A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise

F. Ros, Serge Guillaume

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise

Frédéric Ros[a,*], Serge Guillaume[b]

[a] *Laboratory PRISME, Orléans university, France*
[b] *ITAP, Univ Montpellier, Irstea, Montpellier SupAgro, Montpellier, France*

Abstract

Hierarchical clustering is widely used in data mining. The single linkage criterion is powerful, as it allows for handling various shapes and densities, but it is sensitive to noise. Two improvements are proposed in this work to deal with noise. First, the single linkage criterion takes into account the local density to make sure the distance involve core points of each group. Second, the hierarchical algorithm forbids the merging of representative clusters, higher than a minimum size, once identified. The experiments include a sensitivity analysis to the parameters and a comparison of the available criteria using datasets known in the literature. The latter proved that local criteria yield better results than global ones. Then, the three single linkage criteria were compared in more challenging situations that highlighted the complementariness between the two levels of improvement: the criterion and the clustering algorithm.

Keywords: Agglomerative, dissimilarity, density

1. Introduction

Data reduction plays an important role in data mining, either for knowledge discovery or information summary, and clustering is a popular way to achieve this goal. Many techniques are available. Hierarchical clustering, refer

*Corresponding author
Email addresses: frederic.ros@univ-orleans.fr (Frédéric Ros), serge.guillaume@irstea.fr (Serge Guillaume)

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

to Murtagh & Contreras (2012) for an overview, includes a family of algorithms that yield a set of nested partitions between the trivial two extremes: the one in which the clusters are made up of singletons and the unique cluster with all the items.

Two strategies are possible, top-down or bottom-up approaches: divisive algorithms start from the whole set while agglomerative ones start from the singletons. From the data representation point of view two schemes are possible: central and pairwise clustering (Murtagh & Contreras, 2012). In central clustering, the data are described by their explicit coordinates in the feature space and each cluster is represented by a prototype. This group includes the centroid, median and minimum variance methods. The latter is also called the Ward criterion. It agglomerates the two clusters such that the within-class variance of the whole partition thereby obtained is minimum. The minimum variance method produces clusters which satisfy compactness and isolation criteria and hierarchies are also more balanced which is often of practical advantage.

In pairwise clustering, the data are indirectly represented by a dissimilarity matrix, which provides the pairwise comparison between different elements. In this family, the agglomerative linkage criterion combines the dissimilarities between items. The complete linkage selects the pair of clusters whose merge has the smallest diameter, i.e. the two clusters for which the maximum dissimilarity is minimum, the single linkage selects the ones for which the minimum dissimilarity between items in the two clusters is minimum. Between these extremes, the average linkage criterion is computed as the average dissimilarity. The average and complete linkage have the advantage of clustering compact clusters and yield well localized classes.

A simple example in the next section shows that dealing with well separated clusters, without noise, the above-mentioned techniques fail with the noticeable exception of the single linkage criterion. This criterion generalizes the nearest neighbor concept to sets. This is useful to tackle thin clusters but this local criterion also fails with a small amount of noise.

Two approaches are proposed to improve the behavior of the single linkage

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

criterion, which is known to be prone to yield undesirable elongated clusters especially in presence of noise. This drawback is also called "the chaining effect".

The first one deals with the criterion itself by taking into account the local density. According to the local density, items may be labeled as noise. The single linkage criterion is applied until the two closest points, one in each of the clusters, are not labeled as noise. The final value is the average of the distances of all the points in between weighted by the local densities. The result depends on the amount of noise. The value is higher than the single linkage one, as it is affected by noise, but lower than the distance between the two points that are not labeled as noise. The idea is to propose a criterion that partially inherits the properties of *global* linkages without their limitations.

The second approach comes to propose a modified agglomerative clustering algorithm that gives the expected number of representative clusters. A clustering algorithm may be used with different objectives. When the aim is knowledge or data structure discovery, the number of clusters cannot be a parameter. Instead, the algorithm itself has to propose a suitable partition. In this work, the number of desired clusters is known and the goal for the agglomerative algorithm is to yield the number of representative clusters, i.e. with a minimum size, to avoid isolated points. The agglomerative process is first carried out until the number of representative clusters is reached. In the second step, an important restriction constraints the same process: the merging between the identified clusters is forbidden.

The ambition of this paper is thus to improve the single linkage criterion while promoting a modified version of the hierarchical clustering algorithm to better manage noisy data. The rest of the paper is organized as follows. Section 2 recalls the basics, shows the interest and properties of the single linkage criterion but also illustrates its failure in presence of noise using a toy example. Section 3 is dedicated to the improvement of the single linkage criterion. The main idea is illustrated using a simple example, then the local density estimation, and noise labeling, is detailed and, finally, the global behavior of the proposal is studied. The second part of the proposal, the hierarchical algorithm

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

to deal with noise is described in Section 4. The impact of the main parameter, the proportion of the data for considering the partition as representative, is also studied. Numerical experiments are carried out in Section 5. They show the complementariness of the two approaches. The final remarks and open perspectives are stated in Section 6.

2. Single linkage is powerful

Given a set of elements, S, a function d is called a distance if it satisfies the following properties, $\forall i, j \in S$:

1. $d(i,j) \geq 0$, non-negativity
2. $d(i,j) = d(j,i)$, symmetry
3. $d(i,j) = 0 \iff i = j$
4. $d(i,j) \leq d(i,l) + d(j,l)$, $\forall l \in S$

The function d is called a pseudo-metric when the property (3) is weakened as follows: $i = j \Longrightarrow d(i,j) = 0$.

When the triangular inequality, property (4), is not fulfilled, the function d is called a dissimilarity.

Let c_i and c_j be two groups of items, $i \in c_i$ and $j \in c_j$ and d a distance function. The most popular agglomerative methods are:

- Single link: $d_s(c_i, c_j) = \min_{i,j} d(i,j)$
- Complete link: $d_c(c_i, c_j) = \max_{i,j} d(i,j)$
- Average link: $d_a(c_i, c_j) = \dfrac{1}{|c_i||c_j|} \sum_{i,j} d(i,j)$
- Centroid: $d_g(c_i, c_j) = d(g_i, g_j)$, with $g_i = \dfrac{\sum_i i}{|c_i|}$
- Median: $d_m(c_i, c_j) = median\ \{d(i,j)\}$
- Ward: $d_w(c_i, c_j) = \dfrac{|c_i||c_j|}{|c_i| + |c_j|} d(g_i, g_j)$

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

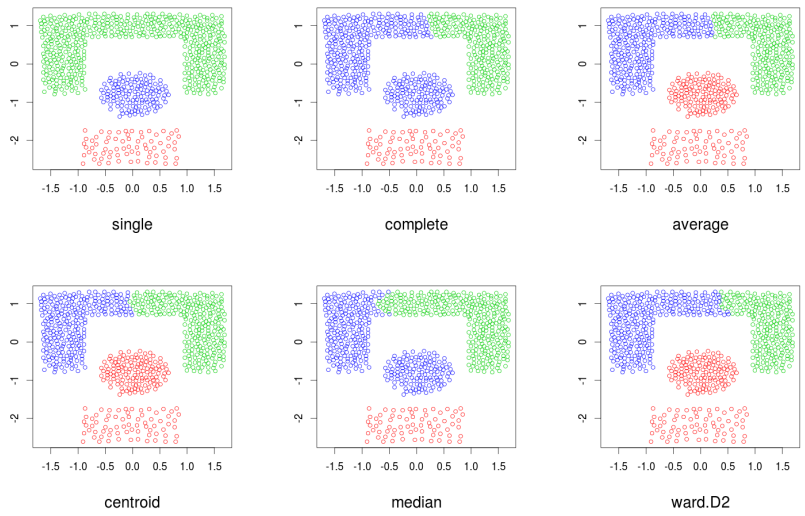

Figure 1: Illustrative behavior of the agglomerative methods. The axis labels are the x and y coordinates.

