Evidence Accumulation Clustering Based on the K-Means Algorithm

Ana Fred[1] and Anil K. Jain[2]

[1] Instituto de Telecomunicações
Instituto Superior Técnico, Lisbon, Portugal
afred@lx.it.pt
[2] Department of Computer Science and Engineering
Michigan State University, USA
jain@cse.msu.edu

Abstract. The idea of evidence accumulation for the combination of multiple clusterings was recently proposed [7]. Taking the K-means as the basic algorithm for the decomposition of data into a large number, k, of compact clusters, evidence on pattern association is accumulated, by a voting mechanism, over multiple clusterings obtained by random initializations of the K-means algorithm. This produces a mapping of the clusterings into a new similarity measure between patterns. The final data partition is obtained by applying the single-link method over this similarity matrix. In this paper we further explore and extend this idea, by proposing: (a) the combination of multiple K-means clusterings using variable k; (b) using cluster lifetime as the criterion for extracting the final clusters; and (c) the adaptation of this approach to string patterns. This leads to a more robust clustering technique, with fewer design parameters than the previous approach and potential applications in a wider range of problems.

1 Introduction

Clustering algorithms can be categorized into hierarchical methods and partitional methods [3,12]. A partitional structure organizes patterns into a small number of clusters. The K-means is one of the simplest clustering algorithms in this class: it is computationally efficient and does not require the specification of many parameters. Hierarchical methods propose a nesting of clusterings, providing additional information about data structure, represented graphically as a dendrogram. A particular partition is obtained by cutting the dendrogram at some level. The single link algorithm is one of the most popular methods in this class [12].

A large number of clustering algorithms exist [12,13]. Examples of different classes of algorithms are model-based techniques [8,18,23], non-parametric density estimation based methods [21], central clustering [2], square-error clustering [19], and graph theoretical based [4,26] methods. Each handles differently the issues related to cluster validity [1,10,20,8], number of clusters [15,25], and

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 442–451, 2002.

structure imposed on the data [6,24,16]; yet, no single algorithm can adequately handle all sorts of cluster shapes and structures.

Inspired by the work in sensor fusion and classifier combination techniques in pattern recognition [14], Fred [7] proposed a combination of clusterings in order to devise a consistent data partition. It follows a split and merge strategy. First, the data is split into a large number of small clusters, using the K-means algorithm; with fixed k, different clusterings are produced by an arbitrary initialization of cluster centers. The clustering results are combined using a voting mechanism, leading to a new similarity matrix between patterns. The final clusters are obtained by applying the single-link (SL) method on this matrix, thus merging small clusters produced in the first stage of the method.

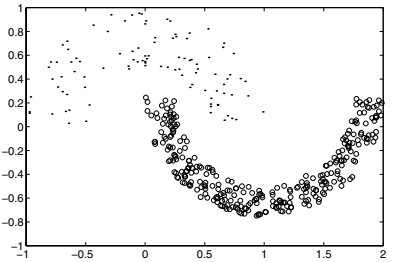
In this paper we further analyze the above method and propose three main refinements/extensions: the use of cluster lifetime as a criterion for the identification of the final data partition from the dendrogram produced by the SL method, instead of fixed level thresholding; the combination of clusterings with different values of k in a reasonably large range; adaptation of this approach to process string patterns. These modifications improve the previous strategy in terms of robustness and simplicity of the method, with fewer parameters to be defined.

Section 2 discusses the method in [7]. Refinements and extensions of the method are proposed in section 3. The performance of the new method is illustrated through a set of experimental results given in section 4, followed by the conclusions.

