Clustering Under a Hypothesis of Smooth Dissimilarity Increments

Ana L.N. Fred
Instituto Superior Técnico
Instituto de Telecomunicações
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
afred@lisboa.lx.it.pt

José M. N. Leitão
Instituto Superior Técnico
Instituto de Telecomunicações
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
jleitao@red.lx.it.pt

Abstract

The problem of cluster defining criteria has been addressed in various forms. In this paper, a new cluster isolation criterion is proposed, underlying a hypothesis of smooth dissimilarity increments between neighboring patterns within a cluster. This isolation criterion is merged in a hierarchical agglomerative clustering algorithm, producing a data partitioning and simultaneous accessibility to the intrinsic data inter-relationships in terms of a dendrogram-type graph. By defining adequate dissimilarity measures, the new algorithm is applied to vector based pattern analysis and to categorization of structural patterns. Both simulated data and real applications, in the context of automatic analysis of contour images, are presented to illustrate and evaluate the method. Examples demonstrate the versatility of the method in identifying arbitrary shape and size clusters, intrinsically finding the number of clusters.

1. Introduction

Clustering algorithms play an important role in exploratory data analysis and datamining, providing a means to ascertain structure within the data. Two main strategies are used for clustering: hierarchical methods and partitional methods [3, 7]. Partitional structure organizes patterns into a small number of clusters.

Hierarchical methods propose a nesting of clusterings, providing additional information about data structure, represented graphically as a dendrogram. A particular algorithm can be obtained by the definition of the similarity measure between patterns and clusters [4], the later ultimately conditioning the structure of the clusters identified. The single link algorithm is one of the most popular methods in this class [7]. Data partitioning is usually obtained by setting a threshold on the dendrogram; cluster validity studies have also been proposed [2, 1] for the *a posteriori* analysis of structures, in order to evaluate the clustering results and define meaningful clusters.

In this paper we propose a new criterion for cluster isolation based on a hypothesis of smooth dissimilarity increments between neighboring patterns within a cluster. The integration of this criterion in a hierarchical clustering framework produces a partitioning of the data, while exhibiting data inter-relations in terms of a dendrogram type graph. The structure of the obtained dendrogram, unlike conventional hierarchical clustering methods, is isolation criterion dependent, expanding the range of pattern structures handled by these methods. Additionally, the problem of deciding the number of clusters is subsumed and intrinsically dictated by the criterion.

Section 2 outlines the new cluster isolation criterion. A hierarchical agglomerative algorithm adopting this criterion is described in section 3. The characteristics of the new method are analysed and illustrated in based on examples, covering both simulated (section 4) and real data, the later in the context of automatic analysis of contour images (section 5). Conclusions are drawn in a last section.

2. Smoothness Hypothesis and Cluster Isolation Criterion

The proposed cluster isolation criterion derives from the following intuitive concepts and assumptions:

- A cluster is a set of patterns sharing important characteristics in a given context;
- A dissimilarity measure encapsulates the notion of pattern resemblance;

- Higher resemblance patterns are more likely to belong to the same cluster and should be associated first;
- Dissimilarity between neighboring patterns within a cluster should not occur with abrupt changes;
- The merging of well separated clusters incur in abrupt changes in dissimilarity values.

The first two assumptions emphasize the fact that the way clustering algorithms address patterns interrelationships are context dependent, context being asserted in terms of dissimilarity functions.

