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

Clustering Ensembles

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

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

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

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

- 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:
  (i) construction of the clustering ensemble;
  (ii) evidence accumulation of pair-wise associations;
  (iii) extraction of the final consensus partition.
- The combination step (ii) yields the co-occurrence matrix $\mathbf{C}$:

$$C_{i,j} = \text{“number of times objects } i \text{ and } j \text{ co-occurred”}$$

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 and Co-Occurrence Matrix

- $\mathcal{S}$ – sequence of all pairs of objects co-occurring in a common cluster over the clustering ensemble $\mathcal{E}$
- A co-occurrence pair $\mathbf{s} \in \mathcal{S}$ is defined as:

$$\mathbf{s}_m = (y_m, z_m) \in \mathcal{X} \times \mathcal{X}, \text{ for } m = 1, \ldots, |\mathcal{S}|$$
$$\text{where } y_m \neq z_m,\ y_m \in \mathcal{C}_k^i \text{ and } z_m \in \mathcal{C}_k^i.$$

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

$$C_{y,z} = \sum_{m=1}^{|\mathcal{S}|} \mathbb{I}\big((y_m, z_m) = (y, z)\big), \quad \text{for } y, z \in \mathcal{X}$$

Generative Model

- Hypothesis:
  - Underlying clusters revealed by the observations $\mathcal{S}$
- Generative model for $\mathcal{S}$:
  - Interpret $\mathcal{S}$ as i.i.d. samples of a pair of r.v. $(Y, Z) \in \mathcal{X} \times \mathcal{X}$
  - Introduce $R \in \{1, ..., L\}$: 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 $r = 1, ..., L$: the distribution of the latent variables R;
  - $\mathbb{P}(Y = y|R = r) = \mathbb{P}(Z = y|R = r)$, for $y = 1, ..., N$ and $r = 1, ..., L$: the conditional distributions of Y and Z given the latent variables R.

Mixture Model

- We write these distributions compactly as:
  - $\mathbf{p} = (p_1, ..., p_L)$: an L-vector,where $p_r = \mathbb{P}(R = r)$
  - $\mathbf{B} = [B_{r,j}]$: an $L \times N$ matrix, where

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

  of course, $\mathbf{B}$ 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 $\mathcal{S} = \{(y_m, z_m),\ m = 1, ..., |\mathcal{S}|\}$ contains $|\mathcal{S}|$ i.i.d. samples of (Y, Z),

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

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

$$\begin{aligned} \mathbb{P}(\mathcal{S}, \mathcal{R}|\mathbf{p}, \mathbf{B}) &= \prod_{m=1}^{|\mathcal{S}|} p_{r_m} B_{r_m,y_m} B_{r_m,z_m} \\ \log \mathbb{P}(\mathcal{S}, \mathcal{R}|\mathbf{p}, \mathbf{B}) &= \sum_{m=1}^{|\mathcal{S}|} \sum_{r=1}^{L} \mathbb{I}(r_m = r) \log\big(p_r\, B_{r,y_m}\, B_{r,z_m}\big). \end{aligned}$$

Maximum Likelihood Estimate

- The EM algorithm yields maximum marginal likelihood estimates of $\mathbf{p}$ and $\mathbf{B}$:

$$(\widehat{\mathbf{p}}, \widehat{\mathbf{B}}) = \arg\max_{\mathbf{p},\mathbf{B}} \mathbb{P}(\mathcal{S}|\mathbf{p}, \mathbf{B})$$

  - (E-Step) Compute

$$Q(\mathbf{p}, \mathbf{B}; \widehat{\mathbf{p}}, \widehat{\mathbf{B}}) = \mathbb{E}_{\mathcal{R}}\left[\log \mathbb{P}(\mathcal{S}, \mathcal{R}|\mathbf{p}, \mathbf{B})|\widehat{\mathbf{p}}, \widehat{\mathbf{B}}\right]$$

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

E-Step

- The Q-function is given by

$$Q(\mathbf{p}, \mathbf{B}; \widehat{\mathbf{p}}, \widehat{\mathbf{B}}) = \sum_{m=1}^{|\mathcal{S}|} \sum_{r=1}^{L} \widehat{R}_{m,r} \log\left(p_r \, B_{r,y_m} \, B_{r,z_m}\right)$$

where

$$\widehat{R}_{m,r} \equiv \mathbb{E}\left[\mathbb{I}(R_m = r) \,\middle|\, \mathcal{S}, \widehat{\mathbf{p}}, \widehat{\mathbf{B}}\right] = \mathbb{P}\left[R_m = r \,\middle|\, (y_m, z_m), \widehat{\mathbf{p}}, \widehat{\mathbf{B}}\right],$$

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

$$\widehat{R}_{m,r} = \frac{\widehat{p}_r \, \widehat{B}_{r,y_m} \, \widehat{B}_{r,z_m}}{\sum_{s=1}^{L} \widehat{p}_s \, \widehat{B}_{s,y_m} \, \widehat{B}_{s,z_m}}$$

M-Step

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

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

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

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

where

$$\widehat{C}_{y,z}^{r} = \sum_{i=1}^{|\mathcal{S}|} \widehat{R}_{m,r} \, \mathbb{I}\left((y_m, z_m) = (y, z)\right)$$

is a weighted version of the co-association matrix.

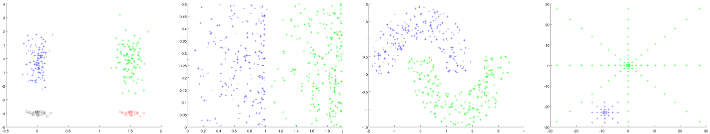
Interpretation of the estimates

- The parameter estimates returned by the algorithm have clear interpretations:
  - $\widehat{p}_1, \ldots, \widehat{p}_L$ are the cluster probabilities;
  - $\widehat{B}_{r,y}$ is the "degrees of ownership" of object y by cluster r.
- The estimate of probability that object y belongs to cluster r (denoted as $\widehat{V}_{y,r}$), can be obtained by applying Bayes law:

$$\widehat{\mathbb{P}}(R = r|Y = y) = \frac{\widehat{\mathbb{P}}(R = r, Y = y)}{\widehat{\mathbb{P}}(Y = y)} = \frac{\widehat{B}_{r,y}\,\widehat{p}_r}{\sum_{s=1}^{L} \widehat{B}_{s,y}\,\widehat{p}_s}.$$

Experimental Setup

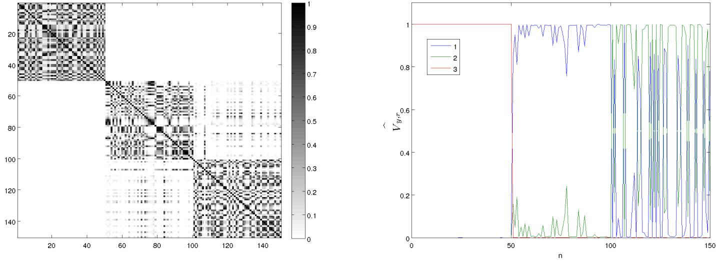
- We evaluate PEnCA on several UCI benchmark datasets.
- The synthetic two-dimensional datasets used for this study are



(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



(e) Co-Occurence Matrix

(f) Soft assignments

Figure: Example of co-occurrence matrix matrix and soft assignments $\widehat{\mathbb{P}}(R = r|Y = y)$ obtained by PEnCA for the *Iris* dataset (with $L = 3$).

Results

Comparison with baseline [Topchy, Jain, Punch, 2004], another mixture model (MM) for clustering ensembles

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

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?