A Generative Dyadic Model for Evidence Accumulation Clustering

André Lourenço*, Ana Fred†, and Mário Figueiredo†

* † Instituto Superior Técnico

* Instituto Superior de Engenharia de Lisboa    * † Instituto de Telecomunicações

Lisboa, Portugal

First International Workshop on Similarity-Based Pattern Analysis and Recognition
Outline

1. Introduction
   - Clustering Ensembles and Evidence Accumulation
   - Dyadic Data Analysis

2. Probabilistic Ensemble Clustering Algorithm (PEnCA)
   - Generative Mixture Model
   - Estimation

3. Experimental Results
   - Experimental Setup
   - Examples and Discussion

4. Conclusions and Future Work
Clustering Ensembles

- Notation: \( \mathcal{X} = \{1, \ldots, N\} \): set of \( N \) objects to be clustered;

\[ \mathcal{E} = \{\mathcal{P}^1, \ldots, \mathcal{P}^M\} : \text{ensemble of clusterings}, \]

\[ \mathcal{P}^i = \{C^i_1, \ldots, C^i_{K_i}\} : \text{clustering with } K_i \text{ clusters} \]

\[ C^i_j \subseteq \mathcal{X}, \quad \bigcup_{j=1}^{K_i} C^i_j = \mathcal{X}, \quad j \neq l \Rightarrow C^i_j \cap C^i_l = \emptyset \]

- Different clustering algorithms: different pattern organization.

- Clustering combination methods aim at “better” / “more robust” partitioning by combining an ensemble of clusterings.
Evidence Accumulation Clustering (EAC)

- EAC: [Fred and Jain, 2001, 2005]
  - clustering ensemble method
  - each clustering provides evidence of pair-wise relationships

- Major Steps:
  1. construction of the clustering ensemble;
  2. evidence accumulation of pair-wise associations;
  3. extraction of the final consensus partition.

- The combination step (ii) yields the co-occurrence matrix $C$:

  \[ C_{i,j} = \text{“number of times objects } i \text{ and } j \text{ co-occurred”} \]
Dyadic Data Analysis

- Dyadic data: each datum is a dyad (a pair of objects) [Hofmann, Puzicha, Jordan, 1998, 1999].

- The **co-occurrence matrix** can be seen as a summary of the information in an observed set of pairs of objects: a **dyadic dataset**.
Dyadic Data Analysis

Dyadic Data and Co-Occurrence Matrix

- \( S \) – sequence of all pairs of objects co-occurring in a common cluster over the clustering ensemble \( \mathcal{E} \)

- A co-occurrence pair \( s \in S \) is defined as:

\[
\mathbf{s}_m = (y_m, z_m) \in \mathcal{X} \times \mathcal{X}, \text{ for } m = 1, \ldots, |S|
\]

where \( y_m \neq z_m \), \( y_m \in C^i_k \) and \( z_m \in C^i_k \).

- The co-occurrence matrix, \( \mathbf{C} = [C_{y,z}] \), is a \((N \times N)\) matrix which collects a statistical summary of \( S \):

\[
C_{y,z} = \sum_{m=1}^{|S|} \mathbb{I}((y_m, z_m) = (y, z)), \text{ for } y, z \in \mathcal{X}
\]
Hypothesis:
Underlying clusters revealed by the observations $S$

Generative model for $S$:
- Interpret $S$ as i.i.d. samples of a pair of r.v. $(Y, Z) \in \mathcal{X} \times \mathcal{X}$
- Introduce $R \in \{1, \ldots, L\}$: a multinomial latent class variable.
- $Y$ and $Z$ are i.i.d. given $R$:

$$
\mathbb{P}(Y = y, Z = z | R = r) = \mathbb{P}(Y = y | R = r) \mathbb{P}(Z = z | R = r)
$$

and

$$
\mathbb{P}(Z = z | R = r) = \mathbb{P}(Y = z | R = r),
$$
Mixture Model

- The joint distribution of $(Y, Z)$,

$$
\mathbb{P}(Y = y, Z = z) = \sum_{r=1}^{L} \mathbb{P}(Y = y|R = r) \mathbb{P}(Y = z|R = r) \mathbb{P}(R = r),
$$

is parameterized by:

