Finding Multiple Stable Clusterings

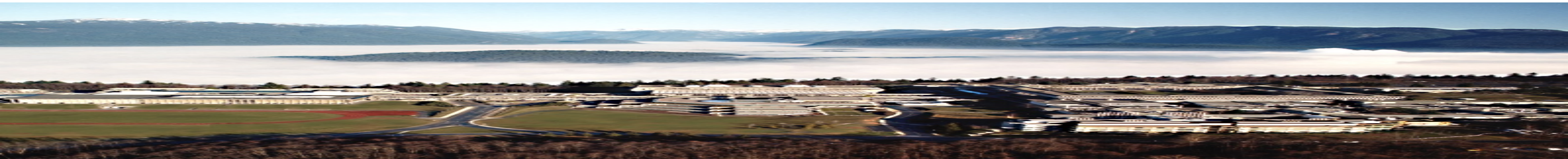
Juhua Hu*, Qi Qian^, Jian Pei*, Rong Jin^, and Shenghuo Zhu’

*Simon Fraser University
Burnaby, BC, Canada

^Michigan State University
East Lansing, MI, USA

’Alibaba Group
Seattle, WA, USA

ICDM 2015
IEEE International Conference on Data Mining



All About (Multi-)Clustering

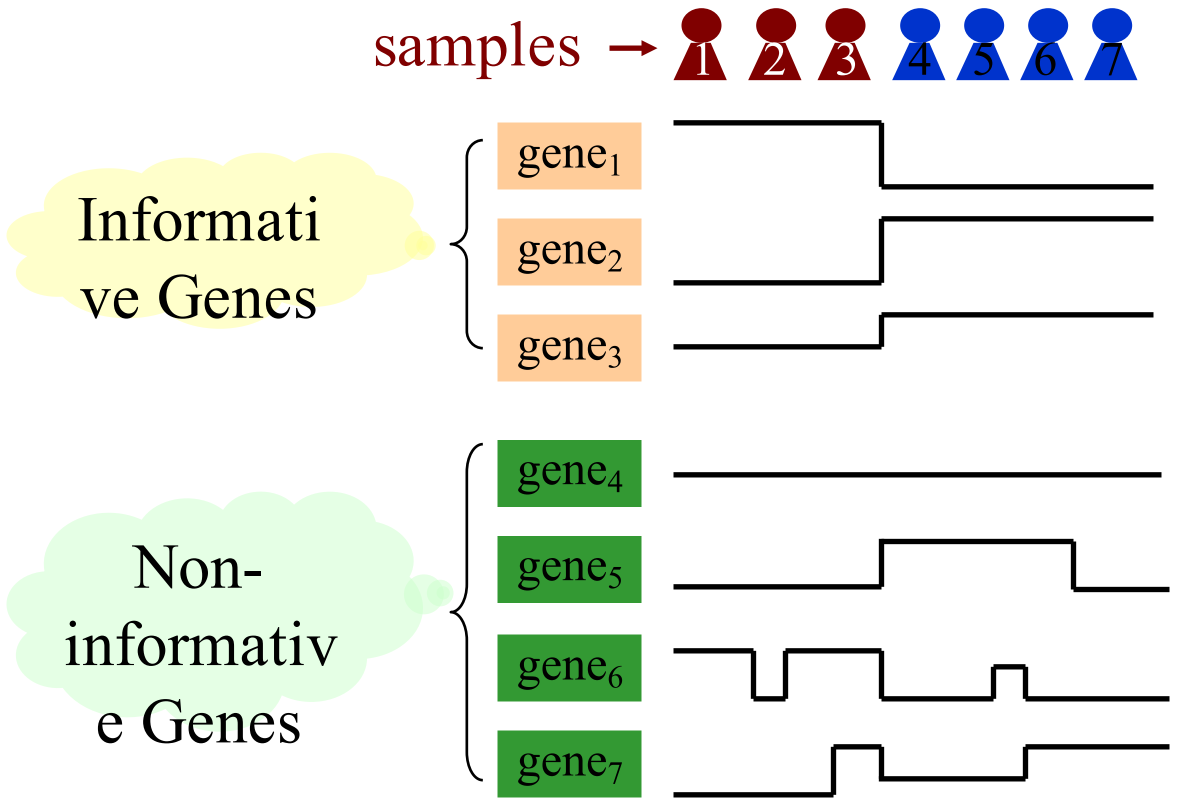
IDEAL
Intelligent Data Engineering and Analytics Lab

- Explorative
- Iterative
- Subjective
- “Every model is wrong, but some are more useful than the others”
- “If you torture the data long enough, it will confess” – Ronald H. Coase

Why Multiple Clusterings?

