On the Scalability of Evidence Accumulation Clustering

André Lourenço*†‡, Ana L. N. Fred†‡ and Anil K. Jain§
*Instituto Superior de Engenharia de Lisboa
†Instituto Superior Técnico
‡Instituto de Telecomunicações, Lisboa, PORTUGAL
§Michigan State University, USA
alourenco@deetc.isel.ipl.pt afred@lx.it.pt jain@msu.edu

***Abstract*—This work focuses on the scalability of the Evidence Accumulation Clustering (EAC) method. We first address the space complexity of the co-association matrix. The sparseness of the matrix is related to the construction of the clustering ensemble. Using a split and merge strategy combined with a sparse matrix representation, we empirically show that a linear space complexity is achievable in this framework, leading to the scalability of EAC method to clustering large data-sets.**

***Keywords*-Cluster analysis; combining clustering partitions; cluster fusion, evidence accumulation; large data-sets.**

I. Introduction

Clustering combination techniques are a recent and promising trend in clustering [1], [2], [3], [4], [5], [6]. Combining the information provided by a set of N different partitions (the *clustering ensemble* (CE) - $\mathbb{P}$) of a given data set, clustering combination results typically outperform the result of a single clustering algorithm, achieving better and more robust partitioning of the data.

The Evidence Accumulation Clustering (EAC) method, proposed by Fred and Jain [1], [2], seeks to find consistent data partitions by considering pair-wise relationships. The method can be decomposed into three major steps: (a) construction of the clustering ensemble, $\mathbb{P}$; (b) ensemble combination, through evidence accumulation; and (c) extraction of the final partition.

In the combination step (b), the clustering ensemble, $\mathbb{P}$, is transformed into a learned pair-wise similarity, summarized in a $n_s \times n_s$ co-association matrix, $\mathcal{C}$

$$\mathcal{C}(i,j) = \frac{n_{ij}}{N}, i,j \in 1,\ldots,N, \qquad (1)$$

where n_s is the number of objects to be clustered, and n_{ij} represents the number of times a given object pair (i,j) is placed in the same cluster over the N partitions of the ensemble.

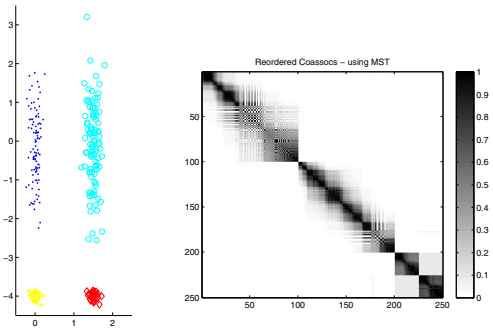
In order to recover the "natural"clusters, a clustering algorithm is applied to the learned similarity matrix, $\mathcal{C}$, yielding the combined data partition, P^*. Although it is mostly the hierarchical agglomerative methods that have been applied in step (c) [2], any clustering algorithm can be used, either taking $\mathcal{C}$ as a pair-wise similarity matrix, or deriving a feature space from it using multi-dimensional scaling (MDS).

EAC is a powerful and robust method, but direct or naive implementation of its basic steps can, however, limits the scalability of the EAC method, namely due to the $O(n_s^2)$ space complexity required to store the co-association matrix [6], [5].

In this paper we address the scalability of EAC, theoretically analyzing the method from a space complexity perspective. We propose: (1) a compact representation of the co-association matrix, $\mathcal{C}$, exploring its intrinsic sparseness; and (2) guidelines for the construction of the clustering ensemble $\mathbb{P}$, that further increases the sparseness of $\mathcal{C}$, leading to an overall split and merge strategy for the EAC. Experimental results, on several benchmark data-sets, confirm that this strategy leads to a linear space complexity. We show that this significant space complexity improvement does not compromise, and may even lead to increased performance of clustering combination results.

II. Co-association Matrix Representation

Typically, the co-association matrix, $\mathcal{C}$, generated by the EAC method, is very sparse. This is illustrated with the synthetic 2D cigar data set (figure 1(a)), and the corresponding $\mathcal{C}$ matrix (figure 1(b)). The color scheme ranges from white ($\mathcal{C}(i,j) = 0$) to black ($\mathcal{C}(i,j) = 1$), corresponding to the magnitude of similarity.



(a) Cigar data. (b) Co-association Matrix.

Figure 1. Synthetic 2-D data set (a) with $n_s = 250$ objects, and the corresponding co-association matrix (b).

Under the *working hypothesis* of well separated and

balanced clusters, the structure of the co-association matrix resembles a perfect block diagonal matrix, where each block corresponds to a cluster, and the number of co-associations (non-zero elements of the matrix $\mathcal{C}$), is given by:

$$N_{assocs} = \sum_{k=1}^{K} (n_{sk})^2, \tag{2}$$

where K is the number of "natural"clusters in the data-set, and n_{sk} is the number of samples in cluster k.

