Context-Dependent Clustering based on Dissimilarity Increments

Ana L. N. Fred

Instituto de Telecomunicações / Instituto Superior Técnico
IST-Torre Norte, Av. Rovisco Pais, 1049-001, Lisboa, Portugal

***Abstract* – We explore the idea of context dependent clustering under a hierarchical agglomerative framework. Inter-pattern relationships (within a cluster) are modelled by the statistical distribution of dissimilarity increments between neighboring patterns. This distribution characterizes context, forming the basis of a new cluster isolation criterion. The integration of this criterion into a hierarchical agglomerative clustering framework produces a partitioning of the data, while exhibiting pattern structure in terms of a dendrogram-type graph. We further extend the applicability of the method to large data sets by proposing the integration of sampling techniques into the clustering process.**

***Keywords* – Clustering, hierarchical methods, dissimilarity increments, context, scalability.**

I. Introduction

Clustering techniques require the definition of a similarity measure between patterns. Directly using dissimilarity values or exploring point densities for the patterns, either emphasizing compactness or connectedness in feature space, two main strategies are adopted: hierarchical methods and partitional methods [1]. Partitional structure organizes patterns into a small number of clusters; a data partition is obtained as the result of an optimization process or by exploring local structure. Examples of techniques in this class include mixture decomposition [2, 3, 4], non-parametric density estimation based methods [5], central clustering [6], square-error clustering [7], shape fitting approaches [8], geometrical approaches [9]. The K-means is a very popular algorithm in this category. Assuming *a priori* knowledge about the number of classes and based on the square-error criterion, it is a computationally efficient clustering technique that identifies hyper-spherical clusters [1].

Hierarchical methods produce a nesting of data clusterings in a hierarchical structure, that can be represented graphically as a dendrogram. Mostly inspired by graph theory [10], both agglomerative [1, 11] and divisive approaches [12] have been attempted, the first starting with many clusters that are successively merged in accordance with inter cluster similarity, and the later working in the opposite direction. Variations of the algorithms can be obtained by the definition of a similarity measure between patterns and clusters [13]. The single link algorithm is one of the most popular methods in this class [1]. Data partitioning is usually obtained by setting a threshold on the dendrogram; cluster validity studies have also been proposed [14, 15] for the *a posteriori* analysis of structures, in order to evaluate the clustering results and define meaningful clusters.

In this paper we explore the idea of context dependent clustering under a hierarchical agglomerative framework. Assuming a hypothesis of smooth dissimilarity increments between neighboring patterns within a cluster [16], the statistics of dissimilarity first derivative is modelled by an exponential distribution. This statistical model characterizes context, a cluster isolation criterion being derived based on a pair-wise context analysis (section II). Introduction of this criterion into a hierarchical agglomerative technique (section II-A) leads to the partitioning of the data without requiring ad-hoc specification of parameters (such as the number of clusters or threshold on the dendrogram). The characteristics of the method are illustrated through a set of examples. A major difficulty with hierarchical methods concerns its complexity, both in terms of time and space, limiting its range of applicability. We here propose an extension of the method by integrating sampling techniques into the clustering process, in order to be able to process larger data sets (section III). The performance of the method and its extension is evaluated through a set of examples in section IV.

II. Cluster Isolation Criterion and Hierarchical Clustering

Let X be a set of patterns and $x_i \in \mathcal{R}^d$ represent an element in this set. Given pattern x_i and some dissimilarity measure, $d(.,.)$, between patterns, let (x_i, x_j, x_k) be the triplet of nearest neighbors:

$$\begin{array}{ll} (x_i, x_j, x_k) & - \text{ nearest neighbors} \\ & x_j : j = arg\ \min_l \{d(x_l, x_i) \quad , l \neq i\} \\ & x_k : k = arg\ \min_l \{d(x_l, x_j) \quad , l \neq i, \neq j\}\ . \end{array}$$