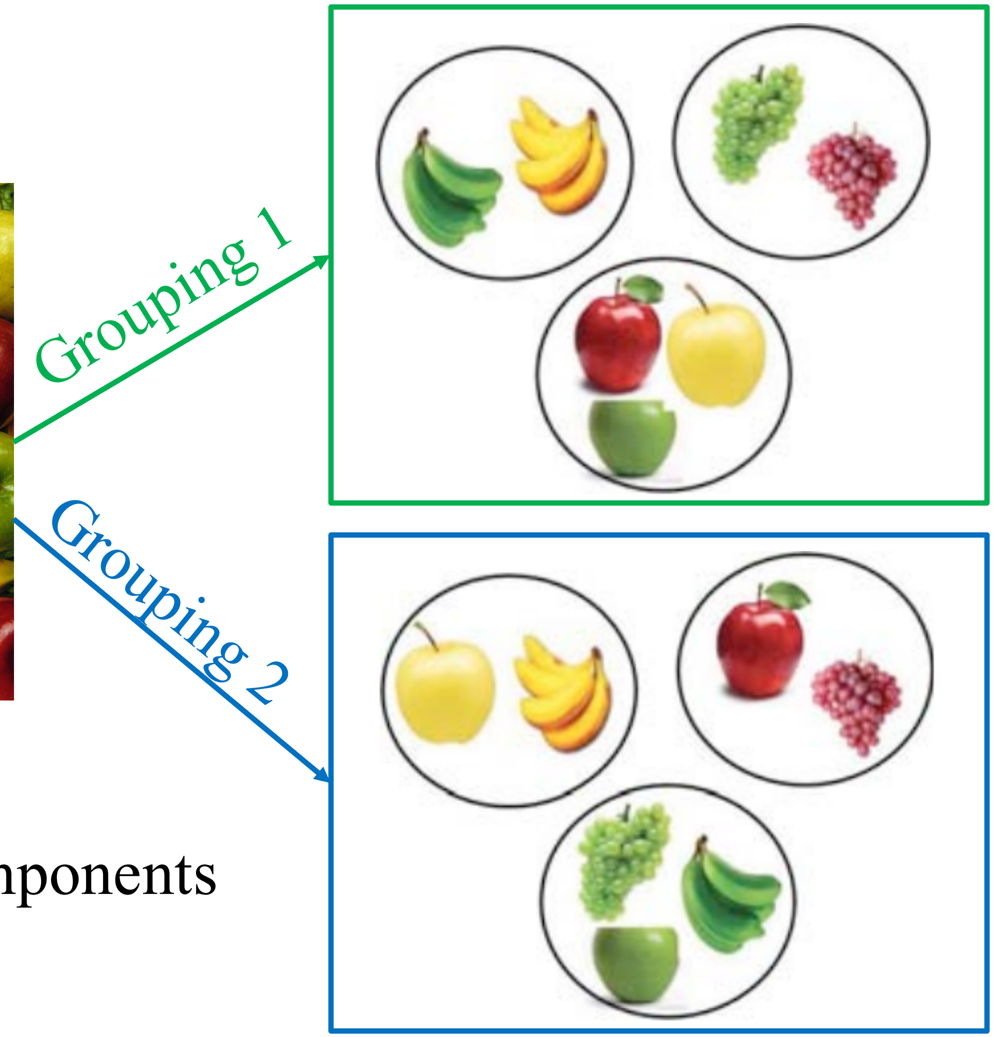
IDEAL
Intelligent Data Engineering and Analytics Lab

- Phenotype finding



Why Multiple Clusterings?