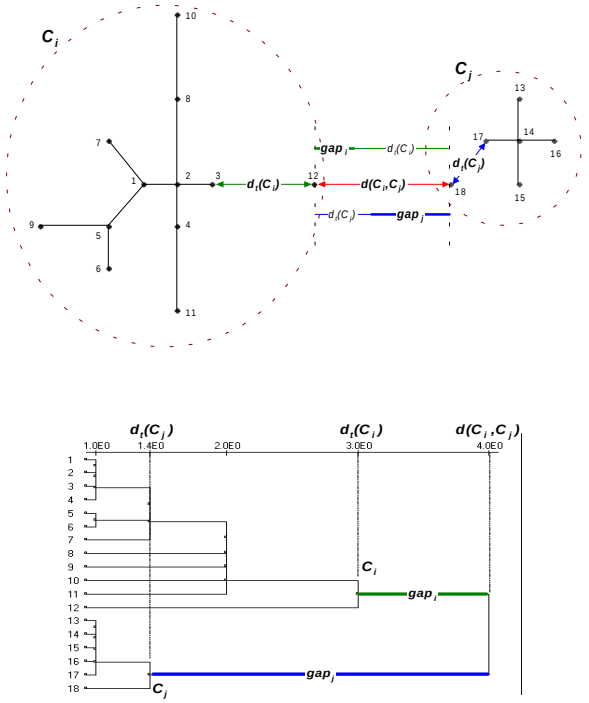

Figure 1. Definition of gap in an example of 18 2-dimensional points associated using the single-link method, adopting the Euclidean distance. On top: plot of data, grouped in two clusters. Bottom: corresponding dendrogram.

The above considerations suggested the definition of a new cluster isolation criterion for aglomerative type algorithms, based on the analysis of increments of dissimilarity measures between neighboring patterns. Let C_i, C_j be two sets of patterns, candidate for merging, and assume that patterns are included in clusters with increasing order of dissimilarity. Let $d_t(C_i)$ ($d_t(C_j)$) represent the minimum dissimilarity for the formation of cluster i (j, respectively), that is, the value of the dissimilarity of the latest pattern association, and $d(C_i, C_j)$ the dissimilarity between the two clusters. We define *dissimilarity increment* or *gap* between cluster i and j as the asymmetric increase in the dissimilarity value, needed in order to allow the data association into a single cluster:

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

Figure 1 illustrates the concept using the single link method for cluster association.

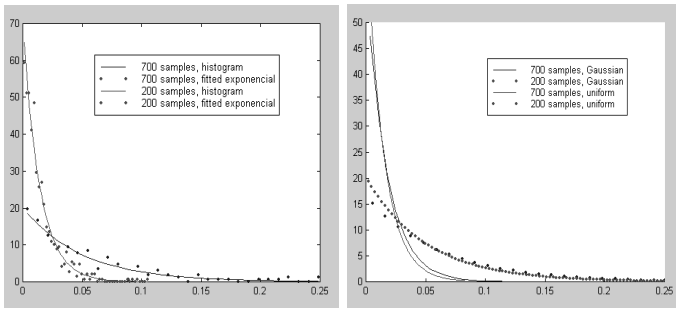

Figure 2. Distribution of $gaps$. The plots on the left represent histograms (dotted lines) and fitted exponentials for the gaps computed between neighboring patterns generated randomly from a uniform distribution in the interval $[0, 10]$; on the right, the curves represent estimated gap distributions for data generated from the same uniform distribution or from a gaussian $N(0, 5)$. As shown, higher data densities (corresponding, for each data model, to a higher number of generated patterns) lead to narrower distributions.

The statistical distribution of these increments within a cluster has a smooth evolution, and can be modelled by an exponential probability density function, $p(x) = \beta exp^{-\beta x}, x > 0$, the parameter β characterizing the data dispersion. Figure 2 shows histograms and estimated exponentials for *gaps* computed from data randomly generated according to Gaussian and uniform distributions. This shows that different data generating models lead to almost coincident curves, the parameter β essentially reflecting data sparseness. Tails of these distributions correspond to patterns in frontier or borderline positions between clusters. On the other hand, increments computed for elements in distinct clusters will, in general, have high values, located on the tail of the statistic for each cluster.

