ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

Jerzy Korzeniewski[*]

A PROPOSAL OF MODIFICATION OF AGGLOMERATIVE CLUSTERING ALGORITHMS

Abstract. In the paper, a modification of agglomerative clustering algorithms is proposed which can be applied to any kind of agglomeraitve algorithm. The idea of the modification is to stress the local density of observations' distribution, while performing clustering based on the dissimilarity matrix. The following clustering algorithms are examined: single link, complete link, group average link and centroid link. The quality of clustering is assessed by means of the silhouette indices on subsets generated with the Milligan's Clustgen software. The results prove that the Author's modifications almost always improve the standard methods.

Key words: cluster analysis, agglomerative algorithms, silhouette indices.

I. MODIFICATION PROPOSAL

As it is well known hierachical agglomerative clustering is characterized by the following features:

- we start from *n* one-element classes (i.e as many as the number of observations);
- at every agglomerative step the number of classes is reduced by one by pooling together two classes;
- after *n*-1 steps we obtain one class containing all observations.

Linking or pooloing classes together is done through the following algorithm.

- In the distance between classes matrix (dissimilarity matrix) we look for two most similar classes in the sense of an established criterion (e.g. two closest classes). Let us say that such classes will be the classes denoted by *i, j*.
- We reduce the number of classes by one by pooling together classes *i, j*.
- We transform the distance between classes matrix so that all pairs of distances would be defined. again(we define the distance between the new class and all other classes).

[*] Ph.D., Department of Statistical Methods, University of Łódź.

- The above three steps are repeated until all observations belong to one class.

The basic drawback of algorithms of this kind is the “chain disease”. It consists in the tendency to link the closest classes and, as a result, one class may contain very different observations but the ones linked with a chain of observations, out of which every two consecutive observations are very similar. We may try to eliminate this drawback by putting more stress at every step of algorithm on linking classes from regions in which the density of observations distributions is higher. Let us investigate the idea in the following illustration.

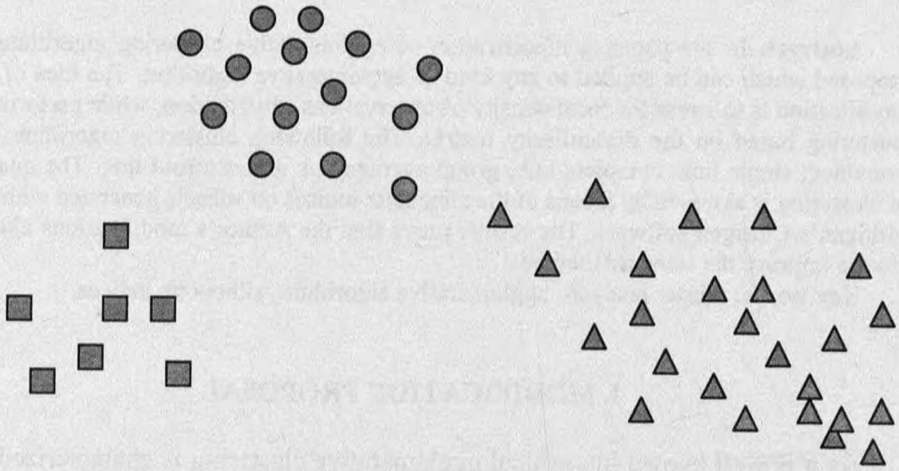

Fig. 1. Three clusters of observations of two dimensional Euclidean space – each denoted by observations of different shape.

Let us assume that at some step of an agglomerative procedure we are to link two out of three clusters presented in Figure 1. If we took a centroid link algorithm i.e. the one linking two clusters with the closest centres, we would have to link circles with squares. If we modify this algorithm by putting more stress on higher density of observations we would rather benefit from linking circles with triangles because ”between” these two centroids the density of observations is higher than in the case of circles-squares or squares-triangles. The local density of observations distributions may be taken into account in various ways – the natural one seems to be the way in which we relate the number of observations in a specified subset of the Euclidean space to the volume of this subset. In terms of the example presented in Figure 1 it would look as follows : we link the two clusters for which the smallest is the distance of their centroids divided by the number of observations which are closer to each of the two centroids than the distance between the centroids. Furthermore, we should relate i.e. divide the number of such points by the volume of the pertinent subset of the two dimensional plane – in this case by an expression proportional to the squared distance of the centroids. The choice of the pertinent subset is sometimes a matter of relatively arbitrary choice because in some agglomerative algorithms (e.g. mean cluster distance) there are no natural points of reference as in the case of the centroid distance. In such cases we should propose some points of reference. Precise definitions of the modifications of four agglomerative algorithms are given below.