Taking into account the symmetric nature of this matrix, and the fact that the principal diagonal elements have value 1, the required elements that need to be retained consist only of the upper (or lower) triangular matrix, with a total number of co-associations given by:

$$N_{assocs\triangle} = \sum_{k=1}^{K} n_{sk} \times (n_{sk} - 1)/2. \tag{3}$$

We propose to use a **sparse representation** of the co-association matrix, $\mathcal{C}$, storing only the upper triangular non-zero elements of this matrix, substantially reducing the space complexity of the method and making it more attractive for large data sets.

III. Building CEs: Split and Merge Strategy

The sparseness of a matrix can be quantified by its *density*, or normalized ℓ^0 norm, defined by $\| \mathcal{C} \|_0 = nnz/n_s^2$, where nnz is the number of non-zero elements. The density of a perfect K block diagonal matrix $\mathcal{C}$ is given by

$$\| \mathcal{C} \|_0 = \frac{\sum_{k=1}^{K} (n_{sk})^2}{(n_s)^2} \tag{4}$$

Considering balanced clusters, each cluster has $\frac{n_s}{K}$ elements, and the density becomes $\| \mathcal{C} \|_0 = \frac{1}{K}$. Empirically, the value of $\| \mathcal{C} \|_0$ becomes smaller than $1/K$, as the number of co-associations becomes less than N_{assocs}. This number depends on the strategy used for generating the clustering ensemble.

The splitting of "natural"clusters into smaller clusters induces micro-blocks (smaller than the perfect block diagonal structures) in the $\mathcal{C}$ matrix, resulting in an increased sparseness (lower density). In order to achieve this, we propose the following strategy for building clustering ensembles:

CE construction rule: Apply several clustering algorithm(s) with many different values of K, the number of clusters in each partition of the clustering ensemble; K is randomly chosen in an interval $[K_{min}, K_{max}]$.

A large value of K_{min}, in addition to inducing high granularity partitioning and consequently reduced space complexity, is important in order to prevent the existence of clusters in the CE with samples from different "natural"clusters. Overall, this follows a split & merge strategy [2], with the split "natural"clusters being combined in $\mathcal{C}$ during the combination step (b) of the EAC method; they are eventually recovered during the merging step (c) produced by clustering the matrix $\mathcal{C}$. One possible choice for K_{min} is to base it on the minimum number of gaussians in a gaussian mixture decomposition of the data [7].

We propose and analyze two alternative criteria for determining $\{K_{min}, K_{max}\}$, as a function of n_s, the number of samples in the data set:

(A) Sqrt: $\{K_{min}, K_{max}\} = \{\lceil \sqrt{n_s}/2 \rceil, \lceil \sqrt{n_s} \rceil\}$;

(B) Linear: $\{K_{min}, K_{max}\} = \{\lceil n_s/A \rceil, \lceil n_s/B \rceil\}$, with $A > B$

The number of non-zero elements in $\mathcal{C}$ is related to the number of associations within each "natural"cluster over each partition of the ensemble, i.e., the partitioning granularity. According to the working hypothesis of well separated and balanced clusters, each cluster C_m, from a data partition with K clusters, should have $n_{sm} = n_s/K$ objects, contributing to $(n_{sm})^2$ entries in $\mathcal{C}$; overall, a single partition produces $K(n_{sm})^2 = (n_s)^2/K$ non-zero values in $\mathcal{C}$. Over the N partitions of the clustering ensemble, a random partitioning of the "natural"clusters leads to the construction of partially overlapping clusters. Notice that shared elements of two overlapping clusters produce exactly the same co-associations in the matrix $\mathcal{C}$; new entries in the co-association matrix are the result of the non-overlapping elements. The density of $\mathcal{C}$ is thus larger for smaller values of K (with K_{min} giving the minimum value) and lower cluster overlap. On average, we consider that the overall contribution of the clustering ensemble (including unbalanced clusters) duplicates the co-associations produced in a single balanced clustering with K_{min} clusters, leading to the following estimate of the number of associations using the proposed clustering ensemble construction rule:

$$N_{assocs\_S\&M} = \frac{2(n_s)^2}{K_{min}} \overset{\text{(SqrtT)}}{=} 4n_s\sqrt{n_s} \tag{5}$$

