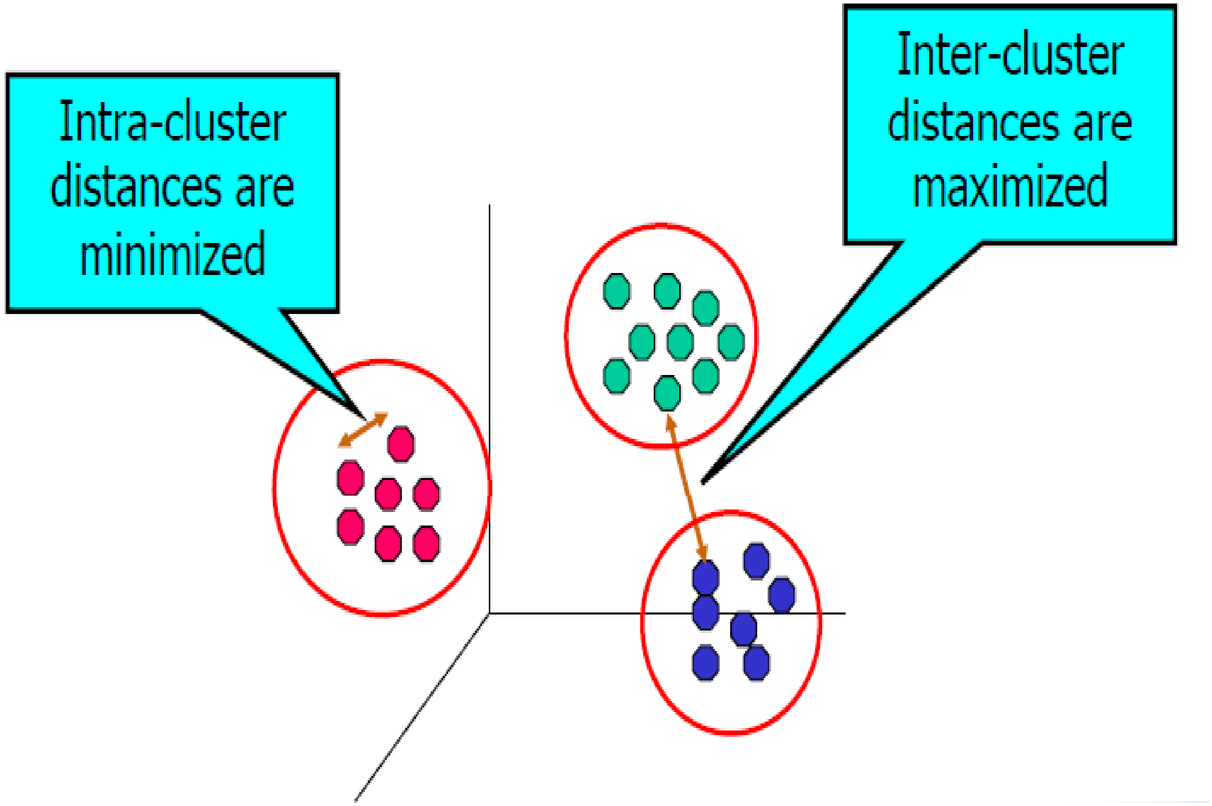
K-means Based Consensus Clustering

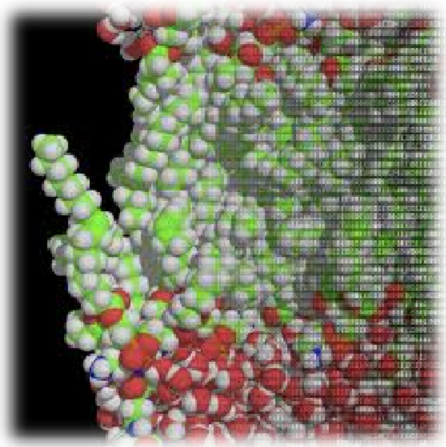
Dr. Junjie Wu
Beihang University

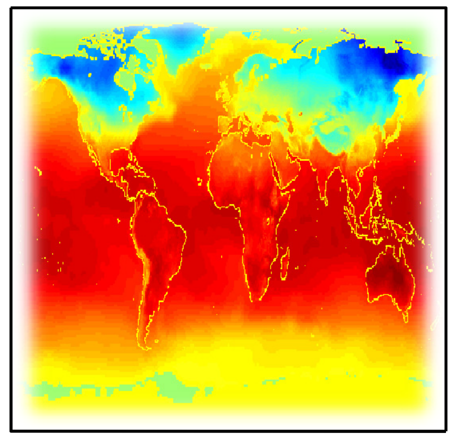
Outline

- **Motivations**
- Point-to-Centroid Distance
- Utility Functions for KCC
- Experimental Results
- Concluding remarks

Cluster Analysis

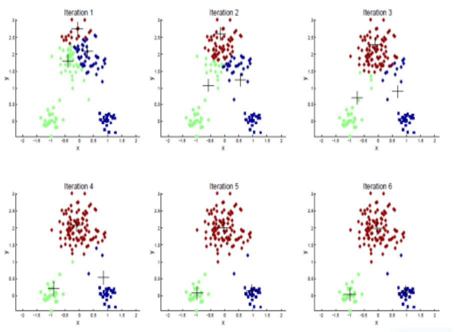


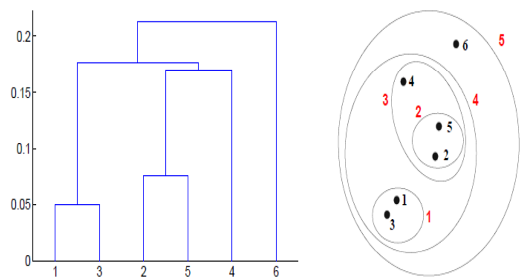


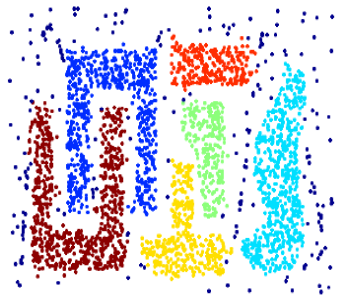


Clustering Algorithms

- Prototype-based: K-means, FCM
- Density-based: DBSCAN, CLIQUE
- Graph-based: AHC, MinMaxCut
- Hybrid: K-means + AHC







