Subspace Clustering Ensembles

Carlotta Domeniconi

Department of Computer Science
George Mason University

Joint work with: Francesco Gullo and Andrea Tagarelli

Third MultiClust Workshop
April 28, 2012
Anaheim, California

Data Clustering: challenges and advanced approaches

Data Clustering challenges in real-life domains:

1. High dimensionality
2. Ill-posed nature

Advances in data clustering:

- Subspace Clustering (handles issue 1)
- Clustering Ensembles (handles issue 2)
- Subspace Clustering Ensembles (handles both issues 1 and 2)

Subspace Clustering (1)

Subspace clustering: discovering clusters of objects that rely on the type of information (feature subspace) used for representation

- In high dimensional spaces, finding compact clusters is meaningful only if the assigned objects are projected onto the corresponding subspaces

Subspace Clustering (2)

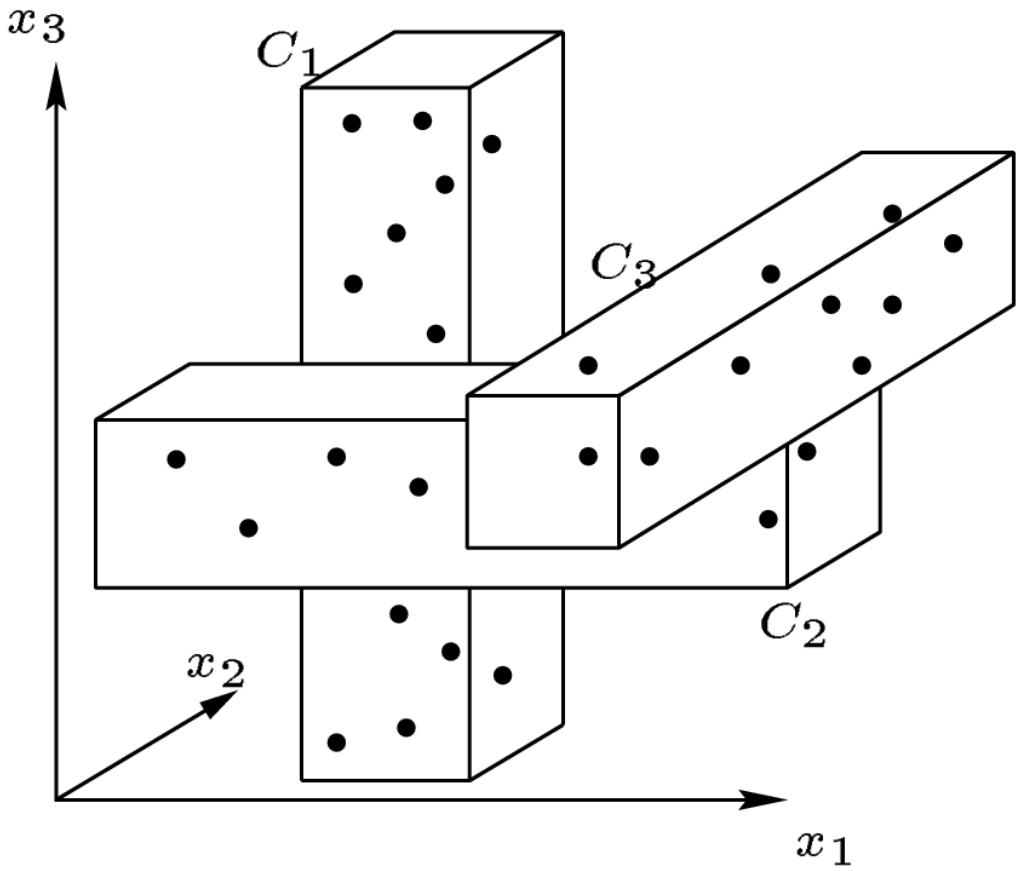


figure borrowed from [Procopiuc et Al., SIGMOD'02]

Subspace Clustering (3)

