Pairwise Probabilistic Clustering using Evidence Accumulation

Samuel Rota Bulò[a], André Lourenço[c], Ana Fred[b,c] and Marcello Pelillo[a]

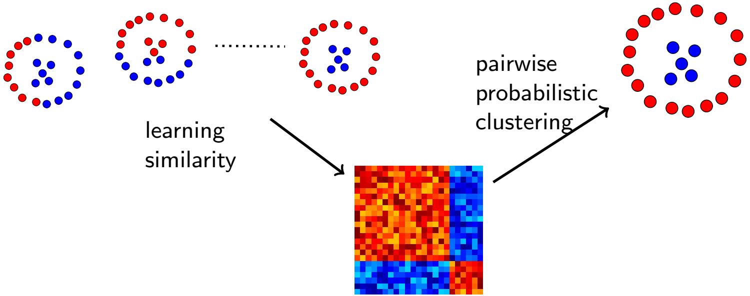
[a] CVPR Lab @ University of Venice, Italy [b] Instituto Superior Técnico
[c] Instituto de Telecomunicações, Lisbon, Portugal

SSPR 2010

Outline

Goal

- Development of a principled (soft) clustering approach built upon the evidence accumulation framework.
- Evidence accumulation allows to combine the results of multiple clusterings into a single similarity matrix (co-association matrix) by viewing each clustering result as an independent evidence of pairwise data organization.



Problem setting

- $O = \{1, \dots, n\}$ is the set of data object to cluster.
- K is the number of desired classes.
- $\mathcal{E} = \{cl_i\}_{i=1}^{N}$ is the esemble of N clusterings of O.
- each clustering is a function $cl_i : O_i \to \{1, \dots, K_i\}$ from the set of objects $O_i \subseteq O$ to a class label.
- $\Omega_{ij} = \{p = 1 \dots N \ : \ i, j \in O_p\}$ is the set of indices of clusterings where i and j have been classified.
- $N_{ij} = |\Omega_{ij}|$.

Goal

Learn from the ensemble of clustering $\mathcal{E}$ how to cluster the objects into K classes.

Outline

Introduction

Pairwise Probabilistic Clustering

Experiments

Conclusions

Assumptions

- We start from the assumption that objects can be softly assigned to clusters.
- For each object $i \in O$ we want to estimate an *unknown assignment* $\mathbf{y}_i$, which is a probability distribution over the set of cluster labels $\{1, \ldots, K\}$.
- Each assignment is a point of the *standard simplex*, i.e., $\mathbf{y}_i \in \Delta_K$, where

$$\Delta_K = \left\{\mathbf{x} \in \mathbb{R}_+^K : \|\mathbf{x}\|_1 = 1\right\}.$$

- Assuming independent cluster assignments, the probability of objects i and j to occur in a same cluster can be derived as $\mathbf{y}_i^\top \mathbf{y}_j$.

Observations

- Let $Y^\top Y$ be the $n \times n$ matrix of object co-occurrence probabilities, where $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \in \Delta_K^n$.
- For each $i, j \in O$, let X_{ij} be a Bernoulli r.v. indicating whether objects i and j occur in a same cluster.
- Note that $E[X_{ij}] = \mathbf{y}_i^\top \mathbf{y}_j$ according to our model.
- From the clusterings ensemble we collect N_{ij} independent realizations $x_{ij}^{(p)}$ (where $p \in \Omega_{ij}$), which are given by

$$x_{ij}^{(p)} = \begin{cases} 1 & \text{if } cl_p(i) = cl_p(j)\,, \\ 0 & \textit{otherwise}\,. \end{cases}$$

Co-association matrix

- By taking the mean of the x_{ij}'s, we obtain the empirical probability of co-occurrence c_{ij}, which is the fraction of times objects i and j have been assigned to a same cluster:

$$c_{ij} = \frac{1}{N_{ij}} \sum_{p \in \Omega_{ij}} x_{ij}^{(p)} .$$

- The matrix $C = (c_{ij})$ is known as co-association matrix within the evidence accumulation-based framework for clustering .
- The matrix C (empirical co-association matrix) is the maximum likelihood estimate of $Y^\top Y$ (true co-association matrix) given the observations from the clusterings ensemble $\mathcal{E}$.

