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 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: 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: 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 $\mathcal{E}_{CE} = \{C_{CE}^{(1)}, \ldots, C_{CE}^{(m)}\}$ of clustering solutions defined over the same set $\mathcal{D}$ of data objects

**output** a consensus clustering $C_{CE}^*$ 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, ...
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

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

**input** a *projective ensemble*, i.e., a set $\mathcal{E} = \{C_1, \ldots, C_{|\mathcal{E}|}\}$ of projective clusterings defined over the same set $\mathcal{D}$ of data objects

**output** a *projective consensus clustering* $C^*$ that aggregates the information from $\mathcal{E}$ by optimizing a consensus function $f(\mathcal{E})$
Two formulations of PCE are proposed in [Gullo et al., ICDM ’09]:

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

**Major results:**

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

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

\[ \Psi_o(C, \mathcal{E}) = \sum_{\hat{C} \in \mathcal{E}} \overline{\psi}_o(C, \hat{C}), \quad \Psi_f(C, \mathcal{E}) = \sum_{\hat{C} \in \mathcal{E}} \overline{\psi}_f(C, \hat{C}) \]

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

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

\[ J(\bar{u}, \bar{v}) = (\bar{u} \cdot \bar{v}) / (\|\bar{u}\|_2^2 + \|\bar{v}\|_2^2 - \bar{u} \cdot \bar{v}) \in [0, 1] \text{ (Tanimoto coefficient)} \]
Issues in the early two-objective PCE

Example

Ensemble:

\[ \mathcal{E} = \{\hat{C}\}, \text{ where } \hat{C} = \{\hat{C}', \hat{C}''\} \rightarrow \begin{cases} \hat{C}' = \langle \vec{\Gamma}', \vec{\Delta}' \rangle \\ \hat{C}'' = \langle \vec{\Gamma}'', \vec{\Delta}'' \rangle \end{cases} (\vec{\Delta}' \neq \vec{\Delta}'') \]

Candidate projective consensus clustering:

\[ \mathcal{C} = \{C', C''\} \rightarrow \begin{cases} C' = \langle \vec{\Gamma}', \vec{\Delta}'' \rangle \\ C'' = \langle \vec{\Gamma}'', \vec{\Delta}' \rangle \end{cases} \]

\[ \Rightarrow \mathcal{C} \text{ minimizes both the objectives of the earlier two-objective PCE formulation } (\Psi_o(C, \mathcal{E}) = \Psi_f(C, \mathcal{E}) = 0): \text{ it is mistakenly recognized as ideal!} \]
Cluster-based PCE: formulation

**Idea:** avoiding to keep functions $\Psi_o$ and $\Psi_f$ separated

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

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

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

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

$$\psi_{of}(C', C'') = \sum_{C'' \in C''} \left( 1 - \max_{C'' \in C''} \tilde{J}(X_{C'}, X_{C''}) \right)$$

$$X_C = \tilde{\Gamma}^T \tilde{\Delta} = \begin{pmatrix} \Gamma_{C,\bar{o}_1} \Delta_{C,1} & \ldots & \Gamma_{C,\bar{o}_1} \Delta_{C,|\mathcal{F}|} \\ \vdots & \ddots & \vdots \\ \Gamma_{C,\bar{o}_{|\mathcal{D}|}} \Delta_{C,1} & \ldots & \Gamma_{C,\bar{o}_{|\mathcal{D}|}} \Delta_{C,|\mathcal{F}|} \end{pmatrix}$$

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

F. Gullo, C. Domeniconi, A. Tagarelli

Advancing Data Clustering via Projective Clustering Ensembles
Cluster-based PCE: heuristics

The proposed formulation is very close to standard CE formulations

⇒ 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 \( C^* \)

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

- \( \Phi_\mathcal{E} = \bigcup_{C \in \mathcal{E}} C \) is the set of the clusters contained in all the solutions of the ensemble \( \mathcal{E} \)

- Key points: deriving \( \hat{\Gamma}_{\mathcal{M}}^* \) and \( \hat{\Delta}_{\mathcal{M}}^* \)
The CB-PCE Algorithm: deriving $\vec{\Gamma}^*_M$

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

$$\{\vec{\Gamma}^*_M | M \in M\} = \arg\min_{\{\vec{\Gamma}_M | M \in M\}} Q$$

subject to:

$$\sum_{M \in M} \Gamma_{M, \bar{o}} = 1, \quad \forall \bar{o} \in D$$

$$\Gamma_{M, \bar{o}} \geq 0, \quad \forall M \in M, \quad \forall \bar{o} \in D$$

where

$$Q = \sum_{M \in M} \sum_{\bar{o} \in D} \Gamma^\alpha_{M, \bar{o}} A_{M, \bar{o}}, \quad A_{M, \bar{o}} = \frac{1}{|M|} \sum_{M' \in M} 1 - \Gamma_{M, \bar{o}}$$

**Theorem**

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

$$\Gamma^*_M, \bar{o} = \left[ \sum_{M' \in M} \left( \frac{A_{M, \bar{o}}}{A_{M', \bar{o}}} \right)^{\frac{1}{\alpha - 1}} \right]^{-1}$$
The CB-PCE Algorithm: deriving $\vec{\Delta}^*_M$

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

$$\{\vec{\Delta}^*_M | M \in \mathcal{M}\} = \arg\min_{\{\vec{\Delta}_M | M \in \mathcal{M}\}} \sum_{M \in \mathcal{M}} \sum_{f \in \mathcal{F}} \Delta^\beta_{M,f} \ B_{M,f}$$

s.t.

