2010 International Conference on Pattern Recognition

Probabilistic Clustering using the Baum-Eagon Inequality

Samuel Rota Bulò and Marcello Pelillo
DSI - University of Venice - Italy
{srotabul,pelillo}@dsi.unive.it

Abstract

The paper introduces a framework for clustering data objects in a similarity-based context. The aim is to cluster objects into a given number of classes without imposing a hard partition, but allowing for a soft assignment of objects to clusters. Our approach uses the assumption that similarities reflect the likelihood of the objects to be in a same class in order to derive a probabilistic model for estimating the unknown cluster assignments. This leads to a polynomial optimization in probability domain, which is tackled by means of a result due to Baum and Eagon. Experiments on both synthetic and real standard datasets show the effectiveness of our approach.

1. Introduction

Clustering is the unsupervised learning task of organizing a set of data objects (or simply objects) into groups. Commonly, clustering methods work under the assumption that objects are explicitly described in terms of features. However, a more challenging and appealing trend, which has became popular in the last few years, considers a similarity-based scenario, where the information about the objects to be clustered is expressed in terms of their *similarities*.

Unfortunately, the clustering problem is ill-posed, as there is no commonly accepted notion of a cluster. Indeed, there is a large variety of approaches, which tackle this problem by making more or less restrictive assumptions about the result they are aiming to. The most popular one forces objects to be clustered into a fixed number k of classes and, typically, this is also coupled with the requirement that each data object belongs to a single cluster, yielding a hard partition of the data. This last assumption is, however, too restrictive for many important applications such as clustering micro-array gene expression data, text categorization, perceptual grouping, labeling of visual scenes and medical diagnosis.

Inspired by a recent work due to Zass and Sashua [8], we introduce a probabilistic framework for clustering in a similarity-based context. The aim is to cluster objects into a given number of classes without forcing crisp partitions, but allowing for soft assignments of objects to clusters. To this end, we first design a statistical model for the similarities parametrized by the unknown cluster assignments. We derive the model from the assumption that the similarity between two objects follows a Gaussian distribution centered around the likelihood of them to occur in a same cluster, which in turn depends on the unknown cluster assignments of the two objects. We use then the model to estimate the unknown parameters from the similarities by adopting a maximum likelihood approach. This reduces the clustering problem to a polynomial optimization in probability domain, which is solved by means of the Baum-Eagon inequality [1]. This result, indeed, provides us with a class of nonlinear transformations that serve our purpose. Experiments conducted on both synthetic and real standard datasets show the effectiveness of our approach.

2. A probabilistic model for k-clustering

Let $O = \{1, \dots, n\}$ be a set of data objects (or simply objects) to cluster into k classes and consider a scenario in which objects are not explicitly described in terms of feature vectors, but a $n \times n$ nonnegative real matrix $W = (w_{ij})$ is given, whose entries provide a measure of the likelihood that two objects occur in the same cluster.

In this paper we take a probabilistic perspective by allowing objects to belong to mixtures of clusters, i.e., cluster memberships are discrete distributions over the set $\{1, \dots, k\}$ of clusters or, technically, points of the *standard simplex* Δ_k, which is given by

$$\Delta_k = \left\{ \mathbf{x} \in \mathbb{R}^k_+ : \|\mathbf{x}\|_1 = 1 \right\} .$$

Let $\mathbf{y}_1, \dots, \mathbf{y}_n \in \Delta_k$ be the unknown cluster memberships of the objects in O. Then the likelihood of objects i and j to be clustered together (under independence assumption) is given by $\alpha \mathbf{y}_i^\top \mathbf{y}_j$, where α is a positive real value. Suppose also $Y = (\mathbf{y}_1, \dots, \mathbf{y}_n) \in \Delta_k^n$ to be the matrix obtained by stacking the $\mathbf{y}_i$'s in a row. Then, the matrix with the "true" likelihoods is given by $\alpha Y^\top Y$.

We want now to find the values of Y and α such that the true likelihoods $\alpha Y^\top Y$ best represent the empirical ones W. To this end, suppose that each entry w_{ij} has a measurement error that is independently random and Gaussian distributed around the "true" likelihood $\alpha \mathbf{y}_i^\top \mathbf{y}_j$. Assume also that the standard deviations σ are

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.353

the same for all these distributions. Then we can estimate α, Y and σ^2 from W by maximizing the likelihood of the data W, which is given by

$$\begin{aligned} p(W|Y,\alpha,\sigma^2) &= \prod_{i,j} p(w_{ij}|\mathbf{y}_i,\mathbf{y}_j,\alpha,\sigma^2) \\ &= \prod_{i,j} \mathcal{N}(w_{ij}|\alpha\mathbf{y}_i^\top\mathbf{y}_j,\sigma^2) \end{aligned} \tag{1}$$

where $\mathcal{N}(x|\mu,\sigma^2)$ denotes the Gaussian probability density function with mean μ and variance σ^2. Since maximizing (1) is the same as minimizing its negative logarithm, we obtain the following minimization problem

$$\begin{aligned} \min \quad & \frac{1}{\sigma^2}\|W-\alpha Y^\top Y\|^2 + n^2 \log\sigma \\ \text{s.t.} \quad & Y \in \Delta_k^n,\ \alpha,\sigma \in \mathbb{R}\,. \end{aligned} \tag{2}$$