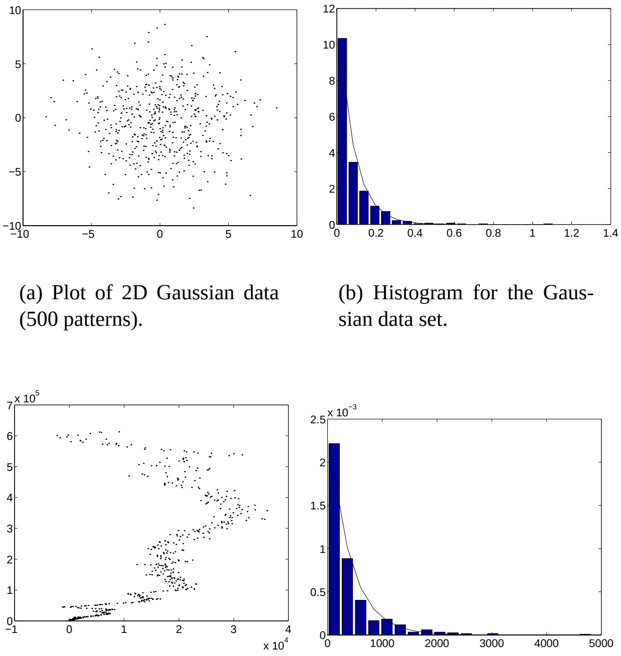
We define *dissimilarity increment* between the neighboring patterns by

$$d_{inc}(x_i, x_j, x_k) = |d(x_i, x_j) - d(x_j, x_k)|,$$

which can be seen as the first derivative of the dissimilarity function at the first point of the ordered list of neighboring samples. The dissimilarity increments between neighboring patterns within a natural cluster typically exhibit an exponential distribution [16], as illustrated in figure 1. Each cluster is hence characterized by a parametrical model (exponential distribution), which defines a context. According

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. Contacts: E-mail: afred@lx.it.pt

to this parameterization of clusters, two clusters are distinguishable if they exhibit distinct distributions and/or if they are well separated, in which case patterns in distinct clusters are placed far in the tail of the other cluster distribution. This constitutes the basis of the proposed cluster isolation criterion: given two clusters candidate for merging, evaluate each cluster from its context point of view; if the dissimilarity increments between neighboring patterns from the first to the second cluster are inconsistent with the first cluster statistics, isolate this cluster.



(a) Plot of 2D Gaussian data (500 patterns).

(b) Histogram for the Gaussian data set.

(c) Directional expanding data model.

(d) Histogram for the expanding data set.

Figure 1 - Histograms (bar graphs) and fitted exponential distributions (solid line curves) of the dissimilarity increments computed over neighboring patterns in the data. (a)- 2D gaussian distribution ($N([0,0],[10\ 0; 0\ 10])$). (c)- Data generated by the model: $x(k+1) = x(k) + n_s(k)k, y(k+1) = y(k) + n(k)$, where n_s and $n(k)$ represent uniform noise in the range $[-10; 10]$ and $[0; 10]$, respectively.

Adopting a hierarchical agglomerative strategy, with dissimilarity between clusters being defined as the minimum dissimilarity between inter-cluster pattern pairs (dissimilarity between nearest neighbor patterns in either cluster, as with the single link method), the concept of dissimilarity increments between patterns is easily extended to the context of clusters. We define *gap* between two clusters as the dissimilarity increment between nearest neighbor patterns in opposite clusters. This leads to a context-dependent definition of *gap*. Let C_i, C_j be two clusters candidate for merging and let $d_t(C_i)$ and $d_t(C_j)$) represent the value of the dissimilarity on the latest pattern association in cluster C_i and C_j, respectively. Let $d(C_i, C_j)$ be the dissimilarity between the two clusters. We define *dissimilarity increment* or *gap* between cluster i and cluster j as the asymmetric increase in the dissimilarity value, needed in order to allow the data association into a single cluster, as seen from C_i context:

$$gap_i = d(C_i, C_j) - d_t(C_i). \qquad (1)$$

Graphically, these gaps correspond to intervals between successive cluster associations in the dendrogram, seen from each cluster point of view (see figure 2).

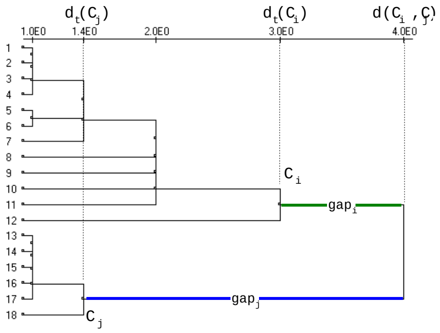


Figure 2 - Dendrogram and the definition of gap.

