TIANJIN UNIVERSITY

Mixed Sparsity Regularized Multi-view Unsupervised Feature Selection

Kennedy W. Wangila Ke Gao Pengfei Zhu*

Qinghua Hu Changqing Zhang

School of Computer Science and Technology
Tianjin University, China
zhupengfei@tju.edu.cn

September 20, 2017

Outlines

1. Introduction
2. Multi-view Feature Selection
3. Mixed Sparsity Regularized Feature Selection
4. Experiment
5. Conclusion

Outlines

1. **Introduction**
2. **Multi-view Feature Selection**
3. **Mixed Sparsity Regularized Feature Selection**
4. **Experiment**
5. **Conclusion**

1. Introduction

text mining

computer vision

video understanding

bio-informatics

(a) *UC IRVINE MACHINE LEARNING REPOSITORY*

APPLICATION DOMAIN	DATA NAME	DIMENSION	YEAR
ACOUSTICS	ISOLET	617	1994
GAME	CHESS	36	1989
	CONNECT-4	42	1995
IMAGE	LETTER	16	1991
	COREL	89	1999
LIFE SCIENCE	SOYBEAN	35	1987
	MOLECULAR	58	1990
	MAMMALS	72	1992
	SPECTF	44	2001
	P53	5,409	2010
MULTI-VIEW	INTERNET AD	1,558	1998
PHYSICS	SPECTROMETER	102	1988
	H_2O TREAT. PLANT	38	1993
SOCIOLOGY	INSURANCE	86	2000
TEXT	BAG OF WORDS	100,000	2008
	URL	**3,231,961**	2009
TIME-SERIES	PEMS-SF	138,672	2011
	GAS SENSOR	**1,950,000**	2013
VIDEO	YOUTUBE MVG	**1,000,000**	2013

(c) *LIBSVM DATABASE*

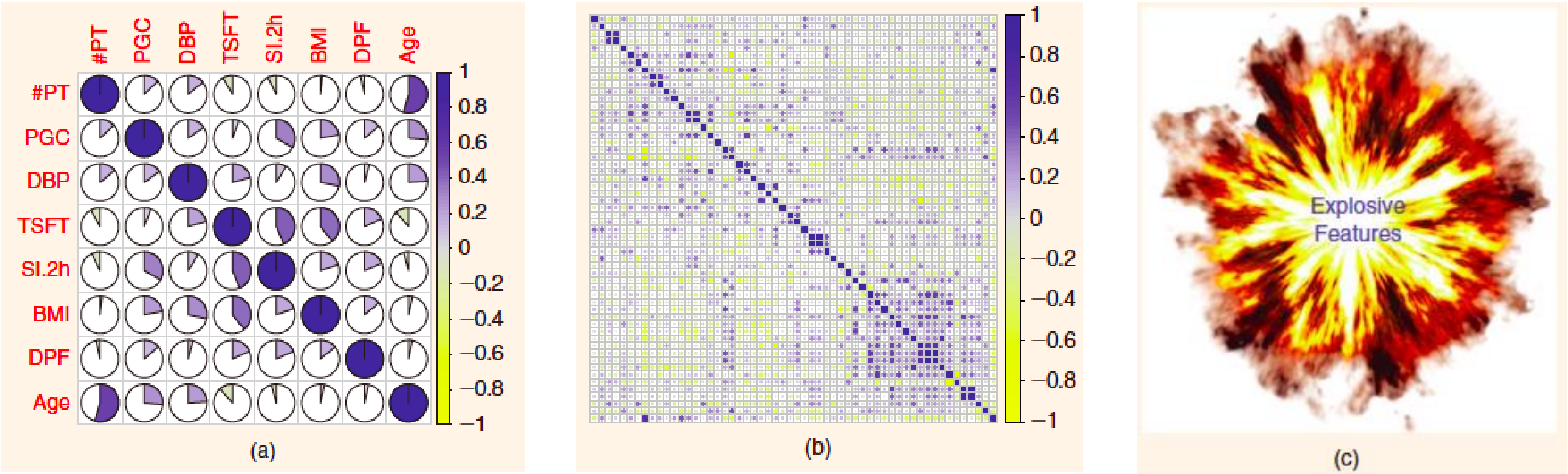
APPLICATION DOMAIN	DATA NAME	DIMENSION	YEAR
IMAGE	USPS	256	1994
	GISETTE	5,000	2003
LIFE SCIENCE	LEUKEMIA	7,129	1999
	COLON-CANCER	2,000	1999
	BREAST-CANCER	7,129	2001
TEXT	NEWS20	62,061	1995
	REAL-SIM	20,958	1998
	SECTOR	55,197	1998
	RCV1	47,236	2004
	NEWS20.BINARY	**1,355,191**	2005
	WEBSPAM	**16,609,143**	2006
	SIAM	30,438	2007
	LOG1P	**4,272,227**	2009
EDUCATION	KDD2010	**29,890,095**	2010

