Pairwise Probabilistic Clustering Using Evidence Accumulation

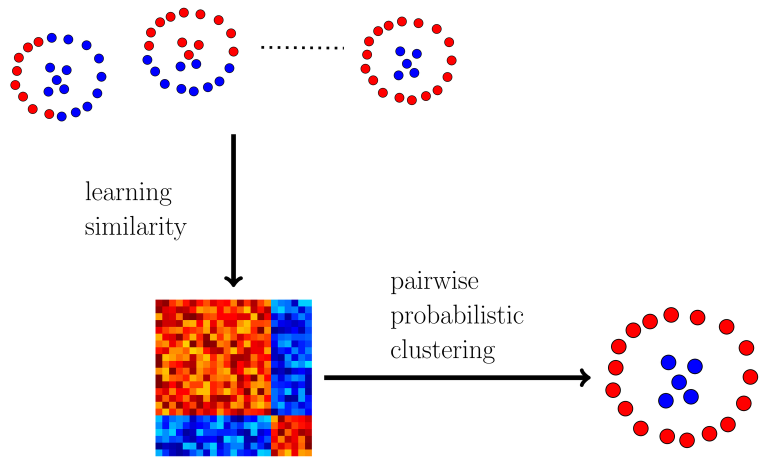
Samuel Rota Bulò[a], André Lourenço[c], Ana Fred[b,c] and Marcello Pelillo[a]

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

http://simbad-fp7.eu/

1. 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, called co-association matrix, by viewing each clustering result as an independent evidence of pairwise data organization.



2. 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. The ensemble is obtained by running different algorithms with different. parametrizations on (possibly) sub-sampled versions of the original dataset.
- 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.

3. Proposed Solution

- 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 $\{1, \dots, K\}$.
- Each assignment is a point of the standard simplex, i.e., $\mathbf{y}_i \in \Delta^K$, where

$$\Delta_K = \{\mathbf{x} \in \mathbb{R}_+^K \; : \; \|\mathbf{x}\|_1 = 1\} . \qquad \text{(standard simplex)}$$

- Assuming independent cluster assignments, the probability of objects i and j to occur in a same cluster can be derived as $\mathbf{y}_i^\top \mathbf{y}_j$.
- Let $Y^\top Y$ be the $n \times n$ matrix of object co-occurrence probabilities, where $Y = (\mathbf{y}_1, \dots, \mathbf{y}_n) \in \Delta_K^n$.
- 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

$$x_{ij}^{(p)} = \begin{cases} 1 & \text{if } cl_p(i) = cl_p(j) \, , \\ 0 & otherwise \, . \end{cases} \qquad \text{(observation of co-occurrence)}$$

- By taking their mean, 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:

$$c_{ij} = \frac{1}{N_{ij}} \sum_{p \in \Omega_{ij}} x_{ij}^{(p)} . \qquad \text{(empirical co-occurrence probability)}$$

- The matrix $C = (c_{ij})$ is known as co-association matrix within the evidence accumulation-based framework for clustering [3, 4].
- 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 $\mathcal{E}$.
- 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:

Solution

$$Y^* = \arg\min \; \|C - Y^\top Y\|_F^2 \quad \text{s.t. } Y \in \Delta_K^n . \qquad (1)$$

4. Baum-Eagon Inequality

Baum-Eagon Inequality [1]

Theorem 1. *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:*

$$z_{ri} = x_{ri} \frac{\partial Q(X)}{\partial x_{ri}} \Bigg/ \sum_{s=1}^{k} x_{si} \frac{\partial Q(X)}{\partial x_{si}} , \qquad (2)$$

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 and Eagon [1] introduced a class of nonlinear transformations in probability domain.
- This result applies also to inhomogeneous polynomials [2].
- The Baum-Eagon inequality provides an effective iterative means for maximizing polynomial functions in probability domains.

5. The Algorithm

- We can not use the Baum-Eagon Inequality for optimizing (1) directly, as we need a maximization of a polynomial with nonnegative coefficients.
- Consider the following optimization program:

$$\max \; 2Tr(CY^\top Y) + \|Y^\top E_K Y\|^2 - \|Y^\top Y\|^2 \quad \text{s.t. } Y \in \Delta_K^n , \qquad (3)$$

Equivalence

Theorem 2. *The maximizers of (3) are minimizers of (1) and vice versa. Moreover, the objective function of (3) is a polynomial with nonnegative coefficients in the variables y_{ki}, which are elements of Y.*

- We can now use the Baum-Eagon inequality to locally optimize (3). This leads to the following updating rule for $Y = (y_{ki})$:

Updating rule

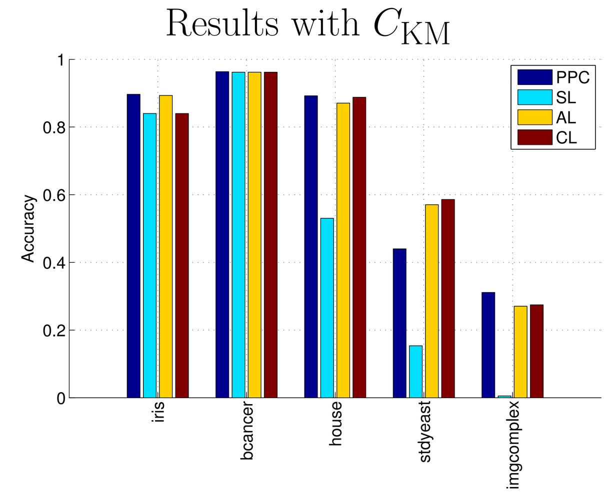
$$y_{ki}^{(t+1)} = y_{ki}^{(t)} \frac{n + [Y(C - Y^\top Y)]_{ki}}{n + \sum_k y_{ki}^{(t)} [Y(C - Y^\top Y)]_{ki}}$$

- The computational complexity of the proposed dynamics is $O(\gamma k n^2)$, where γ is the average number of iterations required to converge (in the experiments we kept γ fixed).

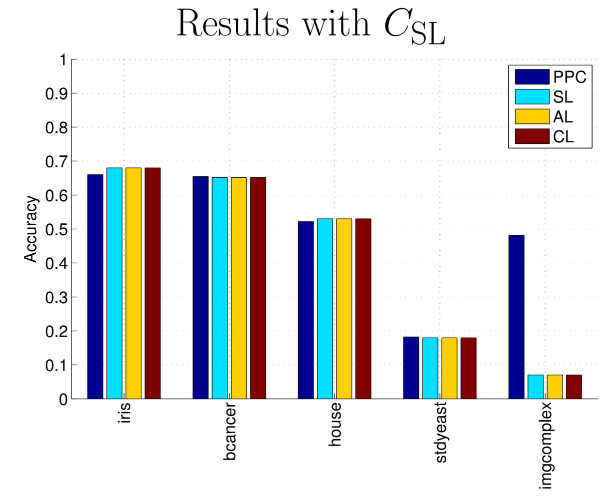
6. Experiments

- We conducted experiments on different real data-sets from the UCI Machine Learning Repository: iris, house-votes, std-yeast-cell and breast-cancer.
- We considered also 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 (the sampling rate was fixed to 0.9).
- The clustering algorithms used to produce the ensemble were the following [5]: Single Link (SL), Complete Link (CL), Average Link (AL) and K-means (KM).
- We run each algorithm several times in order to produce clusterings with different number of classes. For each algorithm and parametrization we generated 100 data partitions from the subsampled versions of the data-set; we name such set of partitions as base ensemble.
- Overall, we formed four base ensembles, namely $\mathcal{E}_{SL}$, $\mathcal{E}_{AL}$, $\mathcal{E}_{CL}$ and $\mathcal{E}_{KM}$.
- We created a large ensemble $\mathcal{E}_{All}$ from the union of all of the base esembles.
- 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 the agglomerative hierarchical algorithms SL, AL and CL. Each method was provided with the optimal number of classes as input parameter.