The first three methods are representative of the linkage methods while in the last three ones the cluster centers are specified. The Ward method is often referred to as the minimum variance method.

The elementary distance, $d(i, j)$, is usually the Euclidean distance but the linkage methods, single, complete or average, are not restricted to the Euclidean distance. Other metrics can be used, among them the Manhattan, Chebyshev or Angular ones.

The methods' behavior is illustrated using a simple example with 3 well separated clusters of different shapes and densities. The hierarchical clustering is performed using the *hclust* R function (R Core Team, 2013). In R, the Ward criterion is implemented in the Ward.D2 method.

Figure 1 shows that the only method that yields the expected result in this illustrative case is the single linkage one.

Single linkage generalizes the nearest neighbor concept to sets. As it takes into account only one element in each set, it allows for the identification of thin clusters of various shapes, e.g. lines or spirals, which may correspond to roads

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

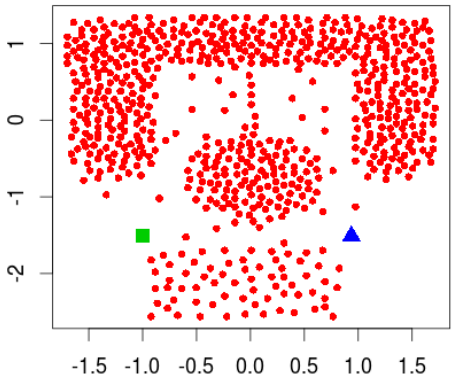

Figure 2: Illustrative behavior of the single linkage criterion using a basic hierarchical algorithm with the presence of noise. The axis labels are the x and y coordinates.

or cyclones in image analysis.

Moreover, it is shown in Basalto et al. (2008) that the single linkage has another strong asset: it is a dissimilarity for all $c_i, c_j \subset S$ such $c_i \cap c_j = \emptyset$. This is the case in clustering: the groups are separated. Whenever two groups share a point, the criterion is zero and property (3) is no more satisfied. A simple counter example from this work (Basalto et al., 2008) shows that the triangular inequality is not fulfilled.

But, real world data often include noise that makes the cluster less separated. The standard hierarchical clustering algorithm using the single linkage with a cut at three clusters yields two isolated points and all the remaining points in the same group as shown in Figure 2.

This result is clearly not satisfactory. Two ways for noise handling are proposed in the following: an improvement of the single linkage criterion in Section 3 and of the hierarchical algorithm in Section 4.

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

3. Taking into account the local density in single linkage

The single linkage is chosen due to its ability to deal with various shapes, spheres, rectangles or even lines or spirals. To overcome its main drawback, the undesirable chaining effect, a noise management process based on the local density distribution in each group is added.

The main idea is to reach the core of the cluster according to the local densities. The distribution allows to identify noisy items. Then instead of considering only the closest items whatever their label, the process consists in iteratively selecting the neighbors until the closest points which are not labeled as noise are found. When a noisy pattern is met during this iterative process it is removed for the next iterations. The criterion is computed as the sum of the distances between the closest points in between, including the noise, weighted by their local densities.

The algorithm of the Single linkage with noise, *sln*, is shown in Algorithm 1.

The algorithm computes the agglomerative criterion between the two groups, G_i and G_j. The input parameters are the groups, the local density for each point that includes a noise label.

The single linkage criterion is computed in line 6 and weighted by the density (lines 17-18). If one of the involved points is labeled as noise it is removed from the group for the next iteration (lines 7-8 and 12-13). The algorithm stops when the two points connected by the single linkage have a density higher than the threshold (lines 10, 15 and 19).

Before going into details with the local density computation and the noise labeling, the main idea of the proposal is illustrated with a toy example.

3.1. Illustrative example

The synthetic data to illustrate the proposal are plotted in Figure 3. There are made up of two symmetrical groups of items, the blue and the red groups. Each of them include regular circle items and two noise ones, shown as a square and a triangle.

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

Algorithm 1 *sln*: the single linkage with noise

Input: G_i, G_j, $dens$
Output: $d(G_i, G_j)$, between group agglomerative criterion
Stop=**false**, $d = 0$, $w = 0$
while Stop==**false** **do**
  Stop1=**false**, Stop2=**false**
  $sl = d(x_i, x_j) = \min_{x_l \in G_i, x_m \in G_j} d(x_l, x_m)$
  if $(noise(x_i)$ ==**true**$)$ **then**
    $G_i = G_i \setminus \{x_i\}$
  else
    Stop1=**true**
  end if
  if $(noise(x_j)$ ==**true**$)$ **then**
    $G_j = G_j \setminus \{x_j\}$
  else
    Stop2=**true**
  end if
  $d = d + sl(dens(x_i) + dens(x_j))$
  $w = w + dens(x_i) + dens(x_j)$
  Stop = Stop1 **and** Stop2
end while
return d/w

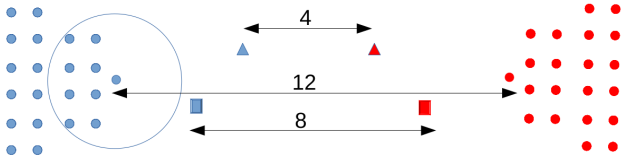

Figure 3: Main idea illustration

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

The single-link distance between the groups of circle-shaped items is 12, the corresponding distances for the square and triangle shaped items are respectively of 8 and 4. The density is estimated by the number of items in a given volume. This number is 9 for the closest circle-shaped items and drops to 1 for the noise points.

According to the proposal, the single linkage with noise is computed as follows:

1. At the first step the single linkage involves the two triangle-shaped items. The algorithm updates: $sl = d = 4(1 + 1)$ and $w = 2$. The triangles are removed for the next iteration as they are labeled as noise.
2. Then the single linkage is between the square-shaped items: $sl = 8$. The current parameters are: $d = 4 \cdot 2 + 8 \cdot 2$ and $w = 2 + 2$. The squares are removed for the next iteration.
3. The single linkage is now between two circle shaped items, with $sl = 12$ and the maximum relative density (1). The values are updated as follows: $d = 4 \cdot 2 + 8 \cdot 2 + 12 \cdot 18$ and $w = 2 + 2 + 18$
4. The algorithms ends end returns: $d/w = 10.9$.