Curse of Dimensionality

- Zhai Y, Ong Y S, Tsang I W. The Emerging" Big Dimensionality"[J]. IEEE Computational Intelligence Magazine, 2014, 9(3): 14-26.
- http://archive.ics.uci.edu/ml/

1. Introduction

Correlations between feature pairs using a 2D correlation matrix



The evolution (rise) of feature dimensionality in correlation matrices.

(a) Diabetes (8 features)

(b) Lung Cancer (56 features)

(c) Psoriasis (529,651 features)

- Zhai Y, Ong Y S, Tsang I W. The Emerging" Big Dimensionality"[J]. IEEE Computational Intelligence Magazine, 2014, 9(3): 14-26.

1. Introduction

Problem

- Time and storage

The high-dimensionality data always contain a plenty of redundant data and noise, which may leads to high time complexity, large storage burden.

- Generalization ability

The high-dimensional data may increase the number of parameters of the learning machines, and therefore easily lead to over-fitting and degradation of the generalization ability.

1. Introduction

Solution

- **Subspace learning:**

Subspace learning is one of the most effective ways to eliminate the curse of dimensionality by projecting the data to a low-dimensional feature subspace.

- **Feature selection:**

Feature selection directly selects a subset of relevant and most representative features. And it is also an effective technique to reduce storage burden and time complexity, and improve generalization ability of the learned.

1. Introduction

Feature selection

Filter Methods

Filter method based on general features like the correlation with the variable to predict.

e.g. variance, Laplacian Score, consistence

Wrapper Methods

Wrapper Methods evaluate subsets of variables which allows to detect the possible interactions between variables.

e.g. R-SVM, SVM-RFE

Embedded Methods

Embedded methods takes advantage of its own variable selection process and performs feature selection and classification simultaneously.

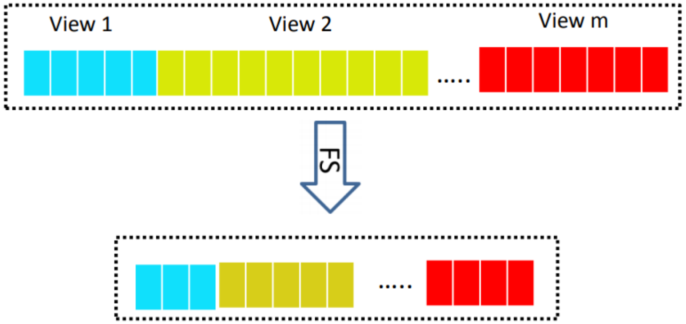
e.g. EUFS

- Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine learning, 2002, 46(1): 389-422.

Outlines

1. Introduction
2. **Multi-view Feature Selection**
3. Mixed Sparsity Regularized Feature Selection
4. Experiment
5. Conclusion