The idea for cluster isolation is therefore to define a limit on the dissimilarity increments such that

most of the patterns densely connected are included in the same cluster,while all others, not conveying this smoothness hypothesis, are rejected. The setting of this threshold is motivated by the following fact, characteristic of the exponential distribution: the crossing of the tangential line, at points multiple of the distribution mean value, $i \times \frac{1}{\beta}$, with the x axis, is given by $(i+1) \times \frac{1}{\beta}$, as shown in figure 3. This suggests the choice of a threshold as a multiple of the mean value of the distribution. The isolation criterion can be stated as:

- *Let C_i, C_j be two clusters, candidate for merging, and μ_i, μ_j be the respective mean values of dissimilarity increments in each cluster. Compute the increments for each cluster, gap_i, gap_j, as in expression 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 data. If neither cluster exceeds the gap limit, join them.*

The design parameter, α, constrains the degree of isolation; values in the range $[3, 5]$ provide reasonable choices, as illustrated in figure 3.

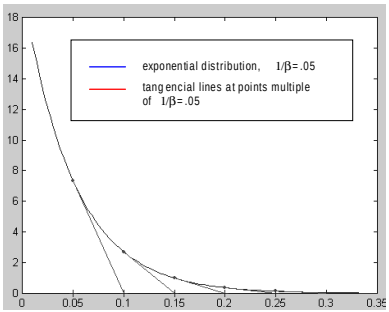

Figure 3. Defining a threshold on the *gaps* value (x axis). Dots are located on points multiple of the distribution mean, $\frac{1}{\beta}$, and red lines are tangents at those points. The crossing at the x axis occur at points $\frac{i}{\beta}$, i being a positive integer. Values for i in the range $[3, 5]$ cover the significant part of the distribution, being reasonable choices.

3. Hierarchical Agglomerative Partitioning Algorithm

The following describes a hierarchical aglommerative clustering algorithm that incorporates the proposed cluster isolation criterion.

Input: N_s samples; α.
Output: Data partitioning.
Steps:

1. Set: $Final\_clusters = \Phi$; $n = N_s$;
   $d_t[i] = \mu[i] = jumps[i] = 0,\ i = 1, \ldots, n$;
   Put each sample in a cluster C_i;
   $Clusters = \bigcup_i C_i$;
2. If $Clusters = \Phi$ or $n = 1$ stop, returning the clusters found in $Final\_clusters \bigcup Clusters$; 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 $(gap_j < \alpha\mu[j]))$
   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 $(gap_i \geq \alpha\mu[i])$ set
   $Final\_clusters = Final\_clusters \bigcup C_i$;
   Remove C_i from $Clusters$;
   $n = n - 1$.
   If $(gap_j \geq \alpha\mu[j])$ set
   $Final\_clusters = Final\_clusters \bigcup C_j$;
   Remove C_j from $Clusters$;
   $n = n - 1$.
   Go to step 2.

As the estimates of the mean values $\mu[i]$ are not reliable for very small samples sizes premature isolation of clusters may occur. In order to overcome this situation, inhibition of cluster isolation actions may be adopted when clusters have very low dimensions; in the examples provided next, inhibition of cluster isolation was only implemented when both clusters under analysis had less than 10 samples.

This algorithm, while producing a partitioning of the data, also provides relevant information on the relations between patterns, that can be graphically displayed as a dendrogram. It is important to note that,

unlike classical hierarchical clustering algorithms, the partitions obtain do not correspond to cutting the dendrogram at a given threshold, since, according to the isolation criterion proposed, clusters obeying the criterion will be frozen in the dendrogram, the remaining clusters continuing in the pursuit for possible merging. Therefore the structure of the resulting dendrogram is conditioned by the cluster isolation criterion. This and other characteristics are illustrated and analyzed in the next sections in light of examples.

4. Clustering of 2-D Patterns with Variable Structure

In order to evaluate the ability of the algorithm, and underlying criterion, in the identification of arbitrary shape, variable sized clusters, a complex test was built comprising the following simulated data: a 10, equally spaced, points circle; two concentric rings of dense (200 samples each), uniform distributed points; two parallel, bar-shaped, random sets with distinct densities (100 samples each); two neighboring, 50 point, Gaussian distributions with distinct covariances; a star-shaped pattern formed by 29 points. The above patterns were joined in an intermixed structure, where concentric and crossing patterns were configured.