The standard single linkage is then 4. Without the noisy items, the distance would have been 12.

3.2. Local density estimation and noise labeling

Several techniques are available for local density estimation based on kernel or neighborhood, either the number of nearest neighbors or the number of neighbors within a given hyper-volume. They are not studied in the work, the reader may refer to Ros & Guillaume (2016, 2017) for a recent survey.

In this work, the density is computed using an hyper-volume whose radius, the same for all the dimensions, is chosen for each group: its value is the minimum for which the average number of neighbors is higher than a given proportion of the group size, e.g. $p = 2\%$.

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

The distance parameter can be defined at the scale of the whole dataset if the density is homogeneous. A local setting, as proposed above, allows for coping with groups of different densities.

The density for a given point, x, is thus the number of items that fall within the hyper-volume centered on x.

The whole distribution is taken into account to identify noise items. The noise detection is based upon the interquartile range. A data item, x, is labeled as noise when its density $dens(x) < Q_1 - \alpha(Q_3 - Q_1)$.

For displaying the outliers in the boxplot, the value of $\alpha = 1.5$ was proposed by Tukey (1977). The objective in this work is not outlier identification, but to ensure the points that are not labeled as noise are part of the cluster. A typical value of $\alpha = 0.1$ is studied in this paper.

3.3. Behavior

Figure 4 shows that the results are correct with the toy example used in the previous section. The three clusters are well identified and isolated points appear in blue ink. The results are identical for the three values of α studied in this work: $0.05, 0.10, 0.15$.

The behavior of the Single linkage with noise with respect to the standard single linkage is illustrated using the example of three Gaussian distributions. Two of them have fixed parameters while the third one is moving. Two configurations are plotted in Figure 5. The means of the fixed groups are $(1, 1)$ and $(1.5, 1)$ for the black and green clusters. The mean of the red group is (m, m), with m ranging from 1.5 to 5 by step of 0.5. The two configurations illustrated in the figure correspond to $m = 2.5$, first row, and $m = 3.5$ in the second row.

The groups are made up of 600 random points from the Gaussian distribution $\mathcal{N}(mean, \sigma = 0.25)$. Moreover, two of the groups, the ones centered in $(1, 1)$, black, and the moving one, red, also include 120 random points from the $\mathcal{N}(mean, 3\sigma)$ distribution.

Figure 5 illustrates the impact of the parameter α: fewer points are labeled as noise as α increases. The noise points are plotted using a disk in a light

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

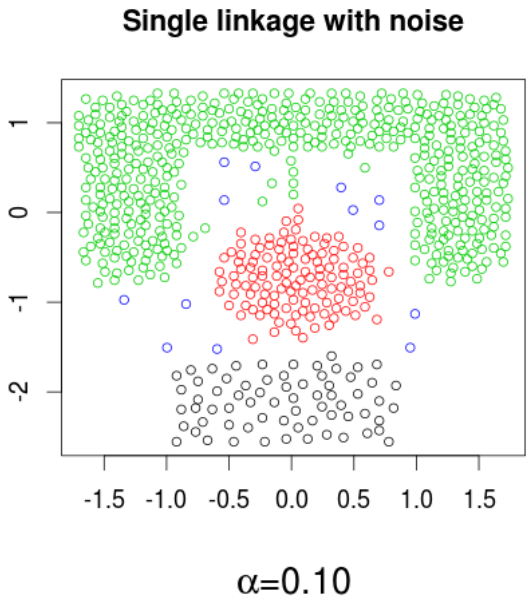

Figure 4: Single linkage with noise and three significant clusters

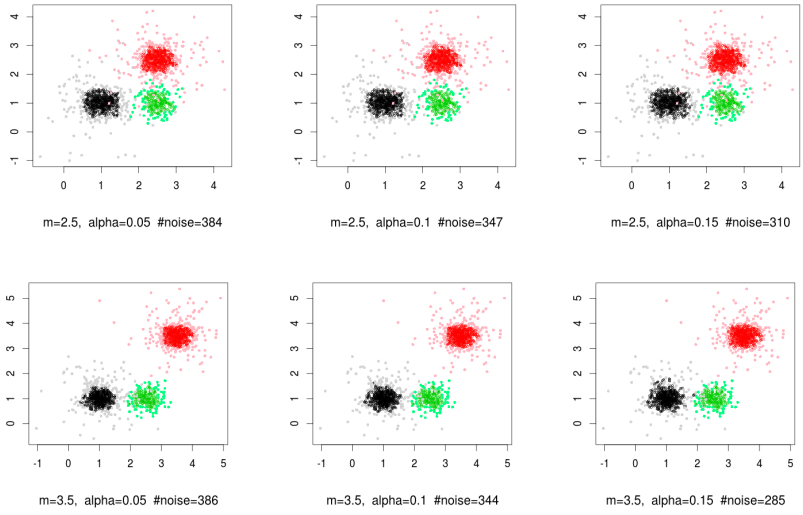

Figure 5: Illustrative behavior of the agglomerative methods with the presence of noise for two values of m. The number of points labeled as noise decreases with increasing values of α. The axis labels are the x and y coordinates.

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

Table 1: Distance between the two fixed groups

m	2.5		3.5	
α	sl	sln	sl	sln
0.05	0.012 (0.008)	0.255 (0.081)	0.014 (0.008)	0.243 (0.056)
0.10	0.012 (0.008)	0.246 (0.076)	0.014 (0.008)	0.238 (0.057)
0.15	0.012 (0.008)	0.222 (0.067)	0.014 (0.008)	0.197 (0.067)

Table 2: Distance between the red and green groups

m	2.5		3.5	
α	sl	sln	sl	sln
0.05	0.014 (0.011)	0.305 (0.074)	0.329 (0.213)	1.625 (0.089)
0.10	0.014 (0.011)	0.283 (0.066)	0.329 (0.213)	1.582 (0.093)
0.15	0.014 (0.011)	0.237 (0.062)	0.323 (0.213)	1.467 (0.098)

variation of the cluster color.

The experiment consists of 100 random trials for each value of m.

Tables 1 and 2 show the average value of the criterion between the two groups according to three values of α for the two configurations plotted in Figure 5. The corresponding standard deviation is given in parenthesis.

In Table 1 the two fixed groups are considered and thus the result does not depend on m. The observed variations between the two values of m for a given α are due to the random generation of the data. Two comments have to be made. First the distance computed using the single linkage (*sl* column) is always lower than the one computed using the single linkage with noise method (*sln*). Second, the *sln* decreases with increasing values of α. This is also expected as the higher α the fewer the number of points that are are labeled as noise.

In Table 2, the moving group is involved and the differences in the results increase with m. The standard deviation is significantly lower using *sln* than *sl*. This is also illustrated in Figure 6.