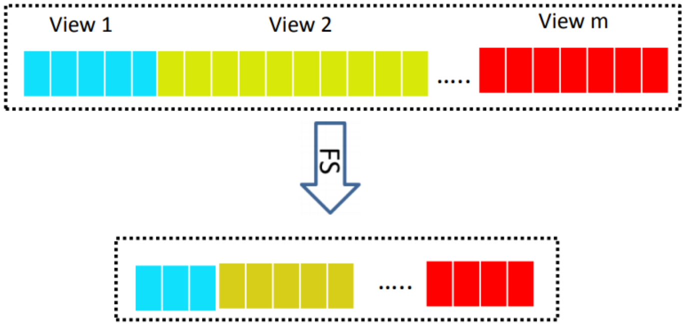
2. Multi-view Feature Selection



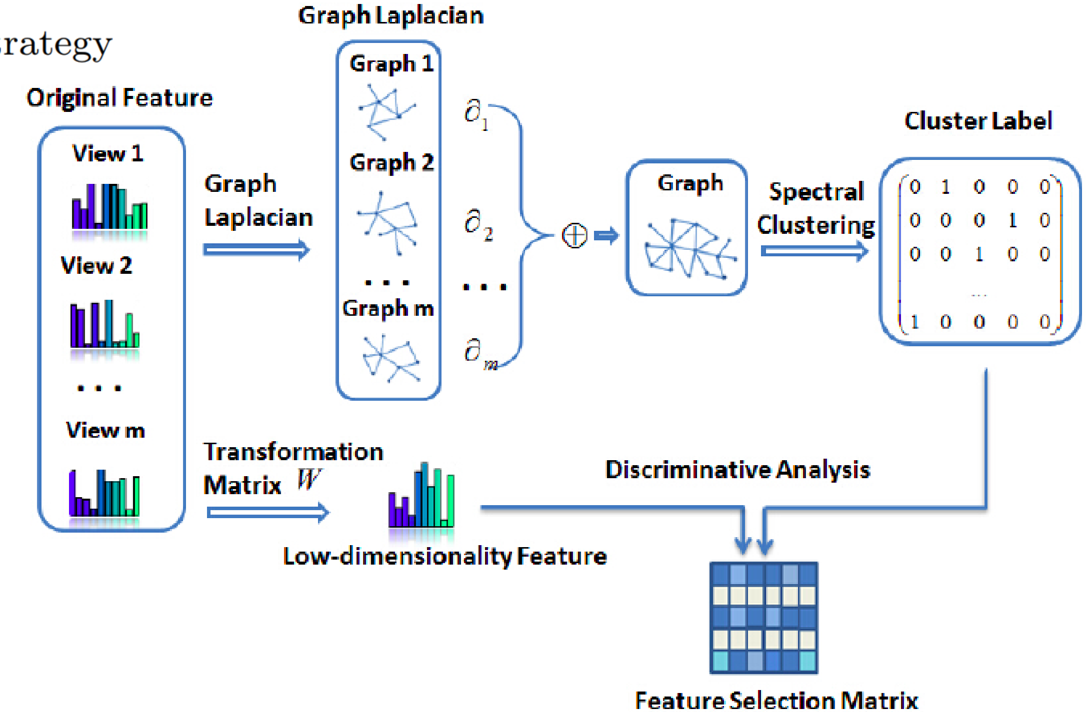
(a) The Concatenating Strategy

- Tang J, Hu X, Gao H, et al. Unsupervised Feature Selection for Multi-View Data in Social Media[M]// Proceedings of the 2013 SIAM International Conference on Data Mining. 2013.

2. Multi-view Feature Selection

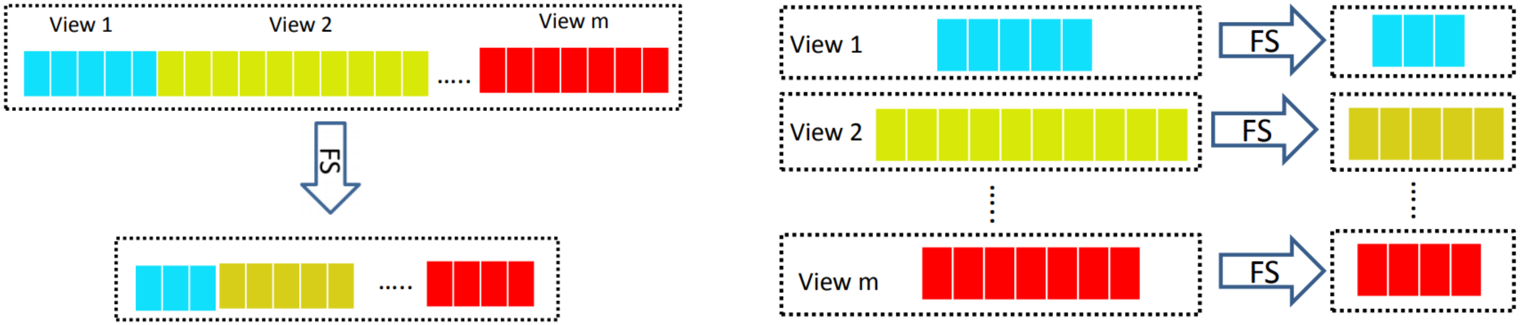


(a) The Concatenating Strategy



- Shi H, Li Y, Han Y, et al. Cluster structure preserving unsupervised feature selection for multi-view tasks[J]. Neurocomputing, 2016, 175(PA):686-697.