- Categorization



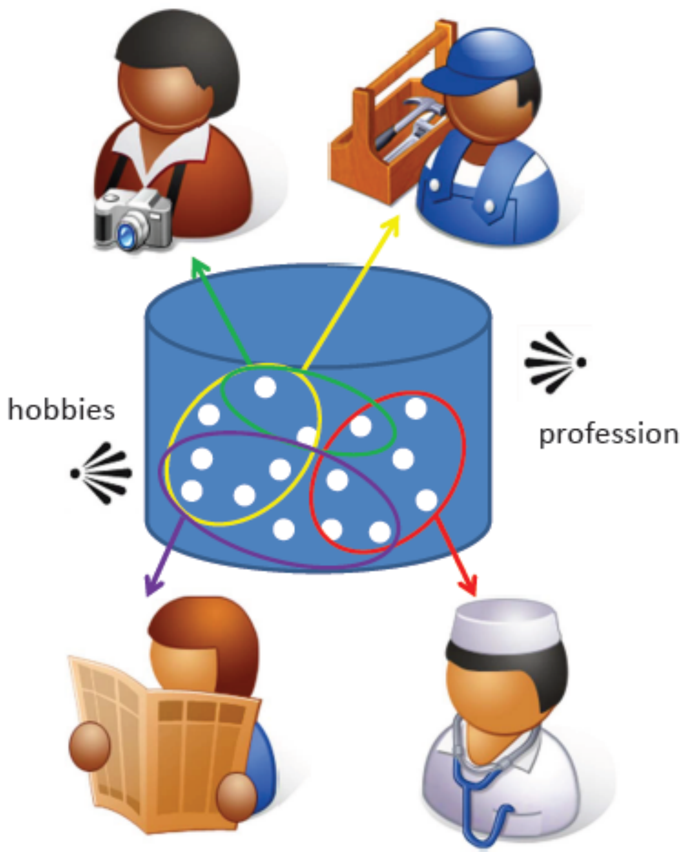
Grouping 3: nutrition components

…

Why Multiple Clusterings?

IDEAL
Intelligent Data Engineering and Analytics Lab

- Customer relation management



Given: profiles of customers

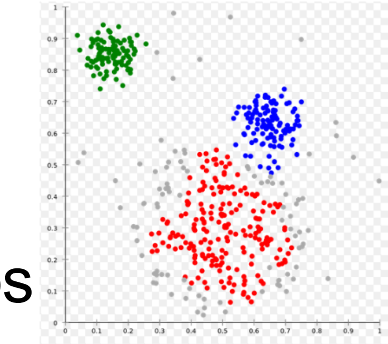
Task: product recommendation

Possible way: Group customers with similar behavior

Grouping 1: profession
Grouping 2: hobbies
Grouping 3: gender
…

Traditional Clustering

IDEAL
Intelligent Data Engineering and Analytics Lab



- Goal
  - Group similar objects in one group
  - Separate dissimilar objects in different groups
- Examples: k-means, PAM, DBScan, …
- BUT, only a single clustering solution is given
  - A clustering consists of multiple clusters
- Challenges
  - How to find multiple independent clusterings?
  - How to measure the independency among different clusterings?

Related work

- Alternative clustering (e.g., COALA [Bae & Bailey, ICDM'06])
  - Given a clustering
  - Dissimilarity + Quality
  - Highly sensitive to the input clustering
- Meta-clustering[Caruana et al., ICDM'06]
  - Generate many clusterings
  - Dissimilarity
  - High computational cost
- Subspace multi-clustering
  - Different subspaces reflect different perspectives
  - Exponential number of subspaces & overwhelming results
  - E.g., CLIQUE[Agrawal et al., SIGMOD'98], grid-based

Challenges Remained

IDEAL
Intelligent Data Engineering and Analytics Lab

- Too many clusterings – overwhelming
- Stable clusterings – not sensitive to initialization and noise

Problem Formulation

IDEAL
Intelligent Data Engineering and Analytics Lab

- Input
  - Data $X \in \mathbb{R}^{d \times n}$
  - The number of clusters (in each clustering)
- Output
  - A clustering $c = \{X_1, X_2, \ldots, X_k\}$ is an exclusive partitioning of the input data
  - Multiple clusterings
- Feature subspaces within the simplex

$$\Delta^d = \{w_1\mathbf{q}_1 + w_2\mathbf{q}_2 \cdots + w_d\mathbf{q}_d | w_m \geq 0, \sum_{m=1}^{d} w_m = 1\}$$

$$\mathbf{q}_1 = (1, 0, 0, \cdots, 0), \mathbf{q}_2 = (0, 1, 0, \cdots, 0), \cdots, \mathbf{q}_d = (0, 0, 0, \cdots, 1)$$