The coefficient of variation is significantly lower using *sln*, highlighting the

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

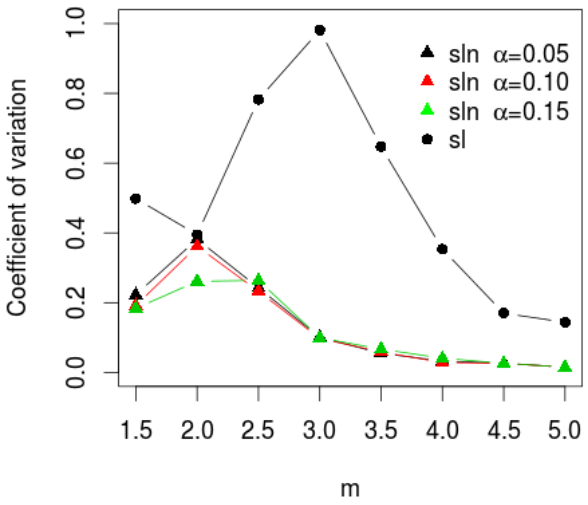

Figure 6: Comparison of the coefficient of variation for *sl* and *sln* using the eight values of m.

robustness of the calculations with respect to the random noise.

4. The hierarchical clustering algorithm to deal with noise

The previous section showed that noise can be managed at the criterion level. In this one it is shown that it can also be handled by the clustering algorithm algorithm. The number of desired clusters is known and what is at stake for the agglomerative algorithm is to yield the number of representative clusters, meaning clusters with a minimum size to avoid the phenomenon observed in Section 2, and illustrated in Figure 2.

4.1. Description of the algorithm

The standard agglomerative algorithm does not account for noise. To make a fair comparison between the agglomerative criteria, only representative clusters are taken into account. The proposed implementation is shown in Algorithm 2.

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

Algorithm 2 The hierarchical clustering algorithm

Input: data (size n), nclust, criterion, prop
Output: data with a cluster or a noise label
$MinSize = max\big(2, min(0.02n, n/10nclust)\big)$
repeat
  MergeClosestCluster(criterion)
  UpdateLocalDensity(new cluster)
  PropSize=0, nCrep=0 {number of clusters higher than $MinSize$}
  for (i in clusters) **do**
    if ($|c_i| > MinSize$) **then**
      $nCrep++$, $PropSize+ = |c_i|$
    end if
  end for
  if ($PropSize < prop \cdot n$) **then**
    $nCrep = 0$ {Clusters are not representative enough}
  end if
until $\big((nCrep > 0)$ AND $(nCrep \leq nclust)\big)$
Tag the $nCrep$ representative clusters
BuildFinalPartition(nclust, criterion, clusters, labels, tag)
return cluster labels

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

Two thresholds are used in the algorithm. The first one is the number of items for a cluster to be representative: it can be set at 2% of the data but it can also be defined according to the number of clusters, *nclust*. The minimum of these values is chosen. In case of a large number of clusters, the minimum size is set at 2. Hence the proposed formula in line 3: $MinSize = max\big(2, min(0.02n, n/10nclust)\big)$.

The second one makes sure the proportion of the data in representative clusters is high enough. This important parameter, $0 \leq prop \leq 1$, is studied in the following.

To speed up the first steps of the algorithm the *sln* criterion is only used when at least one of the clusters is higher than $MinSize$. Otherwise the standard single linkage criterion is used.

To account for density variation, the local density estimation is performed using a radius, the same for all the dimensions, specific to a group. An update for the new cluster is required after each merging (line 6).

Once the representative clusters are identified, they are tagged (line 16) before calling the *BuildFinalPartition* function. The main characteristic of this function is to avoid the merging of two clusters tagged as representative, line 6 of Algorithm 3 while the agglomerative clustering is carried out until the end, i.e. all the remaining items are labeled.

When calling the *BuildFinalPartition*, to cases may occur. First, the number of representative clusters is exactly equal to the final number of clusters, i.e. $nCrep = nclust$ in Algorithm 2. In this case, the remaining items are assigned one of the $nCrep$ labels. The second case occurs when $nCrep < nclust$. In this situation, additional clusters may appear in the *BuildFinalPartition* function, when several sub clusters not enough representative by themselves are merged together, to reach the desired number of clusters. Otherwise, the final number of cluster remains lower than *nclust*.

This algorithm prevents the hierarchical process from generating clusters that only include isolated points. It is driven by the proportion parameter.

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

Algorithm 3 The Build Final Partition function

Input: nclust, criterion, clusters, labels, tag
Output: labels
while $(nClusters > nclust)$ **do**
  $min = \infty$
  for $(c_i, c_j$ in clusters$)$ **do**
    if $\big(not(tag(c_i)\ AND\ tag(c_j))\big)$ **then**
      if $(criterion(c_i, c_j) < min)$ **then**
        $c_{w1} = c_i,\ c_{w2} = c_j$
      end if
    end if
  end for
  $Merge(c_{w1}, c_{w2})$
  UpdateLocalDensity$(c_{w1} \cup c_{w2})$
  $nClusters - -$
end while

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

4.2. Behavior according to the proportion parameter

The difference with the basic version of the hierarchical clustering algorithm can be illustrated using the behavior according to the proportion parameter. A basic version is equivalent to a value of $prop = 1$.

First the very challenging $S4$ data[1] shown in Figure 7 are used. The 1800 data points are organized in 15 clusters with a high level of overlap. The algorithm is run with $nclust = 15$ which yields $MinSize = 12$. A low value of $prop$ allows for the density peaks to be identified but when this value increases the number of representative clusters decreases. When $prop = 0.5$ then $nCrep = 13$ and only one can be identified with $prop = 0.7$ as shown in the right plot of Figure 8. In the last step of the algorithm the remaining points would be assigned to the unique representative group.

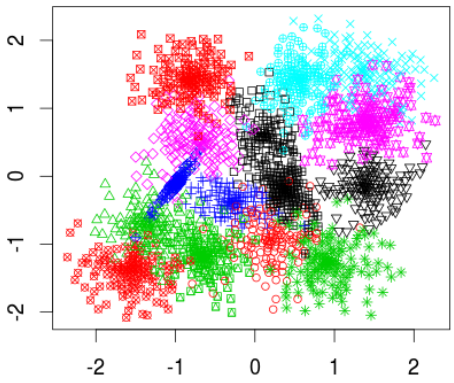

Figure 7: The challenging $S4$ data. The axis labels are the x and y coordinates.

[1] https://cs.joensuu.fi/sipu/datasets/

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

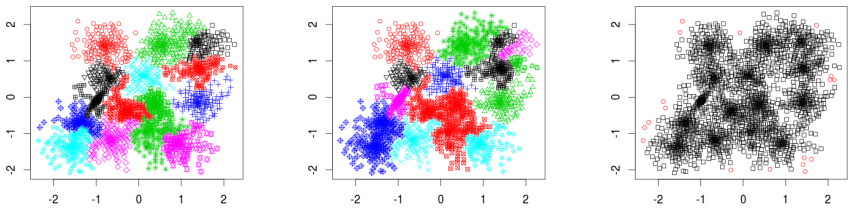

Figure 8: The $S4$ data processed with $prop = \{0.4, 0.5, 0.7\}$ from left to right. The axis labels are the x and y coordinates.

