Data Clustering Using Evidence Accumulation

Ana L.N. Fred
Telecommunications Institute
Instituto Superior Técnico, Portugal
afred@lx.it.pt

Anil K. Jain
Dept. of Computer Science and Engineering
Michigan State University, USA
jain@cse.msu.edu

Abstract

We explore the idea of evidence accumulation for combining the results of multiple clusterings. Initially, n $d-$dimensional data is decomposed into a large number of compact clusters; the K-means algorithm performs this decomposition, with several clusterings obtained by N random initializations of the K-means. Taking the co-occurrences of pairs of patterns in the same cluster as votes for their association, the data partitions are mapped into a co-association matrix of patterns. This $n \times n$ matrix represents a new similarity measure between patterns. The final clusters are obtained by applying a MST-based clustering algorithm on this matrix. Results on both synthetic and real data show the ability of the method to identify arbitrary shaped clusters in multidimensional data.

1. Introduction

Data clustering is an important but an extremely difficult problem. Clustering techniques require the definition of a similarity measure between patterns, which is not easy to specify in the absence of any prior knowledge about cluster shapes. A large number of clustering algorithms exist [7], yet no single algorithm can adequately handle all sorts of cluster shapes and structures. Each algorithm has its own approach for handling cluster validity [1, 6, 12, 5], number of clusters [8, 10], and structure imposed on the data [2, 13, 11]. The K-means algorithm is one of the simplest clustering algorithms: it is computationally efficient and does not require the user to specify many parameters. Its major limitation is the inability to identify clusters with arbitrary shapes, ultimately imposing hyper-spherical clusters on the data.

We explore the idea of evidence accumulation for combining the results of multiple clusterings. The idea of combining multiple sources has been addressed in areas like sensor fusion and supervised learning techniques in pattern recognition - known as classifier combination [9]. A recent work on the combination of multiple clusterings is reported in [4].

There are several possible ways to accumulate evidence in the context of unsupervised learning: (1) combine results of different clustering algorithms; (2) produce different results by resampling the data, such as in bootstrapping techniques (like bagging) and boosting; (3) running a given algorithm many times with different parameters or initializations. In this paper we take the last approach, using the well known K-means algorithm as the underlying clustering algorithm to produce clustering ensembles. First, the data is split into a large number of compact and small clusters; different decompositions are obtained by random initializations of the K-means algorithm. The data organization present in the multiple clusterings is mapped into a co-association matrix which provides a measure of similarity between patterns. The final data partition is obtained by clustering this new similarity matrix, corresponding to the merging of clusters.

2. Evidence Accumulation

The idea of evidence accumulation-based 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.

Given n $d-$dimensional patterns, the proposed strategy follows a split-and-merge approach:

Split Decompose multidimensional data into a large number of small, spherical clusters. The K-means algorithm performs this decomposition, with various clustering results obtained by random initializations of the algorithm.

Combine In order to cope with partitions with different numbers of clusters, we propose a voting mechanism to combine the clustering results, leading to a new measure of similarity between patterns. The underlying assumption is that patterns belonging to a "natural "cluster are very likely to be co-located in the same cluster in different clusterings. Taking the co-occurrences of pairs of patterns in the same cluster as

1051-4651/02 $17.00 © 2002 IEEE

votes for their association, the data partitions produced by multiple runs of K-means are mapped into a $n \times n$ co-association matrix:

$$co\_assoc(i,j) = \frac{votes_{ij}}{N},$$

where N is the number of clusterings and $votes_{ij}$ is the number of times the pattern pair (i,j) is assigned to the same cluster among the N clusterings.

Merge In order to recover natural clusters, we emphasize neighborhood relationship and apply a minimum spanning tree (MST) algorithm, cutting weak links at a threshold of t; this is equivalent to cutting the dendrogram produced by the single link (SL) method over this similarity matrix at the threshold t, thus merging clusters produced in the splitting phase.