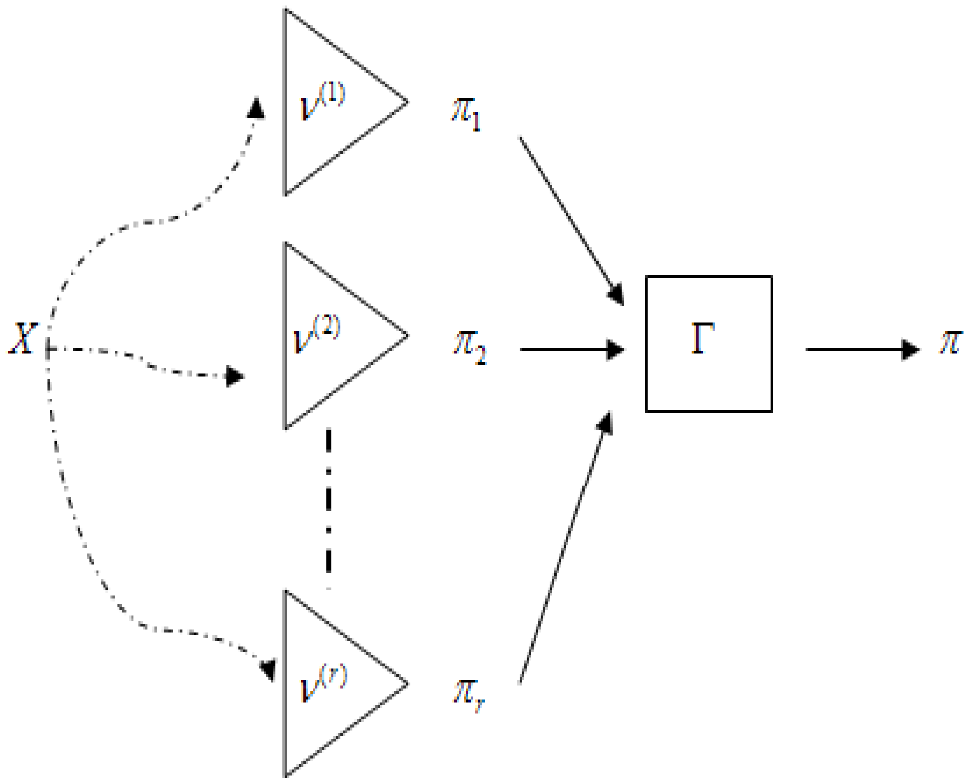
Problems with Single Clusterer

- No perfect one!
- Sensitive to data factors
- Hard to set proper parameters
- May converge to bad saddle points
- Or just too heuristic …

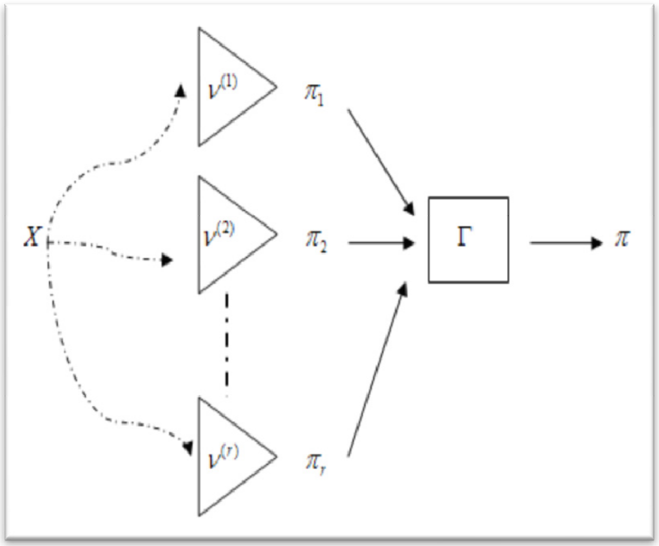
Can we find a new way?

Consensus Clustering

- To find a best partitioning from multiple basic partitionings (an ensemble-classifier thinking)



Consensus Clustering, cont'd



$$\Gamma(\pi, \Pi) = \sum_{i=1}^{r} w_i U(\pi, \pi_i)$$

- U: utility function
- w_i: weight of π_i
- Γ: consensus function

$$X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$$

$$\pi_1 = (1,1,1,2,2,3,3)^T$$

$$\pi_2 = (2,2,2,3,3,1,1)^T$$

$$\pi_3 = (1,1,2,2,2,3,3)^T$$

$$\pi_4 = (1,1,1,1,2,2,2)^T$$

$$\Pi = \{\pi_1, \pi_2, \pi_3, \pi_4\}$$

Consensus Clustering, cont'd

- Advantages
  - **Robust**: lower the risks from weird data, improper algorithms and parameters, etc.
  - **Novelty**: may help find a better structure
  - **Concurrency**: run in parallel
  - **A Must**: only have partitioning episodes
- Challenge
  - NP-complete problem

Related Work

- Graph-based algorithms (Strehl et al., JMLR, 2002)
- Co-association matrix method (Fred and Jain, PAMI, 2005)
- Binary matrix method (Topchy et al., ICDM, 2003)
- Other heuristics (Lu et al., AAAI, 2008; …)

Why K-means Consensus Clustering

- Simple
- Robust
- Efficient

	π_1	π_2	π_3	π_4
x_1	1	2	1	1
x_2	1	2	1	1
x_3	1	2	2	?
x_4	2	3	2	1
x_5	2	3	2	2
x_6	3	1	3	?
x_7	3	1	3	?

