Pairwise Probabilistic Clustering Using Evidence Accumulation

Samuel Rota Bulò[1], André Lourenço[3], Ana Fred[2,3], and Marcello Pelillo[1]

[1] Dipartimento di Informatica - University of Venice - Italy
{srotabul,pelillo}@dsi.unive.it
[2] Instituto Superior Técnico - Lisbon - Portugal
[3] Instituto de Telecomunicações - Lisbon - Portugal
{arlourenco,afred}@lx.it.pt

Abstract. In this paper we propose a new approach for consensus clustering which is built upon the evidence accumulation framework. Our method takes the co-association matrix as the only input and produces a soft partition of the dataset, where each object is probabilistically assigned to a cluster, as output. Our method reduces the clustering problem to a polynomial optimization in probability domain, which is attacked by means of the Baum-Eagon inequality. Experiments on both synthetic and real benchmarks data, assess the effectiveness of our approach.

1 Introduction

There is a close connection between the concepts of pairwise similarity and probability in the context of unsupervised learning. It is a common assumption that, if two objects are similar, it is very likely that they are grouped together by some clustering algorithm, the higher the similarity, the higher the probability of co-occurrence in a cluster. Conversely, if two objects co-occur very often in the same cluster (high co-occurrence probability), then it is very likely that they are very similar. This duality and correspondence between pairwise similarity and pairwise probability within clusters forms the core idea of the clustering ensemble approach known as evidence accumulation clustering (EAC) [1].

Evidence accumulation clustering combines the results of multiple clusterings into a single data partition by viewing each clustering result as an independent evidence of pairwise data organization. Using a pairwise frequency count mechanism amongst a clustering committee, the method yields, as an intermediate result, a co-association matrix that summarizes the evidence taken from the several members in the clustering ensemble. This matrix corresponds to the maximum likelihood estimate of the probability of pairs of objects being in the same group, as assessed by the clustering committee. One of the main advantages of EAC is that it allows for a big diversification within the clustering committee. Indeed, no assumption is made about the algorithms used to produce the data partitions, it is robust to incomplete information, i.e., we may include partitions over sub-sampled versions of the original data set, and no restriction is made on the number of clusters of the partitions.

E.R. Hancock et al. (Eds.): SSPR & SPR 2010, LNCS 6218, pp. 395–404, 2010.

Once a co-association matrix is produced according to the EAC framework, a consensus clustering is obtained by applying a clustering algorithm, which typically induces a hard partition, to the co-association matrix. Although having crisp partitions as baseline for the accumulation of evidence of data organization is reasonable, this assumption is too restrictive in the phase of producing a consensus clustering. This is for instance the case for many important applications such as clustering micro-array gene expression data, text categorization, perceptual grouping, labeling of visual scenes and medical diagnosis. In fact, the importance of dealing with overlapping clusters has been recognized long ago [2] and recently, in the machine learning community, there has been a renewed interest around this problem [3,4]. Moreover, by inducing hard partitions we loose important information like the level of uncertainty of each label assignment. It is also worth considering that the underlying clustering criteria of ad hoc algorithms do not take advantage of the probabilistic interpretation of the computed similarities, which is an intrinsic part of the EAC framework.

In this paper we propose a new approach for consensus clustering which is built upon the evidence accumulation framework. Our idea was inspired by a recent work due to Zass and Sashua [5]. Our method takes the co-association matrix as the only input and produces a soft partition of the data set, where each object is probabilistically assigned to a cluster, as output. In order to find the unknown cluster assignments, we fully exploit the fact that each entry of the co-association matrix is an estimation of the probability of two objects to be in a same cluster, which is derived from the ensemble of clusterings. Indeed, it is easy to see that under reasonable assumptions, the probability that two objects i and j will occur in the same cluster is a function of the unknown cluster assignments of i and j. By minimizing the divergence between the estimation derived from the co-association matrix and this function of the unknowns, we obtain the result of the clustering procedure. More specifically, our method reduces the clustering problem to a polynomial optimization in the probability domain, which is attacked by means of the Baum-Eagon inequality [6]. This inequality, indeed, provides us with a class of nonlinear transformations that serve our purpose. In order to assess the effectiveness of our findings we conducted experiments on both synthetic and real benchmark data sets.