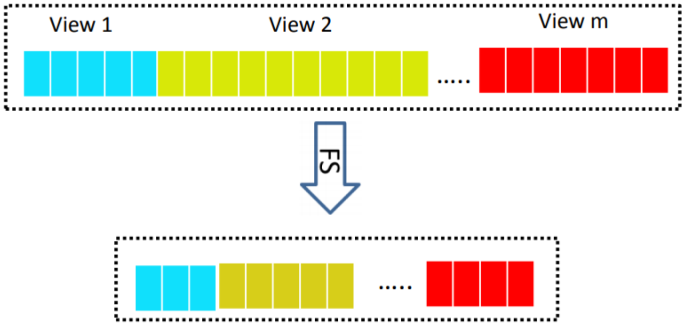
2. Multi-view Feature Selection



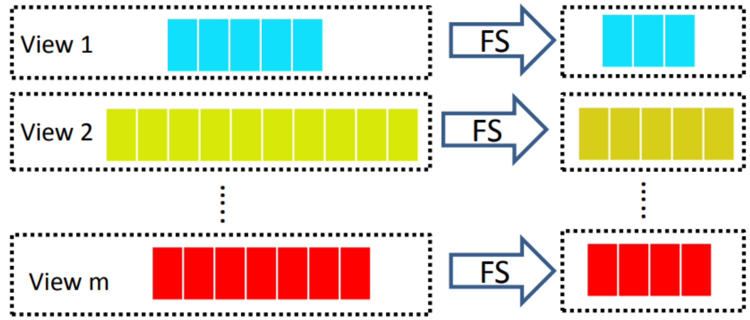
(a) The Concatenating Strategy (b) The Separation Strategy

- Tang J, Hu X, Gao H, et al. Unsupervised Feature Selection for Multi-View Data in Social Media[M]// Proceedings of the 2013 SIAM International Conference on Data Mining. 2013.

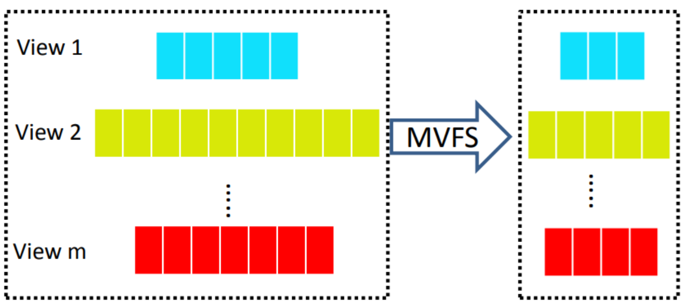
2. Multi-view Feature Selection



(a) The Concatenating Strategy



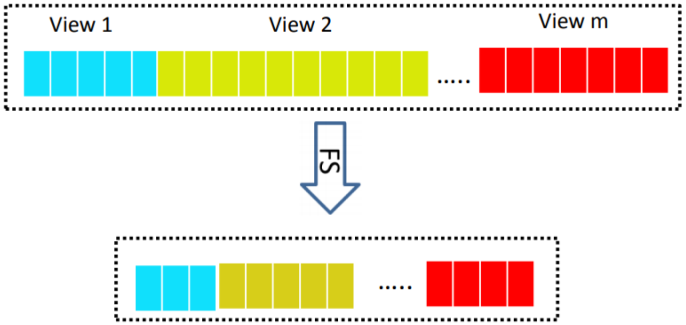
(b) The Separation Strategy



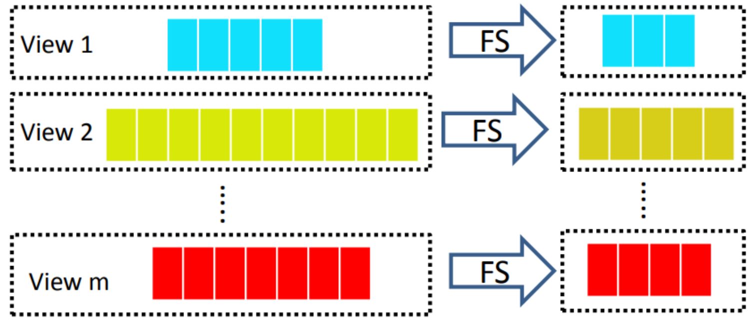
(c) Multi-view Feature Selection

- Tang J, Hu X, Gao H, et al. Unsupervised Feature Selection for Multi-View Data in Social Media[M]// Proceedings of the 2013 SIAM International Conference on Data Mining. 2013.