$\Leftrightarrow$

X^b

π_1			π_2			π_3			π_4	
1	0	0	0	1	0	1	0	0	1	0
1	0	0	0	1	0	1	0	0	1	0
1	0	0	0	1	0	0	1	0	0	0
0	1	0	0	0	1	0	1	0	1	0
0	1	0	0	0	1	0	1	0	0	1
0	0	1	1	0	0	0	0	1	0	0
0	0	1	1	0	0	0	0	1	0	0

NP-complete

U?

Roughly linear
(K-means)

Main Contributions

- Theory for KCC utility functions
- KCC algorithm for inconsistent samples
- Some empirical findings
  - U_H is a good KCC utility function
  - RFS strategy is useful in some circumstances
  - Some mutual funds have unethical behaviors

Outline

- Motivations
- **Point-to-Centroid Distance**
- Utility Functions for KCC
- Experimental Results
- Concluding remarks

K-means Clustering

- Objective function

$$\min \sum_{k=1}^{K} \sum_{x \in C_k} w_x f(x, m_k)$$

- Two phase iterations
  - Assign *x* to the nearest centroid
  - Update centroids
- The arithmetic mean centroid is preferred

Point-to-Centroid Distance

- What if fix centroid type to arithmetic mean?
- A definition

(Point-to-Centroid Distance): Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a nonempty open convex set. A twice continuously differentiable function $f : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}_+$ is called a point-to-centroid distance, if there exists some higher-order continuously differentiable convex function $\phi : \mathcal{S} \mapsto \mathbb{R}$ such that $f(x,y) = \phi(x) - \phi(y) - (x-y)^T \nabla \phi(y)$.

- A theorem

Let $\mathcal{S}$ be a nonempty open convex set. Assume any data set to be clustered is a subset of $\mathcal{S}$, i.e., $\mathcal{X} \subset \mathcal{S}$. Then a distance function $f : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}_+$ fits K-means directly if and only if f is a point-to-centroid distance.

Examples

- $f(x,y) = \phi(x) - \phi(y) - (x-y)^T \nabla \phi(y)$

$\phi(x)$	f(x,y)	Name
$\lVert x \rVert^2$	$\lVert x-y \rVert^2$	Sqrt Euc. Dist.
-H(x)	$D(x \parallel y)$	KL-divergence
$\lVert x \rVert$	$\lVert x \rVert - x^t y / \lVert y \rVert$	Cosine Dist.

- Interestingly, f is not a metric!

Outline

- Motivations
- Point-to-Centroid Distance
- **Utility Functions for KCC**
- Experimental Results
- Concluding remarks

Definition and Simplification

DEFINITION 1 (KCC UTILITY FUNCTION). *A utility function U is called a KCC utility function, if there exists a distance function f such that*

$$\max \sum_{i=1}^{r} w_i U(\pi, \pi_i) \Leftrightarrow \min \sum_{k=1}^{K} \sum_{x_l^b \in C_k} f(x_l^b, m_k)$$

A Critical Fact

- The contingency table for π and π_i

π_i

π		$C_1^{(i)}$	$C_2^{(i)}$	$\cdots$	$C_{K_i}^{(i)}$	$\sum$
	C_1	$n_{11}^{(i)}$	$n_{12}^{(i)}$	$\cdots$	$n_{1K_i}^{(i)}$	n_{1+}
	C_2	$n_{21}^{(i)}$	$n_{22}^{(i)}$	$\cdots$	$n_{2K_i}^{(i)}$	n_{2+}
	$\cdot$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$
	C_K	$n_{K1}^{(i)}$	$n_{K2}^{(i)}$	$\cdots$	$n_{KK_i}^{(i)}$	n_{K+}
	$\sum$	$n_{+1}^{(i)}$	$n_{+2}^{(i)}$	$\cdots$	$n_{+K_i}^{(i)}$	n

n $\longrightarrow$ p

- The centroid is right the normalized row!

$$m_{k,ij} = \frac{\sum_{x_l^{(b)} \in C_k} x_{l,ij}^{(b)}}{n_{k+}} = \frac{n_{kj}^{(i)}}{n_{k+}} = \frac{p_{kj}^{(i)}}{p_{k+}}, 1 \leq j \leq K_i$$

A Sufficient Condition

THEOREM 2. *A utility function U is a KCC utility function if there exists a continuously differentiable convex function ϕ such that*

$$\sum_{i=1}^{r} w_i U(\pi, \pi_i) = \sum_{k=1}^{K} p_{k+} \phi(m_k)$$

where $\Pi = \{\pi_1, \cdots, \pi_r\}$ and π are arbitrary partitionings of $\mathcal{X}$, $p_{k+} = |C_k|/|\mathcal{X}|$ is the relative size of cluster C_k in π, and m_k is the centroid (i.e., the arithmetic mean of cluster members) of C_k when applying π to $\mathcal{X}^{(b)}$.

$$\min \sum_{k=1}^{K} \sum_{x_l^b \in C_k} f(x_l^b, m_k) \overset{P2C-D}{\Leftrightarrow} \min \sum_l \phi(x_l^b) - \sum_{k=1}^{K} p_{k+} \phi(m_k) \Leftrightarrow \max \sum_{k=1}^{K} p_{k+} \phi(m_k)$$

A Sufficient Condition, cont'd

THEOREM 3. *If U is a KCC utility function satisfying Theorem 2 then there exists a continuously differentiable convex function* φ *such that*

$$\phi(m_k) = \sum_{i=1}^{r} w_i \varphi(m_{k,i}), 1 \le k \le K$$

COROLLARY 1. *If U is a KCC utility function satisfying Theorem 2 then there exists a continuously differentiable convex function* φ *such that* $\forall\, i$

$$U(\pi, \pi_i) = \sum_{k=1}^{K} p_{k+} \varphi(< p_{k1}^{(i)} / p_{k+}, \cdots, p_{kj}^{(i)} / p_{k+}, \cdots, p_{kK_i}^{(i)} / p_{k+} >)$$