$$\overset{\text{(LinearT)}}{=} 2A \cdot n_s \tag{6}$$

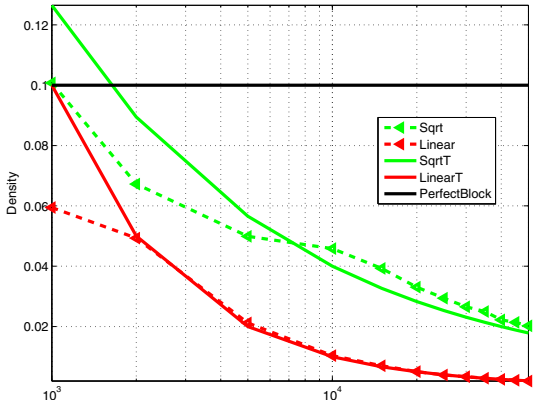
where (5) and (6) represent, respectively, the estimates for the criteria (A) and (B), the latter corresponding to linear space complexity. The corresponding estimated densities are $\| \mathcal{C} \|_0 = \frac{4}{\sqrt{n_s}}$ and $\| \mathcal{C} \|_0 = \frac{2A}{n_s}$.

For the sake of simplicity, in the next section we illustrate and evaluate this strategy using K-means clusterings for constructing the ensemble.

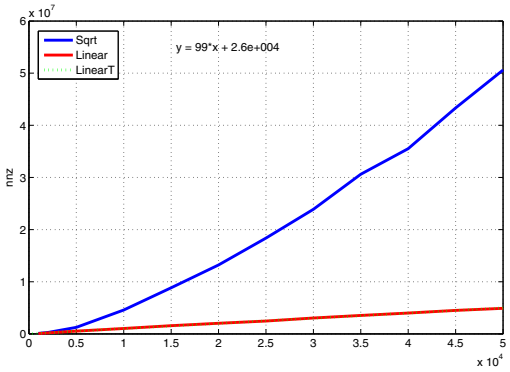
IV. Experimental Evaluation

To illustrate the performance of the proposed sparse representation, consider a mixture of Gaussians composed

of 10 2D-gaussians, with equal number of samples ($n_s/10$), means $\mu_i = [0, 12i]$ and covariances $\Sigma_i = [1, 0; 0, 1]$, with n_s in the interval $[10^3; 5 \times 10^4]$. For each n_s, we create a clustering ensemble with N=150 partitions, produced using the K-means algorithm with random number of clusters, following the criteria proposed in section III (with A=50, and B=20). Figure 2 plots the evolution of the density of $\mathcal{C}$ and the number of non-zero elements in $\mathcal{C}$, as a function of the number of samples (n_s), for criteria (A) Sqrt and (B) Linear.



(a) Density of co-association matrix (n_s is in logarithmic scale).



(b) Number of Non-Zero Elements.

Figure 2. Density (a) and number of non-zero elements (b) of co-association matrices as a function of n_s for a mixture of gaussians.

As shown in figure 2(a), both criteria lead to a decrease in density as n_s increases. Theoretical estimates (curves SqrtT and LinearT) seem reasonable for large n_s values, providing a good match with the corresponding experimental results (curves Sqrt and Linear), in particular for the Linear criterion (B). Both criteria lead to experimental density values far below the curve PerfectBlock, corresponding to the density of a perfect block diagonal matrix, as per equation (4).

Figure 2(b) plots the number of non-zero elements of $\mathcal{C}$ as a function of n_s. The line LinearT represents a linear regression over the empirical results using criteria (B). The

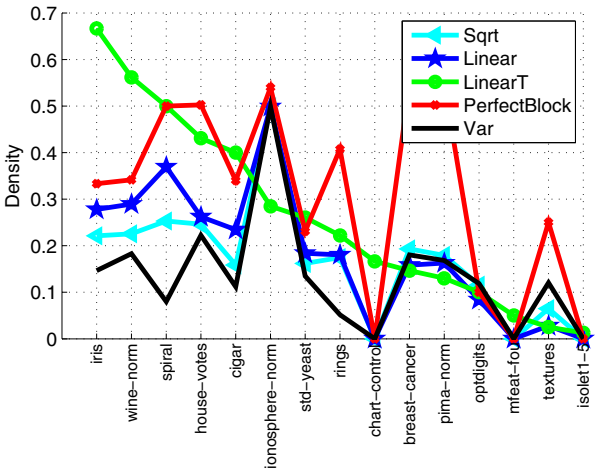


Figure 3. Density of co-association matrices for benchmark data-sets.

plot is consistent with the proposed theoretical estimate, reinforcing the observation that the use of criteria (B) enables linear space complexity.

Table I characterizes several benchmark data-sets (synthetic and other from the UCI repository [8]), values of k_{min} and k_{max} used for building clustering ensembles, and accuracy of corresponding combined partitions using the EAC method (columns CI – represent the Consistency Index [1], obtained by matching the clusters in the combined partition with the ground truth labels, corresponding to the percentage of correct labeling). In addition to criteria (A) and (B), the columns (Var) represent K intervals with $k_{min} \geq 10$, ensuring that k_{min} is always larger than the minimum number of components in a gaussian mixture decomposition [7].