The cluster isolation criterion can be stated as follows:

- *Let C_i, C_j be two clusters which are candidates for merging, and let μ_i, μ_j be the respective mean values of the dissimilarity increments in each cluster. Compute the increments for each cluster, gap_i and gap_j, as defined in equation (1). If $gap_i \geq \alpha\mu_i$ ($gap_j \geq \alpha\mu_j$), isolate cluster C_i (C_j) and proceed the clustering strategy with the remaining patterns. If neither cluster exceeds the gap limit, merge them.*

Notice that the above criterion can be regarded as a context-dependent cluster isolation rule where the context is modelled by the parametric distribution of dissimilarity increments. The isolation rule consists of comparing the value of the dissimilarity increment, seen from the context of each cluster, with a dynamic threshold, $\alpha\mu_i$, computed from this context; inconsistency of *gap* values in a given context (cluster) determines the isolation of that cluster.

The design parameter, α, constrains the degree of isolation; values in the range 3–5 provide reasonable choices, corresponding to the rejection of atypical patterns [16].

A. Hierarchical Clustering Algorithm

The schematic description in table I incorporates the cluster isolation criterion described in the previous section into a hierarchical agglomerative type clustering algorithm.

B. Illustrative Example

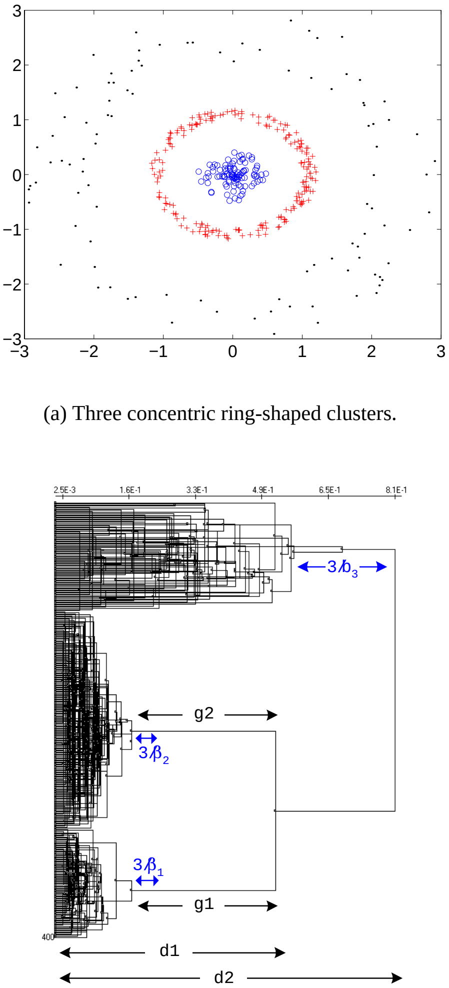
Figure 3 illustrates the clustering algorithm on three concentric 2D clusters (figure 3(a)). The k-means algorithm, imposing spherical clusters on the data, is unable to correctly identify the natural clusters. The single link method doesn't perform better (see the dendrogram in figure 3(b)): setting a threshold on the dendrogram will either lead to the merging of the inner clusters or to the splitting of the outer cluster.

According to the proposed method, dissimilarity increments are compared with a dynamic, cluster dependent threshold. For instance, the gaps, $g1$ and $g2$, are compared