Examples

	$\varphi(m_{k,i})$	$U(\pi,\pi_i)$	$f(x_l^b, m_k)$
U_C	$\sum_{j=1}^{K_i} m_{k,ij}^2 - \sum_{j=1}^{K_i} (p_{+j}^{(i)})^2$	$\sum_{k=1}^{K} p_{k+} \sum_{j=1}^{K_i} (p_{kj}^{(i)} / p_{k+})^2 - \sum_{j=1}^{K_i} (p_{+j}^{(i)})^2$	$\sum_{i=1}^{r} w_i \parallel x_{l,i}^{(b)} - m_{k,i} \parallel^2$
U_H	$\sum_{j=1}^{K_i} m_{k,ij} \log m_{k,ij} - \sum_{j=1}^{K_i} p_{+j}^{(i)} \log p_{+j}^{(i)}$	$MI(C, C^{(i)})$	$\sum_{i=1}^{r} w_i D(x_{l,i}^{(b)} \parallel m_{k,i})$
$U_{\cos}$	$\parallel m_{k,i} \parallel - \parallel < p_{+1}^{(i)}, \cdots, p_{+K_i}^{(i)} > \parallel$	$\sum_{k=1}^{K} p_{k+} \sqrt{\sum_{j=1}^{K_i} (p_{kj}^{(i)} / p_{k+})^2} - \sqrt{\sum_{j=1}^{K_i} (p_{+j}^{(i)})^2}$	$\sum_{i=1}^{r} w_i (1 - \cos(x_{l,i}^{(b)}, m_{k,i}))$
U_{Lp}	$\parallel m_{k,i} \parallel_p - \parallel < p_{+1}^{(i)}, \cdots, p_{+K_i}^{(i)} > \parallel_p$	$\sum_{k=1}^{K} p_{k+} \sqrt[p]{\sum_{j=1}^{K_i} (p_{kj}^{(i)} / p_{k+})^p} - \sqrt[p]{\sum_{j=1}^{K_i} (p_{+j}^{(i)})^p}$	$\sum_{i=1}^{r} w_i \left(1 - \frac{\sum_{j=1}^{K_i} x_{l,ij}^{(b)} (m_{k,ij})^{p-1}}{(\parallel m_{k,i} \parallel_p)^{p-1}}\right)$

U_C U_H $U_{\cos}$ U_{L_5} U_{L_8}

Properties

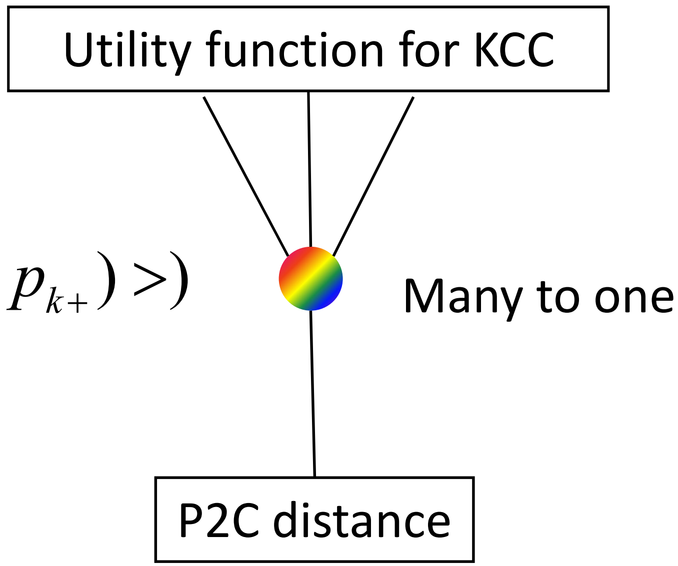
- Non-uniqueness of U_φ

$$\varphi_s(m_{k,i}) = \varphi(m_{k,i}) - \underbrace{\varphi(<p_{+1}^{(i)},\cdots,p_{+K_i}^{(i)}>)}_{\alpha}$$

$$U_{\varphi_s}(\pi,\pi_i) = \sum_{k=1}^{K} p_{k+}\varphi_s(<(p_{k1}^{(i)}/p_{k+}),\cdots,(p_{kK_i}^{(i)}/p_{k+})>)$$

$$= U_\varphi(\pi,\pi_i) - \varphi(<p_{+1}^{(i)},\cdots,p_{+K_i}^{(i)}>)$$

standard form, or utility gain



Properties, cont'd

- Non-negativity of Utility Gain

$$\because U_{\varphi}(\pi,\pi_i) = \sum_{k=1}^{K} p_{k+}\varphi(m_{k,i}) \geq \varphi(\sum_{k=1}^{K} p_{k+}m_{k,i})$$

$$= \varphi(\sum_{k=1}^{K} < p_{k1}^{(i)},\cdots,p_{kK_i}^{(i)} >) = \varphi(< p_{+1}^{(i)},\cdots,p_{+K_i}^{(i)} >)$$

$$\therefore U_{\varphi_s}(\pi,\pi_i) = U_{\varphi}(\pi,\pi_i) - \varphi(< p_{+1}^{(i)},\cdots,p_{+K_i}^{(i)} >) \geq 0$$