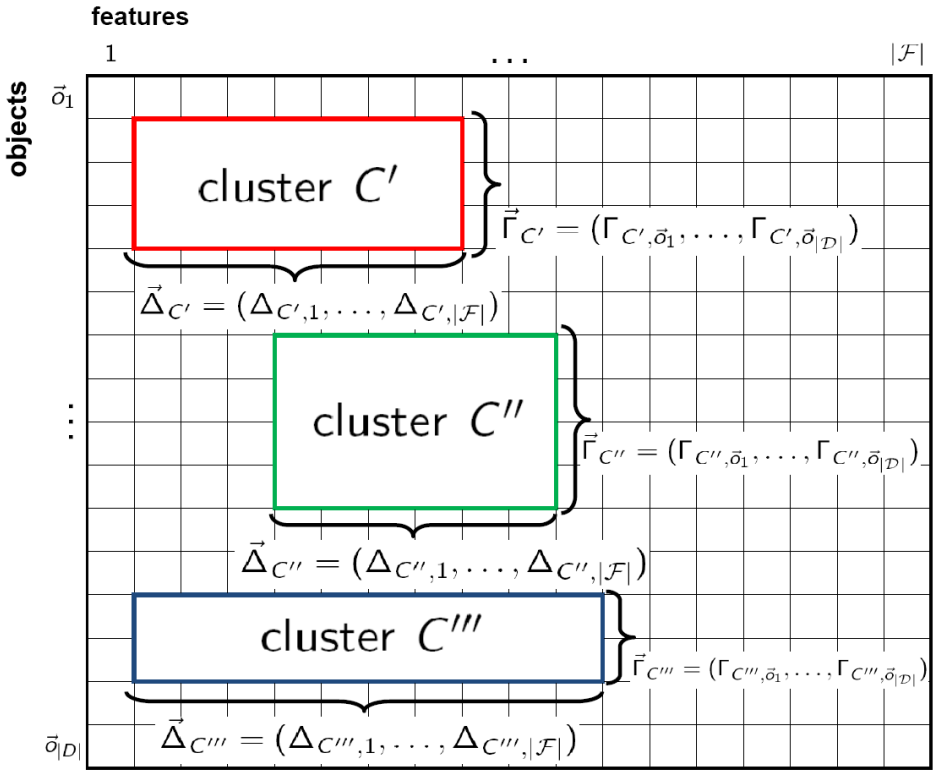
input a set $\mathcal{D}$ of data objects defined on a feature space $\mathcal{F}$

output a *subspace clustering*, i.e., a set of *subspace clusters*

A subspace cluster $C = \langle \vec{\Gamma}_C, \vec{\Delta}_C \rangle$:

- $\vec{\Gamma}_C$ is the *object-to-cluster* assignment vector ($\Gamma_{C,\vec{o}} = \Pr(\vec{o} \in C), \forall \vec{o} \in \mathcal{D}$)
- $\vec{\Delta}_C$ is the *feature-to-cluster* assignment vector ($\Delta_{C,f} = \Pr(f \in C), \forall f \in \mathcal{F}$)