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

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

where

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

**Theorem**

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

$$\Delta^*_M,f = \left[ \sum_{f' \in \mathcal{F}} \left( \frac{B_{M,f'}}{B_{M,f}} \right)^{\frac{1}{\beta-1}} \right]^{-1}$$
Speeding-up CB-PCE: the FCB-PCE algorithm

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

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

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

- Proposed methods
  - CB-PCE: $O(K^2|E|^2|D||F|)$
  - FCB-PCE: $O(K^2|E|^2(|D| + |F|))$

- Earlier methods
  - MOEA-PCE (two-objective): $O(ltK^2|E|(|D| + |F|))$
  - EM-PCE (single-objective): $O(K|E||D||F|)$
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 $\tilde{C}$):
      $$\Theta(C) = \text{NMI}(C, \tilde{C}) - \text{avg}_{\hat{C} \in E} \text{NMI}(\hat{C}, \tilde{C})$$
    - internal evaluation (w.r.t. the ensemble solutions):
      $$\Upsilon(C) = \text{avg}_{\hat{C} \in E} \text{NMI}(C, \hat{C}) / \text{avg}_{\hat{C}', \hat{C}'' \in E} \text{NMI}(\hat{C}', \hat{C}'')$$

- **efficiency**

- Competitors: earlier two-objective PCE (MOEA-PCE) and single-objective PCE (EM-PCE)
### Accuracy Results: external evaluation

<table>
<thead>
<tr>
<th></th>
<th>$\Theta_{of}$</th>
<th></th>
<th>$\Theta_o$</th>
<th></th>
<th>$\Theta_f$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MOEA</td>
<td>+.049</td>
<td>EM</td>
<td>+.019</td>
<td>CB</td>
<td>+.092</td>
<td>FCB</td>
</tr>
<tr>
<td>PCE</td>
<td></td>
<td>PCE</td>
<td>PCE</td>
<td>PCE</td>
<td>PCE</td>
<td>FCB-PCE</td>
</tr>
<tr>
<td>min</td>
<td>+.032</td>
<td>+.011</td>
<td>+.027</td>
<td>+.051</td>
<td>-0.007</td>
<td>-0.095</td>
</tr>
<tr>
<td>max</td>
<td>+.164</td>
<td>+.204</td>
<td>+.345</td>
<td>+.276</td>
<td>+.319</td>
<td>+.228</td>
</tr>
<tr>
<td>avg</td>
<td>+.115</td>
<td>+.110</td>
<td>+.185</td>
<td>+.171</td>
<td>+.142</td>
<td>+.116</td>
</tr>
</tbody>
</table>

- Evaluation in terms of **object-based representation only** ($\Theta_o$), **feature-based representation only** ($\Theta_f$), **object- and feature-based representations altogether** ($\Theta_{of}$)

- 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

<table>
<thead>
<tr>
<th></th>
<th>$\gamma_{of}$</th>
<th>$\gamma_{o}$</th>
<th>$\gamma_{f}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOEA PCE</td>
<td>EM PCE</td>
<td>CB PCE</td>
<td>FCB PCE</td>
</tr>
<tr>
<td>min</td>
<td>.993</td>
<td>.851</td>
<td>.98</td>
</tr>
<tr>
<td>max</td>
<td>1.170</td>
<td>1.207</td>
<td>1.305</td>
</tr>
<tr>
<td>avg</td>
<td>1.048</td>
<td>.996</td>
<td>1.110</td>
</tr>
</tbody>
</table>

- Evaluation in terms of **object-based representation only** ($\gamma_{o}$), **feature-based representation only** ($\gamma_{f}$), **object- and feature-based representations altogether** ($\gamma_{of}$)

- 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)

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

- FCB-PCE always faster than CB-PCE and MOEA-PCE
- FCB-PCE generally slower than EM-PCE, even if the difference decreases as $|D| + |F|$ (resp. $K$) increases (resp. decreases)
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!
<table>
<thead>
<tr>
<th>dataset</th>
<th># objects</th>
<th># attributes</th>
<th># classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>150</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Glass</td>
<td>214</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Ecoli</td>
<td>327</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Yeast</td>
<td>1,484</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Image</td>
<td>2,310</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>Abalone</td>
<td>4,124</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>Letter</td>
<td>7,648</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>Tracedata</td>
<td>200</td>
<td>275</td>
<td>4</td>
</tr>
<tr>
<td>ControlChart</td>
<td>600</td>
<td>60</td>
<td>6</td>
</tr>
</tbody>
</table>