A new cluster isolation criterion based on dissimilarity increments

Ana L.N. Fred, José M.N. Leitão
PAMI 2003
Can you find the 8 clusters?

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

(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.
How? Dissimilarity increments

- Define: \((x_i, x_j, x_k)\) — nearest neighbors
  \[x_j : j = \arg \min_l \{d(x_l, x_i) \mid l \neq i\}\]
  \[x_k : k = \arg \min_l \{d(x_l, x_j) \mid l \neq i, l \neq j\}\].

- then the dissimilarity increment is:
  \[d_{inc}(x_i, x_j, x_k) = |d(x_i, x_j) - d(x_j, x_k)|,\]

It characterises the local density
How?

- Dissimilarity increments in a cluster typically have an exponential distribution:

\[ p(x) = \frac{1}{\lambda} \exp\left(-\frac{x}{\lambda}\right), \quad x > 0, \]

where \( \lambda \) is the scale parameter. Experimental evidence shows that the increments of the dissimilarity measure between neighboring patterns, as defined above, typically exhibit an exponential distribution. This is illustrated in Fig. 1, which plots histograms and fitted distributions of dissimilarity increments for various data sets. Examples include uniformly distributed patterns, Gaussian-distributed patterns, random patterns, patterns generated according to a stochastic model, and grid corrupted by Gaussian noise.
Use this in hierarchical clustering

- Define distance between two clusters $C_i$ and $C_j$:

$$gap_i = |d(C_i, C_j) - d_t(C_i)|.$$ 

- Check if both gaps are small (compared to distr. of dist.)

- If **small**: merge clusters

- If only one gap is **large**: freeze cluster
Conclusions

- Define clusters based on local density
- Freezing clusters allows for clusters with different densities

Fig. 18. Dendrograms produced by the single-link and the proposed method. A plot of the clustering obtained is overlayed on the graph. (a) Single-link method. Clustering: cut at level 2. (b) Proposed method: clustering obtained with $\rho < 10^{-6}$. 

\[ p(x) = \frac{1}{\lambda} \exp \left( -\frac{x}{\lambda} \right) \]

For simplicity, only the circle and the star-shaped clusters are considered. Fig. 18a shows the dendrogram produced by the single-link method. Due to spatial proximity, a few points of the circle are associated with the star-shaped cluster. The proposed method (Fig. 18b) changes the way the dendrogram is produced by eliminating the association of the star with its nearest point on the circle (pink cross on the plot in Fig. 18a). Since this association is not possible according to the statistics of dissimilarities in the star pattern, the later is frozen in the dendrogram; further associations continue with the remaining data, making it possible to connect the circle (plot in Fig. 18b).