Table I

HIERARCHICAL CLUSTERING ALGORITHM BASED ON DISSIMILARITY INCREMENTS.

```
Input: $N$ samples; $\alpha$ (default value is 3).
Output: Data partitioning.
Steps:
  1. Set: $Final\_clusters = \phi; n = N$;
       Put the $i$th sample in cluster $C_i,\ i = 1, \ldots, n$;
       $Clusters = \bigcup_i C_i,\ i = 1, \ldots, n$;
       $d_t[i] = \mu[i] = jumps[i] = 0,\ i = 1, \ldots, n$;
  2. If $(Clusters == \phi)$ or $(n == 1)$
       then stop, returning $Final\_clusters \bigcup Clusters$;
      else continue.
  3. Choose the most similar pair of clusters $(C_i, C_j)$ from $Clusters$. Let
       $gap_i = d(C_i, C_j) - d_t[i]$
       $gap_j = d(C_i, C_j) - d_t[j]$
  4. If $((\mu[i] == 0)$ or $(gap_i < \alpha\mu[i]))$ and
          $((\mu[j] == 0)$ or $(gap_j < \alpha\mu[j]))$
       then
          join the clusters $C_i, C_j$ into cluster $C_{i,j} : C_{i,j} = C_i \bigcup C_j$
          Let $I$ be the index for the merged cluster;
          Replace $C_i, C_j$ by $C_{i,j}$ in $Clusters$;
          $d_t[I] = d(C_i, C_j)$;
          $jumps[I] = jumps[i] + jumps[j] + 2$;
          $\mu[I] = \mu[i]\frac{jumps[i]}{jumps[I]} + \mu[j]\frac{jumps[j]}{jumps[I]} + \frac{gap_i + gap_j}{jumps[I]}$;
          Go to step 2.
       else continue.
  5. If $(gap_i \geq \alpha\mu[i])$
       then set $Final\_clusters = Final\_clusters \bigcup C_i$;
          Remove $C_i$ from $Clusters$;
          $n = n - 1$.
       end if
       If $(gap_j \geq \alpha\mu[j])$
       then set $Final\_clusters = Final\_clusters \bigcup C_j$;
          Remove $C_j$ from $Clusters$;
          $n = n - 1$.
       end if
      Go to step 2.
```



(a) Three concentric ring-shaped clusters.

(b) Dendrogram produced by the single link method. $d1$ and $d2$ represent distances between clusters; $g1$ and $g2$ represent gaps seen from each cluster. These gaps are compared with a dynamic threshold, which is context dependent ($\alpha = 3$ is shown). The proposed method isolates the inner clusters, thus recovering the true cluster structure in (a).

Figure 3 - Clustering of concentric patterns.

with the corresponding cluster threshold ($3/\beta_1 = 0.053$ and $3/\beta_2 = 0.035$, respectively). As a result, two clusters are isolated and frozen in the dendrogram; merging steps continue with the remaining data, thus leading to a third cluster, as gaps are smaller than the cluster threshold, $3/\beta_3 = 0.21$, the true clusters being identified.

III. SAMPLING TECHNIQUES IN CLUSTERING

A major difficulty with the hierarchical approaches concerns its computational complexity, both in terms of time and space. Hierarchical methods operate on $n(n-1)/2$ entries of a symmetric $n \times n$ proximity matrix, with n being the number of patterns. This limits the usage of this technique to relatively small data sets.

In this section we propose the combination of sampling techniques with the previous clustering method in order to extend the range of applicability to larger data sets. The idea is to map the high dimensional data set into a reasonable small number of prototypes, easily handled by the hierarchical clustering method. Simple random selection among the training patterns does not provide a good solution, as total randomness may lead to a distortion of inter- and intra-cluster relationships. We therefore partition the data sets into a large number of small and compact clusters, representing each cluster by its centroid; centroids are then clustered using the hierarchical clustering technique; the corresponding data partition is obtained by joining all the patterns represented by the prototypes gathered in the same cluster. The K-means algorithm is elected to perform this mixture decomposition of data, due to its simplicity and computational efficiency. It may happen that small natural clusters are merged in this sampling+clustering step. Each formed cluster should therefore go through a more detailed analysis, either by subsequent application of the sampling technique or by direct clustering using the hierarchical method (depending on the number of patterns present), for detection of finer grained clusters.

The overall method can be described schematically as follows:

- Step1. If the data set is small, use each sample as a centroid and go to step 2. Otherwise, decompose the data into a large number of small hyper-spherical clusters using the K-means algorithm.