The data shown in the left plot of Figure 9 illustrate the case of clusters with varying density. The black group is made up of two peaks that are denser than the one of the red group. The algorithm is run with $nclust = 2$. When the parameter is set at a small value, the first two representative groups belong to the same cluster. This is illustrated for $prop = 0.3$ in the central plot of the figure. Then the remaining points are processed and the final result is plotted in the right part of the figure.

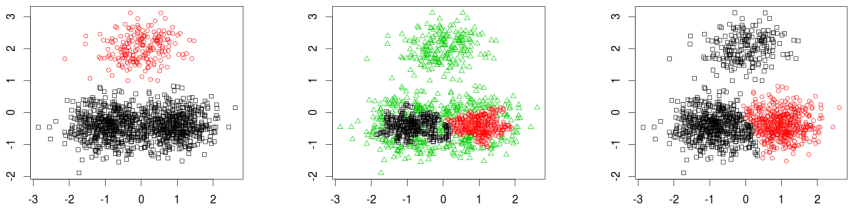

Figure 9: The data, the two representative clusters and the final partition from left to right. The axis labels are the x and y coordinates.

When some more structured noise is added, the usual single linkage criterion fails to build the two expected clusters as shown in the left part of Figure 10 while the improvement proposed in this work is successful (right plot). These results hold for $0.6 \leq prop < 1$.

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

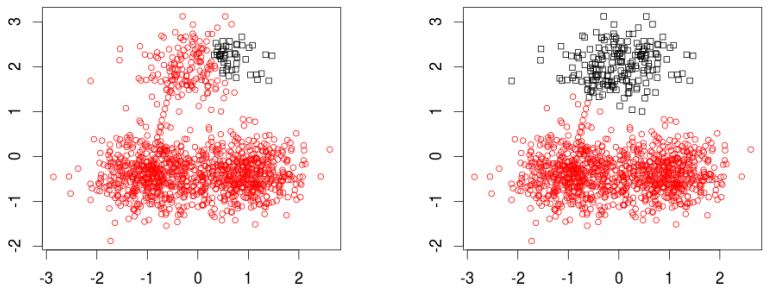

Figure 10: The data processed using the single linkage (left) and the proposed improvement (right). These results hold for $0.6 \leq prop < 1$. The axis labels are the x and y coordinates.

This example shows that even if the two proposals to deal with noise, the criterion and the agglomerative clustering algorithm, may be redundant to some extent, there exist situations when the two of them are needed to reach the expected result.

5. Numerical experiments

First nine criteria are compared using twelve synthetic datasets. Then the three of them that yield the best results are tested against more difficult data, including noise and overlap.

5.1. Comparison of criteria using twelve illustrative datasets

The criteria used for this comparison are:

- the ones already used in a previous section: Single link (C1), Complete link (C2), Average link (C3), Centroid (C4), Ward (C5);
- three others ones that are introduced in the following: Hausdorff (C6), Eq. (1), Full with neighborhood (C7), Eq. (3), Single with neighborhood (C8) Eq. (4);

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

- the proposed single link with noise (C9).

The Hausdorff criterion, based on the Hausdorff distance, was proposed in Basalto et al. (2008). It is defined as:

$$d_h(c_i, c_j) = max\{\sup_i \inf_j d(i,j), \sup_j \inf_i d(i,j)\} \tag{1}$$

A method that bridges the gap between two kinds of agglomerative criteria, the ones based on a distance in the input space and those based on the connectivity in the neighborhood graph, was proposed in Lee & Olafsson (2011). The connectivity is penalized by the distance as follows:

$$c^d(c_i, c_j) = \frac{\sum_{i \in c_i} \sum_{j \in c_j} \frac{b_{ij} + b_{ji}}{d(i,j)}}{|c_i| \ |c_j|}, \ b_{ij} = \begin{cases} 1 & if \quad j \in N^k(i) \\ 0 & otherwise \end{cases} \tag{2}$$

where $N^k(i)$ is the set of the k nearest neighbors of i, defined as:
$N^k(i) = \{x_{(1)}, x_{(2)}, \ldots, x_k)\}$, with $||x_{(1)} - i|| \leq ||x_{(2)} - i|| \leq \ldots ||x_{(n-1)} - i||$.

Two new criteria are inspired from this method: a global one that averages the local configurations, d_n^a in Eq. (3), and a local, one that corresponds to the minimum, d_n^s in Eq. (4). The latter is another single linkage. In these formulas the distance is penalized by the connectivity, hence the difference with the connectivity in Eq. (2).

$$d_n^a(c_i, c_j) = \frac{\sum_{i \in c_i} \sum_{j \in c_j} \frac{d(i,j)}{b_{ij} + b_{ji} + 1}}{|c_i| \ |c_j|}, \ b_{ij} = \begin{cases} 1 & if \quad j \in N^k(i) \\ 0 & otherwise \end{cases} \tag{3}$$

$$d_n^s(c_i, c_j) = \min_{i,j} \frac{d(i,j)}{b_{ij} + b_{ji} + 1}, \ b_{ij} = \begin{cases} 1 & if \quad j \in N^k(i) \\ 0 & otherwise \end{cases} \tag{4}$$

The neighborhood involved in these criteria is defined by the number of neighbors, k. A small value would be in the spirit of the local criterion but the result would be sensitive to noise or outliers. Large values would yield smoother results but are likely to hide local differences. According to the literature (Duda

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

et al., 2012), the number of neighbors is defined as: $k = \max(3, \lambda\sqrt{n})$, $\lambda \in [0.1; 1]$. In this expression, n can be the whole data size (global approach) or the group one (local approach). The local approach is preferred to deal with cluster size or density variations. In the experiment the value $\lambda = 0.2$ is used.

The twelve datasets

Twelve 2−dimensional datasets, representative of the diversity of situations a clustering algorithm has to cope with, were selected. They are plotted in Figure 11.

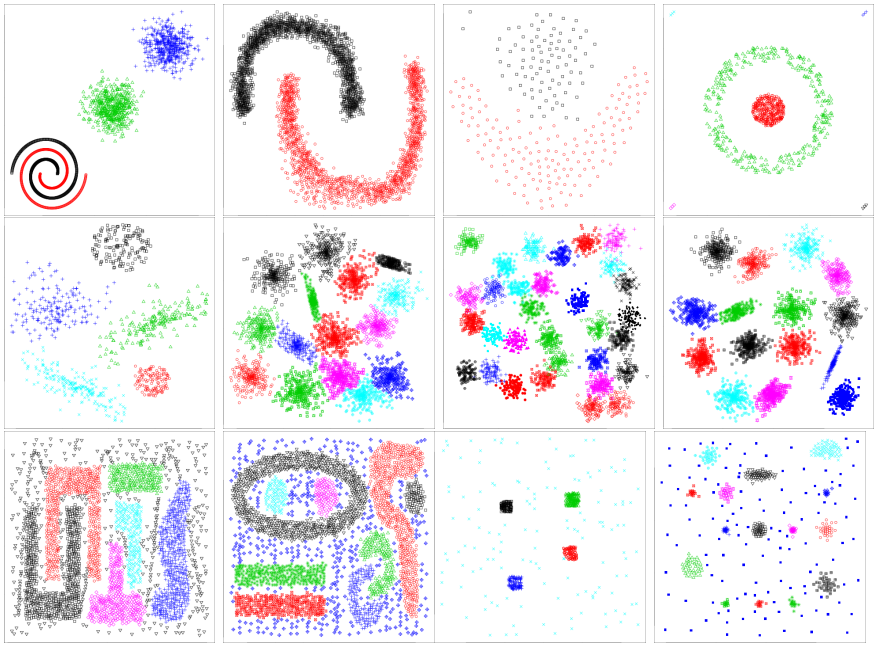

Figure 11: The twelve datasets. The axis labels are the x and y coordinates.