2 A Probabilistic Model for Clustering

Let $O = \{1, \dots, n\}$ be a set of data objects (or simply objects) to cluster into K classes and let $\mathcal{E} = \{cl_i\}_{i=1}^{N}$ be an ensemble of N clusterings of O obtained by running different algorithms with different parameterizations on (possibly) sub-sampled versions of the original data set O. Data sub-sampling is herein put forward as a most general framework for the following reasons: it favors the diversification of the clustering ensemble; it models situations of distributed clustering where local clusterers have only partial access to the data; by using this type of data perturbation, the co-association matrix has an additional interpretation of pairwise stability that can further be used for the purpose of cluster validation [7].

Each clustering in the ensemble $\mathcal{E}$ is a function $cl_i : O_i \to \{1, \ldots, K_i\}$ from the set of objects $O_i \subseteq O$ to a class label. For the afore-mentioned reasons, O_i is a subset of the original data set O and, moreover, each clustering may assume a different number of classes K_i. We denote by Ω_{ij} the indices of the clusterings where i and j have been classified, which is given by

$$\Omega_{ij} = \{p = 1 \ldots N \,:\, i, j \in O_p\} \,.$$

Consider also $N_{ij} = |\Omega_{ij}|$, where $| \cdot |$ provides the cardinality of the argument, which is the number of clusterings where i and j have been both classified.

The aim of our work is to learn, from the ensemble of clusterings $\mathcal{E}$, how to cluster the objects into K classes, without having, in principle, any other information about the objects we are going to cluster. To this end, we start from the assumption that objects can be softly assigned to clusters. Hence, the clustering problem consists in estimating, for each object $i \in O$, an unknown assignment $\mathbf{y}_i$, which is a probability distribution over the set of cluster labels $\{1, \ldots, K\}$, or, in other words, an element of the *standard simplex* Δ_K given by

$$\Delta_K = \{\boldsymbol{x} \in \mathbb{R}_+^K \,:\, \|\mathbf{x}\|_1 = 1\} \,,$$

where $\mathbb{R}_+$ is the set of nonnegative reals, and $\| \cdot \|_1$ is the ℓ^1-norm. The kth entry of $\mathbf{y}_i$ thus provides the probability of object i to be assigned to cluster k. Given the unknown cluster assignments $\mathbf{y}_i$ and $\mathbf{y}_j$ of objects i and j, respectively, and assuming independent cluster assignments, the probability of them to occur in a same cluster can be easily derived as $\mathbf{y}_i^\top \mathbf{y}_j$. Suppose now $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \in \Delta_K^n$ to be the matrix formed by stacking the $\mathbf{y}_i$'s, which in turn form the columns of Y. Then, the $n \times n$ matrix $Y^\top Y$ provides the co-occurrence probability of any pair of objects in O.

For each pair of objects i and j, let X_{ij} be a Bernoulli distributed random variable (r.v.) indicating whether objects i and j occur in a same cluster. Note that, according to our model, the mean (and therefore the parameter) of X_{ij} is $\mathbf{y}_i^\top \mathbf{y}_j$, i.e., the probability of co-occurrence of i and j. For each pair of objects i and j, we collect from the clusterings ensemble N_{ij} independent realizations $x_{ij}^{(p)}$ of X_{ij}, which are given by:

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

for $p \in \Omega_{ij}$. 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:

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

The matrix $C = (c_{ij})$, derived from the empirical probabilities of co-occurrence of any pair of objects, is known as the *co-association matrix* within the evidence accumulation-based framework for clustering [8,1]. Since C is the maximum likelihood estimate of $Y^\top Y$ given the observations from the clustering ensemble $\mathcal{E}$, we will refer to the former as the *empirical co-association matrix*, and to the latter as the *true co-association matrix*.

At this point, by minimizing the divergence, in a least-square sense, of the true co-association matrix from the empirical one, with respect to Y, we find a solution Y^* of the clustering problem. This leads to the following optimization problem:

$$\begin{aligned} Y^* = \arg\min \quad & \|C - Y^\top Y\|_F^2 \\ \text{s.t.} \quad & Y \in \Delta_K^n \,. \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm. Note that Y^* provides us with soft assignments of the objects to the K classes. Indeed, y^*_{ki} gives the probability of object i to be assigned to class k. If a hard partition is needed, this can be forced by assigning each object i to the highest probability class, which is given by: $\arg\max_{k=1\dots K}\{y^*_{ki}\}$. Moreover, by computing the entropy of each $\mathbf{y}_i$, we can obtain an indication of the uncertainty of the cluster assignment for object i.