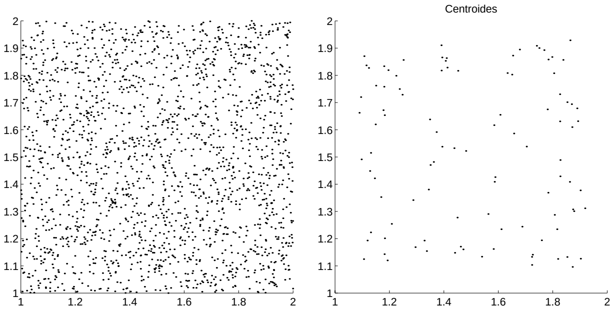
- Step2. Apply the hierarchical method to the centroids representing the clusters in the previous mixture decomposition and obtain a partition.
- Step3. Get a data partition by assigning each sample to the cluster where its representative (centroid) belongs. If a single cluster was obtained, stop the procedure; otherwise, repeat steps 1 to 3 with each cluster in the partition.

IV. Experimental Results

The hierarchical clustering method and its extension are tested in a set of examples.

A. Uni-Modal Random Data

It is known that most of the clustering algorithms impose structure on data. We here evaluate the performance of the proposed method on random data drawn from a uniform distribution. Figure 4(a) shows a 2D projection of 2000 patterns uniformly distributed in a 5-dimensional hypercube. Applying the hierarchical clustering algorithm to this data set, using the Euclidean distance and $\alpha = 3$ (default value), a single cluster is obtained. The combined K-means+hierarchical clustering of centroids technique was tested with $k = 50$, 100 and 200. Figure 4(b) shows the centroids given by the K-means algorithm, with $k = 100$. The extended algorithm led to the identification of a single cluster with a considerable increase in efficiency.



(a) 2-D projection of random data.

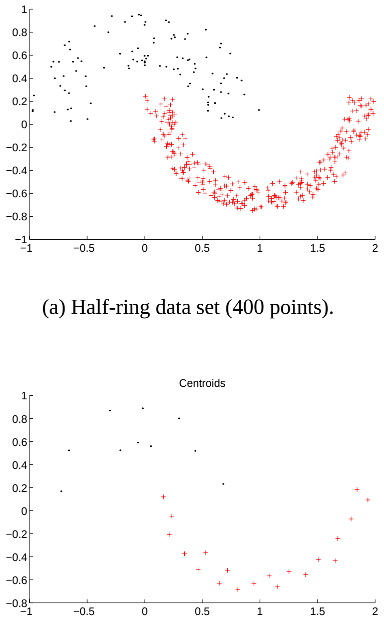
(b) Sampling at 100 points.

Figure 4 - Random data and prototypes obtained by the K-means algorithm.

B. Half-Rings Data Set

The half-rings data set, represented in figure 5(a), constitutes an example of well separated clusters easily handled by the proposed techniques, that are not adequately handled by the single link (see figure 6) or the k-means algorithm.

Direct application of the hierarchical method based on dissimilarity increments identifies the two natural clusters in figure 5(a). The combined K-means+hierarchical clustering technique led to the consistent identification of two clusters based on centroids (tests included $k = 20$, 30 and 50), as illustrated in figure 5(b). Application of the technique to each cluster produced on the first phase of the extended technique led to no further partitioning, the true cluster structure being once again recovered.



(a) Half-ring data set (400 points).

(b) Hierarchical clustering of 30 centroids.

Figure 5 - Half-rings data set and prototypes using the k-means algorithm.