Properties , cont'd

- Utility Gain Ratio

$$\varphi_n(m_{k,i}) = \varphi_s(m_{k,i}) / |\varphi(< p_{+1}^{(i)}, \cdots, p_{+K_i}^{(i)} >)|$$

$$U_{\varphi_n}(\pi, \pi_i) = \sum_{k=1}^{K} p_{k+} \varphi_n(< (p_{k1}^{(i)} / p_{k+}), \cdots, (p_{kK_i}^{(i)} / p_{k+}) >)$$

$$= \frac{U_\varphi(\pi, \pi_i) - \varphi(< p_{+1}^{(i)}, \cdots, p_{+K_i}^{(i)} >)}{|\varphi(< p_{+1}^{(i)}, \cdots, p_{+K_i}^{(i)} >)|}$$

normalized form, or utility gain ratio

Inconsistent Samples

- Often we have basic partitionings from inconsistent sample sets
- Adjustments for KCC

$$m_{k,ij} = \frac{\sum_{x_l^b \in C_k \cap C_k^{(i)}} x_{l,ij}^b}{n_k - \tilde{n}_k^{(i)}}$$

$$p_{k+}^{(i)} = \frac{n_k - \tilde{n}_k^{(i)}}{n - \tilde{n}_+^{(i)}}$$

$$U(\pi, \pi_i) = \sum_{k=1}^{K} p_{k+}^{(i)} \varphi(< p_{k1}^{(i)} / p_{k+}^{(i)}, \cdots, p_{kK_i}^{(i)} / p_{k+}^{(i)} >)$$

Inconsistent Samples, cont'd

- The proof for convergence

$$\begin{aligned}
\Delta &= \sum_{i=1}^{r}\sum_{k=1}^{K}\sum_{x_l^b \in C_k \cap C_k^{(i)}} f(x_{l,i}^b, \underline{y}) - \sum_{i=1}^{r}\sum_{k=1}^{K}\sum_{x_l^b \in C_k \cap C_k^{(i)}} f(x_{l,i}^b, \underline{m_{k,i}}) \\
&= \sum_{i=1}^{r}\left[\sum_{k=1}^{K}\sum_{x_l^b \in C_k \cap C_k^{(i)}} f(x_{l,i}^b, y) - \sum_{k=1}^{K}\sum_{x_l^b \in C_k \cap C_k^{(i)}} f(x_{l,i}^b, m_{k,i})\right] \\
&= \sum_{i=1}^{r}\left[\sum_{k=1}^{K}\sum_{x_l^b \in C_k \cap C_k^{(i)}} \phi(m_{k,i}) - \phi(y) - (x_{l,i}^b - y)\nabla\phi(y)\right] \\
&= \sum_{i=1}^{r}\sum_{k=1}^{K}(n_k - \tilde{n}_k^{(i)}) f(m_{k,i}, y) \\
&\geq 0
\end{aligned}$$

P2C distance

Outline

- Motivations
- Point-to-Centroid Distance
- Utility Functions for KCC
- **Experimental Results**
- Concluding remarks

Data

Table 4: Some Characteristics of Real-World Data Sets.

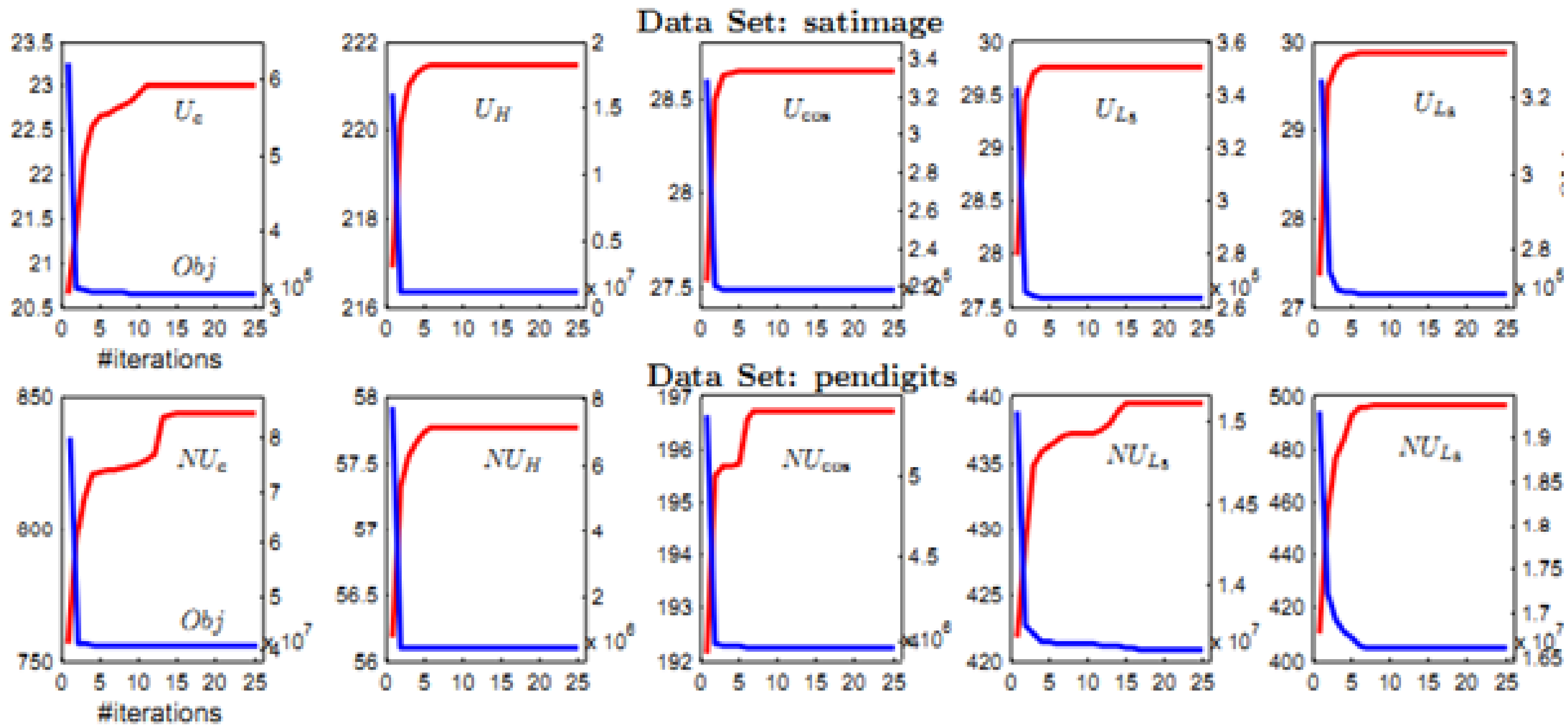
Data Set	Source	#Objects	#Attributes	#Classes	MinClassSize	MaxClassSize	CV
breast_w	UCI	699	9	2	241	458	0.439
breastTissues	UCI	106	9	6	14	22	0.185
ecoli	UCI	336	7	8	2	143	1.160
iris	UCI	150	4	3	50	50	0.000
pendigits	UCI	10992	16	10	1055	1144	0.042
satimage	UCI	4435	36	6	415	1072	0.425
wine	UCI	178	13	3	48	71	0.194
yeast	UCI	1484	8	10	5	463	1.170
k1b	WebACE	2340	21839	6	60	1389	1.316
sports	TREC	8580	126373	7	122	3412	1.022
tr45	TREC	690	8261	10	18	160	0.669