3 Related Work

In [5] a similar approach is proposed for pairwise clustering. First of all, a preprocessing on the similarity matrix W looks for its closest doubly-stochastic matrix F under ℓ_1 norm, or Frobenius norm, or relative entropy [9]. The k-clustering problem is then tackled by finding a completely-positive factorization of $F = (f_{ij})$ in the least-square sense, i.e., by solving the following optimization problem:

$$\begin{aligned} G^* = \arg\min \quad & \|F - G^\top G\|_F^2 \\ \text{s.t.} \quad & G \in \mathbb{R}_+^{k\times n} \,. \end{aligned} \tag{2}$$

Note that this leads to an optimization program, which resembles (1), but is inherently different. The elements g_{ri} of the resulting matrix G provide an indication of object i to be assigned to class r. However, unlike our formulation, these quantities are not explicit probabilities and it may happen for instance that $g_{ri} = 0$ for all $r = 1\dots k$, i.e., some objects may remain in principle unclassified.

The approach proposed to find a local solution of (2) consists in iterating the following updating rule:

$$g_{ri} \leftarrow \frac{g_{ri}\sum_{j\neq i}^n g_{rj} f_{ij}}{\sum_{s=1}^k g_{si}\sum_{j\neq i}^n g_{sj} g_{rj}} \,.$$

The computational complexity for updating all entries in G once (complete iteration) is $O(kn^2)$, while we expect to find a solution in $O(\gamma kn^2)$, where γ is the average number of complete iterations required to converge. A disadvantage of this iterative scheme is that updates must be sequential, i.e., we cannot update all entries of G in parallel.

4 The Baum-Eagon Inequality

In the late 1960s, Baum and Eagon [6] introduced a class of nonlinear transformations in probability domain and proved a fundamental result which turns out to be very useful for the optimization task at hand. The next theorem introduces what is known as the Baum-Eagon inequality.

Theorem 1 (Baum-Eagon). *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:*

$$z_{ri} = x_{ri} \frac{\partial Q(X)}{\partial x_{ri}} \Big/ \sum_{s=1}^{k} x_{si} \frac{\partial Q(X)}{\partial x_{si}} , \tag{3}$$

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.

This result applies to homogeneous polynomials, however in a subsequent paper, Baum and Sell [10] proved that Theorem 1 still holds in the case of arbitrary polynomials with nonnegative coefficients, and further extended the result by proving that $\mathcal{M}$ increases Q homotopically, which means that for all $0 \leq \eta \leq 1$, $Q(\eta\mathcal{M}(X) + (1-\eta)X) \geq Q(X)$ with equality if and only if $\mathcal{M}(X) = X$.

The Baum-Eagon inequality provides an effective iterative means for maximizing polynomial functions in probability domains, and in fact it has served as the basis for various statistical estimation techniques developed within the theory of probabilistic functions of Markov chains [11]. It is indeed not difficult to show that, by starting from the interior of the simplex, the fixed points of the Baum-Eagon dynamics satisfy the first-order Karush-Kuhn-Tucker necessary conditions for local maxima and that we have a strict local solution in correspondence to asymptotically stable point.

5 The Algorithm

In order to use the Baum-Eagon theorem for optimizing (1) we need to meet the requirement of having a polynomial to maximize with nonnegative coefficients in the simplex-constrained variables. To this end, we consider the following optimization program, which is proved to be equivalent to (1):

$$\begin{array}{ll} \max & 2Tr(CY^\top Y) + \|Y^\top E_K Y\|^2 - \|Y^\top Y\|^2 \\ \text{s.t.} & Y \in \Delta_K^n , \end{array} \tag{4}$$

where E_K is the $K \times K$ matrix of all 1's, and $Tr(\cdot)$ is the matrix trace function.

Proposition 1. *The maximizers of* (4) *are minimizers of* (1) *and vice versa. Moreover, the objective function of* (4) *is a polynomial with nonnegative coefficients in the variables y_{ki}, which are elements of Y.*

Proof. Let $P(Y)$ and $Q(Y)$ be the objective functions of (1) and (4), respectively.

To prove the second part of the proposition note that trivially every term of the polynomial $\|Y^\top Y\|^2$ is also a term of $\|Y^\top E_K Y\|^2$. Hence, $Q(Y)$ is a polynomial with nonnegative coefficients in the variables y_{ki}.