The overall method for evidence accumulation-based clustering is summarized below.

Data clustering using Evidence Accumulation:
Input: n $d-$dimensional patterns; k - initial number of clusters; N - number of clusterings. t - threshold. *Output:* Data partitioning. *Initialization:* Set co_assoc to a null $n \times n$ matrix. **1.** Do N times: **1.1.** Randomly select k cluster centers. **1.2.** Run the K-means algorithm with the above initialization and produce a partition P. **1.3.** 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 a SL technique: **2.1.** Find majority voting associations: For each pattern pair, (i,j), such that $co\_assoc(i,j) > t$, merge the patterns in the same cluster; if the patterns were in distinct previously formed clusters, join the clusters; **2.2.** For each remaining pattern not included in a cluster, form a single element cluster;

The proposed technique has two design parameters: k- the number of clusters for the K-means algorithm; and t, the threshold on the MST.

The K-means algorithm can be seen as performing a decomposition of the data into a mixture of spherical Gaussians. Low values of k are not adequate to identify distinct components while large values may produce an over-fragmentation of the data (in the limit, each sample forming a cluster). Intuitively, k should be greater than the true number of clusters; the minimum value of k, however, is not directly related to the true number of clusters, as a cluster may itself be a combination of several components. The value of k may be specified by identifying the number of components in the mixture of gaussians model [3]; alternatively, a rule of thumb, $k = \sqrt{n}$ may be used, with n being the number of input patterns, or several values for k may be evaluated.

Concerning the threshold parameter, typically the value $t = 0.5$ is selected, meaning that patterns to be placed in a cluster in the final partition must have been co-located in a cluster at least 50% of the times over the N clustering ensembles. In exploratory data analysis, we recommend that clusterings obtained for several values for t should be analyzed.

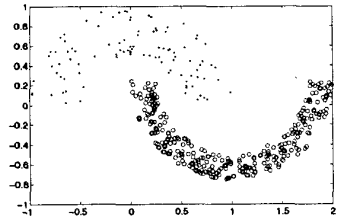
3. Experimental Results

We illustrate the characteristics of the proposed technique with several artificial and real data sets. In particular, we show that the proposed method can identify complex cluster shapes (spiral data set), even in the presence of uneven data sparseness (half-rings data set); treatment of random data (section 3.3); gaussian data with varying cluster separability (section 3.4); and the Iris data set. Results presented here are based on the combination of 200 K-means clusterings ($N = 200$), a value high enough to ensure that convergence of the method is achieved.

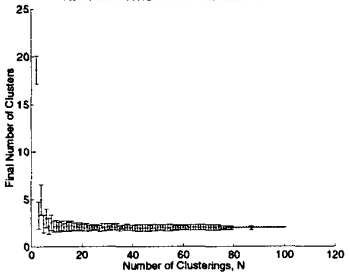
3.1 Half-Rings Data Set

The half-rings data set, as shown in figure 1(a) consists of two clusters with uneven sparseness (upper cluster - 100 patterns; lower cluster - 300 patterns). The K-means algorithm by itself is unable to identify the two natural clusters here, imposing a spherical structure on the data. The single-link method does not perform much better, as shown in figure 1(c). In order to apply the evidence accumulation technique, the initial value of k must be specified. The mixture decomposition method reported in [3] identifies 10 gaussian components; the rule of thumb $k = \sqrt{n}$ gives $k = 20$. Figure 1(b) plots the evolution of the number of clusters identified by the proposed method with $k = 10$, as a function of the number of clusterings, N (error bars were calculated over 25 experiments); convergence to a 2-cluster solution is obtained for $N \approx 90$. As the K-means is a very fast algorithm, we shall use $N = 200$ hereafter. Table 1 shows the number of clusters identified by the proposed method for several values of k and t. Results with varying t are consistent; higher t values reduce the range of k that identify the natural clusters. The single cluster obtained with $k = 5$ is justified by the use of an insufficient number of components; at $k = 20$ we begin to observe excessive granularity of the initial partitions; these results agree with the number of gaussian components given by [3] for this data set.