Strategies for Basic Clusterings

- For UCI data sets:
  - Random Parameter Selection (RPS) with K-means clustering: $K_i \in [K, \lceil \sqrt{n} \rceil], \forall\, i$
  - Random Feature Selection (RFS) with K-means clustering: two features for a basic clustering
- For text data sets:
  - Multiple Clustering Algorithms with CLUTO (5 clustering methods $\times$ 5 objective functions)
- Validation measure: normalized Rand index R_n

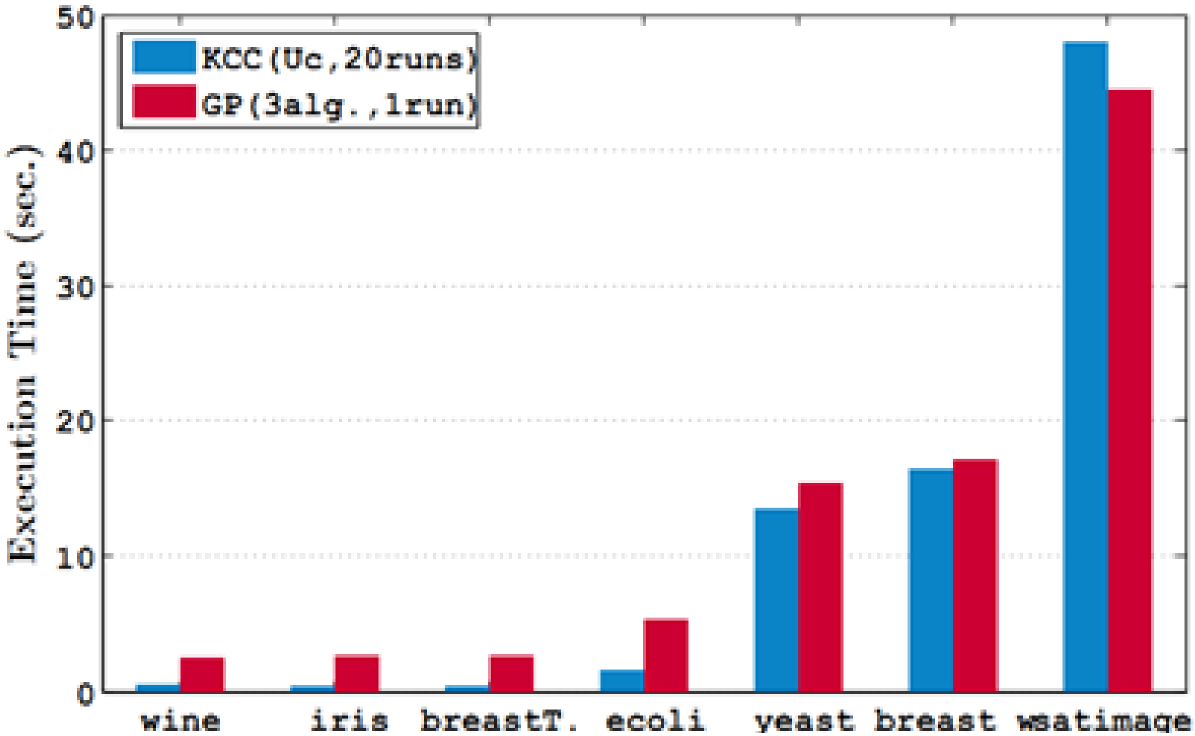
Experimental Results, #1

- Convergence property of KCC



Experimental Results, #1

- Convergence property of KCC



Experimental Results, #2

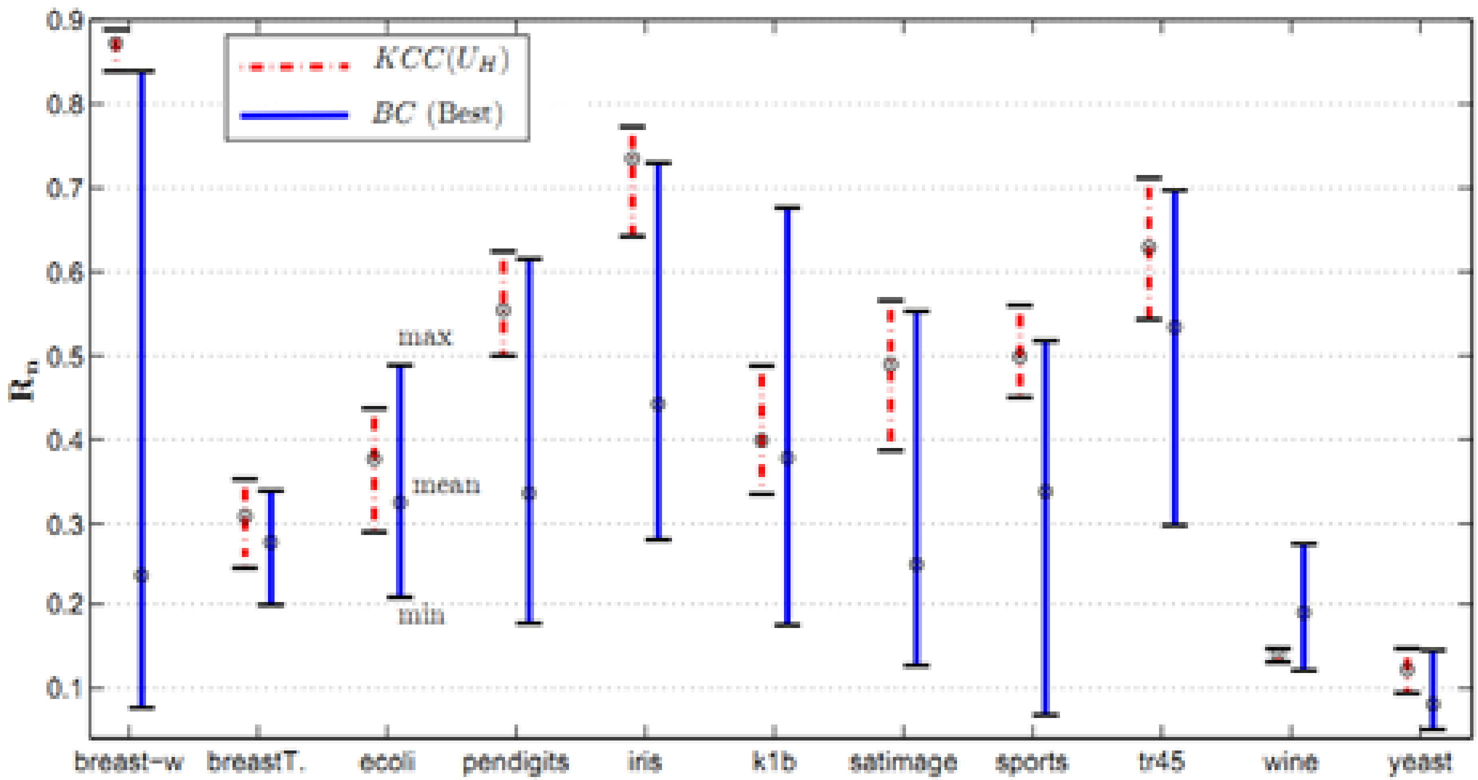
- Clustering accuracy of KCC

dataset	U_C	U_H	$U_{\cos}$	U_{L_5}	U_{L_8}	NU_C	NU_H	$NU_{\cos}$	NU_{L_5}	NU_{L_8}	GP	BC_AVG
breast_w	0.196	**0.872**	0.642	0.135	0.134	0.037	0.862	0.133	0.134	0.133	0.492	0.264
ecoli	0.356	**0.377**	0.358	0.364	0.364	0.360	0.377	0.367	0.353	0.358	0.351	0.310