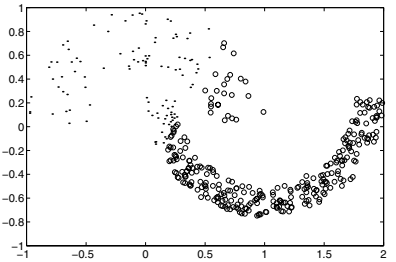
2 Evidence Accumulation Clustering

The idea of evidence accumulation clustering is to combine the results of multiple clusterings into a single data partition, by viewing each clustering result as an independent evidence of data organization. Fred [7] used the K-means algorithm as the basic algorithm for decomposing the data into a large number, k, of compact clusters; evidence on pattern association is accumulated, by a voting mechanism, over N clusterings obtained by random initializations of the K-means algorithm. This produces a mapping of the clusterings into a new similarity measure between patterns, summarized in the matrix co_assoc, where $co\_assoc(i, j)$ indicates the fraction of times the pattern pair (i, j) is assigned to the same cluster among N clusterings. The final data partition is obtained by applying the single-link method over this similarity matrix, using a fixed threshold, t.

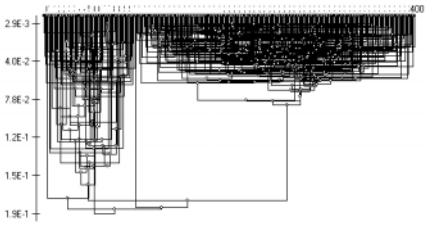
The method has two design parameters: k, the number of clusters for the K-means algorithm; and t, the threshold on the dendrogram produced by the SL method. We discuss these parameters using the half-rings data set example, depicted in figure 1(a). This data set is composed of 400 two-dimensional patterns (upper cluster - 100 patterns; lower cluster - 300 patterns). Due to the particular cluster shapes, the K-means algorithm by itself is unable to identify the two natural clusters (see figure 1(b)). The uneven data sparseness of the two



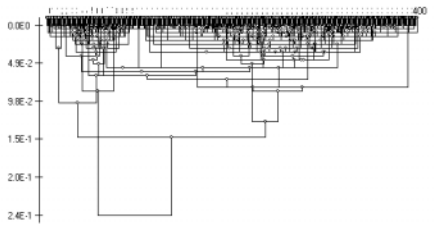
(a) Half-rings shaped clusters



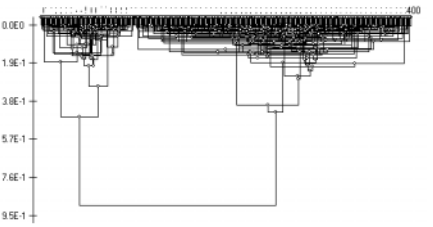
(b) K-means clustering, $K = 2$



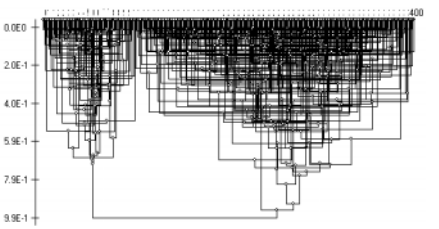
(c) Single-link method. Thresholding this graph splits the upper ring cluster into several small clusters



(d) Evidence accumulation clustering, $k = 5$



(e) Evidence accumulation clustering, $k = 15$



(f) Evidence accumulation clustering, $k = 80$

Fig. 1. Half-rings data set. Vertical axis on dendrograms (d) to (f) corresponds to distances, $d(i, j)$, with $d(i, j) = 1 - co\_assoc(i, j)$

clusters also prevents the SL method to produce the correct data partition, as shown by the associated dendrogram (figure 1(c)). Figures 1(d)- 1(f) plot the dendrograms produced by the evidence accumulation algorithm after 200 runs ($N = 200$) of the K-means algorithm, for different values of k. The K-means algorithm can be seen as performing a decomposition of the data into a mixture of Gaussians. k is the critical parameter in this decomposition: low values of k are not enough to capture the complexity of the data, while large values may produce an over-fragmentation of the data (in the limit, each pattern forming a