$\vec{\Gamma}$ and $\vec{\Delta}$ may handle both soft and hard assignments

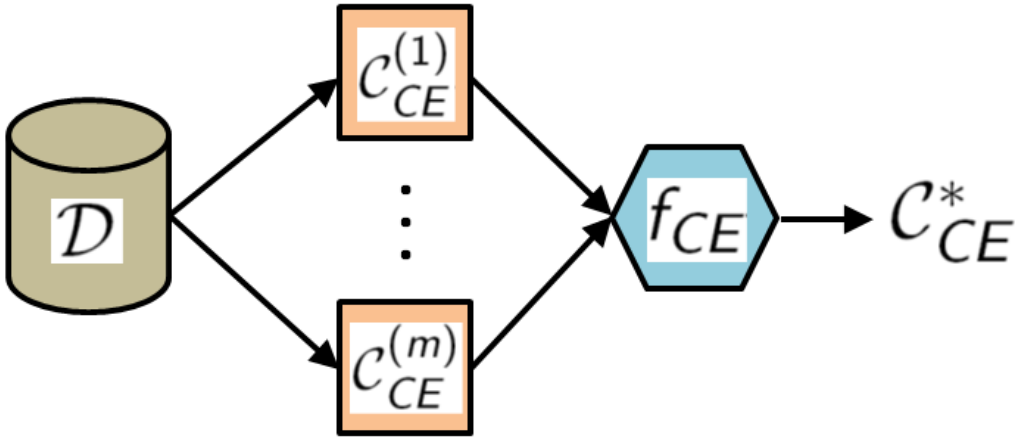


Applications: biomedical data (e.g., microarray data), recommendation systems, text categorization, ...

Clustering Ensembles (1)

Clustering Ensembles: combining multiple clustering solutions to obtain a single consensus clustering

Clustering Ensembles (2)



input an *ensemble*, i.e., a set $\mathcal{E}_{CE} = \{\mathcal{C}_{CE}^{(1)}, \ldots, \mathcal{C}_{CE}^{(m)}\}$ of clustering solutions defined over the same set $\mathcal{D}$ of data objects

output a *consensus clustering* $\mathcal{C}_{CE}^{*}$ that aggregates the information from $\mathcal{E}_{CE}$ by optimizing a *consensus function* $f_{CE}(\mathcal{E}_{CE})$

Applications: proteomics/genomics, text analysis, distributed systems, privacy preserving systems, ...

Clustering Ensembles (3)

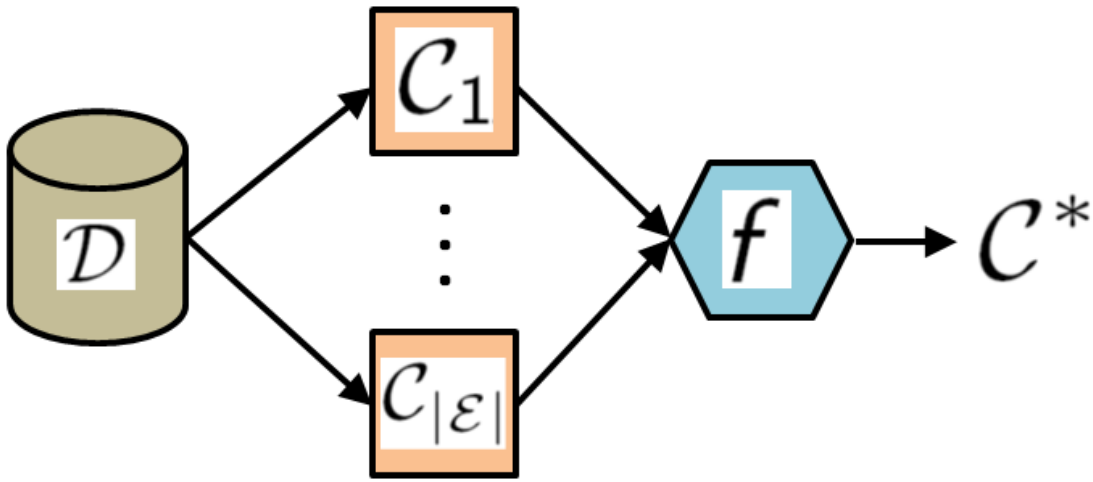
Approaches:

- *Instance-based CE* :
  direct comparison between data objects based on the *co-association* matrix
- *Cluster-based CE* :
  (1) groups clusters (to form *metaclusters*) and (2) object-to-metacluster assignments
- *Hybrid CE* :
  combination of instance-based CE and cluster-based CE

Subspace Clustering Ensembles

[Gullo et al., ICDM '09]

Goal: addressing **both** the ill-posed nature of clustering and the high dimensionality of data



input a *subspace ensemble*, i.e., a set $\mathcal{E} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{E}|}\}$ of subspace clusterings defined over the same set $\mathcal{D}$ of data objects

output a *subspace consensus clustering* $\mathcal{C}^*$ that aggregates the information from $\mathcal{E}$ by optimizing a *consensus function* $f(\mathcal{E})$