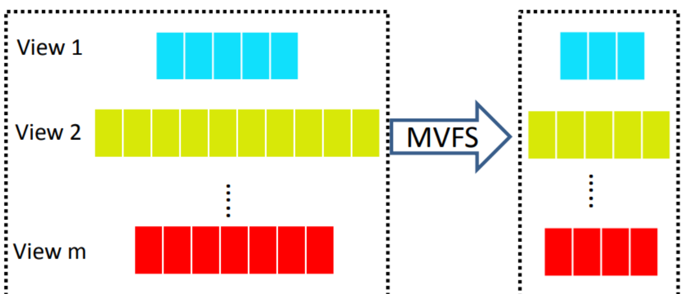
2. Multi-view Feature Selection



(a) The Concatenating Strategy



(b) The Separation Strategy



(c) Multi-view Feature Selection

$$\min_{\mathcal{W},\mathbf{Z}} \quad \mathcal{J}(\mathcal{W},\mathbf{Z}) = \sum_{i=1}^{m} \boxed{\lambda_i}\left(Tr(\mathbf{Z}^\top \mathbf{L}_i \mathbf{Z}) + \alpha(\|\mathbf{X}_i^\top \mathbf{W}_i - \mathbf{Z}\|_F^2 + \beta\|\mathbf{W}_i\|_{2,1})\right)$$

$$s.t. \quad \mathbf{Z}^\top \mathbf{Z} = \mathbf{I}, \qquad \mathbf{Z} \geq 0.$$

- Tang J, Hu X, Gao H, et al. Unsupervised Feature Selection for Multi-View Data in Social Media[M]// Proceedings of the 2013 SIAM International Conference on Data Mining. 2013.

Outlines

1. Introduction
2. Multi-view Feature Selection
3. **Mixed Sparsity Regularized Feature Selection**
4. Experiment
5. Conclusion

3. Mixed Sparsity Regularized Feature Selection

Mixed Sparsity Regularized Learning

$\mathbf{X}=[\mathbf{X}_1,\dots,\mathbf{X}_i,\dots,\mathbf{X}_m] \in \mathbb{R}^{d\times n}$ $\quad \mathbf{X}_i \in \mathbb{R}^{d_i\times n}$ $\quad \mathbf{P} \in \mathbb{R}^{d\times r}$

$$\min_{\mathbf{P}^T\mathbf{P}=\mathbf{I}} \sum_{i=1}^{m} \sqrt{Tr(\mathbf{P}^T\mathbf{X}\mathbf{L}_i\mathbf{X}^T\mathbf{P})} + \lambda R(\mathbf{P})$$

- **Feature level:**

$$\|\mathbf{P}\|_{2,1} = \sum_{i=1}^{d} \|\mathbf{p}_i\|_2 = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{r} p_{ij}^2}$$

- **View level:**

$$\|\mathbf{P}^i\|_F = \sqrt{\sum_{j=1}^{d_i} \|\mathbf{p}_j^i\|_2^2} = \sqrt{\sum_{j=1}^{d_i}\sum_{k=1}^{r} (p_{jk}^i)^2}$$

$$\min \sum_{i=1}^{m} \sqrt{Tr(\mathbf{P}^T\mathbf{X}\mathbf{L}_i\mathbf{X}^T\mathbf{P})} + \lambda_1\|\mathbf{P}\|_{2,1} + \lambda_2 \sum_{j=1}^{m} \|\mathbf{P}^j\|_F$$
$$s.t.\ \mathbf{P}^T\mathbf{P} = \mathbf{I}$$

3. Mixed Sparsity Regularized Feature Selection

Parameter-free Multi-view Learning

$$\min \sum_{i=1}^{m} \sqrt{Tr(\mathbf{P}^T\mathbf{X}\mathbf{L}_i\mathbf{X}^T\mathbf{P})} + \lambda_1\|\mathbf{P}\|_{2,1} + \lambda_2\sum_{j=1}^{m}\left\|\mathbf{P}^j\right\|_F$$