cluster). By using the method in [5] the data set is decomposed into 10 gaussian components. This should be a lower bound on the value of k to be used with the K-means, as this algorithm imposes spherical shaped clusters, and therefore a higher number of components may be needed for evidence accumulation. This is in agreement with the dendrograms in figures 1(d)- 1(f). As shown in figure 1(d), although the two-cluster structure starts to emerge in the dendrogram for $k = 5$, the two natural clusters cannot yet be identified. A clear cluster separation is present in the dendrogram for $k = 15$ (fig. 1(e)). As k increases, similarity values between pattern pairs decrease, and links in the dendrograms progressively form at higher levels, causing the two natural clusters to be less clearly defined (see fig. 1(f) for $k = 80$). The same conclusions can be drawn by analyzing table 1, showing the number of clusters identified for different values of k and of t. The lifetime of a cluster in the dendrogram for a given k (distance gap between two successive merges) can be evaluated on the corresponding line in this table. As shown, using a fixed threshold, the range of k values for which the true number of clusters is identified is limited and depends on t. Using the longest lifetime (clusters persisting for the largest range of t) as the criterion for identifying the final number of clusters, leads to the values on the rightmost column of table 1, with the identification of the true clusters for a larger k range.

Table 1. Number of clusters identified as a function of k and t for the half-rings data set ($N = 200$). The 2^* notation indicates that, although the correct number of clusters is identified, this does not correspond to the correct data partition. The rightmost column indicates the final number of clusters according to the largest lifetime criterion

$k\backslash t$	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95	NC
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2^*	5	7	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2^*	2^*	5	20	2^*
10	1	1	1	1	1	2	2	2	2	2	2	2	2	3	5	7	8	18	71	2
15	1	2	2	2	2	2	2	2	2	2	3	5	5	6	8	13	33	64	134	2
20	1	2	2	2	2	2	2	2	3	5	6	6	7	9	18	34	61	94	171	2
25	1	2	2	2	2	2	2	2	4	6	7	10	14	19	31	54	84	121	215	2
30	1	1	2	2	2	2	3	5	6	8	9	14	20	25	40	71	99	145	252	2
40	1	2	2	2	2	2	3	7	9	11	18	25	34	46	67	95	137	197	279	2
80	3	4	4	5	7	14	19	26	31	47	65	85	97	130	157	188	227	276	334	4

3 Evidence Accumulation Clustering with Varying k and Dynamic Threshold

As noted in the previous section, cluster lifetime is a better criterion for identifying the natural clusters than a fixed threshold, as the dendrogram scales up with increasing values of k. On the other hand, in order to determine an adequate value or range for k, one should use some *a priori* information (for instance, by applying a mixture decomposition method for determining the number of

components in the mixture). Otherwise, several values of k should be tested, the final number of clusters being the most stable solution found.

The evidence of a clear cluster separability on the dendrograms associated with a large range for k (see figures 1(e), 1(f)) suggests a combination of K-means clusterings with variable k. Our hypothesis is that the combined evidence will reinforce the intrinsic data structure, diluting the effect produced by low values of k (while combined with other values, low k values contribute to a scaling up of similarity measures - lower values on the dendrograms); high values of k produce random, high granularity data partitions, so they should also not be disruptive of the structure imposed by more adequate k values, scaling down the similarity values. We therefore propose a combination of multiple K-means clusterings with varying k, the final data partition being obtained as the cluster configuration with the highest lifetime in the dendrogram produced by the SL method over the similarity matrix, co_assoc. The proposed evidence accumulation clustering method is summarized below:

Data clustering using Evidence Accumulation.

Input:

n $d-$dimensional patterns;
k_min - minimum initial number of clusters;
k_max - maximum initial number of clusters;
N - number of clusterings.

Output: Data partitioning.

Initialization: Set co_assoc to a null $n \times n$ matrix.

1. Do N times:
- **1.1.** Randomly select k in the interval $[k\_min; k\_max]$.
- **1.2.** Randomly select k cluster centers.
- **1.3.** Run the K-means algorithm with the above k and initialization, and produce a partition P.
- **1.4.** Update the co-association matrix: for each pattern pair, (i, j), in the same cluster in P, set $co\_assoc(i, j) = co\_assoc(i, j) + \frac{1}{N}$.