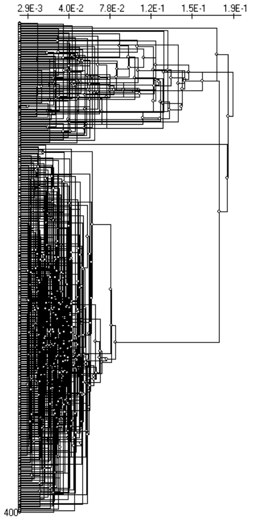
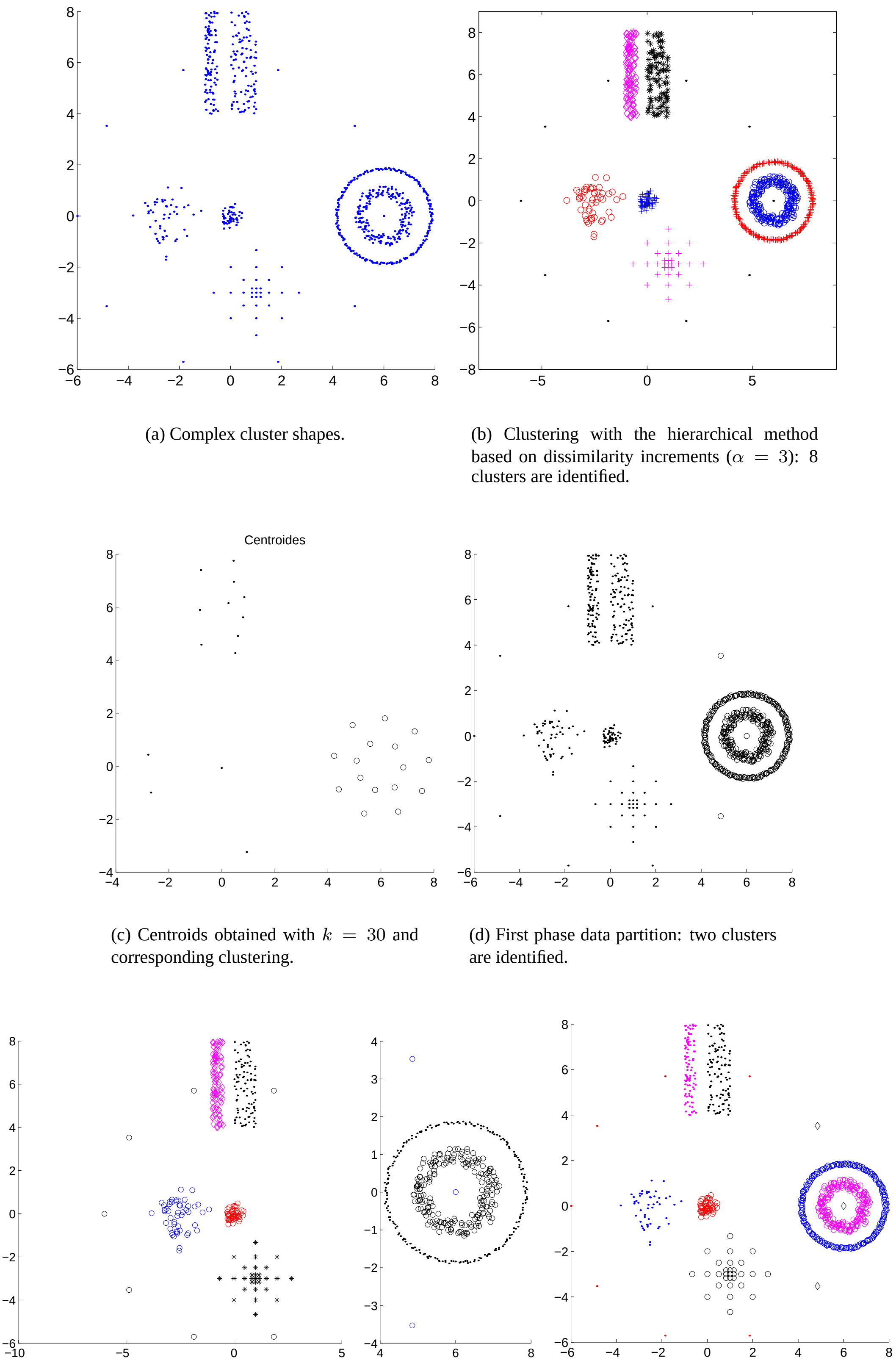


Figure 6 - Single-link method on the half-ring data. Thresholding this graph splits the upper ring cluster into several small clusters.

C. Complex Image

The final example constitutes a complex structure of clusters (see figure 7(a)). The hierarchical clustering method based on dissimilarity increments leads to the identification of 8 clusters, as plotted in figure 7(b).



(a) Complex cluster shapes.

(b) Clustering with the hierarchical method based on dissimilarity increments ($\alpha = 3$): 8 clusters are identified.

(c) Centroids obtained with $k = 30$ and corresponding clustering.

(d) First phase data partition: two clusters are identified.

(e) Second phase of the clustering procedure: further partitioning of the first sub-cluster.

(f) Partitioning of the second sub-cluster.

(g) Final data partition.

Figure 7 - Clustering of complex clusters based on dissimilarity increments.

The extended version was tested with $k = 30$, 50 and 70. The combined strategy now required three phases of analysis, the clusters identified being consistent with the ones represented in figure 7(b), except for the outer circle which was split into a variable number of clusters, depending on k and on the K-means initialization. It should be noticed that this is an extremely sparse structure intermingled with the remaining clusters, which justifies the variability of associations made.