Some are from the data clustering repository of the computing school of Eastern Finland University[2], while others come from a benchmark data for clustering repository [3] or were proposed in the published literature. These datasets,

[2]https://cs.joensuu.fi/sipu/datasets/
[3]https://github.com/deric/clustering-benchmark/blob/master/src/main/resources/datasets/

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

detailed in Table 3, are usually considered for testing new clustering algorithms but they do not represent the diversity of cases a clustering algorithm has to tackle. One homemade dataset, with well separated clusters but of different sizes, has been added to complete this diversity.

Table 3: The twelve datasets

	Size	$\#clust$	Name	Origin
D1	2000	4	2Sp2glob	Piantoni et al. (2015)
D2	4811	2	BANANA	Footnote 3
D3	240	2	FLAME	Fu & Medico (2007)
D4	770	2	TARGET	Footnote 3
D5	850	4	DS850	Footnote 3
D6	5000	15	S3	Footnote 2
D7	3100	31	D31	Veenman et al. (2002)
D8	5000	15	S2	Footnote 2
D9	8000	6	Chameleon	Karypis et al. (1999)
D10	10000	9	cluto-t7.10k	Footnote 3
D11	622	4	Zelnik4	Footnote 3
D12	588	16	Home	Homemade

The selected data include some variability in cluster shape, size, density, amount of noise and degree of separation. The datasets plotted in the first row of Figure 11, from $D1$ to $D4$, show a shape diversity. In the second row, from $D5$ to $D8$, the shapes are quite simple, with different elongation and some overlap between groups. In the last row, the datasets include some noise with a diversity of shapes, $D9$ and $D10$, or well separated clusters, $D11$ and $D12$.

Results

In a first step, the sampling algorithm *ProTraS* (Ros & Guillaume, 2018) was run in order to limit the dataset size and to speed up the agglomerative

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

algorithm. The approximation level was set at 0.015 to ensure a precise representation of the original data.

Then the hierarchical algorithm was run with the desired number of clusters for each dataset and the final partition was compared to the ground truth using three indices: the Mutual Information index (Cover & Thomas, 2012), the F-measure (Makhoul et al., 1999) and the Rand index (Rand, 1971). When noise is identified using a specific label in the ground truth, noisy points are not taken into account for index computation.

The results are given in Table 4. The first rows display the Mutual Information index values for the 12 datasets. Then for each of the indices and criteria, the distributions are summarized by the mean and minimum values.

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

Table 4: Mutual Information Index for the 12 datasets and the 9 criteria and summary for the three indices.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
D1	1.000	0.809	0.754	0.771	0.761	0.792	0.754	1.000	1.000
D2	1.000	0.293	0.574	0.556	0.440	0.219	0.434	1.000	1.000
D3	1.000	0.390	0.404	0.112	0.374	0.171	0.396	0.936	1.000
D4	0.939	0.283	0.220	0.155	0.204	0.264	0.213	0.939	0.939
D5	1.000	0.806	0.952	0.972	0.826	0.828	0.964	0.813	1.000
D6	0.932	0.846	0.936	0.918	0.889	0.870	0.921	0.920	0.937
D7	0.935	0.925	0.936	0.942	0.929	0.919	0.941	0.921	0.938
D8	0.986	0.933	0.981	0.982	0.979	0.919	0.983	0.974	0.986
D9	1.000	0.682	0.666	0.658	0.622	0.594	0.695	0.877	1.000
D10	0.948	0.621	0.607	0.624	0.607	0.618	0.631	0.878	0.993
D11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
D12	1.000	0.996	1.000	1.000	0.984	1.000	1.000	1.000	1.000
	Mutual Index								
Mean	**0.978**	0.715	0.752	0.724	0.718	0.683	0.744	**0.938**	**0.983**
Min	0.932	0.283	0.220	0.112	0.204	0.171	0.213	0.813	0.937
	F-measure								
Mean	**0.979**	0.813	0.842	0.838	0.815	0.797	0.831	**0.933**	**0.985**
Min	0.921	0.569	0.523	0.496	0.524	0.591	0.534	0.815	0.926
	Rand Index								
Mean	**0.994**	0.858	0.877	0.860	0.855	0.841	0.867	**0.968**	**0.996**
Min	0.973	0.590	0.551	0.499	0.543	0.520	0.548	0.873	0.983

The maximum value over the 12 datasets is 1 for all the criteria and the three indices. It is not shown in the summaries. The mean is higher than 0.9 for the three single linkage criteria according to the three indices. But the minimum is above the same threshold only for the usual single linkage (C1) and the proposed single linkage with noise (C9). This experiment confirms that single linkage based methods are more suitable than global ones to deal with a

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

diversity of shapes and densities.

5.2. Experiments with noisy data

The three criteria that stand out from the previous experiment are now compared in more critical situations.

The *genRandomClust* R package[4] is used for partition generation. This is an implementation of the method proposed in Qiu & Joe (2006a). The degree of separation between any cluster and its nearest neighboring cluster can be set at a specified value regarding the separation index proposed in Qiu & Joe (2006b). The cluster covariance matrices can be arbitrary positive definite matrices. The *eigen* method is used in the experiment. It first randomly generates eigenvalues $(\lambda_1, \ldots, \lambda_p)$ for the covariance matrix then uses columns of a randomly generated orthogonal matrix, $Q = (\alpha_1, ..., \alpha_p)$, as eigenvectors. The covariance matrix is then built as $Q \cdot diag(\lambda_1, ..., \lambda_p) \cdot Q^T$.

The package uses the basic parameters for cluster generation such as the number of clusters, the space dimension and their respective sizes but also allows for variability management. A ratio between the upper and the lower bound of the eigenvalues can be specified. The default value is 10, but 30 was used in all the experiments to produce more variation in the elongated shapes. The range of variances in the covariance matrix was set at $rangeVar = [1, 30]$. This value is chosen greater than the default one, $[1, 10]$, in order to yield a higher variation in the cluster densities. The only parameter used in this experiment is the value of the separation index between two neighboring clusters, $SepVal$. It ranges from -1 to 1. The closer to 1 the value, the more separated the clusters.