Similarity between Two Objects

IDEAL
Intelligent Data Engineering and Analytics Lab

- Under a feature weight vector $\mathbf{w} = (w_1, w_2, \cdots, w_d)$

$$S_{i,j} = e^{-\|\mathbf{x}'_i - \mathbf{x}'_j\|_2^2}$$

  - Where $\mathbf{x}'_i = \mathbf{w} \odot \mathbf{x}_i$
- Similarity matrix S

Clustering Stability

IDEAL
Intelligent Data Engineering and Analytics Lab

- Normalized Laplacian matrix $L = D^{-1/2} S D^{-1/2}$
  - Where D is a diagonal matrix formed by

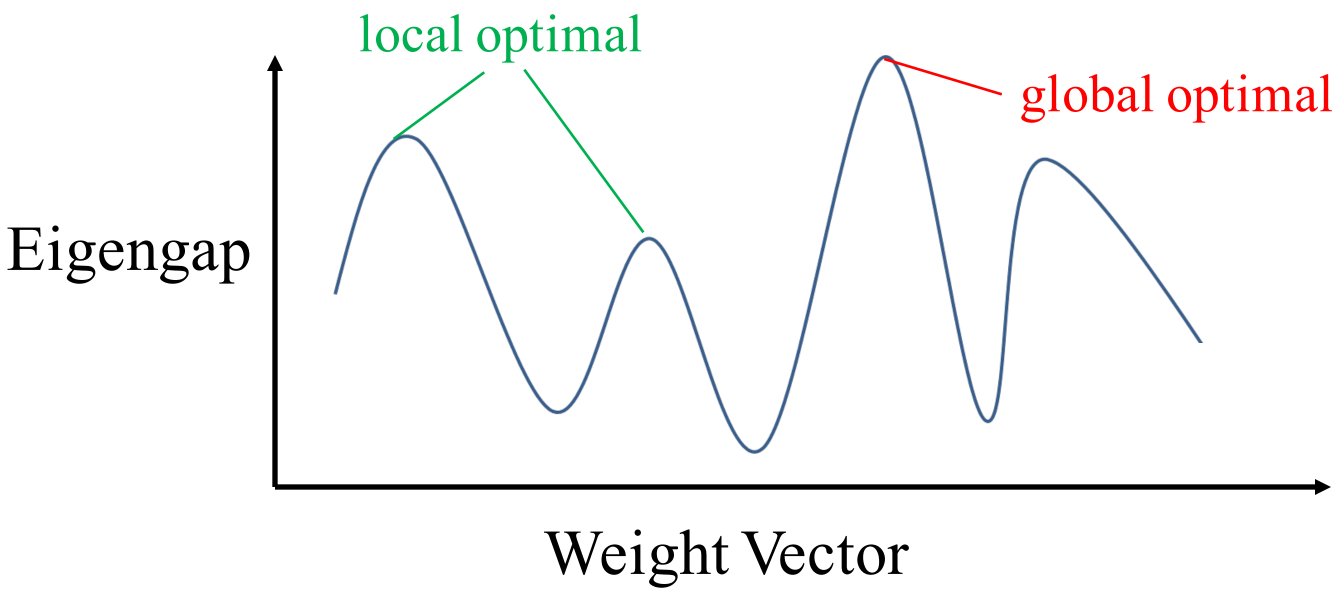
$$D_i = \sum_{j=1}^{n} S_{i,j}, \qquad i = 1, 2, \ldots, n$$

Given a Laplacian matrix L, if the eigengap $\lambda_k(L) - \lambda_{k+1}(L)$ is large enough, the top k eigenvectors of $L_{perb} = L + \epsilon$ are the same as those of L, where ϵ is a symmetric perturbation matrix of small spectral norm $\|\epsilon\|_2$.

- In spectral clustering, if the top k eigenvectors are the same for L and L_{perb}, the clusterings based on the same eigenvectors are the same

Finding one stable clustering

$$\arg\max_{\mathbf{w}\in\Delta^d} \lambda_k(L) - \lambda_{k+1}(L)$$



1. Randomly initialize w
2. Iterative gradient ascent

Multiple stable clusterings

$$\arg\max_{\mathbf{w}\in\Delta^d} \lambda_k(L) - \lambda_{k+1}(L) + \frac{\delta}{2}\frac{1}{|W|}\sum_{\boxed{\mathbf{w}_p\in W}} \|\mathbf{w} - \mathbf{w}_p\|_2^2$$

