Enhancing Single-Objective Projective Clustering Ensembles

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

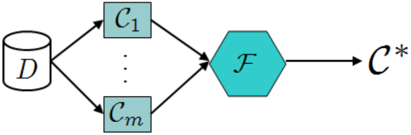
* Dept. of Electronics, Computer and Systems Science
University of Calabria, Italy

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

10th IEEE International Conference on Data Mining (ICDM)
December 14-17, 2010
Sydney, Australia

Projective Clustering Ensembles (PCE)

[Gullo et Al., ICDM '09]

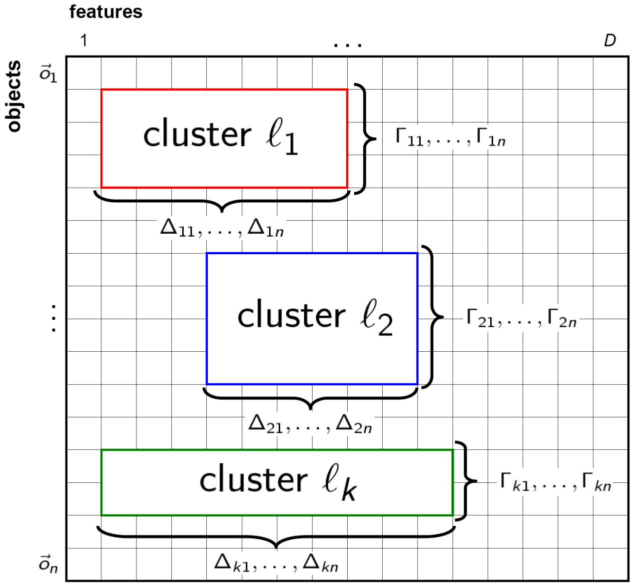


input a *projective ensemble* $\mathcal{E}$, i.e., a set of *projective clustering solutions*

output a *projective consensus clustering* $\mathcal{C}^*$ computed according to a *consensus function* $\mathcal{F}$

A *projective clustering solution* $\mathcal{C}$ is a triple $\langle \mathcal{L}, \Gamma, \Delta \rangle$:

- $\mathcal{L}$: cluster labels $\{\ell_1, \ldots, \ell_K\}$
- Γ: *object-based representation* (Γ_{kn} gives the probability $\Pr(\ell_k|\vec{o}_n)$ that object $\vec{o}_n$ belongs to cluster ℓ_k, $\forall \vec{o}_n$, $\forall \ell_k$)
- Δ: *feature-based representation* (Δ_{kd} gives the probability $\Pr(d|\ell_k)$ that the d-th feature is a relevant dimension for cluster ℓ_k, $\forall d$, $\forall \ell_k$)


features
1
...
D
objects
$\vec{o}_1$
$\vdots$
$\vec{o}_n$
cluster ℓ_1
$\Gamma_{11}, \ldots, \Gamma_{1n}$
$\Delta_{11}, \ldots, \Delta_{1n}$
cluster ℓ_2
$\Gamma_{21}, \ldots, \Gamma_{2n}$
$\Delta_{21}, \ldots, \Delta_{2n}$
cluster ℓ_k
$\Gamma_{k1}, \ldots, \Gamma_{kn}$
$\Delta_{k1}, \ldots, \Delta_{kn}$

Projective Clustering Ensembles: Early Methods

Two formulations of PCE are described 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

Goal

Improving accuracy of single-objective PCE, while maintaining the advantages in terms of efficiency w.r.t. the two-objective counterpart

Single-Objective PCE: Early Formulation

- **Objective function**:

$$Q(\hat{\mathcal{C}}, \mathcal{E}) = \sum_{k=1}^{K} \sum_{n=1}^{N} \hat{\Gamma}_{kn}^{\alpha} \sum_{h=1}^{H} \gamma_{hn} \sum_{d=1}^{D} \left(\hat{\Delta}_{kd} - \delta_{hd}\right)^2$$

- **Solution**:

$$\Gamma_{kn}^{*} = \left[\sum_{k'=1}^{K} \left(\frac{X_{kn}}{X_{k'n}} \right)^{\frac{1}{\alpha-1}} \right]^{-1} \quad \text{and} \quad \Delta_{kd}^{*} = \frac{Z_{kd}}{Y_k}$$

where

$$X_{kn} = \sum_{h=1}^{H} \gamma_{hn} \sum_{d=1}^{D} \left(\hat{\Delta}_{kd} - \delta_{hd}\right)^2 \qquad Y_k = \sum_{n=1}^{N} \hat{\Gamma}_{kn}^{\alpha} \sum_{h=1}^{H} \gamma_{hn} = M \sum_{n=1}^{N} \hat{\Gamma}_{kn}^{\alpha}$$

$$Z_{kd} = \sum_{n=1}^{N} \hat{\Gamma}_{kn}^{\alpha} \sum_{h=1}^{H} \gamma_{hn} \ \delta_{hd}$$

Single-Objective PCE: Major Issue

The single-objective PCE objective function

$$Q(\hat{\mathcal{C}}, \mathcal{E}) = \sum_{k=1}^{K} \sum_{n=1}^{N} \hat{\Gamma}_{kn}^{\alpha} \sum_{h=1}^{H} \gamma_{hn} \sum_{d=1}^{D} \left(\hat{\Delta}_{kd} - \delta_{hd}\right)^2$$

estimates the distance between any pair of data objects only considering their feature-based representation given by:

$$\sum_{h=1}^{H} \gamma_{hn} \sum_{d=1}^{D} \left(\hat{\Delta}_{kd} - \delta_{hd}\right)^2$$