As for the second part, by simple algebra, we can write $Q(Y)$ in terms of $P(Y)$ as follows:

$$\begin{aligned} Q(Y) &= \|C\|^2 - P(Y) + \|Y^\top E_K Y\|^2 \\ &= \|C\|^2 - P(Y) + 1\,, \end{aligned}$$

where we used the fact that $\|Y^\top E_K Y\| = 1$. Note that the removal of the constant terms from $Q(Y)$ leaves its maximizers over Δ_K^n unaffected. Therefore, maximizers of (4) are also maximizers of $-P(Y)$ over Δ_K^n and thus minimizers of (1). This concludes the proof.

By Proposition 1 we can use Theorem 1 to locally optimize (4). This allows us to find a solution of (1). Note that, in our case, the objective function is not a homogeneous polynomial but, as mentioned previously, this condition is not necessary [10]. By applying (3), we obtain the following updating rule for $Y = (y_{ki})$:

$$y_{ki}^{(t+1)} = y_{ki}^{(t)} \frac{n + [Y(C - Y^\top Y)]_{ki}}{n + \sum_k y_{ki}^{(t)} [Y(C - Y^\top Y)]_{ki}}\,, \tag{5}$$

where we abbreviated $Y^{(t)}$ with Y and any non-constant iteration of (5) strictly decreases the objective function of (1).

The computational complexity of the proposed dynamics is $O(\gamma k n^2)$, where γ is the average number of iterations required to converge (note that in our experiments we kept γ fixed). One remarkable advantage of this dynamics is that it can be easily parallelized in order to benefit from modern multi-core processors. Additionally, it can be easily implemented with few lines of Matlab code.

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. Additionally, we considered also the image-complex synthetic data-set, shown in figure 1. 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 [12]: Single Link (SL), Complete Link (CL), Average Link (AL) and K-means (KM).

Table 1 summarizes the experimental setting that has been considered. For each data-set, we report the optimal number of clusters K and the size n of the data-set, respectively. As for the ensemble, each algorithm was run several times in order to produce clusterings with different number of classes, K_i. For each clustering approach and each parametrization of the same we generated $N = 100$ different subsampled versions of the data-set.

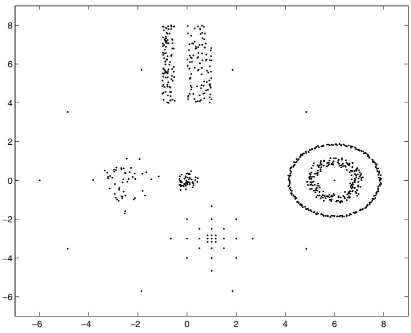


Fig. 1. Image Complex Synthetic data-set

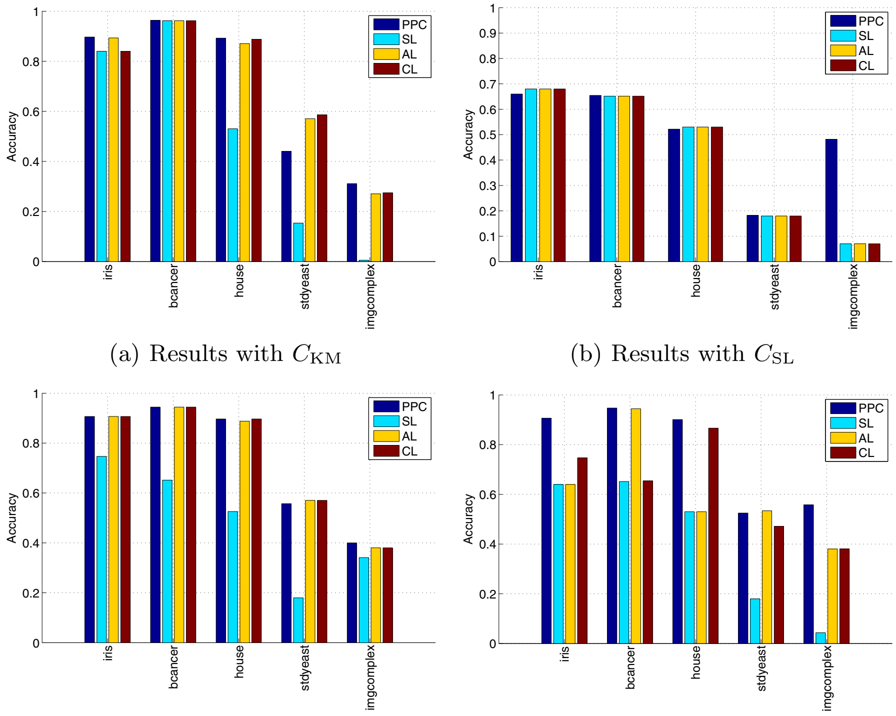
Table 1. Benchmark data-sets and parameter values used with different clustering algorithms (see text for description)