Pairwise Probabilistic Clustering

- We find a solution Y^* of the clustering problem by minimizing the divergence in the least-square sense of the true co-association matrix from the empirical one with respect to Y:

$$\begin{aligned} Y^* = \arg\min \quad & \|C - Y^\top Y\|^2 \\ \text{s.t.} \quad & Y \in \Delta_K^n \, . \end{aligned}$$

- This can be rewritten in terms of a maximization of a polynomial with nonnegative coefficients:

$$\begin{aligned} Y^* = \arg\max \;\; & 2Tr(CY^\top Y) + \|Y^\top E_K Y\|^2 - \|Y^\top Y\|^2 \\ \text{s.t.} \;\; & Y \in \Delta_K^n \, , \end{aligned}$$

where E_K is a $K \times K$ matrix of 1s.

Algorithm

Baum-Eagon Inequality

Let $X = (x_{ri}) \in \Delta_K^n$ and $Q(X)$ be a homogeneous polynomial in the variables x_{ri} with nonnegative coefficients. Define the mapping $Z = (z_{ri}) = \mathcal{M}(X)$ as follows:

$$z_{ri} = x_{ri} \frac{\partial Q(X)}{\partial x_{ri}} \Big/ \sum_{s=1}^{k} x_{si} \frac{\partial Q(X)}{\partial x_{si}}, \quad (1)$$

for all $i = 1 \dots n$ and $r = 1 \dots k$. Then $Q(\mathcal{M}(X)) > Q(X)$, unless $\mathcal{M}(X) = X$. In other words $\mathcal{M}$ is a growth transformation for the polynomial Q.

- Baum-Sell show that the same holds for inhomogeneous polynomial.
- By applying the Baum-Eagon inequality, we obtain the following update rule:

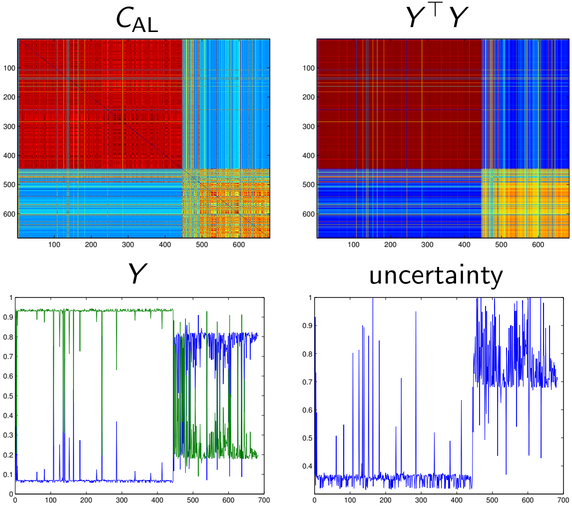
$$y_{ki}^{(t+1)} = y_{ki}^{(t)} \frac{n + [Y(C - Y^\top Y)]_{ki}}{n + \sum_k y_{ki}^{(t)} [Y(C - Y^\top Y)]_{ki}}$$