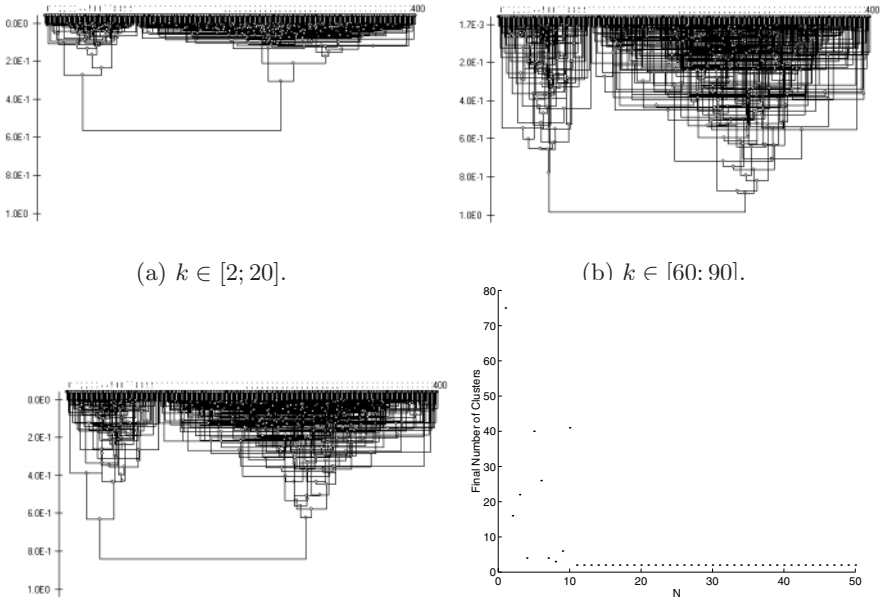
2. Detect consistent clusters in the co-association matrix using the SL technique: compute the SL dendrogram and identify the final clusters as the ones with the highest lifetime.

---

4 Experimental Results

4.1 Vector Representations: Artificial Data Sets

The proposed evidence accumulation clustering method was applied to the half-rings data set, used as the illustrative example in section 2. Several ranges for k were tested in order to evaluate the robustness of the method. Dendrograms for some of these tests are plotted in figure 2. The number of clusterings used were $N = 600$; experiments with $N = 200$ and lower values led to similar results, since the method converges for values of N around 50 (see figure 2(d)).



(a) $k \in [2; 20]$.

(b) $k \in [60; 90]$.

(c) $k \in [2; 80]$.

(d) Convergence curve ($k \in [2; 80]$).

Fig. 2. Combining 600 K-means clusterings, with varying k for data in fig 1(a). Dendrograms (a) to (c) illustrate the wide range of k values with a clear cluster separation, showing the robustness of the combination technique. (d)- Convergence curve of the final number of clusterings as a function of N, the number of clusterings, for $k \in [2; 80]$

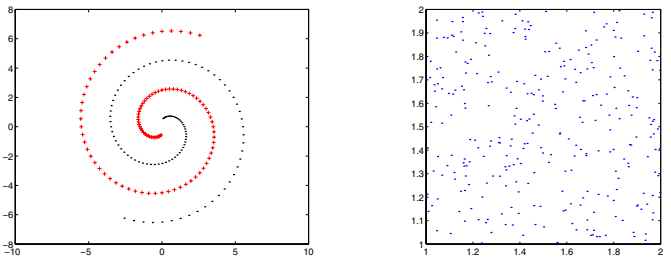
Table 4 summarizes the experiments and the number of clusters (NC) obtained. As shown, all ranges for k, except the ones completely below the minimum number of mixture components, 10, (first two columns), lead to the correct identification of the natural clusters, demonstrating the robustness of the method.