$\Longrightarrow$ **objects belonging to distinct clusters that share similar feature-based representation may be wrongly recognized as similar by Q!**

Enhancing Single-Objective PCE: Proposal

Two new heuristics

1. E-EM-PCE
2. E-2S-PCE

First Proposal: E-EM-PCE

Idea

"Completing" function Q by adding a term for computing dissimilarity between objects according to their object-based representation too

Considering the events:

- $A_{nn'}$: "$\vec{o}_n$ *and* $\vec{o}_{n'}$ *are clustered together in* the ensemble $\mathcal{E}$"
- $B_{n'}$: "$\vec{o}_{n'}$ *belongs to* $\hat{\ell}_k$"

the term to be added to function Q is:

$$X'_{kn} = \sum_{\forall n' \neq n} (1 - \Pr(A_{nn'}) \; \Pr(B_{n'})) = \sum_{\forall n' \neq n} 1 - \frac{\hat{\Gamma}_{kn'}}{M} \sum_{h=1}^{H} \gamma_{hn} \; \gamma_{hn'}$$

Second Proposal: E-2S-PCE (1)

Motivation

In E-EM-PCE, the object-to-cluster assignments of the output consensus clustering still depend on the feature-based representation of data objects

Idea

Computing object-to-cluster (Γ^*) and feature-to-cluster (Δ^*) assignments of the consensus clustering sequentially

Second Proposal: E-2S-PCE (2)

- First step (computing Γ^*): resorting to standard clustering ensembles by exploiting a *co-occurrence* matrix properly re-defined
- Second step (computing Δ^* as a kind of centroid):

$$\Delta^* = \arg\min_{\hat{\Delta}} \sum_{k=1}^{K} \sum_{n=1}^{N} \Gamma^*_{kn} \sum_{h=1}^{H} \gamma_{hn} \sum_{d=1}^{D} \left(\hat{\Delta}_{kd} - \delta_{hd}\right)^2$$

$$\Longrightarrow \quad \Delta^*_{kd} = \left(M \sum_{n=1}^{N} \Gamma^*_{kn} \right)^{-1} \sum_{n=1}^{N} \Gamma^*_{kn} \sum_{h=1}^{H} \gamma_{hn}\, \delta_{hd}, \ \forall k, \forall d$$

Evaluation Methodology

- Benchmark datasets from UCI (Iris, Wine, Glass, Ecoli, Yeast, Image, Abalone, Letter) and UCR (Tracedata, ControlChart)
- Evaluation in terms of:
  - **accuracy** (w.r.t. reference classifications according to *Normalized Mutual Information* (NMI))
  - **efficiency**
- Competitors: earlier two-objective PCE (MOEA-PCE) and single-objective PCE (EM-PCE)

Accuracy Results

	NMI_{of}				NMI_o				NMI_f			
	MOEA PCE	*EM PCE*	*E-EM PCE*	*E-2S PCE*	*MOEA PCE*	*EM PCE*	*E-EM PCE*	*E-2S PCE*	*MOEA PCE*	*EM PCE*	*E-EM PCE*	*E-2S PCE*
min	+.049	+.019	+.036	+.057	+.032	+.011	+.033	+.027	-.007	-.095	-.092	-.017
max	+.164	+.204	+.209	+.220	+.319	+.228	+.252	+.294	+.233	+.416	+.416	+.416
avg	+.115	+.110	+.129	+.137	+.142	+.116	+.129	+.138	+.093	+.093	+.092	+.120

- Evaluation in terms of object-based representation only (NMI_o), feature-based representation only (NMI_f), object- and feature-based representations altogether (NMI_{of})
- The proposed E-EM-PCE and E-2S-PCE were on average more accurate than EM-PCE, up to 0.019 (E-EM-PCE) and 0.027 (E-2S-PCE)
- Gap from MOEA-PCE drastically reduced, even achieving gains up to 0.014 (E-EM-PCE) and 0.027 (E-2S-PCE)
- E-2S-PCE generally better than E-EM-PCE

Efficiency Results

data	*MOEA-PCE*	*EM-PCE*	*E-EM-PCE*	*E-2S-PCE*
Iris	17,223	55	250	353
Wine	21,098	184	477	522
Glass	61,700	281	1,257	939
Ecoli	94,762	488	2,354	2,291
Yeast	1,310,263	1,477	5,459	80,158
Segm.	1,250,732	11,465	37,048	154,720
Abal.	13,245,313	34,000	312,485	1,875,968
Letter	7,765,750	54,641	451,453	2,057,187
Trace	86,179	4,880	4,138	2,285
Contr.	291,856	2,313	2,900	9,874

- The proposed E-EM-PCE and E-2S-PCE maintained a large efficiency gain w.r.t. MOEA-PCE (up to 2 orders of magnitude)
- The advantage of EM-PCE w.r.t. E-EM-PCE and E-2S-PCE was noticeable only when the ratios K/D and N/D increase

Conclusions

- Improving accuracy of the single-objective formulation of the newly emerged Projective Clustering Ensembles (PCE) problem, while maintaining high the efficiency:
  - Adjusting early objective function $\Longrightarrow$ E-EM-PCE heuristic
  - Performing two sequential steps for object- and feature-to-cluster assignments $\Longrightarrow$ E-2S-PCE heuristic
- Both accuracy and efficiency claims 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