Outline

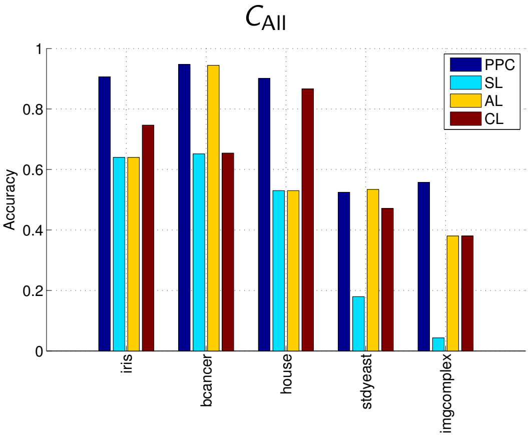
Experimental setting

- Experiments on iris, house-votes, std-yeast-cell and breast-cancer datasets (from the UCI ML repo), and the image-complex synthetic data-set.
- For each data-set, we produced the clustering ensemble $\mathcal{E}$ by running different clustering algorithms, with different parameters, on subsampled versions of the original data-set (sampling rate 0.9).
- Clustering algorithms used : Single Link (SL), Complete Link (CL), Average Link (AL) and K-means (KM).
- Partitions obtained by a single algorithm form a base ensemble.
- Overall, we formed four base ensembles, namely $\mathcal{E}_{\mathsf{SL}}$, $\mathcal{E}_{\mathsf{AL}}$, $\mathcal{E}_{\mathsf{CL}}$ and $\mathcal{E}_{\mathsf{KM}}$, and a large ensemble $\mathcal{E}_{\mathsf{All}}$ from the union of all of them.
- For each ensemble we created a corresponding co-association matrix, namely C_{SL}, C_{AL}, C_{CL}, C_{KM} and C_{All}.
- For each of these co-association matrices, we applied our Pairwise Probabilistic Clustering (PPC) approach, and compared it against the performances obtained with the same matrices by SL, AL and CL.
- Each method was provided with the optimal number of classes as input parameter.

Breast-cancer

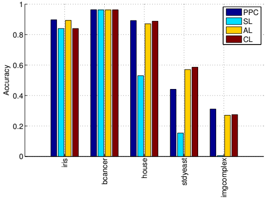


Results

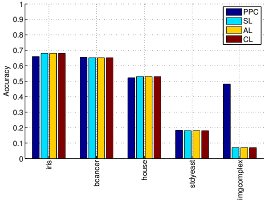

C_{All}
Accuracy
1
0.8
0.6
0.4
0.2
0
iris
bcancer
house
stdyeast
imgcomplex
PPC
SL
AL
CL

Results

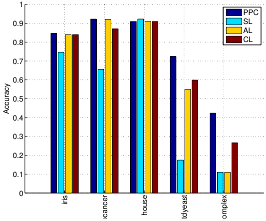
C_{KM}



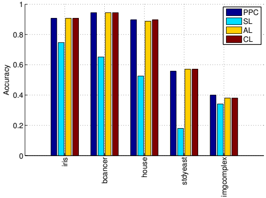
C_{SL}



C_{CL}



C_{AL}



Outline

Conclusions

- Taking advantage of the probabilistic interpretation of the computed similarities of the the co-association matrix, derived from the ensemble of clusterings, using the Evidence Accumulation Clustering, we proposed a principled soft clustering method.
- Our method reduces the clustering problem to a polynomial optimization in probability domain, which is attacked by means of the Baum-Eagon inequality.
- The new method produces a soft partition of the data. Nevertheless, when converting these soft labels into a crisp partition the method leads to better results than the competitors.

Question

- We have noise in the data, but also noise in the clustering ensemble! Can the information about uncertainty related to the soft cluster assignments in Y be used to identify "noisy" partitions in the ensemble $\mathcal{E}$?
- Can we use Evidence Accumulation as a means of learning pairwise similarities from multiple sources with different data representations ?

L. E. Baum and J. A. Eagon.
An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology.
Bull. Amer. Math. Soc., 73:360–363, 1967.

A. Fred and A. K. Jain.
Data clustering using evidence accumulation.
In *Int. Conf. Patt. Recogn. (ICPR)*, pages 276–280, 2002.

A. Fred and A. K. Jain.
Combining multiple clustering using evidence accumulation.
IEEE Trans. Pattern Anal. Machine Intell., 27(6):835–850, 2005.

A. K. Jain and R. C. Dubes.
Algorithms for data clustering.
Prentice-Hall, 1988.