[4] https://www.r-project.org/

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

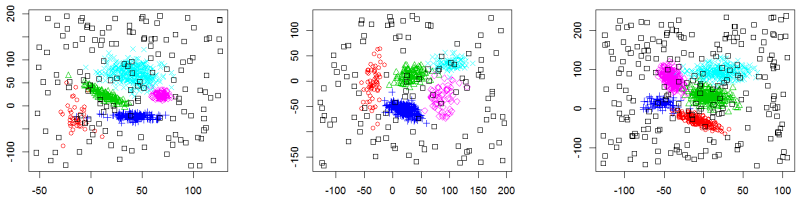

Figure 12: Three examples in dimension 2 with $SepVal = 0.1$ and addition of 20% of random noise points. The axis labels are the x and y coordinates.

Figure 12 shows examples of 2-dimensional data organized in 5 clusters of size randomly chosen between 100 and 500. A random noise has been added, the number of points is 20% of the whole size. The clusters are similar in density and shape and include some internal variance: ellipsoids that are more or less elongated.

Table 5 shows the statistics of 30 runs for 5-dimensional data, $SepVal = \{0.3, 0.2\}$ and $prop = 0.8$.

Table 5: Summary of 30 runs with random noise for 5-dimensional data, for two values of $SepVal$ and $prop = 0.8$.

	C1	C8	C9	C1	C8	C9
SepVal	0.3			0.2		
Rand Index						
Mean	0.839	0.839	0.999	0.368	0.450	0.870
σ	0.26	0.26	0.001	0.21	0.25	0.12
Mutual Information Index						
Mean	0.796	0.796	0.995	0.212	0.331	0.839
σ	0.32	0.32	0.003	0.28	0.33	0.16
F-measure						
Mean	0.866	0.866	0.999	0.432	0.490	0.832
σ	0.21	0.21	0.001	0.12	0.16	0.16

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

The scores for $C1$ and $C8$ are rather high for $SepVal = 0.3$ (between 0.8 and 0.9) but they drop to less than 0.5 for $SepVal = 0.2$. If the performance of $C9$ decreases between $SepVal = 0.3$ and $SepVal = 0.2$, $C9$ remains highly more robust to smaller degrees of separation.

The results are similar with $prop = 0.6$ and $prop = 0.7$, i.e., $C9$ performs better than its two competitors. The difference between the three criteria is however smaller. As an example, the means for the Mutual Information index for $SepVal = 0.2$ and $prop = 0.7$ for the three criteria are: 0.904, 0.894 and 0.985. The same values for $prop = 0.6$ become: 0.952, 0.931 and 0.975.

These results illustrate that the proposed criterion, $C9$, yields better results than its two competitors under the different configurations.

The difference can be highlighted by adding a more structured noise to the previous configuration. One hundred pairs of points are randomly selected and the pair, $P1$ and $P2$, for which the product of the distance and the local densities, $p = d(P1, P2) \cdot dens(P1) \cdot dens(P2)$, is maximum is kept. The number of points along the line defined by $P1$ and $P2$ is set at 0.005% of the whole size. They are equally spaced and then a random noise in the range of $\pm 10\%$ of the coordinate in each dimension is generated for each data point. In this experiment, the range of variances in the covariance matrix was set at the default value, $rangeVar = [1, 10]$.

Figure 13 illustrates the result in dimension 2. In the plot the random component around the structured noise points was not added to simplify the representation.

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

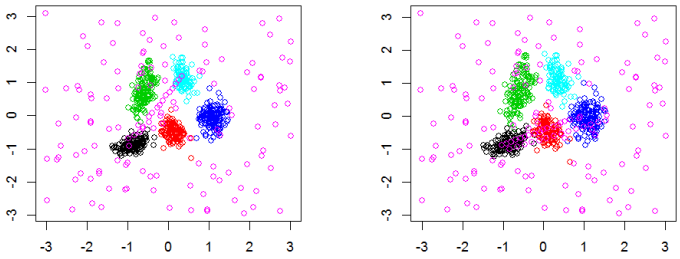

Figure 13: Two examples that illustrate the structured noise.

Tests were carried out in various dimensions and with several degrees of separation. The smallest separation degree that yielded a correct result for at least one criterion is reported for the dimensions considered. Tables 6, 7 and 8 summarize the distributions of the three indices over 30 random sets for each configuration.

The Silhouette index (Rousseeuw, 1987) is used to characterize the partitions. It ranges in $[-1, +1]$: the higher the value the more separated the clusters. Due the large amount of noise, the index values are quite low, even for $2-$dimensional data.

As expected the minimum separation degree to get at least one correct result increases with the dimension space. When the criterion is the single linkage ($C1$) or the single linkage weighted by the local neighborhood ($C8$), the results in dimension 2 highly depend on the parameter *prop* of the clustering algorithm. When the noise accounts for a significant part of the data, the tuning of this parameter may result difficult. This is not the case for the single linkage with noise ($C9$). Using this criterion a wide range of values yield similar results.

When the space dimension increases, $dim = 5$, $C1$ and $C8$ yield poor results whatever the proportion parameter value, while $C9$ is still efficient. In higher dimensions, differences tend to get smaller as the clusters are less separated even when generated with the same set of parameters.

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

Table 6: Summary of the 30 runs per configuration with structured noise for $C1$

Dim	Sep	prop	Rand		Mutual		F-measure		Silhouette	
			μ	σ	μ	σ	μ	σ	μ	σ
	0.1	0.6	0.884	0.13	0.801	0.16	0.871	0.11		
2	0.1	0.7	0.766	0.30	0.651	0.39	0.800	0.24	0.427	0.02
	0.1	0.8	0.241	0.02	$< 10^{-3}$	0.00	0.381	0.02		
	0.2	0.6	0.509	0.37	0.381	0.46	0.384	0.46		
5	0.2	0.7	0.430	0.33	0.290	0.41	0.531	0.27	0.308	0.06
	0.2	0.8	0.410	0.32	0.270	0.39	0.521	0.26		
	0.3	0.6	0.982	0.04	0.971	0.045	0.970	0.06		
8	0.3	0.7	0.942	0.1	0.92	0.12	0.923	0.13	0.252	0.02
	0.3	0.8	0.685	0.2	0.630	0.25	0.646	0.14		

This phenomenon is referred to as the *scourge of dimension* and *the surprising behavior of distance metrics in high dimensional space* was studied by Aggarwal et al. (2001). The main effect is the *concentration of the measure* in some specific space areas that prevents the fine discrimination that was possible in lower dimension spaces. $C8$ which is less efficient than $C9$ in low dimension spaces, become more resistant to the increase in the dimension. This can be explained by the way the local density is estimated: $C9$ uses a radius and may be more sensitive to the concentration of the measure than $C8$ that uses a given number of neighbors. Anyway, dealing with high dimensional data requires specific procedures.