Figures 7(c) to 7(g) illustrate the process for $k = 30$. On the first phase of the combined method, 30 centroids are defined by the k-means algorithm, and organized into two clusters by the hierarchical method (fig. 7(c)); this leads to a first division of the patterns as shown in figure 7(d). A second round of the procedure is then run on each isolated cluster, leading to the partitions in figures 7(e) and 7(f). Application of the clustering technique to each of the clusters identified on the second phase does not produce further partitioning. The resulting clusters are therefore the ones plotted in figure 7(g).

For $k = 50$ and $k = 70$ comparable results were obtained, with an initial partition into 3 clusters; the outer circle was split into tree clusters, and the remaining clusters were correctly identified.

V. Conclusions

A hierarchical clustering algorithm, exploring the idea of context dependent clustering, was presented. According to the proposed method, inter-pattern relationships are modelled by the statistical distribution of dissimilarity increments between neighboring patterns. This distribution is used to characterize each cluster, forming the basis of a context-based cluster isolation criterion. Introduction of this criterion into a hierarchical agglomerative technique leads to the partitioning of the data without requiring ad-hoc specification of parameters.

The method was further extended by proposing the integration of sampling techniques into the clustering process in a combination of the K-means algorithm with the hierarchical method.

Experimental results showed the ability of the method to identify arbitrarily shaped, well separated clusters. The extended method provided an efficient way to cluster the data, without or with minor degradation of the clustering results, thus expanding the range of applicability of the clustering strategy to larger data sets.

References

[1] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988.

[2] G. McLachlan and K. Basford, *Mixture Models: Inference and Application to Clustering*, Marcel Dekker, New York, 1988.

[3] S. Roberts, D. Husmeier, I. Rezek, and W. Penny, “Bayesian approaches to gaussian mixture modelling”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, November 1998.

[4] M. Figueiredo, J. Leitão, and A. K. Jain, “On fitting mixture models”, in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, E. Hancock and M. Pellilo, Eds. 1999, pp. 54–69, Springer Verlag.

[5] 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*, San Jose, January 1999, vol. SPIE Vol. 3656, pp. 501–510.

[6] J. Buhmann and M. Held, “Unsupervised learning without overfitting: Empirical risk approximation as an induction principle for reliable clustering”, in *International Conference on Advances in Pattern Recognition*, Sameer Singh, Ed. 1999, pp. 167–176, Springer Verlag.

[7] B. Mirkin, “Concept learning and feature selection based on square-error clustering”, *Machine Learning*, vol. 35, pp. 25–39, 1999.

[8] D. Stanford and A. E. Raftery, “Principal curve clustering with noise”, Tech. Rep., University of Washington, http://www.stat.washington.edu/raftery, 1997.

[9] J. A. Garcia, J. Valdivia, F. J. Cortijo, and R. Molina, “A dynamic approach for clustering data”, *Signal Processing*, vol. 2, pp. 181–196, 1995.

[10] C. Zahn, “Graph-theoretical methods for detecting and describing gestalt structures”, *IEEE Trans. Computers*, vol. C-20, no. 1, pp. 68–86, 1971.

[11] Y. El-Sonbaty and M. A. Ismail, “On-line hierarchical clustering”, *Pattern Recognition Letters*, pp. 1285–1291, 1998.

[12] M. Chavent, “A monothetic clustering method”, *Pattern Recognition Letters*, vol. 19, pp. 989–996, 1998.

[13] A. L. Fred and J. Leitão, “A comparative study of string dissimilarity measures in structural clustering”, in *International Conference on Advances in Pattern Recognition*, Sameer Singh, Ed., pp. 385–384. Springer, 1998.

[14] R. Dubes and A. K. Jain, “Validity studies in clustering methodologies”, *Pattern Recognition*, vol. 11, pp. 235–254, 1979.

[15] T. A. Bailey and R. Dubes, “Cluster validity profiles”, *Pattern Recognition*, vol. 15, no. 2, pp. 61–83, 1982.

[16] 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*, Barcelona, 2000, vol. 2, pp. 190–194.