Subspace Clustering Ensembles

- Desirable requirements for the objective function:
  - independence from the original feature values of the input data
  - independence from the specific clustering ensemble algorithms used
  - ability to handle hard as well as *soft* data clustering in a subspace setting
  - ability to allow for *feature weighting* within each cluster

Early two-objective SCE formulation

Motivation:

A subspace consensus clustering C^* derived from an ensemble $\mathcal{E}$ should meet two requirements. C^* should capture the underlying clustering structure of the data:

- through the data clustering of the solutions in $\mathcal{E}$

AND

- through the assignments of features to clusters of the solutions in $\mathcal{E}$

$\Longrightarrow$ SCE can be naturally formulated considering two objectives

Subspace Clustering Ensembles: Early Methods

Two formulations have been introduced in [Gullo et al., ICDM'09]:

- Two-objective SCE $\Longrightarrow$ Pareto-based multi-objective evolutionary heuristic algorithm *MOEA-PCE*
- Single-objective SCE $\Longrightarrow$ EM-like heuristic algorithm *EM-PCE*

Major results:

- Two-objective SCE: high accuracy, expensive
- Single-objective SCE: lower accuracy, high efficiency

Early two-objective SCE formulation

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \ \{\Psi_o(\mathcal{C},\mathcal{E}),\ \Psi_f(\mathcal{C},\mathcal{E})\}$$

$$\Psi_o(\mathcal{C},\mathcal{E}) = \sum_{\hat{\mathcal{C}}\in\mathcal{E}} \overline{\psi}_o(\mathcal{C},\hat{\mathcal{C}}), \qquad \Psi_f(\mathcal{C},\mathcal{E}) = \sum_{\hat{\mathcal{C}}\in\mathcal{E}} \overline{\psi}_f(\mathcal{C},\hat{\mathcal{C}})$$

$$\overline{\psi}_o(\mathcal{C}',\mathcal{C}'') = \frac{\psi_o(\mathcal{C}',\mathcal{C}'')+\psi_o(\mathcal{C}'',\mathcal{C}')}{2} \quad \psi_o(\mathcal{C}',\mathcal{C}'') = \frac{1}{|\mathcal{C}'|}\sum_{C'\in\mathcal{C}'}\left(1-\max_{C''\in\mathcal{C}''} J(\vec{\Gamma}_{C'},\vec{\Gamma}_{C''})\right)$$

$$\overline{\psi}_f(\mathcal{C}',\mathcal{C}'') = \frac{\psi_f(\mathcal{C}',\mathcal{C}'')+\psi_f(\mathcal{C}'',\mathcal{C}')}{2} \quad \psi_f(\mathcal{C}',\mathcal{C}'') = \frac{1}{|\mathcal{C}'|}\sum_{C'\in\mathcal{C}'}\left(1-\max_{C''\in\mathcal{C}''} J(\vec{\Delta}_{C'},\vec{\Delta}_{C''})\right)$$

$$J(\vec{u},\vec{v}) = (\vec{u}\cdot\vec{v})/(\|\vec{u}\|_2^2+\|\vec{v}\|_2^2-\vec{u}\cdot\vec{v}) \in [0,1] \text{ (Tanimoto coefficient)}$$

Issues in the early two-objective SCE

Example

Ensemble:

$$\mathcal{E} = \{\hat{\mathcal{C}}\}, \text{ where } \hat{\mathcal{C}} = \{\hat{C}', \hat{C}''\} \longrightarrow \begin{cases} \hat{C}' = \langle \vec{\Gamma}', \vec{\Delta}' \rangle \\ \hat{C}'' = \langle \vec{\Gamma}'', \vec{\Delta}'' \rangle \end{cases} \quad (\vec{\Delta}' \neq \vec{\Delta}'')$$

Candidate subspace consensus clustering:

$$\mathcal{C} = \{C', C''\} \longrightarrow \begin{cases} C' = \langle \vec{\Gamma}', \vec{\Delta}'' \rangle \\ C'' = \langle \vec{\Gamma}'', \vec{\Delta}' \rangle \end{cases}$$

$\Longrightarrow$ $\mathcal{C}$ minimizes both the objectives ($\Psi_o(\mathcal{C}, \mathcal{E}) = \Psi_f(\mathcal{C}, \mathcal{E}) = 0$): $\mathcal{C}$ is mistakenly recognized as ideal!