Figure 3 presents the densities of the co-association matrices obtained for these data-sets for several clustering ensemble building criteria. By cross analysis of table I and figure 3, we can observe that higher k_{min} values lead to lower density. As a result, the criterion Var achieves lower densities for most data sets, corresponding to situations of higher k_{min} values than those produced by the other criteria. It is also evident that the proposed criteria induce lower densities, when compared to the PerfectBlock curve. Experimental results with criterion (B) in these data sets show typically lower densities than the theoretical estimate, LinearT.

In table I, for each data set we have marked maximal k_{min} and CI values. Analysis of these results show that the granularity of the clustering ensemble, dictated by the value of k_{min}, positively influences the quality of the clustering results. In general, higher k_{min} values do not compromise and may even lead to higher CI values.

Data-Sets	K	n_s	Var			(A)			(B)		
			k_{min}	k_{max}	CI	k_{min}	k_{max}	CI	k_{min}	k_{max}	CI
iris	3	150	10	20	0.91	6	12	0.84	3	8	0.84
wine-norm	3	178	10	30	0.96	7	13	0.96	4	9	0.97
spiral	2	200	20	30	0.71	7	14	0.57	4	10	0.54
house.votes	2	232	10	30	0.93	8	15	0.88	5	12	0.88
cigar	4	250	10	30	0.71	8	16	0.71	5	13	0.71
ionosphere.norm	2	351	10	30	0.64	9	19	0.64	7	18	0.64
std-yeast	5	384	10	30	0.69	10	20	0.66	8	19	0.68
rings	3	450	20	50	1.00	11	21	0.60	9	23	0.72
chart-synthetic-control	10	600	13	33	0.57	13	21	0.57	12	30	0.54
breast-cancer	2	683	10	30	0.97	13	26	0.97	14	34	0.97
pima-norm	2	768	10	30	0.65	14	28	0.65	15	38	0.65
optdigits-1000	10	1000	10	30	0.79	16	32	0.80	20	50	0.84
mfeat-fou	4	2000	40	60	0.39	23	45	0.39	40	100	0.39
textures	4	4000	10	30	0.90	32	63	0.97	80	200	0.91
isolet1-5	26	7797	156	176	0.60	44	88	0.61	156	390	0.60

Table I
BENCHMARK DATA-SETS AND CLUSTERING RESULTS, IN TERMS OF CONSISTENCY INDEX, CI, USING THE AVERAGE LINK HIERARCHICAL CLUSTERING TO OBTAIN THE FINAL PARTITION. VAR: MIXTURE OF GAUSSIANS; A: SQRT CRITERION; B: LINEAR CRITERION.

V. CONCLUSIONS

We have addressed the scalability problem of the evidence accumulation clustering method, intrinsically related to the storage of the co-association matrix. Taking advantage of the sparseness of this matrix, we adopted a sparse matrix representation, reducing the space complexity of the method.

In order to further reduce the space complexity, we have proposed a clustering ensemble construction rule, following a split and merge strategy, according to which the clustering algorithms are applied with K, the number of clusters, randomly chosen in the interval $[K_{min}, K_{max}]$. Criteria for the choice of these extreme values were also proposed and analyzed, showing that both space complexity and quality of combination results dependent on the partitioning granularity, dictated by the value of K_{min}.

Experimental results confirm that this strategy leads to linear space complexity of evidence accumulation clustering on several benchmark data, enabling the scalability of this framework to large data-sets. We have shown that this significant space complexity improvements do not compromise, and may even lead to increased performance of clustering combination. The experiments also confirmed linear time complexity. Additional experiments on larger data sets are underway.

ACKNOWLEDGMENT

We acknowledge the financial support from the FET programme, within the EU FP7, under the SIMBAD project (contract no.213250).

REFERENCES

[1] A. Fred, "Finding consistent clusters in data partitions," in *Multiple Classifier Systems*, J. Kittler and F. Roli, Eds., vol. 2096. Springer, 2001, pp. 309–318.

[2] A. Fred and A. Jain, "Combining multiple clustering using evidence accumulation," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 835–850, June 2005.

[3] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *J. of Machine Learning Research 3*, 2002.

[4] A. Topchy, A. Jain, and W. Punch, "A mixture model of clustering ensembles," in *Proc. of the SIAM Conf. on Data Mining*, April 2004.

[5] X. Z. Fern and C. E. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning," in *Proc ICML '04*, 2004.

[6] H. G. Ayad and M. S. Kamel, "Cumulative voting consensus method for partitions with variable number of clusters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 160–173, 2008.

[7] M. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381–396, 2002.

[8] A. Asuncion and D. Newman, "UCI ML repository," 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html