With the view of recovering only the cluster memberships Y, we note that the value of the variance σ^2 in (2) does not affect the value of the optimal Y. Hence, σ can be discarded and thereby (2) can be simplified as follows:

$$\begin{aligned} \min \quad & \|W-\alpha Y^\top Y\|^2 \\ \text{s.t.} \quad & Y \in \Delta_k^n, \quad \alpha \in \mathbb{R}\,. \end{aligned} \tag{3}$$

Note that the optimal solution in the variables Y and α of (3) and (2) are the same. Moreover, the value of Y provides us with *soft assignments* of the objects to the k classes. Indeed, y_{ri} gives the probability of object i to be assigned to class r. If a hard partition is needed, this can be forced by assigning each object i to the most probable class, which is given by: $\arg\max_{r=1\dots k} y_{ri}$.

3. Related works

In [8] a similar approach is proposed. 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:

$$\begin{aligned} \min \quad & \|F-G^\top G\|^2 \\ \text{s.t.} \quad & G \in \mathbb{R}_+^{k\times n}\,. \end{aligned} \tag{4}$$

Note that this leads to an optimization program, which resembles (3), 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 (4) 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 [1] 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:*

$$z_{ri} = x_{ri}\frac{\partial Q(X)}{\partial x_{ri}} \Bigg/ \sum_{s=1}^k x_{si}\frac{\partial Q(X)}{\partial x_{si}}\,, \tag{5}$$

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 [3] 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 [2].

5. The algorithm

In order to use the Baum-Eagon theorem for optimizing (3), assuming α fixed, 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 (3):

$$\begin{aligned} \max \quad & 2Tr(WY^\top Y) + \alpha(\|Y^\top E_k Y\|^2 - \|Y^\top Y\|^2) \\ \text{s.t.} \quad & Y \in \Delta_k^n \,. \end{aligned} \tag{6}$$

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* (6) *are minimizers of* (3) *(assuming $\alpha > 0$ fixed) and vice versa. Moreover, the objective function of* (6) *is a polynomial with nonnegative coefficients in the variables y_{ri}, which are elements of Y.*

Proof. Let $P(Y)$ and $Q(Y)$ be the objective functions of (3) and (6), 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_{ri}.

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

$$\begin{aligned} Q(Y) &= \alpha^{-1}\left[\|W\|^2 - P(Y)\right] + \alpha\|Y^\top E_k Y\|^2 \\ &= \alpha^{-1}\left[\|W\|^2 - P(Y)\right] + \alpha \\ &= -\alpha^{-1}P(Y) + \alpha^{-1}\|W\|^2 + \alpha \,, \end{aligned}$$

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 (6) are also maximizers of $-\alpha^{-1}P(Y)$ over Δ_k^n and thus minimizers of (3). This concludes the proof. □

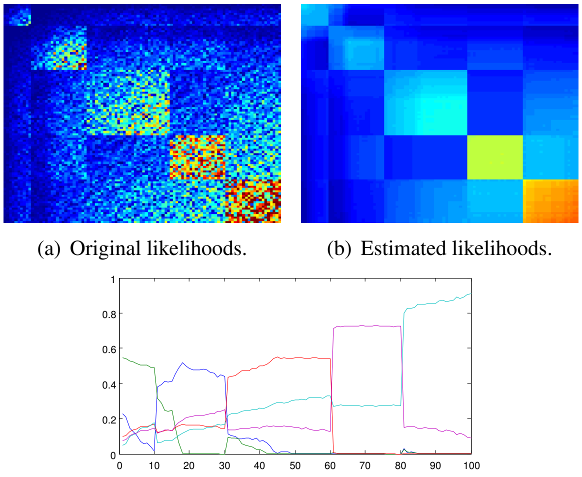
By Proposition 1, assuming $\alpha > 0$ fixed, we can use Theorem 1 to locally optimize (6). The same result guarantees that by so doing we find a solution of (3). Note that, in our case, the objective function is not a homogeneous polynomial but, as mentioned previously, this condition is not necessary [3]. By applying (5), we obtain the following updating rule for $Y = (y_{ri})$:

$$y_{ri}^{(t+1)} = y_{ri}^{(t)} \frac{\alpha n + [Y(W - \alpha Y^\top Y)]_{ri}}{\alpha n + \sum_r y_{ri}^{(t)}[Y(W - \alpha Y^\top Y)]_{ri}} \tag{7}$$

where we abbreviated $Y^{(t)}$ with Y.

The updating rule for the scaling factor α, assuming $Y = Y^{(t+1)}$ fixed, can be derived from (3) by zeroing the first order derivative of the cost function with respect to α, obtaining:

$$\alpha^{(t+1)} = \frac{Tr\left(WY^\top Y\right)}{\|Y^\top Y\|^2} \,, \tag{8}$$



(a) Original likelihoods.