SCE: Limitations and New Formulation

Weaknesses of the earlier SCE methods:

- Conceptual issue intrinsic to two-objective SCE: object- and feature-based cluster representations are treated independently
- Both two- and single-objective SCE do not refer to any instance-based, cluster-based, or hybrid CE approaches: poor versatility and capability of exploiting well-established research

New formulation [Gullo et al., SIGMOD'11]:

- **Goal**: Improving accuracy by solving both the above issues
- New single-objective formulation of SCE
- Two cluster-based heuristics: *CB-PCE* (more accurate) and *FCB-PCE* (more efficient)

Cluster-based SCE: formulation

Idea: avoid keeping functions Ψ_o and Ψ_f separated

$\Longrightarrow$ SCE formulation based on a single objective function:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \ \Psi_{of}(\mathcal{C}, \mathcal{E})$$

$$\Psi_{of}(\mathcal{C}, \mathcal{E}) = \sum_{\hat{\mathcal{C}} \in \mathcal{E}} \overline{\psi}_{of}(\mathcal{C}, \hat{\mathcal{C}})$$

$$\overline{\psi}_{of}(\mathcal{C}', \mathcal{C}'') = \frac{\psi_{of}(\mathcal{C}', \mathcal{C}'') + \psi_{of}(\mathcal{C}'', \mathcal{C}')}{2}$$

$$\psi_{of}(\mathcal{C}', \mathcal{C}'') = \frac{\sum_{C' \in \mathcal{C}'} \left(1 - \max_{C'' \in \mathcal{C}''} \hat{J}(\mathrm{X}_{C'}, \mathrm{X}_{C''})\right)}{|\mathcal{C}'|}$$

$$\mathrm{X}_C = \vec{\Gamma}^T \vec{\Delta} = \begin{pmatrix} \Gamma_{C,\vec{o}_1} \Delta_{C,1} & \cdots & \Gamma_{C,\vec{o}_1} \Delta_{C,|\mathcal{F}|} \\ \vdots & & \vdots \\ \Gamma_{C,\vec{o}_{|\mathcal{D}|}} \Delta_{C,1} & \cdots & \Gamma_{C,\vec{o}_{|\mathcal{D}|}} \Delta_{C,|\mathcal{F}|} \end{pmatrix}$$

$\hat{J}$ is a generalized version of the Tanimoto coefficient operating on real-valued matrices (rather than vectors)

Cluster-based SCE: heuristics

The proposed formulation is very close to standard CE formulations

$\Longrightarrow$ Key idea: developing a cluster-based approach for SCE

Why using a cluster-based approach?

1. It ensures that object- and feature-based representations are considered together
   - Objects maintain their association with the ensemble clusters (and their subspaces), and are finally assigned to meta-clusters (i.e., sets of the original clusters in the ensemble)
2. The other approaches will not work:
   - Instance-based: object- and feature-to-cluster assignments would be performed independently
   - Hybrid: same issue as instance-based SCE

The CB-PCE Algorithm

Require: a subspace ensemble $\mathcal{E}$; the number K of clusters in the output subspace consensus clustering;