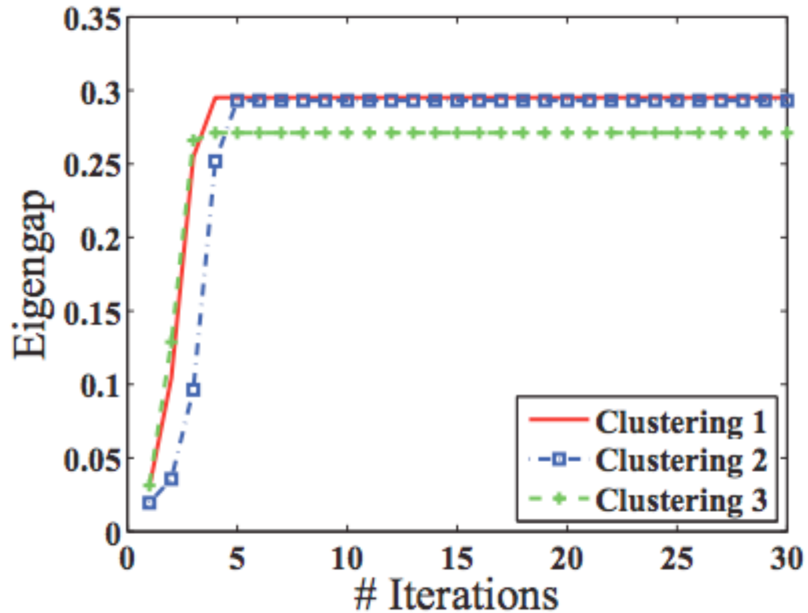
Previously obtained solutions

Sequentially finding stable and different weight vectors

Synthetic Data Sets

IDEAL
Intelligent Data Engineering and Analytics Lab

- $X \in \{0,1\}^{50\times 3}$, k = 2
- Ground truth: each feature itself
- Baselines: k-means and spectral clustering
- Convergence