pendigits	0.545	0.554	**0.591**	0.590	0.565	0.498	0.580	0.576	0.576	0.569	N/A[1]	0.335
satimage	0.338	0.490	0.494	0.484	0.482	0.292	**0.498**	0.454	0.432	0.385	0.385	0.248
yeast	0.127	0.122	0.125	0.133	0.129	0.130	0.119	0.129	**0.134**	0.129	0.119	0.082
k1b	0.350	0.399	0.411	**0.434**	0.408	0.341	0.387	0.351	0.374	0.369	0.423	0.409
sports	0.461	0.499	0.464	0.458	0.481	0.480	0.478	0.495	0.502	**0.510**	0.465	0.397
tr45	0.669	0.629	0.671	0.684	0.670	0.656	0.658	**0.688**	0.652	0.664	0.642	0.536
iris	0.749	0.735	0.746	0.746	0.746	0.702	0.737	0.746	0.746	0.746	**0.915**	0.463
bT[2]	0.301	0.309	0.298	0.286	0.286	0.295	0.299	0.313	0.295	0.282	**0.323**	0.278
wine[3]	0.144	0.140	0.144	0.137	0.137	0.146	0.138	0.145	0.145	0.143	0.147	**0.185**

$$U_H \succ NU_H \succ U_{\cos} \succ U_{L5} \succ U_{L8} \succ NU_{\cos} \succ NU_{L5} \succ NU_{L8} \succ U_C \succ NU_C$$