Ensure: the subspace consensus clustering $\mathcal{C}^*$

1: $\Phi_{\mathcal{E}} \leftarrow \bigcup_{\hat{\mathcal{C}} \in \mathcal{E}} \hat{\mathcal{C}}$
2: $\mathrm{P} \leftarrow \textit{pairwiseClusterDistances}(\Phi_{\mathcal{E}})$
3: $\mathbf{M} \leftarrow \textit{metaclusters}(\Phi_{\mathcal{E}}, P, K)$
4: $\mathcal{C}^* \leftarrow \emptyset$
5: **for all** $\mathcal{M} \in \mathbf{M}$ **do**
6: $\quad \vec{\Gamma}^*_{\mathcal{M}} \leftarrow \textit{object-basedRepresentation}(\Phi_{\mathcal{E}}, \mathcal{M})$
7: $\quad \vec{\Delta}^*_{\mathcal{M}} \leftarrow \textit{feature-basedRepresentation}(\Phi_{\mathcal{E}}, \mathcal{M})$
8: $\quad \mathcal{C}^* \leftarrow \mathcal{C}^* \cup \{\langle \vec{\Gamma}^*_{\mathcal{M}}, \vec{\Delta}^*_{\mathcal{M}} \rangle\}$
9: **end for**

- $\Phi_{\mathcal{E}} = \bigcup_{\mathcal{C} \in \mathcal{E}} \mathcal{C}$ is the set of the clusters contained in all the solutions of the ensemble $\mathcal{E}$
- **Key points**: deriving $\vec{\Gamma}^*_{\mathcal{M}}$ and $\vec{\Delta}^*_{\mathcal{M}}$

Speeding-up CB-PCE: the FCB-PCE algorithm

Using the following (less accurate) measure for comparing clusters during the computation of the meta-clusters:

$$\hat{J}_{fast}(C', C'') = \frac{1}{2}\left(J(\vec{\Gamma}_{C'}, \vec{\Gamma}_{C''}) + J(\vec{\Delta}_{C'}, \vec{\Delta}_{C''})\right)$$

Complexity results:

- CB-PCE: $\mathcal{O}(K^2|\mathcal{E}|^2|\mathcal{D}||\mathcal{F}|)$
- FCB-PCE: $\mathcal{O}(K^2|\mathcal{E}|^2(|\mathcal{D}| + |\mathcal{F}|))$

Evaluation Methodology

- Benchmark datasets from UCI (Iris, Wine, Glass, Ecoli, Yeast, Image, Abalone, Letter) and UCR (Tracedata, ControlChart)
- Evaluation in terms of:
  - **accuracy** (*Normalized Mutual Information* (NMI))
    - external evaluation (w.r.t. the reference classification $\widetilde{\mathcal{C}}$): $\Theta(\mathcal{C}) = NMI(\mathcal{C}, \widetilde{\mathcal{C}}) - avg_{\hat{\mathcal{C}} \in \mathcal{E}} NMI(\hat{\mathcal{C}}, \widetilde{\mathcal{C}})$
    - internal evaluation (w.r.t. the ensemble solutions): $\Upsilon(\mathcal{C}) = avg_{\hat{\mathcal{C}} \in \mathcal{E}} NMI(\mathcal{C}, \hat{\mathcal{C}}) / avg_{\hat{\mathcal{C}}', \hat{\mathcal{C}}'' \in \mathcal{E}} NMI(\hat{\mathcal{C}}', \hat{\mathcal{C}}'')$
  - **efficiency**
- Competitors: earlier two-objective PCE (MOEA-PCE) and single-objective PCE (EM-PCE)

Datasets

dataset	*# objects*	*# attributes*	*# classes*
Iris	150	4	3
Wine	178	13	3
Glass	214	10	6
Ecoli	327	7	5
Yeast	1,484	8	10
Image	2,310	19	7
Abalone	4,124	7	17
Letter	7,648	16	10
Tracedata	200	275	4
ControlChart	600	60	6

Accuracy Results: external evaluation

	Θ_{of}				Θ_{o}				Θ_{f}			
	MOEA PCE	*EM PCE*	*CB PCE*	*FCB PCE*	*MOEA PCE*	*EM PCE*	*CB PCE*	*FCB PCE*	*MOEA PCE*	*EM PCE*	*CB PCE*	*FCB PCE*
min	+.049	+.019	+.092	+.095	+.032	+.011	+.027	+.051	-.007	-.095	+.001	+.009
max	+.164	+.204	+.345	+.276	+.319	+.228	+.309	+.297	+.233	+.416	+.287	+.283
avg	+.115	+.110	+.185	+.171	+.142	+.116	+.185	+.178	+.093	+.093	+.123	+.122