The spiral data set (fig. 3(a)) is another example of complex shaped clusters. Using the method of [7], the two natural clusters are identified for values of k in the interval $[25; 70]$ for $t = 0.5$ or $t = 0.6$. In all the tests with the proposed method, the true clusters were identified for all the intervals considered (values of $k > 90$ were not tested as the number of training patterns is only 200), except for ranges totally in the interval $[2; 20]$, as this is lower than the minimum number of components required to decompose the data (the method in [5] identifies 24 gaussian components).

We also performed tests on uni-modal random data (gaussian and uniform distributions) in order to assess if the proposed clustering technique imposes

Table 4. Evidence accumulation clustering with varying k for the half-rings data set

	k-range							
	[2; 5]	[2; 10]	[2; 20]	[5; 20]	[10; 30]	[30; 60]	[60; 90]	[2; 80]
NC	1	1	2	2	2	2	2	2



(a) (b)

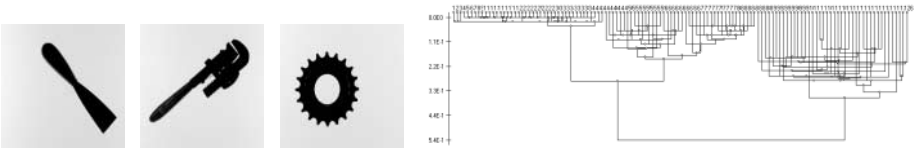
Fig. 3. Artificial data sets. (a)- Spiral data set (100 samples per class). (b)- 2-D projection of 300 patterns uniformly distributed in a 5-dimensional hypercube

some structure on data. In all the tests performed (an example of uniform data set is illustrated in figure 3(b)), a single cluster was identified, no matter what interval for k was considered.

4.2 String Patterns: Clustering of Contour Images

We have applied the proposed technique to the classification of string descriptions of contour images of 2D shapes. The data set consists of 126 images from three types of tools (42 patterns per class); sample images are shown in figures 4(a) to 4(c). Each image was segmented to separate the object from the background and the object boundary was sampled at 50 equally spaced points; object shapes were encoded using an 8-directional differential chain code [9,11]. In order to apply the cluster combination technique, similarity between all pattern pairs was calculated using the Levensthein distance normalized by the length of the editing path [17,22]. The K-means algorithm was adapted in order to handle string patterns: cluster centroids are selected as the training pattern with the minimum average distance to the remaining patterns within a cluster; therefore, the algorithm simply needs a similarity/dissimilarity matrix between pattern pairs as input.

As shown in figure 4(d), a direct application of the SL method to the string patterns using the normalized string edit distance does not produce a correct partitioning of the data. With the proposed method, a good separation of the three clusters is obtained, for instance with $k \in [2; 30]$ and N=200.



(a) T1. (b) T2. (c) T3. (d) SL dendrogram.

Fig. 4. Hardware tools data set

5 Conclusions

We have proposed a novel algorithm for evidence accumulation clustering. The method introduced in [7] was extended/modified by: (1) using cluster lifetime as a criterion for determining the final number of clusters; (2) proposing the formation of clustering ensembles by using the K-means algorithm with random initialization and arbitrary k values within a large interval. Furthermore, the adaptation of the K-means algorithm by using cluster median patterns, and thus simply requiring as input a similarity or dissimilarity matrix, extended the potential use of this technique to a wider range of applications, namely those based on string descriptions. The new method enhances the previous approach in terms of robustness and simplicity of evaluations, with fewer parameters being defined. The ability of the clustering method to correctly identify well separated clusters with complex shapes has been demonstrated on a set of artificial and real data, using both vector and string descriptions of patterns. Moreover, tests on unimodal/uniform data showed that the method does not impose any structure on data, a single cluster being identified for this data. Further tests are needed for touching clusters.

Acknowledgments

This work was partially supported by the Portuguese Foundation for Science and Technology (FCT), Portuguese Ministry of Science and Technology, and FEDER, under grant POSI/33143/SRI/2000, and ONR grant no. N00014-01-1-0266.

