Advancing Data Clustering via Projective Clustering Ensembles

F. Gullo * C. Domeniconi † A. Tagarelli *

* Dept. of Electronics, Computer and Systems Science
University of Calabria, Italy

† Dept. of Computer Science
George Mason University, Virginia (USA)

The 2011 ACM SIGMOD International Conference on Management of Data (SIGMOD'11)
June 12-16, 2011
Athens, Greece

Data Clustering: challenges and advanced approaches

Data Clustering challenges in real-life domains:

1. high-dimensionality, sparsity (in data representation)
2. multiple sets of clusterings

Advances in data clustering:

- Projective Clustering (handles issue 1)
- Clustering Ensembles (handles issue 2)
- Projective Clustering Ensembles (handles both issue 1 and 2)

Projective Clustering (1)

Projective 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

Projective Clustering (2)

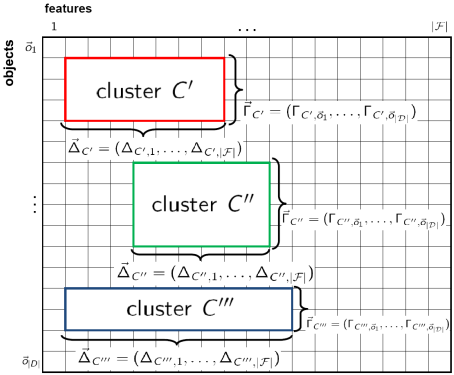
input a set $\mathcal{D}$ of data objects defined on a feature space $\mathcal{F}$

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

A projective 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



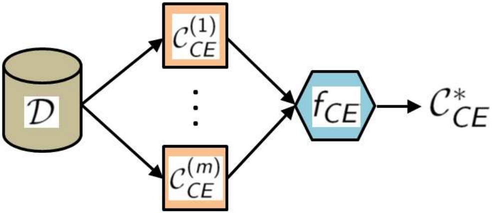
Applications: biomedical data (e.g., microarray data), recommender systems, text categorization, ...

Clustering Ensembles (1)

Clustering Ensembles: combining multiple clustering solutions to present results in the form of a unique solution

- To group objects in different views of the data
- Multiple sets of clusters providing more insights than only one solution

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)

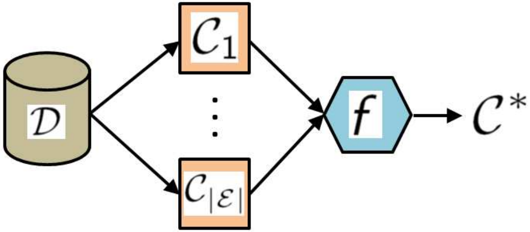
Approaches:

- *Instance-based CE* :
  direct comparison between data objects based on the *co-occurrence* matrix
- *Cluster-based CE* :
  two main steps, i.e., to cluster clusters (to form *metaclusters*) and object-to-metacluster assignment
- *Hybrid CE* :
  combination of instance-based CE and cluster-based CE

Projective Clustering Ensembles

[Gullo et al., ICDM '09]

Goal: addressing both the multi-view nature of clustering and the high-dimensionality in data



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

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

Projective Clustering Ensembles: Early Methods

Two formulations of PCE are proposed in [Gullo et al., ICDM '09]:

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

Major results:

- Two-objective PCE: high accuracy, poor efficiency
- Single-objective PCE: poor accuracy, high efficiency

Goal

Weaknesses of the earlier PCE methods:

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

Goal:

- Improving accuracy by solving both the above issues

Contributions:

- New single-objective formulation of PCE
- Two cluster-based heuristics: *CB-PCE* (more accurate) and *FCB-PCE* (more efficient)

Early two-objective PCE formulation

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

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

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

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

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

Issues in the early two-objective PCE

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 projective 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 of the earlier two-objective PCE formulation ($\Psi_o(\mathcal{C}, \mathcal{E}) = \Psi_f(\mathcal{C}, \mathcal{E}) = 0$): it is mistakenly recognized as ideal!

Cluster-based PCE: formulation

Idea: avoiding to keep functions Ψ_o and Ψ_f separated

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

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

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

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

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

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

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

Cluster-based PCE: heuristics

The proposed formulation is very close to standard CE formulations

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

Why using a cluster-based approach?

1. It ensures that object- and feature-based representations will be kept 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 separately from each other
   - Hybrid: same issue as instance-based PCE (hybrid PCE is a combination of instance-based PCE and cluster-based PCE)

The CB-PCE Algorithm

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

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

$\Phi_{\mathcal{E}} \leftarrow \bigcup_{\hat{\mathcal{C}} \in \mathcal{E}} \hat{\mathcal{C}}$

$\mathrm{P} \leftarrow$ *pairwiseClusterDistances*$(\Phi_{\mathcal{E}})$

$\mathbf{M} \leftarrow$ *metaclusters*$(\Phi_{\mathcal{E}}, P, K)$

$\mathcal{C}^* \leftarrow \emptyset$

for all $\mathcal{M} \in \mathbf{M}$ **do**

  $\vec{\Gamma}^*_{\mathcal{M}} \leftarrow$ *object-basedRepresentation*$(\Phi_{\mathcal{E}}, \mathcal{M})$