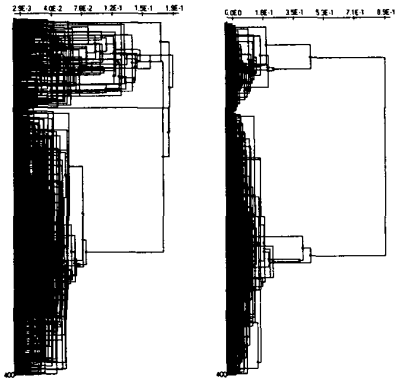
Figure 1(d) shows the dendrogram produced by the single-link method applied to the co-association matrix ob-



(a) Half-rings shaped clusters.



(b) Convergence curve, $k = 10, t = 0.5$.



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

(d) Single-link method on the co-association matrix ($k = 15$). Distance between patterns (i, j) is $1 - co\_assoc(i, j)$.

Figure 1. Half-rings data set (a) and clusterings.

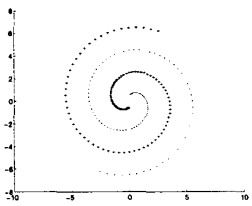
$t \backslash k$	5	10	15	20
0.4	1	2	2	2
0.5	1	2	2	5
0.6	1	2	2	6

Table 1. Number of clusters identified as a function of k and t for the half-rings data set ($N = 200$).

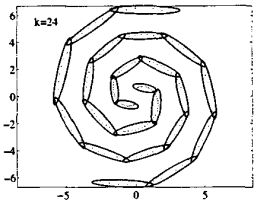
tained by the combination of 200 clusterings generated using the K-means with $k = 15$. The new similarity measure helps in identifying the true structure of the clusters: similarity between patterns within a natural cluster is amplified in comparison with similarity values between patterns in distinct clusters. Using the default value, $t = 0.5$, on the SL clustering over the similarity matrix recovers the natural clusters in figure 1(a).

3.2 Spiral Data

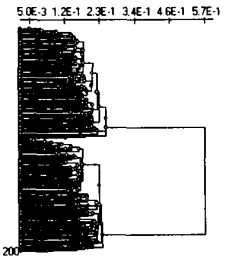
The two spiral patterns, as shown in figure 2(a), demonstrate another example of complex cluster shapes. While the simple K-means algorithm cannot correctly cluster this data, the proposed algorithm easily recovers the true clusters by merging nearby clusters in the decomposition performed in the cluster ensembles, using a sufficiently large value of k.



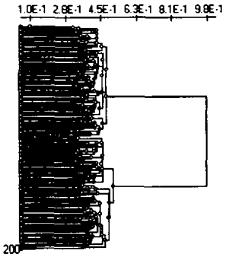
(a) Spiral data (200 samples) with two clusters.



(b) Decomposition of (a) into a mixture of gaussians using the method in [3].



(c) SL dendrogram with evidence accumulation, $k = 30$.



(d) SL dendrogram with evidence accumulation, $k = 80$.

Figure 2. Spiral data (a) and its decomposition using mixture of gaussians (b). (c)-(d): The effect of k on evidence accumulation.

$t \backslash k$	5	10	15	20	25	30	40	50	60	70	80
0.5	1	1	1	1	1	2	2	2	2	2	3
0.6	1	1	1	1	2	2	2	2	3	21	102

Table 2. Number of clusters identified as a function of k and of t for the spiral data.