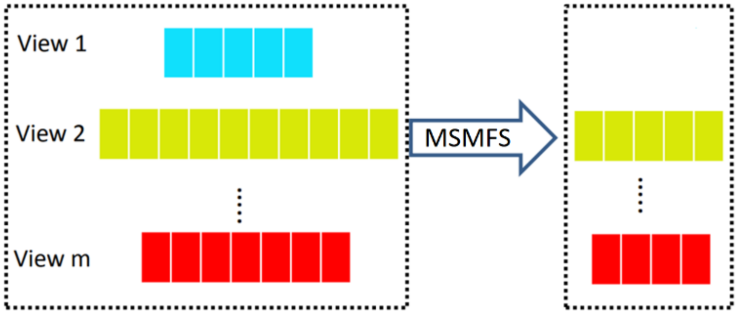
$$s.t.\ \mathbf{P}^T\mathbf{P} = \mathbf{I}$$

Lagrange Multiplier Method

$$\sum_{i=1}^{m} \sqrt{Tr(\mathbf{P}^T\mathbf{X}\mathbf{L}_i\mathbf{X}^T\mathbf{P})} + \lambda_1\|\mathbf{P}\|_{2,1} + \lambda_2\sum_{j=1}^{m}\left\|\mathbf{P}^j\right\|_F + G(\Lambda, F)$$

$$\sum_{i=1}^{m}\alpha_i\frac{\partial Tr(\mathbf{P}^T\mathbf{X}\mathbf{L}_i\mathbf{X}^T\mathbf{P})}{\partial\mathbf{P}} + \lambda_1\frac{\partial\|\mathbf{P}\|_{2,1}}{\partial\mathbf{P}} + \lambda_2\frac{\partial\sum_{j=1}^{m}\left\|\mathbf{P}^j\right\|_F}{\partial\mathbf{P}} + \frac{\partial G(\Lambda, F)}{\partial\mathbf{P}}$$

$$\alpha_i = 1\Big/\left(2\sqrt{Tr(\mathbf{P}^T\mathbf{X}\mathbf{L}_i\mathbf{X}^T\mathbf{P})}\right)$$



(d) Mixed Sparsity Regularized Feature Selection

Outlines

1. Introduction
2. Multi-view Feature Selection
3. Mixed Sparsity Regularized Feature Selection
4. **Experiment**
5. Conclusion

4. Experiment

Settings:

- Clustering accuracy
- Clustering NMI

Datasets:

DATA	Smaples	view	Features	Classes
Caltech10	800	4	200, 512, 59, 680	10
Corel800	800	4	200, 512, 59, 680	10
flickr	1000	4	200, 512, 59, 680	10
mfeat	2000	6	216, 76, 64, 6, 240, 47	10
PPMI	1400	3	200, 200, 200	7
MSRA	210	5	1302, 512, 256, 210, 100	7
Still DB	467	3	200, 200, 200	6

4. Experiment

Experiment Results:

Clustering accuracy result of all data sets

DATA	Laplacian	SPEC	MCFS	UDFS	AUMFS	MSMFS
Caltech10	0.2562	0.2223	0.2873	0.2887	0.3205	**0.3444**
Corel800	0.2986	0.2514	0.2851	0.2702	0.2913	**0.3073**
flickr	0.2146	0.2086	**0.2369**	0.2262	0.2288	0.2360
mfeat	0.5608	0.6416	0.6242	0.6538	0.6129	**0.7105**
PPMI	0.1969	0.2180	0.1987	0.2005	0.1989	**0.2366**
MSRA	0.5099	0.4786	0.5390	0.5155	0.5110	**0.6746**
Still DB	0.3013	0.2857	0.3004	0.3017	**0.3124**	0.3004

Clustering NMI result of all data sets

DATA	Laplacian	SPEC	MCFS	UDFS	AUMFS	MSMFS
Caltech10	0.1461	0.0962	0.1734	0.1767	0.2059	**0.2199**
Corel800	0.2198	0.1235	0.2255	0.1960	0.2302	**0.2400**
flickr	0.0993	0.1026	**0.1353**	0.1184	0.1309	0.1279
mfeat	0.5699	0.5960	0.6157	0.5983	0.5920	**0.6253**
PPMI	0.0224	0.0310	0.0255	0.0194	0.0238	**0.0461**
MSRA	0.4076	0.3902	0.4467	0.4100	0.4122	**0.5915**
Still DB	0.1019	0.0850	0.0930	0.0951	0.1035	**0.1051**

Outlines

1. Introduction
2. Multi-view Feature Selection
3. Mixed Sparsity Regularized Feature Selection
4. Experiment
5. **Conclusion**

5. Conclusion

- Automatically learning the view weights.
- Alleviate the effect of the outlier views and features with noisy information.

Thank you!

Q & A