- $\mathbb{P}(R = r)$, for any $r = 1, \ldots, L$: the distribution of the latent variables $R$;
- $\mathbb{P}(Y = y|R = r) = \mathbb{P}(Z = y|R = r)$, for $y = 1, \ldots, N$ and $r = 1, \ldots, L$: the conditional distributions of $Y$ and $Z$ given the latent variables $R$. 
We write these distributions compactly as:

- \( p = (p_1, \ldots, p_L) \): an \( L \)-vector, where \( p_r = \mathbb{P}(R = r) \)
- \( B = [B_{r,j}] \): an \( L \times N \) matrix, where

\[
B_{r,j} = \mathbb{P}(Y = j | R = r) = P(Z = j | R = r);
\]

of course, \( B \) is a stochastic matrix: \( \sum_j B_{r,j} = 1 \).

With this notation,

\[
\mathbb{P}(Y = y, Z = z, R = r) = p_r B_{r,y} B_{r,z},
\]

and

\[
\mathbb{P}(Y = y, Z = z) = \sum_{r=1}^{L} p_r B_{r,y} B_{r,z}.
\]
Mixture Model

- Assuming $S = \{(y_m, z_m), m = 1, ..., |S|\}$ contains $|S|$ i.i.d. samples of $(Y, Z)$,

  $$P(S|p, B) = \prod_{m=1}^{|S|} \sum_{r=1}^L p_r B_{r,y_m} B_{r,z_m}.$$ 

- The complete likelihood (if $\mathcal{R} = (r_1, ..., r_{|S|})$ was observed) is

  $$P(S, \mathcal{R}|p, B) = \prod_{m=1}^{|S|} p_{r_m} B_{r_m,y_m} B_{r_m,z_m}$$

  $$\log P(S, \mathcal{R}|p, B) = \sum_{m=1}^{|S|} \sum_{r=1}^L \mathbb{I}(r_m = r) \log \left( p_r B_{r,y_m} B_{r,z_m} \right).$$
The EM algorithm yields maximum marginal likelihood estimates of $p$ and $B$:

$$(\hat{p}, \hat{B}) = \arg \max_{p,B} P(S|p, B)$$

- (E-Step) Compute

$$Q(p, B; \hat{p}, \hat{B}) = \mathbb{E}_R \left[ \log P(S, R|p, B)|\hat{p}, \hat{B} \right]$$

- (M-Step) updated the estimates by maximizing the $Q$-function w.r.t. $p$ and $B$. 

**Maximum Likelihood Estimate**
The $Q$-function is given by

$$Q(p, B; \hat{p}, \hat{B}) = \sum_{m=1}^{\|S\|} \sum_{r=1}^{L} \hat{R}_{m,r} \log(p_r B_{r,y}^m B_{r,z}^m)$$

where

$$\hat{R}_{m,r} \equiv \mathbb{E} \left[ \mathbb{I}(R_m = r) \mid S, \hat{p}, \hat{B} \right] = \mathbb{P} \left[ R_m = r \mid (y_m, z_m), \hat{p}, \hat{B} \right],$$

is the conditional probability that the pair $(y_m, z_m)$ was generated by cluster $r$, that is,

$$\hat{R}_{m,r} = \frac{\hat{p}_r \hat{B}_{r,y}^m \hat{B}_{r,z}^m}{\sum_{s=1}^{L} \hat{p}_s \hat{B}_{s,y}^m \hat{B}_{s,z}^m}.$$
M-Step

- maximizing the $Q$-function, w.r.t. $\mathbf{p}$ leads to:

$$\hat{p}_{r}^{\text{new}} = \frac{1}{|S|} \sum_{m=1}^{|S|} \hat{R}_{m,r} \quad \text{for } r = 1, \ldots, L.$$  

- ...with respect to $\mathbf{B}$, yields

$$\hat{B}_{r,y}^{\text{new}} = \sum_{z=1}^{N} \hat{C}_{y,z}^{r} \left( \sum_{t=1}^{N} \sum_{z=1}^{N} \hat{C}_{t,z}^{r} \right)^{-1},$$

where

$$\hat{C}_{y,z}^{r} = \sum_{i=1}^{|S|} \hat{R}_{m,r} \mathbb{I}((y_{m}, z_{m}) = (y, z))$$

is a weighted version of the co-association matrix.
Interpretation of the estimates

- The parameter estimates returned by the algorithm have clear interpretations:
  - $\hat{p}_1, \ldots, \hat{p}_L$ are the cluster probabilities;
  - $\hat{B}_{r,y}$ is the “degrees of ownership” of object $y$ by cluster $r$.