Table 2 shows the number of clusters identified with the evidence accumulation strategy as a function of k for two values of t: 0.5 and 0.6. It shows that low values of k lead to a single cluster being identified; this is to be expected since, when the number of initial components is very small, neighboring patterns in the two spirals are put in the same cluster. The method reported in [3] decomposes this data into 24 gaussian components (fig. 2(b)); since the K-means imposes spherical clusters (as in a unit-covariance gaussian), the value of k should be higher than 24. As shown in Table 2, the true number of clusters is identified for $k \geq 30$, with $t = .5$ (for $k \geq 25$, with $t = .6$). A large value of k scales the dendrogram (see figs. 2(c) and 2(d)), as similarity values decrease due to higher granularity of the partitions produced. This scaling will exceed the fixed threshold, t, after a certain number of components (80, with $t = .5$), and thus the method will give a larger (than true) number of clusters for values of k above this limit. A procedure to identify the true number of clusters, without requiring an external method for determining the number of components, k, may be as follows: run the evidence accumulation method for various values of k and select the "stable" solution found in the plot of the number of clusters as a function of k, just before the curve starts to increase exponentially.

3.3 Random Data

How does the proposed algorithm perform when presented with "random"data that does not contain any natural clusters?

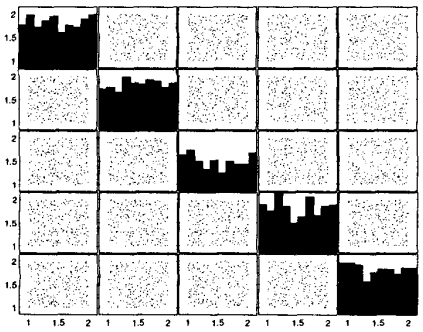


Figure 3. Scatter plots of 5-dimensional random data: rows and columns correspond to features; plots in diagonal positions correspond to histograms of individual features.

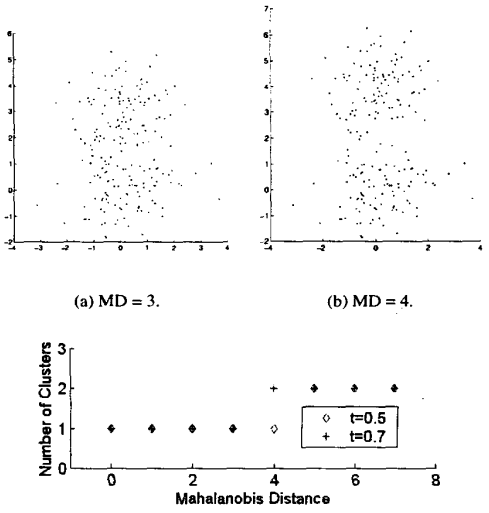
Figure 3 shows 300 patterns uniformly distributed in a 5-dimensional hypercube. The clustering results are shown in Table 3. Notice the consistency of the results obtained for various values of k and t, a single cluster being identified (the 3-cluster solution corresponds to 298 patterns in a single cluster, with two outlier clusters). Similar results are obtained with gaussian distributions.

$t\backslash k$	2	3	4	5	6	7	8	9	10	15	20
0.4	1	1	1	1	1	1	1	1	1	1	2*
0.5	1	1	1	1	1	1	1	1	1	3*	5
0.6	1	1	1	1	3*	3*	3*	3*	3*	11	23

Table 3. Number of clusters as a function of k and t. Elements with the * symbol mean that a cluster is found with all but 1 or two patterns, each of these forming single element clusters.

3.4 2D Gaussian Data

We test the sensitivity of the proposed method on cluster separability with 2-component 2D gaussian data sets (100 patterns per cluster), by varying the Mahalanobis distance (MD) between the two cluster centers. The results are shown in figure 4(c). The method is unable to discern two clusters in the data for Mahalanobis distances below 5, with $t = .5$; by setting a more restrictive threshold, such as $t = .7$, two clusters are identified for MD = 4 (see fig. 4(b). The case of MD = 3 (fig. 4(a)), with the two clusters clearly overlapping, is always identified as a single cluster.



(c) Number of clusters found.