  $\vec{\Delta}^*_{\mathcal{M}} \leftarrow$ *feature-basedRepresentation*$(\Phi_{\mathcal{E}}, \mathcal{M})$

  $\mathcal{C}^* \leftarrow \mathcal{C}^* \cup \{\langle \vec{\Gamma}^*_{\mathcal{M}}, \vec{\Delta}^*_{\mathcal{M}} \rangle\}$

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 $\vec{\Delta}^*_{\mathcal{M}}$

The CB-PCE Algorithm: deriving $\vec{\Gamma}^*_{\mathcal{M}}$

Solving the optimization problem $P_{\vec{\Gamma}^*}$:

$$\{\vec{\Gamma}^*_{\mathcal{M}} | \mathcal{M} \in \mathbf{M}\} = \underset{\{\vec{\Gamma}_{\mathcal{M}} | \mathcal{M} \in \mathbf{M}\}}{\text{argmin}} \; Q$$

$$s.t.$$

$$\sum_{\mathcal{M} \in \mathbf{M}} \Gamma_{\mathcal{M},\vec{o}} = 1, \quad \forall \vec{o} \in \mathcal{D}$$

$$\Gamma_{\mathcal{M},\vec{o}} \geq 0, \quad \forall \mathcal{M} \in \mathbf{M}, \; \forall \vec{o} \in \mathcal{D}$$

where

$$Q = \sum_{\mathcal{M} \in \mathbf{M}} \sum_{\vec{o} \in \mathcal{D}} \Gamma^{\alpha}_{\mathcal{M},\vec{o}} \, A_{\mathcal{M},\vec{o}}\,, \quad A_{\mathcal{M},\vec{o}} = \frac{1}{|\mathcal{M}|} \sum_{M \in \mathcal{M}} 1 - \Gamma_{M,\vec{o}}$$

Theorem

The optimal solution of the problem $P_{\vec{\Gamma}^*}$ is given by ($\forall \mathcal{M}$, $\forall \vec{o}$):

$$\Gamma^*_{\mathcal{M},\vec{o}} = \left[\sum_{\mathcal{M}' \in \mathbf{M}} \left(\frac{A_{\mathcal{M},\vec{o}}}{A_{\mathcal{M}',\vec{o}}} \right)^{\frac{1}{\alpha - 1}} \right]^{-1}$$

The CB-PCE Algorithm: deriving $\vec{\Delta}^*_{\mathcal{M}}$

Solving the optimization problem $P_{\vec{\Delta}^*}$:

$$\{\vec{\Delta}^*_{\mathcal{M}} | \mathcal{M} \in \mathbf{M}\} = \underset{\{\vec{\Delta}_{\mathcal{M}} | \mathcal{M} \in \mathbf{M}\}}{\arg\min} \sum_{\mathcal{M} \in \mathbf{M}} \sum_{f \in \mathcal{F}} \Delta^{\beta}_{\mathcal{M},f} \; B_{\mathcal{M},f}$$

$$s.t.$$

$$\sum_{f \in \mathcal{F}} \Delta_{\mathcal{M},f} = 1, \quad \forall \mathcal{M} \in \mathbf{M}$$

$$\Delta_{\mathcal{M},f} \geq 0, \quad \forall \mathcal{M} \in \mathbf{M}, \; \forall f \in \mathcal{F}$$

where

$$B_{\mathcal{M},f} = |\mathcal{M}|^{-1} \sum_{M \in \mathcal{M}} 1 - \Delta_{M,f}$$

Theorem

The optimal solution of the problem $P_{\vec{\Delta}^*}$ is given by($\forall \mathcal{M}, \forall f$):

$$\Delta^*_{\mathcal{M},f} = \left[\sum_{f' \in \mathcal{F}} \left(\frac{B_{\mathcal{M},f}}{B_{\mathcal{M},f'}} \right)^{\frac{1}{\beta - 1}} \right]^{-1}$$

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

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

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

Complexity results given a set of objects ($\mathcal{D}$), a set of features ($\mathcal{F}$), an ensemble ($\mathcal{E}$), and the number of output clusters (K)

- Proposed methods
  - CB-PCE: $\mathcal{O}(K^2|\mathcal{E}|^2|\mathcal{D}||\mathcal{F}|)$
  - FCB-PCE: $\mathcal{O}(K^2|\mathcal{E}|^2(|\mathcal{D}| + |\mathcal{F}|))$
- Earlier methods
  - MOEA-PCE (two-objective): $\mathcal{O}(ItK^2|\mathcal{E}|(|\mathcal{D}| + |\mathcal{F}|))$
  - EM-PCE (single-objective): $\mathcal{O}(K|\mathcal{E}||\mathcal{D}||\mathcal{F}|)$

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)

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

- Advances on the emerging Projective Clustering Ensembles (PCE) problem have been provided, by improving accuracy of the earlier two-objective PCE formulation
  - The conceptual issues at the basis of two-objective PCE have been solved by proposing an alternative single-objective formulation of PCE
  - Two heuristics (CB-PCE and FCB-PCE) have been proposed
- The claim concerning the improvement of accuracy of two-objective PCE has been confirmed by experimental evidence

Thanks!

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