Data-Sets	K	n	Ensemble
			K_i
iris	3	150	3-10,15,20
house-votes	2	232	2-10,15,20
std-yeast-cell	5	384	5-10,15,20
breast-cancer	2	683	2-10,15,20
image-complex	8	1000	8-15,20,30, 37

Once all the clusterings have been generated, we grouped them by algorithm into several *base ensembles*, namely $\mathcal{E}_{\text{SL}}$, $\mathcal{E}_{\text{AL}}$, $\mathcal{E}_{\text{CL}}$ and $\mathcal{E}_{\text{KM}}$. Moreover, we created a large ensemble $\mathcal{E}_{\text{All}}$ 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 the agglomerative hierarchical algorithms SL, AL and CL. Each method was provided with the optimal number of classes as input parameter.

Figure 2 summarizes the results obtained over the benchmark data-sets. The performances are assessed in terms of accuracy, i.e., the percentage of correct labels. When we consider the base ensembles, i.e., $\mathcal{E}_{\text{SL}}$, $\mathcal{E}_{\text{AL}}$, $\mathcal{E}_{\text{CL}}$ and $\mathcal{E}_{\text{KM}}$, on average our approach achieves the best results, although other approaches, such as the AL, perform comparably well. Our algorithm, however, outperforms the competitors when we take the union $\mathcal{E}_{\text{All}}$ of all the base ensembles into account. Interestingly, the results obtained by PPC on the combined ensemble are as good as the best one obtained in the base ensembles and, in some cases like the image-complex dataset, they are even better.

The different levels of performance obtained by the several algorithms over the different clustering ensembles, as shown in Figures 2(a) to 2(d), are illustrative of the distinctiveness between the underlying clustering ensembles, and the diversity of clustering solutions. It is then clear that the ensemble $\mathcal{E}_{\text{All}}$ has



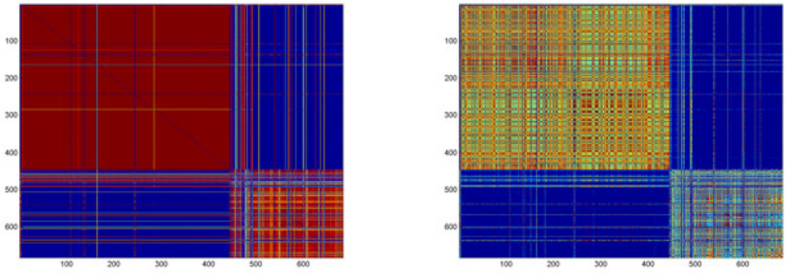
(a) Results with C_{KM}

(b) Results with C_{SL}

(c) Results with C_{AL}

(d) Results with C_{All}

Fig. 2. Experiments on benchmark data-sets

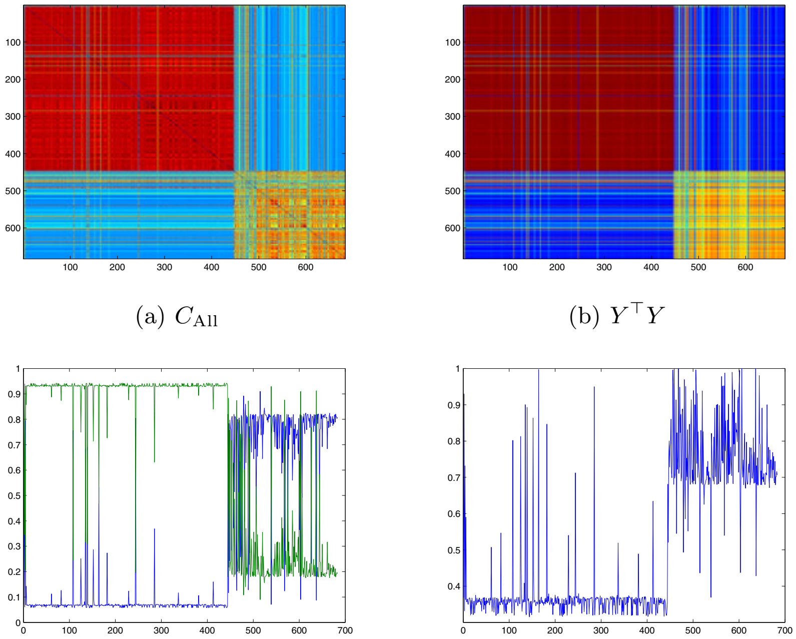


(a) C_{AL}

(b) C_{KM}

Fig. 3. Co-association matrices with ensembles $\mathcal{E}_{AL}$ and $\mathcal{E}_{KM}$