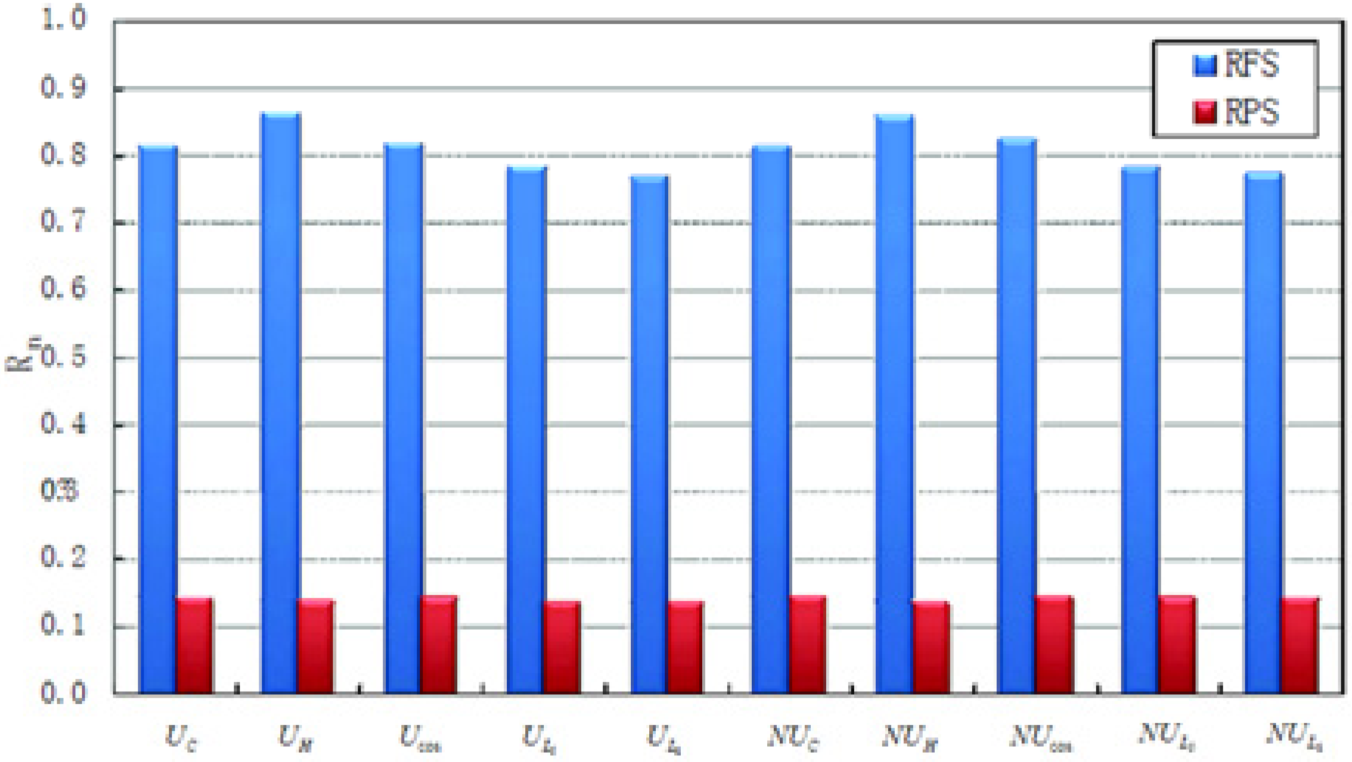
Experimental Results, #2

- Clustering accuracy of KCC



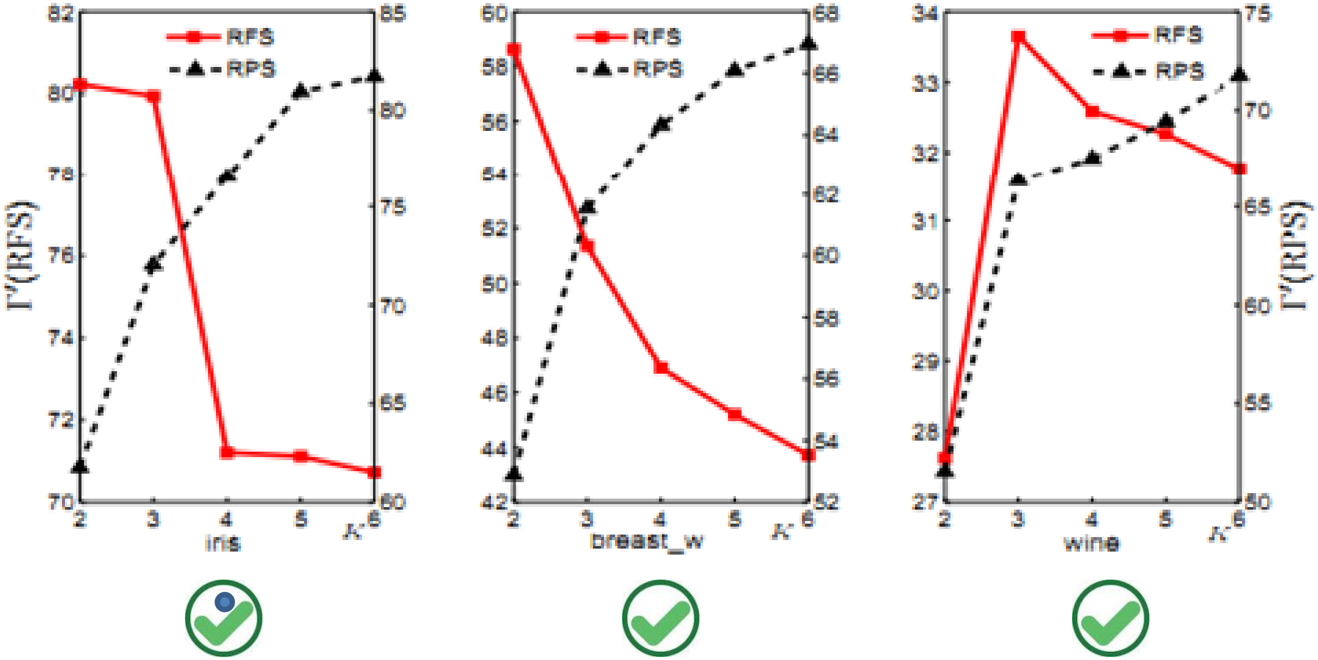
Experimental Results, #3

- Effects of RFS strategy



Experimental Results, #3

- Effects of RFS strategy



Outline

- Motivations
- Point-to-Centroid Distance
- Utility Functions for KCC
- Experimental Results
- **Concluding remarks**

Conclusions

- Study “K-means based consensus clustering”
  - Give a sufficient condition
  - Handle the inconsistent samples
  - Give some empirical results for practical use
- Future work
  - Give the necessary condition (almost done!)
  - Applications

Thank You!

http://datamining.buaa.edu.cn