6. Conclusion

In this work two approaches are introduced for agglomerative processes to deal with noise. The first one is a criterion that improves the single linkage by taking into account the local densities. The criterion is not only the shortest distance, which is sensitive to noise, but the average distance between the nearest neighbors that are not labeled as noise and all the intermediate pairs

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

Table 7: Summary of the 30 runs per configuration with structured noise for $C8$

Dim	Sep	prop	Rand		Mutual		F-measure		Silhouette	
			μ	σ	μ	σ	μ	σ	μ	σ
	0.1	0.6	0.919	0.04	0.822	0.006	0.881	0.06		
2	0.1	0.7	0.795	0.22	0.679	0.28	0.802	0.19	0.427	0.02
	0.1	0.8	0.428	0.25	0.220	0.30	0.431	0.16		
	0.2	0.6	0.589	0.39	0.481	0.48	0.481	0.48		
5	0.2	0.7	0.581	0.38	0.451	0.48	0.640	0.31	0.308	0.06
	0.2	0.8	0.576	0.32	0.472	0.39	0.621	0.29		
	0.3	0.6	0.991	0.009	0.990	0.004	0.990	0.001		
8	0.3	0.7	0.975	0.04	0.972	0.04	0.971	0.06	0.252	0.02
	0.3	0.8	0.846	0.21	0.794	0.23	0.809	0.2		

of items. Noise labeling is done according to the density distribution based on the interquartile range, when the local density is lower than $Q_1 - \alpha(Q_3 - Q_1)$, a typical value of $\alpha = 0.10$ is used in this work. The second is a hierarchical algorithm that yields the desired number of representative, large enough, clusters. The goal is to avoid empty or very small clusters. The algorithm is driven by a proportion parameter. To identify representative clusters, the number of items in all these clusters must be higher than a proportion of the whole data size. Then, the representative clusters cannot be merged and the remaining points are either assigned to one of these clusters or may constitute new groups when the number of representative clusters is lower then the desired number of clusters. The sensitivity to the proportion parameter was studied and showed that the range $0.6 - 0.7$ was identified as suitable for a large variation in the amount of noise.

Experiments were carried out to validate the proposals. First, nine criteria were compared using twelve synthetic datasets, illustrative of the main challenges a clustering algorithm has to tackle: variability in cluster shape, density, size, amount of noise and degree of separation. It clearly showed that local

Author-produced version of the article published in Expert Systems with Applications, 2019, N°128, p. 96-108.
The original publication is available at https://www.sciencedirect.com
Doi: 10.1016/j.eswa.2019.03.031

Table 8: Summary of the 30 runs per configuration with structured noise for $C9$

Dim	Sep	prop	Rand		Mutual		F-measure		Silhouette	
			μ	σ	μ	σ	μ	σ	μ	σ
	0.1	0.6	0.962	0.02	0.893	0.04	0.934	0.05		
2	0.1	0.7	0.975	0.01	0.921	0.02	0.966	0.02	0.427	0.02
	0.1	0.8	0.901	0.13	0.840	0.16	0.918	0.1		
	0.2	0.6	0.981	0.01	0.981	0.01	0.981	0.01		
5	0.2	0.7	0.990	$< 10^{-3}$	0.990	$< 10^{-3}$	0.990	0.001	0.308	0.06
	0.2	0.8	0.938	0.14	0.916	0.17	0.922	0.17		
	0.3	0.6	0.981	0.04	0.975	0.04	0.972	0.03		
8	0.3	0.7	0.961	0.05	0.949	0.07	0.937	0.09	0.252	0.02
	0.3	0.8	0.802	0.22	0.752	0.22	0.774	0.18		

criteria yield better results than global ones. The three local criteria, the well known single linkage, the proposed single linkage with noise and another one where the single linkage is weighted by the mutual neighborhood, outperformed the six other ones according to three partition indices. Then the three criteria were compared with partitions generated using the *genRandomClust* R package with an additional random noise, 20% of the whole size. Different space dimensions and separation degrees were considered. The single linkage with noise stood out. Finally, a more structured noise was added to the previous configuration to highlight the difference between the three criteria. Even if the two competitors of the proposal proved to be efficient with $2-$dimensional data, the single linkage with noise yielded comparable results for a large range of the proportion parameter. When the space dimension increased, the proposal was the only one to yield the expected results. This stresses the complementariness between the two approaches to deal with noise: the criterion and the hierarchical algorithm.

Noise is likely to affect other useful metrics such as cluster validation indices (Liu et al., 2010). Taking into account the local density may help to design a

more robust index, in the same way that it improves the single linkage criterion.

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory* (pp. 420–434). Springer.

Basalto, N., Bellotti, R., De Carlo, F., Facchi, P., Pantaleo, E., & Pascazio, S. (2008). Hausdorff clustering. *Phys. Rev. E*, *78*, 046112. URL: https://link.aps.org/doi/10.1103/PhysRevE.78.046112. doi:10.1103/PhysRevE.78.046112.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Fu, L., & Medico, E. (2007). Flame, a novel fuzzy clustering method for the analysis of dna microarray data. *BMC bioinformatics*, *8*, 3.

Karypis, G., Han, E.-H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, *32*, 68–75.

Lee, J.-S., & Olafsson, S. (2011). Data clustering by minimizing disconnectivity. *Information Sciences*, *181*, 732 – 746. URL: http://www.sciencedirect.com/science/article/pii/S0020025510005335. doi:https://doi.org/10.1016/j.ins.2010.10.028.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (pp. 911–916). IEEE.

Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R. et al. (1999). Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop* (pp. 249–252). Herndon, VA.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*, 86–97. URL: http://dx.doi.org/10.1002/widm.53. doi:10.1002/widm.53.

Piantoni, J., Faceli, K., Sakata, T. C., Pereira, J. C., & de Souto, M. C. (2015). Impact of base partitions on multi-objective and traditional ensemble clustering algorithms. In *International Conference on Neural Information Processing* (pp. 696–704). Springer.

Qiu, W., & Joe, H. (2006a). Generation of random clusters with specified degree of separation. *Journal of Classification*, *23*, 315–334.

Qiu, W., & Joe, H. (2006b). Separation index and partial membership for clustering. *Computational statistics & data analysis*, *50*, 585–603.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: http://www.R-project.org/.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, *66*, 846–850.

Ros, F., & Guillaume, S. (2016). Dendis: A new density-based sampling for clustering algorithm. *Expert Systems with Applications*, *56*, 349–359. doi:10.1016/j.eswa.2016.03.008.

Ros, F., & Guillaume, S. (2017). Dides: a fast and effective sampling for clustering algorithm. *Knowledge and Information Systems*, *50*, 543–56. doi:10.1007/s10115-016-0946-8.

Ros, F., & Guillaume, S. (2018). Protras: A probabilistic traversing sampling algorithm. *Expert Systems with Applications*, *105*, 65–76. doi:10.1016/j.eswa.2018.03.052.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Behavioral Science: Quantitative Methods. Reading, Mass.: Addison-Wesley.

Veenman, C. J., Reinders, M. J. T., & Backer, E. (2002). A maximum variance cluster algorithm. *IEEE Transactions on pattern analysis and machine intelligence*, *24*, 1273–1280.