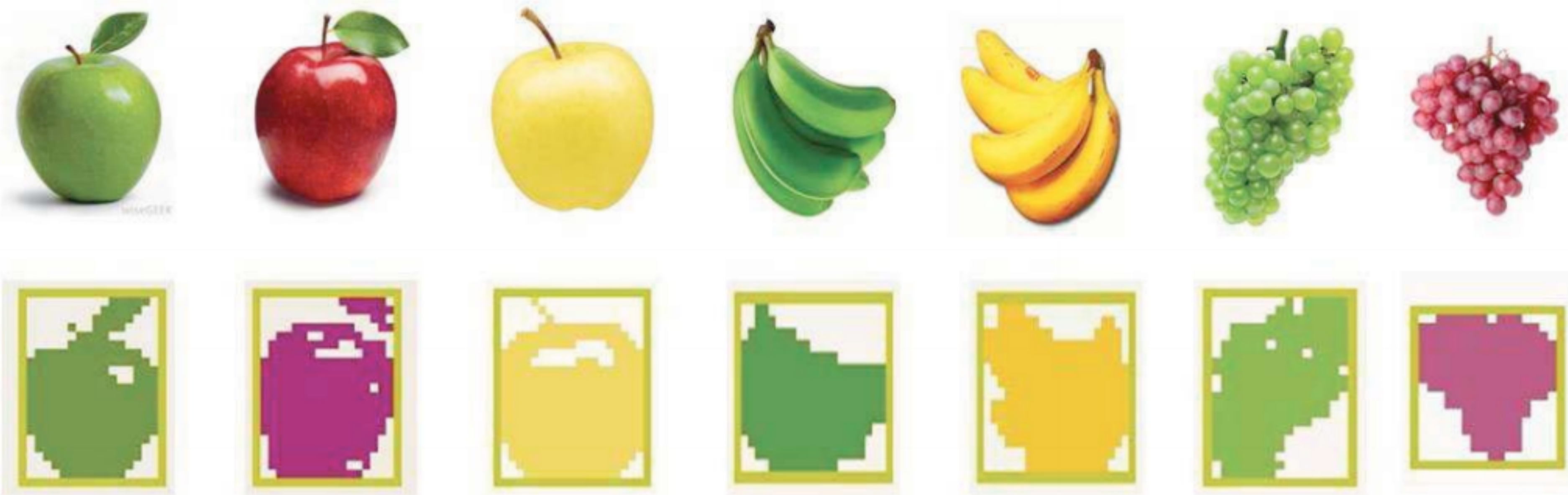
Results on synthetic data

IDEAL
Intelligent Data Engineering and Analytics Lab

Clustering produced by methods	Clustering in ground truth	NMI	RI	AR	MI	HI	Eigengap	Weight vector **w**
k-means	Clustering 1	**1.000**	**1.000**	-	**.0000**	**1.000**	-	(.3333;.3333;.3333)
	Clustering 2	.0043	.4783	-	.5217	-.043		
	Clustering 3	.0039	.4737	-	.5263	-.053		
Spectral	Clustering 1	.0283	.5569	-	.4431	.1138	.0039	(.3333;.3333;.3333)
	Clustering 2	**.7263**	**.7628**	-	**.2372**	**.5257**		
	Clustering 3	.0718	.6491	-	.3509	.2983		
Clustering 1	Clustering 1	**1.000**	**1.000**	-	**.0000**	**1.000**	.2951	(1.000;.0000;.0000)
	Clustering 2	.0043	.4923	-	.5077	-.015		
	Clustering 3	.0039	.5292	-	.4708	.0585		
Clustering 2	Clustering 1	.0043	.4783	-	.5217	-.044	.2931	(.0000;1.000;.0000)
	Clustering 2	**1.000**	**1.000**	-	**.0000**	**1.000**		
	Clustering 3	.0154	.5573	-	.4427	.1146		
Clustering 3	Clustering 1	.0039	.4737	-	.5263	-.053	.2711	(.0000;.0000;1.000)
	Clustering 2	.0154	.5088	-	.4105	.4912		
	Clustering 3	**1.000**	**1.000**	-	**.0000**	**1.000**		

Image data

IDEAL
Intelligent Data Engineering and Analytics Lab



Each image is represented by 6 features

- 3 average color features and 3 shape features

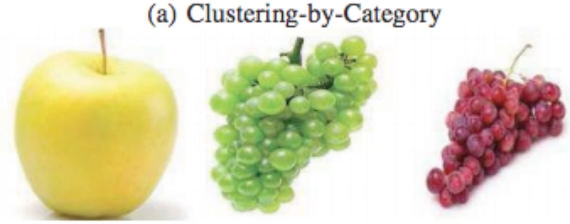
Results on image data

IDEAL
Intelligent Data Engineering and Analytics Lab

Clustering produced by methods	Clusterings in ground truth/by Spectral	NMI	RI	AR	MI	HI	Weight vector **w**
k-means		.1486	.5659	.0684	.4341	.1319	[.1667;.1667;.1667;.1667;.1667;.1667]
Spectral		.1432	.5650	.0611	.4350	.1300	[.1667;.1667;.1667;.1667;.1667;.1667]
Clustering 1	Clustering-by-Category	.1394	.5857	.0818	.4143	.1714	[.2538;.0011;.0765;.0655;.1004;.5027]
Clustering 2		**.1627**	**.6045**	**.1289**	**.3954**	**.2092**	[.3222;.0000;.0000;.0000;.6778;.0000]
Clustering 3		.1449	.5886	.0883	.4114	.1773	-
Clustering 4		.1151	.5716	.0465	.4284	.1432	[.4012;.0000;.0000;.5988;.0000;.0000]
k-means		.5905	.7626	.4905	.2374	.5253	-
Spectral		.5522	.7559	.4730	.2441	.5117	-
Clustering 1	Clustering-by-Color	.6160	.7711	.4926	.2289	.5241	-
Clustering 2		.5564	.7474	.4436	.2526	.4949	-
Clustering 3		**.6886**	**.8051**	**.5681**	**.1949**	**.6103**	[.4468;.0000;.5532;.0000;.0000;.0000]
Clustering 4		.5124	.7291	.3971	.2709	.4582	-
k-means	Spectral	.8839	.9581	.9118	.0419	.9161	-

Cluster representatives that are nearest to the cluster centers for clustering 2 and clustering 3

(a) Clustering-by-Category



(b) Clustering-by-Color

Conclusions

- Contributions
  - Introduce a new measure for multi-clustering
    - Clustering stability
  - Propose a new multi-clustering method MSC
    - Empirically finding all hidden stable clusterings
- Future directions
  - k is not fixed
  - Different stable clusterings have different number of clusters