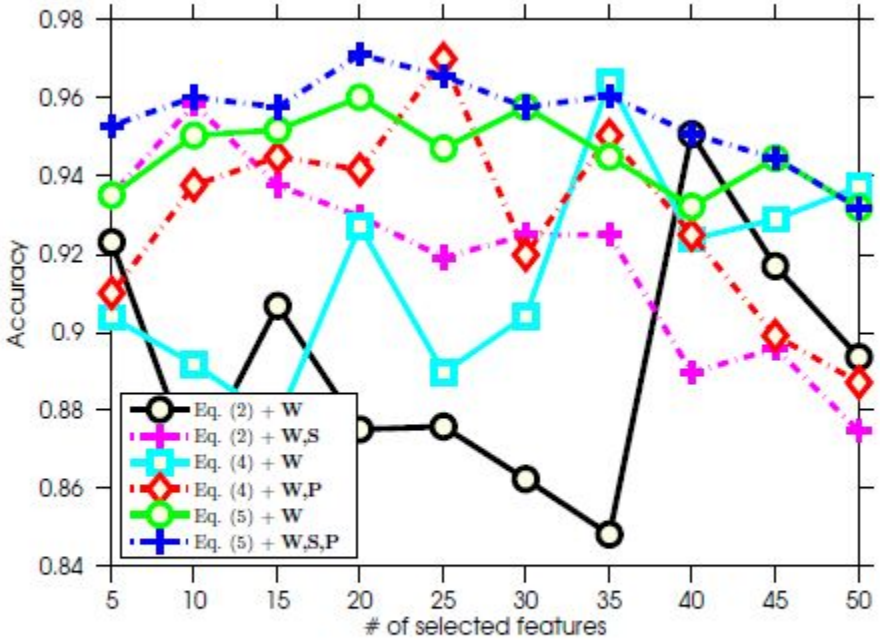


Figure 3: Clustering accuracy w.r.t. 6 different settings of FSASL on USPS200.

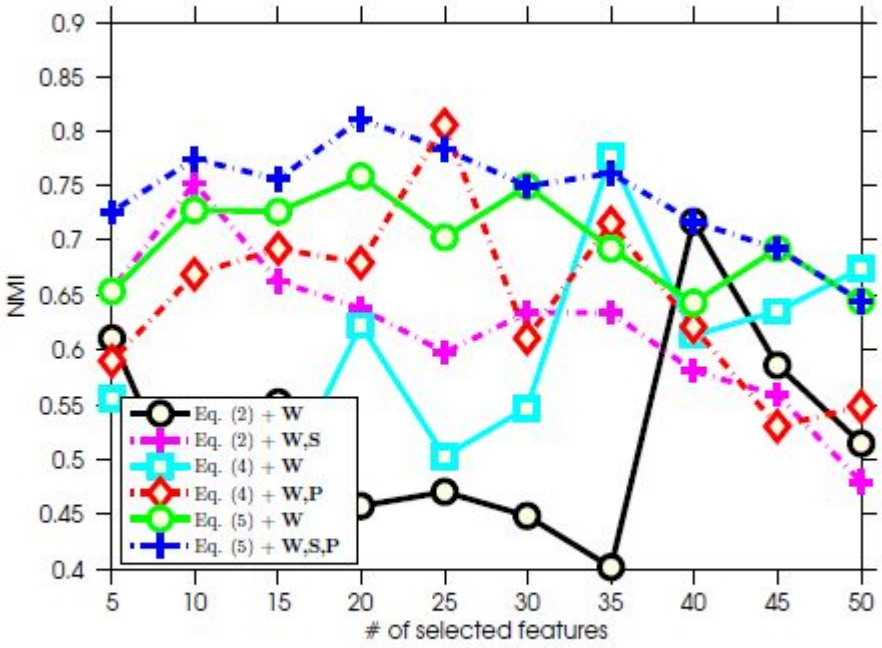


Figure 4: Clustering NMI w.r.t. 6 different settings of FSASL on USPS200.

Q&A