(b) Estimated likelihoods.

(c) Mermberships.

Figure 1. Results on the S-Block dataset.

which is always nonnegative. By iteratively updating Y and α we are able to locally optimize (3). Indeed, any non-fixed iteration of (8) or (7) strictly decreases the objective function of (3).

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.

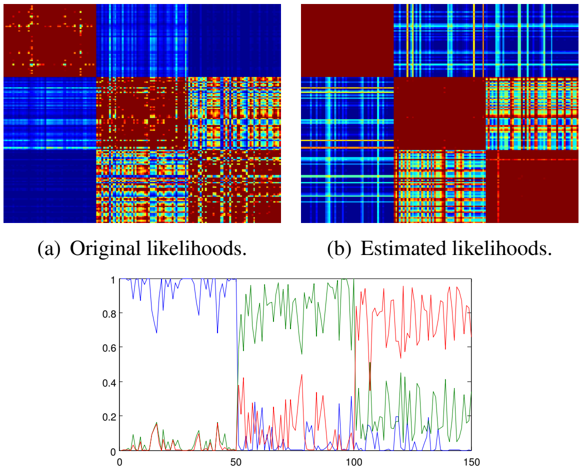
6. Experimental results

We performed experiments on the S100 Block Stochastic (S-Block) synthetic dataset [7], the Iris dataset (Iris), the NIST Handwritten Digits dataset (Digit) and a subset of the SCOP protein dataset (Scop) [4]. We compared our approach based on the Baum-Eagon inequality (BE) against the Copositive Factorization (CP) method [8], which has been described in Section 3, the Normalized Cuts (NCut) [6] and the Ng-Jordan-Weiss (NJW) [5] spectral clustering approaches. Each approach has been executed 10 times and average results in terms of accuracy have been reported.

The qualitative results obtained are shown in Table 1. Our approach outperformed the competitors in the most challenging datasets, namely Digit and Scop. In the Iris dataset all approaches performed comparably well, while in the synthetic one NCut and BE outperformed the other approaches, the latter achieving a slightly lower accuracy than the former.

Dataset	k	n	BE	CP	NCut	NJW
S-Block	5	100	.995±.013	.556±.143	**1.000**±.000	.596±.010
Digit	10	1000	**.700**±.042	.430±.142	.566±.096	.657±.000
Iris	3	150	**.993**±.000	.991±.003	**.993**±.000	**.993**±.000
Scop	5	451	**.706**±.001	.703±.000	.568±.000	.630±.000

Table 1. Clustering results on different datasets.



(a) Original likelihoods.

(b) Estimated likelihoods.

(c) Mermberships.

Figure 2. Results on the Iris dataset.

Figure 1(b) shows the likelihoods $\alpha Y^\top Y$, estimated by our approach on the S-Block dataset. Noise has been considerably smoothed out and the block structure is now well-oulined. In Figure 1(c) we plotted also the cluster memberships of each object, i.e, matrix Y. Here, object indices are on the x-axis and probabilities on the y-axis, and each curve represents the profile of a cluster. As one can see, from the memberships the true cluster assignments can be clearly evinced.

In Figure 2 we present an analoguos analysis on the Iris dataset. This dataset consists of three clusters, two of which are not clearly separated. Our approach is effective also in this case, as from both the estimated likelihoods $\alpha Y^\top Y$ in Figure 2(b) and the cluster memberships in Figure 2(c), the three clusters can be clearly recognized.

7. Conclusion

We introduced a probabilistic framework for clustering in a similarity-based setting. The aim is to cluster objects into a given number of classes, but as opposed to conventional approaches, which induce a hard partition of the data, our algorithm provides soft assignments of objects to clusters. Our approach is based on the reasonable assumption that similarities reflect the likelihood of the objects to be in a same class in order to derive a probabilistic model for estimating the unknown cluster assignments. This reduces the clustering problem to a polynomial optimization in probability domain, which is addressed using the Baum and Eagon inquality. Experiments on both synthetic and real standard datasets show the effectiveness of our approach.

Acknowledgements

We acknowledge financial support from the FET programme within the EU FP7, under the SIMBAD project (contract 213250).

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, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statistics*, 41:164–171, 1970.

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

[4] T. J. P. Hubbard, A. G. Murzin, S. E. Brenner, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Molec. Biology*, 247:536–540, 1995.

[5] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. *Adv. in Neural Inform. Proces. Syst. (NIPS)*, pages 849–856, 2001.

[6] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Machine Intell.*, 22:888–905, 2000.

[7] D. Verma and M. Meila. Comparison of spectral clustering methods. Technical report, University of Washington, 2003.

[8] R. Zass and A. Shashua. A unifying approach to hard and probabilistic clustering. In *Int. Conf. Comp. Vision (ICCV)*, volume 1, pages 294–301, 2005.

[9] R. Zass and A. Shashua. Doubly stochastic normalization for spectral clustering. *Adv. in Neural Inform. Proces. Syst. (NIPS)*, 19:1569–1576, 2006.