References

1. T. A. Bailey and R. Dubes. Cluster validity profiles. *Pattern Recognition*, 15(2):61–83, 1982. 442
2. J. Buhmann and M. Held. Unsupervised learning without overfitting: Empirical risk approximation as an induction principle for reliable clustering. In Sameer Singh, editor, *International Conference on Advances in Pattern Recognition*, pages 167–176. Springer Verlag, 1999. 442

3. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley, second edition, 2001. 442
4. Y. El-Sonbaty and M. A. Ismail. On-line hierarchical clustering. *Pattern Recognition Letters*, pages 1285–1291, 1998. 442
5. M. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002. 445, 447
6. B. Fischer, T. Zoller, and J. Buhmann. Path based pairwise data clustering with application to texture segmentation. In M. Figueiredo, J. Zerubia, and A. K. Jain, editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 2134 of *LNCS*, pages 235–266. Springer Verlag, 2001. 443
7. A. L. Fred. Finding consistent clusters in data partitions. In Josef Kittler and Fabio Roli, editors, *Multiple Classifier Systems*, volume LNCS 2096, pages 309–318. Springer, 2001. 442, 443, 447, 449
8. A. L. Fred and J. Leitão. Clustering under a hypothesis of smooth dissimilarity increments. In *Proc. of the 15th Int'l Conference on Pattern Recognition*, volume 2, pages 190–194, Barcelona, 2000. 442
9. A. L. Fred, J. S. Marques, and P. M. Jorge. Hidden markov models vs syntactic modeling in object recognition. In *ICIP'97*, 1997. 448
10. M. Har-Even and V. L. Brailovsky. Probabilistic validation approach for clustering. *Pattern Recognition*, 16:1189–1196, 1995. 442
11. A. Jain. *Fundamentals of Digital Image Processing.* Prentice-Hall, 1989. 448
12. A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, 1988. 442
13. A.K. Jain, M. N. Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, September 1999. 442
14. J. Kittler, M. Hatef, R. P Duin, and J. Matas. On combining classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998. 443
15. R. Kothari and D. Pitts. On finding the number of clusters. *Pattern Recognition Letters*, 20:405–416, 1999. 442
16. Y. Man and I. Gath. Detection and separation of ring-shaped clusters using fuzzy clusters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(8):855–861, August 1994. 443
17. A. Marzal and E. Vidal. Computation of normalized edit distance and applications. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2(15):926–932, 1993. 448
18. G. McLachlan and K. Basford. *Mixture Models: Inference and Application to Clustering.* Marcel Dekker, New York, 1988. 442
19. B. Mirkin. Concept learning and feature selection based on square-error clustering. *Machine Learning*, 35:25–39, 1999. 442
20. N. R. Pal and J. C. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE Trans. Fuzzy Systems*, 3:370–379, 1995. 442
21. E. J. Pauwels and G. Frederix. Fiding regions of interest for content-extraction. In *Proc. of IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases VII*, volume SPIE Vol. 3656, pages 501–510, San Jose, January 1999. 442
22. E. S. Ristad and P. N. Yianilos. Learning string-edit distance. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(5):522–531, May 1998. 448
23. S. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to gaussian mixture modelling. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(11), November 1998. 442

24. D. Stanford and A. E. Raftery. Principal curve clustering with noise. Technical report, University of Washington, http://www.stat.washington.edu/raftery, 1997. 443
25. H. Tenmoto, M. Kudo, and M. Shimbo. MDL-based selection of the number of components in mixture models for pattern recognition. In Adnan Amin, Dov Dori, Pavel Pudil, and Herbert Freeman, editors, *Advances in Pattern Recognition*, volume 1451 of *Lecture Notes in Computer Science*, pages 831–836. Springer Verlag, 1998. 442
26. C. Zahn. Graph-theoretical methods for detecting and describing gestalt structures. *IEEE Trans. Computers*, C-20(1):68–86, 1971. 442