Figure 4 summarizes the results obtained with the algorithm of section 3, using the Euclidean distance, for several values of the parameter α, in a direct comparison with the single link method.

Dense clusters will form earlier in the dendrogram than sparse data. While this type of structure could be handled by post validation of clusters within the dendrogram produced by the single link algorithm, a distinct feature of the proposed clustering criterion concerns the change in patterns associations due to cluster freezing in the agglomerative process. This is apparent in the comparison of the dendrograms produced by the two approaches. As described previously, when two patterns are confronted for merging (meaning that they are the ones most similar at the time) if the increase of dissimilarity needed for merging is not consistent with one of the cluster's statistic, but being with the other, the first is isolated (frozen in the dendrogram); the second is kept to the continued process of finding possible associations. This enables the gathering, for instance, of the points of the exterior circle (black dots in the middle left picture).

As shown, setting α too small (top left picture) produces spurious clusters due premature thresholding of the exponential density (see section 2). On the other hand, high values of α (picture on the top right), while not refining cluster identifiability, lead to cluster merg-

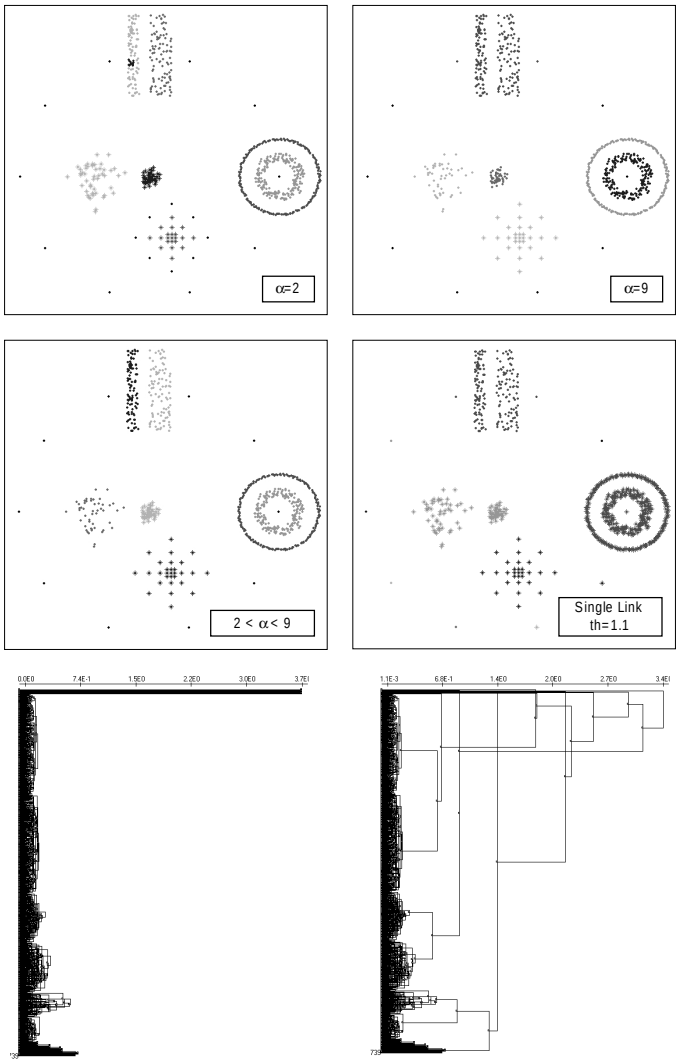

Figure 4. Clustering of complex structures using the new approach. Excessive or low value for the density limiting parameter, α, lead to merging of two clusters (top right) and the creation of a spurious cluster (top left), respectively. The theoretically meaningful choice for α leads to complete separation of clusters (middle left). The single link algorithm (middle right) cannot handle concentric, intermixed or highly disparate density patterns. On the bottom, the dendrograms produced by the new method (left) and the single link (right) are displayed.

ing as a result of tail superposition between neighboring clusters. The results obtained for α in the theoretically meaningful range (middle left picture) comply with intuitive separation of data, being able to cope with the different types of structures and sizes of clusters. The single link method cannot handle the mingled structures or the variability of data densities involved.