Figure 4. 2D Gaussian data with varying Mahalanobis distance (MD) and the number of clusters found by the proposed method for $2 \leq k \leq 5$.

3.5 Iris Data Set

The Iris data set, often used as a benchmark in supervised learning techniques, consists of three types of Iris plants (50 instances per class), represented by 4 features, with one class well separated from the other two, which are intermingled. Table 4 shows the number of clusters found for various values of k and t. The two-cluster solution is the one consistently appearing in most situations, corresponding to the identification of the well separated Setosa class and the merging of the other two in a single cluster. The other frequent solution (for higher values of t) corresponds to the partition of the data into three clusters. Table 5 presents the consistency index [4] which measures the percentage of patterns correctly assigned in the data partitioning, taking as reference the true class labels of the samples. This table reveals the presence of a relatively stable data partition with three clusters (consistency index = .84); the highest consistency index is obtained with $k = 3$, the true number of clusters. It is interesting to note that a direct application of the single-link method to the Iris data set leads to a consistency index of 0.68.

$t\backslash k$	3	4	5	6	7	8	9	10
0.5	2	2	2	2	2	2	2	2
0.6	2	2	2	2	2	3	3	3
0.7	2	2	3	3	5	5	7	9
0.75	3	3	3	3	7	7	10	13

Table 4. Number of clusters as a function of k and t for the Iris data set.

$t\backslash k$	3	4	5	6	7	8	9	10
0.5	.67	.667	.67	.67	.67	.67	.67	.67
0.6	.67	.67	.67	.67	.67	.84	.75	.75
0.7	.67	.67	.84	.84	.75	.67	.63	.53
0.75	.89	.84	.84	.84	.67	.67	.53	.47

Table 5. Consistency index as a function of k and t for the Iris data set.

4. Conclusions

A robust clustering technique based on a combination of multiple clusterings, has been presented. Following a split-and-merge strategy, and based on the idea that smaller clusters are easier to combine, the first step is to decompose complex data into small, compact clusters. The K-means algorithm serves this purpose; an ensemble of clusterings is produced by random initializations of cluster centroids. Data partitions present in these clusterings are mapped into a new similarity matrix between patterns, based on a voting mechanism. This matrix, which is independent of data sparseness, is then used to extract the natural clusters using the single link algorithm. The proposed method has two important parameters; guidelines for setting these parameters are given. The proposed method is able to identify well separated, arbitrarily shaped clusters, as corroborated by experimental results. The method performs poorly, however, in situations of touching clusters, as illustrated by the 2-component gaussian data set example in Figure 4 (a). We are studying ways to overcome this difficulty, namely by combining different clustering algorithms.

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.

[2] J. Buhmann and M. Held. Unsupervised learning without overfitting: Empirical risk approximation as an induction principle for reliable clustering. In S. Singh, editor, *International Conference on Advances in Pattern Recognition*, pages 167–176. Springer Verlag, 1999.

[3] M. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[4] A. L. Fred. Finding consistent clusters in data partitions. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, volume LNCS 2096, pages 309–318. Springer, 2001.

[5] 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.

[6] M. Har-Even and V. L. Brailovsky. Probabilistic validation approach for clustering. *Pattern Recognition*, 16:1189–1196, 1995.

[7] A. Jain, M. N. Murty, and P. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

[8] A. K. Jain and J. V. Moreau. Bootstrap technique in cluster analysis. *Pattern Recognition*, 20(5):547–568, 1987.

[9] 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.

[10] R. Kothari and D. Pitts. On finding the number of clusters. *Pattern Recognition Letters*, 20:405–416, 1999.

[11] 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.

[12] N. R. Pal and J. C. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE Trans. Fuzzy Systems*, 3:370–379, 1995.

[13] D. Stanford and A. E. Raftery. Principal curve clustering with noise. Technical report, University of Washington, http://www.stat.washington.edu/raftery, 1997.