- The estimate of probability that object $y$ belongs to cluster $r$ (denoted as $\hat{V}_{y,r}$), can be obtained by applying Bayes law:

\[
\hat{P}(R = r | Y = y) = \frac{\hat{P}(R = r, Y = y)}{\hat{P}(Y = y)} = \frac{\hat{B}_{r,y} \hat{p}_r}{\sum_{s=1}^{L} \hat{B}_{s,y} \hat{p}_s}.
\]
We evaluate PEnCA on several UCI benchmark datasets.

The synthetic two-dimensional datasets used for this study are:

(a) Cigar data.  
(b) Bars.  
(c) Half Rings.  
(d) Stars.

Clustering ensembles obtained by $K$-means clustering with different numbers of clusters and initializations.
Example and Discussion

Example

(e) Co-Occurrence Matrix

(f) Soft assignments

Figure: Example of co-occurrence matrix matrix and soft assignments \( \hat{P}(R = r | Y = y) \) obtained by PEnCA for the Iris dataset (with \( L = 3 \)).
Comparison with baseline [Topchy, Jain, Punch, 2004], another mixture model (MM) for clustering ensembles

<table>
<thead>
<tr>
<th>Data Set</th>
<th>$N$</th>
<th>$K$</th>
<th>PEnCA</th>
<th>MM</th>
</tr>
</thead>
<tbody>
<tr>
<td>stars</td>
<td>114</td>
<td>2</td>
<td>0.921</td>
<td>0.737</td>
</tr>
<tr>
<td>cigar-data</td>
<td>250</td>
<td>4</td>
<td>0.712</td>
<td>0.812</td>
</tr>
<tr>
<td>bars</td>
<td>400</td>
<td>2</td>
<td>0.985</td>
<td>0.812</td>
</tr>
<tr>
<td>halfrings</td>
<td>400</td>
<td>2</td>
<td>1.000</td>
<td>0.797</td>
</tr>
<tr>
<td>iris-r</td>
<td>150</td>
<td>3</td>
<td>0.920</td>
<td>0.693</td>
</tr>
<tr>
<td>wine-normalized</td>
<td>178</td>
<td>3</td>
<td>0.949</td>
<td>0.590</td>
</tr>
<tr>
<td>house-votes-84-normalized</td>
<td>232</td>
<td>2</td>
<td>0.905</td>
<td>0.784</td>
</tr>
<tr>
<td>ionosphere</td>
<td>351</td>
<td>2</td>
<td>0.724</td>
<td>0.829</td>
</tr>
<tr>
<td>std-yeast-cellcycle</td>
<td>384</td>
<td>5</td>
<td>0.729</td>
<td>0.578</td>
</tr>
<tr>
<td>pima-normalized</td>
<td>768</td>
<td>2</td>
<td>0.681</td>
<td>0.615</td>
</tr>
<tr>
<td>Breast-cancers</td>
<td>683</td>
<td>2</td>
<td>0.947</td>
<td>0.764</td>
</tr>
<tr>
<td>optdigits-r-tra-1000</td>
<td>1000</td>
<td>10</td>
<td>0.876</td>
<td>0.581</td>
</tr>
</tbody>
</table>
Conclusions

- A probabilistic generative model for consensus clustering, based on a dyadic aspect model of evidence accumulation clustering.
- The consensus partition is extracted by solving a maximum likelihood estimation problem via EM.
- The method yields probabilistic assignments of each sample to each cluster.
- Experiments show that the proposed method outperforms another recent probabilistic formulation of ensemble clustering.
- Future work: the probabilistic/generative nature of the approach opens the door to dealing with the model selection problem \( L = ? \): MDL, BIC, non-parametric approaches.
Acknowledgements

- Fundação para a Ciência e Tecnologia (FCT) under the grants SFRH/PROTEC/49512/2009 and PTDC/EIACCO/103230/2008 (Project EvaClue)
- Open Scheme (FET-Open) of the Seventh Framework Programme of the European Commission, under the SIMBAD project (contract 213250)

Questions?
Comments?