5. Application in Automatic Analysis of Contour Images

The problem here concerns the categorization of 181 contour images of 5 types of hardware tools (see figure 5) using string descriptions. Each image was segmented to separate the object from the background and the object boundary was sampled at 50 equally spaced points; object's shapes were encoded using an 8-directional differential chain code [6, 5]. We have shown previously [4] that direct application of hierarchical clustering based on string matching using symbol editing operations to this data does not produce a consistent partitioning of patterns; this is mainly due to non homogeneous distances between patterns in different classes as shown in figure 5, which presents the results of the single link algorithm, using the Levensthein distance normalized by the length of the editing path [8, 9] as dissimilarity measure. With the new algorithm, total separation between classes of tools was achieved by setting $\alpha = 4$. Furthermore, tool t3 was split into two clusters corresponding to distinct poses: open and closed.

6. Conclusions

A new criterion for cluster isolation was proposed based on the assumption of smooth dissimilarity increments between neighboring patterns within a cluster.

A hierarchical agglomerative algorithm adopting the isolation criterion was described and applied to both simulated and real data in the context of object recognition from contour images. As corroborated by the examples, the proposed criterion leads to potentially different dendrograms from the ones obtained with standard hierarchical procedures, based on the pruning effect on dendrogram branches; while conditioning the graph formation and shape, it also provides, at the end, a partitioning of the data.

Examples illustrate the versatility of the method in identifying arbitrary shape and size clusters, intrinsically finding the number of clusters. This has been shown in extensive tests, not presented here due to space limitations

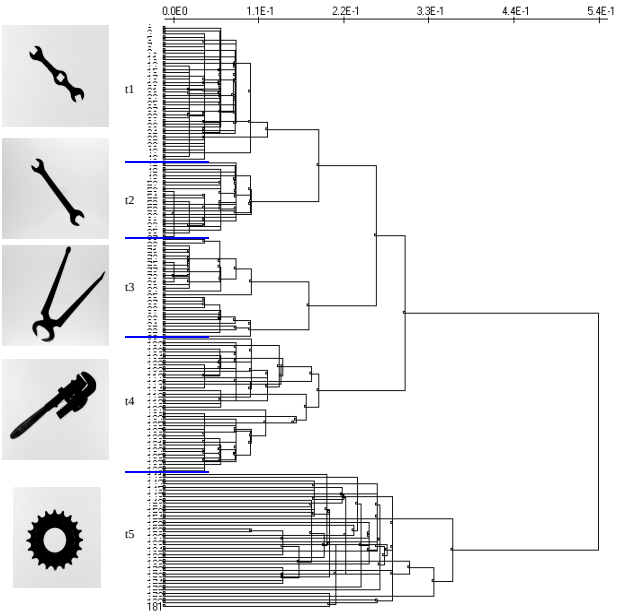

Figure 5. Typical samples of contour images and dendrogram obtained using the single link algorithm.

References

[1] T. A. Bailey and R. Dubes. Cluster validity profiles. *Pattern Recognition*, 15(2):61–83, 1982.

[2] R. Dubes and A. K. Jain. Validity studies in clustering methodologies. *Pattern Recognition*, 11:235–254, 1979.

[3] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[4] A. L. N. Fred and J. M. N. Leitão. A comparative study of string dissimilarity measures in structural clustering. In S. Singh, editor, *International Conference on Advances in Pattern Recognition*, pages 385–384. Springer, 1998.

[5] A. L. N. Fred, J. S. Marques, and P. M. Jorge. Hidden markov models vs syntactic modeling in object recognition. In *ICIP'97*, 1997.

[6] A. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, 1989.

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

[8] A. Marzal and E. Vidal. Computation of normalized edit distance and applications. *IEEE Trans. Pattern Anal. and Machine Intelligence*, 2(15):926–932, 1993.

[9] E. S. Ristad and P. N. Yianilos. Learning string-edit distance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(5):522–531, May 1998.