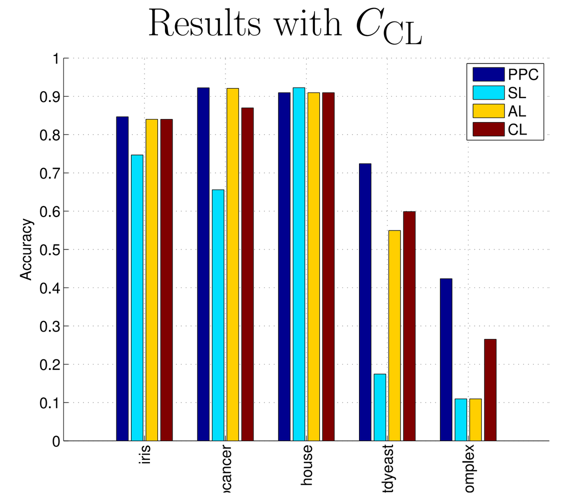
Results with C_{KM}



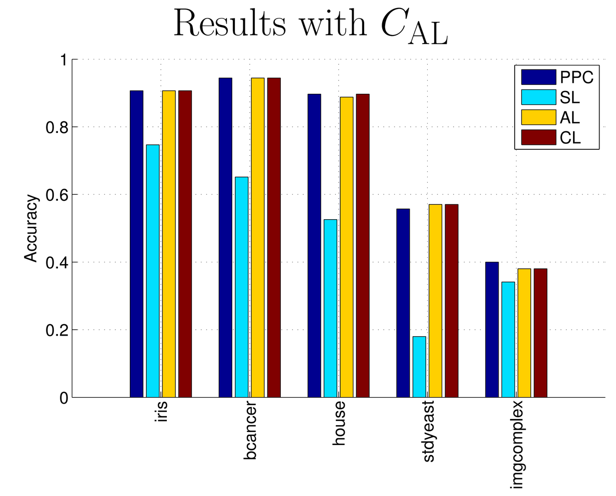
Results with C_{SL}



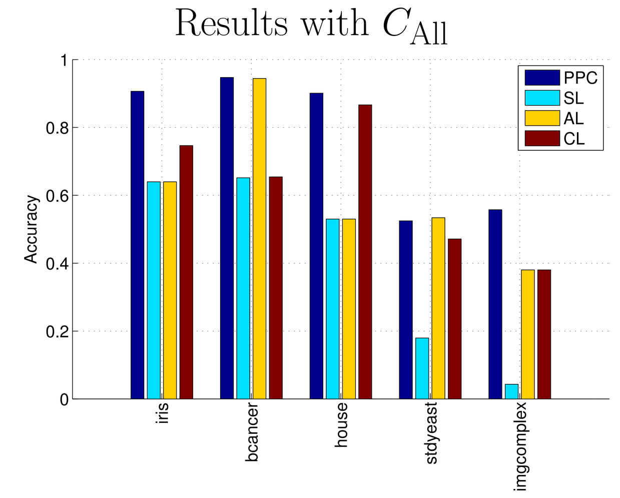
Results with C_{CL}



Results with C_{AL}

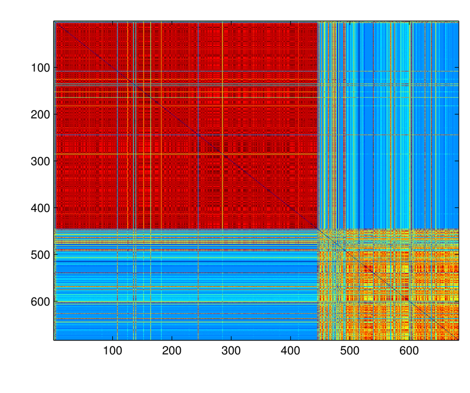


Results with C_{All}

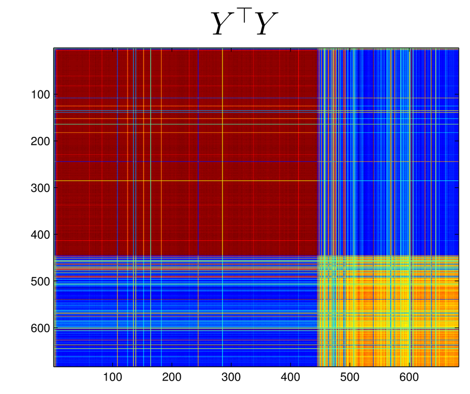


Breast-cancer.

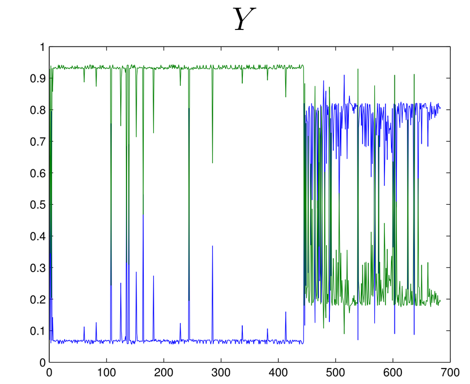
C_{All}



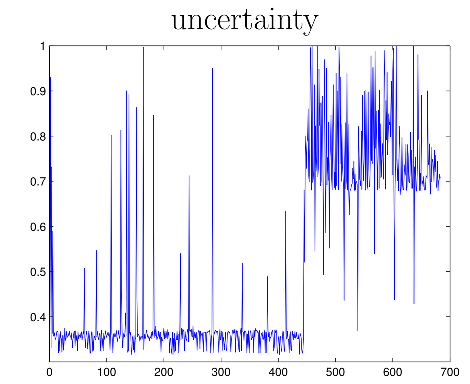
$Y^\top Y$



Y



uncertainty



7. 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.

References

[1] 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.

[2] L. E. Baum and G. R. Sell. Growth transformations for functions on manifolds. *Pacific J. Math.*, 27:221–227, 1968.

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

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

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