- Evaluation in terms of object-based representation only (Θ_o), feature-based representation only (Θ_f), object- and feature-based representations altogether (Θ_{of})
- The proposed CB-PCE and FCB-PCE were on average more accurate than MOEA-PCE, up to 0.070 (CB-PCE) and 0.056 (FCB-PCE)
- The difference was more evident w.r.t. EM-PCE: gains up to 0.075 (CB-PCE) and 0.062 (FCB-PCE)
- CB-PCE generally better than FCB-PCE, as expected

Accuracy Results: internal evaluation

	Υ_{of}				Υ_{o}				Υ_{f}			
	MOEA PCE	*EM PCE*	*CB PCE*	*FCB PCE*	*MOEA PCE*	*EM PCE*	*CB PCE*	*FCB PCE*	*MOEA PCE*	*EM PCE*	*CB PCE*	*FCB PCE*
min	.993	.851	.98	.989	1.025	.971	1.027	1.028	.949	.577	.980	.977
max	1.170	1.207	1.305	1.308	1.367	1.501	1.903	1.903	1.085	1.021	1.234	1.234
avg	1.048	.996	1.110	1.108	1.152	1.141	1.318	1.316	.985	.898	1.049	1.030

- Evaluation in terms of object-based representation only (Υ_o), feature-based representation only (Υ_f), object- and feature-based representations altogether (Υ_{of})
- The overall results substantially confirmed those encountered in the external evaluation
- Gains up to 0.166 (CB-PCE w.r.t. MOEA-PCE), 0.177 (CB-PCE w.r.t. EM-PCE), 0.164 (FCB-PCE w.r.t. MOEA-PCE), 0.175 (FCB-PCE w.r.t. EM-PCE)
- Difference between CB-PCE and FCB-PCE less evident

Efficiency Results (msecs)

dataset	*MOEA PCE*	*EM PCE*	*CB PCE*	*FCB PCE*
Iris	17,223	55	13,235	906
Wine	21,098	184	50,672	993
Glass	61,700	281	110,583	3,847
Ecoli	94,762	488	137,270	4,911
Yeast	1,310,263	1,477	2,218,128	56,704
Segmentation	1,250,732	11,465	6,692,111	47,095
Abalone	13,245,313	34,000	19,870,218	527,406
Letter	7,765,750	54,641	26,934,327	271,064
Trace	86,179	4,880	2,589,899	3,731
ControlChart	291,856	2,313	3,383,936	12,439

- FCB-PCE always faster than CB-PCE and MOEA-PCE
- FCB-PCE generally slower than EM-PCE, even if the difference decreases as $|\mathcal{D}| + |\mathcal{F}|$ (resp. K) increases (resp. decreases)

Conclusions

- Subspace Clustering Ensembles provide a unified framework to address both the curse of dimensionality and the ill-posed nature of clustering
- Cluster-based SCE approach: single-objective formulation
  - it solves the conceptual issues of two-objective SCE
- Future Work: *Alternative* Subspace Clustering Ensembles

References:

- F. Gullo, C. Domeniconi, and A. Tagarelli, *Advancing Data Clustering via Projective Clustering Ensembles*, SIGMOD 2011.
- F. Gullo, C. Domeniconi, and A. Tagarelli, *Enhancing Single-Objective Projective Clustering Ensembles*, ICDM 2010.
- F. Gullo, C. Domeniconi, and A. Tagarelli, *Projective Clustering Ensembles*, ICDM 2009.

Additional pointers:

- P. Wang, K. B. Laskey, C. Domeniconi, and M. I. Jordan, *Nonparametric Bayesian Co-clustering Ensembles*, SDM 2011.
- P. Wang, C. Domeniconi, H. Rangwala, and K. B. Laskey, *Feature Enriched Nonparametric Bayesian Co-clustering*, PAKDD 2012.

Thanks!