the largest diversity when compared to the individual ensembles; this is quantitatively confirmed when computing average pairwise consistency values between partitions in the individual CEs and the one resulting by the merging of these. This higher diversity causes the appearance of noisy-like structure in the



(a) C_{All}

(b) $Y^{\top}Y$

(c) Y

(d) Uncertainty

Fig. 4. Results on the breast-cancer data-set

co-association matrices. This is illustrated in Figures 3(a) and 3(b) corresponding to the co-association matrices C_{AL} and C_{KM}, respectively, when compared to the C_{All} in Figure 4(a). The better performance of the PPC algorithm on the latter CE, can be attributed to a leveraging effect over these local noisy estimates, thus better unveiling the underlying structure of the data. This is illustrated next.

Figures 4(a) and 4(b) show the empirical co-association matrix C_{All} and the true one, respectively, for the breast-cancer data-set. While the block structure of two clusters is apparent in both figures, we can see that the true co-association turns out to be less noisy than the empirical one. In Figure 4(c) we plot the soft cluster assignments, Y. Here, object indices are on the x-axis, and probabilities are on the y-axis, each curve representing the profile of a cluster. As one can see from the cluster memberships, the two clusters can be clearly evinced, although there is a higher level of uncertainty in the assignments of objects belonging to the smallest cluster. Indeed, this can also be seen in Figure 4(d), where we plot the uncertainty h_i in the cluster assignments, which is computed for each object i as the normalized entropy of $\mathbf{y}_i$, i.e.,

$$h_i = -\frac{\sum_{k=1}^{K} y_{ki} \log(y_{ki})}{\log(K)} .$$

7 Conclusion

In this paper we introduced a new approach for consensus clustering. 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 propose 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. Experiments on both synthetic and real benchmarks assess the effectiveness of our approach.

Acknowledgement

We acknowledge financial support from the FET programme within the EU FP7, under the SIMBAD project (contract 213250). This work was partially supported by the Portuguese Foundation for Science and Technology (FCT), Portuguese Ministry of Science and Technology, under grant PTDC/ EIACCO/ 103230/ 2008.

References

1. Fred, A., Jain, A.K.: Combining multiple clustering using evidence accumulation. IEEE Trans. Pattern Anal. Machine Intell. 27(6), 835–850 (2005)
2. Jardine, N., Sibson, R.: The construction of hierarchic and non-hierarchic classifications. Computer J. 11, 177–184 (1968)
3. Banerjee, A., Krumpelman, C., Basu, S., Mooney, R.J., Ghosh, J.: Model-based overlapping clustering. In: Int. Conf. on Knowledge Discovery and Data Mining, pp. 532–537 (2005)
4. Heller, K., Ghahramani, Z.: A nonparametric bayesian approach to modeling overlapping clusters. In: Int. Conf. AI and Statistics (2007)
5. Zass, R., Shashua, A.: A unifying approach to hard and probabilistic clustering. In: Int. Conf. Comp. Vision (ICCV), vol. 1, pp. 294–301 (2005)
6. Baum, L.E., Eagon, J.A.: 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)
7. Fred, A., Jain, A.K.: Learning pairwise similarity for data clustering. In: Int. Conf. Patt. Recogn. (ICPR), pp. 925–928 (2006)
8. Fred, A., Jain, A.K.: Data clustering using evidence accumulation. In: Int. Conf. Patt. Recogn. (ICPR), pp. 276–280 (2002)
9. Zass, R., Shashua, A.: Doubly stochastic normalization for spectral clustering. In: Adv. in Neural Inform. Proces. Syst (NIPS), vol. 19, pp. 1569–1576 (2006)
10. Baum, L.E., Sell, G.R.: Growth transformations for functions on manifolds. Pacific J. Math. 27, 221–227 (1968)
11. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Statistics 41, 164–171 (1970)
12. Jain, A.K., Dubes, R.C.: Algorithms for data clustering. Prentice-Hall, Englewood Cliffs (1988)