Complete link method

As it is known the idea of this method is link at each step of the algorithm the two clusters for which the distance of two most distant points is the smallest. We modify this algorithm in the following way.

- We find the distance r between the two most distant observations for every pair of clusters.
- We find the number x of observations which are closer than r to both most distant observations of both clusters.
- We link the two clusters for which the value of the expression

$$\frac{r^{d+1}}{x} \tag{1}$$

Is the smallest (d – dimension of the Euclidean set space).

Single link method

As it is known the idea of this method is link at each step of the algorithm the two clusters which have the smallest distance of two closest points. We modify this algorithm in the following way.

- We find the distance r between the two most distant observations for every pair of clusters.
- We find the number x of observations which are closer than r to both most distant observations of both clusters.
- We link the two clusters for which the value of the expression

$$\frac{s \cdot r^{d}}{x} \tag{2}$$

Is the smallest (s – distance between two closest observations).

In this modification there is no counting of the observations between two colsest observations because such a modification would not change much as the number of such observations is very small, usually equal to 0. Instead, we propose that x is the number of observations lying "between" two most diatant observations.

Centroid link method

As it is known the idea of this method is link at each step of the algorithm the two clusters for which the distance of two centroids is the smallest. We modify this algorithm in the following way.

- We find the distance r between two centroids for every pair of clusters.
- We find the number x of observations which are closer than r to both centroids.
- We link the two clusters for which the value of the expression

$$\frac{r^{d+1}}{x} \tag{3}$$

is the smallest.

Group average link method

As it is known the idea of this method is to link at each step of the algorithm the two clusters for which the arithmetic mean of all distances is the smallest. We modify this algorithm in the following way.

- We find the distance r between two most distant observations for every pair of clusters.
- We find the number x of observations which are closer than r to both most distant observations of both clusters.
- We link the two clusters for which the value of the expression

$$\frac{s \cdot r^{d}}{x} \tag{4}$$

Is the smallest (s – arithmetic mean of all distances between all pairs of observations).

In this modification, as in the case of the single link method, we propose the two most distant points as the reference points. Other ways are also possible but this one turned out to be most successful.

II. PERFORMANCE ANALYSIS

With the help of the Milligan's CLUSTGEN programme (see. *Milligan* 1985, source *http://www.pitt.edu/~csna/Milligan/readme.html*), 80 data sets were generated, each containing 100 elements, in each of the Euclidean spaces R^4, R^6, R^8 (20 sets with 2, 3, 4 and 5 clusters). Then 120 data sets, each containing 80 elements, were generated, in each of the Euclidean spaces R^4 , R^6 , R^8 (20 sets with 2, 3, 4 and 5 clusters). Every set was divided into the proper (known) number of clusters with each of the 8 investigated clustering algorithms (4 classical algorithms and their 4 modifications). In order to assess the quality of grouping we applied the Rousseeuw's silhouette indices (see e.g. Gordon 1999). The silhouette index for the *i*-th point is given by the formula

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{1}$$

where *a*(*i*) is the average distance between the *i*-th point and all other points in its cluster *b*(*i*) is the average distance to points in the nearest cluster. The Euclidean distance was used. The interpretation of the silhouette index is the following: if a point has negative value of the index it means that it should be rather assigned to some other cluster. Thus, the percentage of points with the negative value of the silhouette index was used as the measure of the quality of grouping.

Table 1. Arithmetic mean percentages of wrongly classified points for sets with 100 elements (well separated clusters) and sets with 120 elements (fuzzy clusters)

Set type Grouping method	100 elements	120 elements
Single link	22,4%	34,5%
Modified single link	11,8%	12,2%
Complete link	4,2%	8,2%
Modified complete link	4,4%	4,9%
Centroid link	43,6%	48,1%
Modified centroid link	9,2%	11,3%
Group average link	24,0%	31,6%
Modified group average link	8,2%	9,5%

Source: own investigations.

III. CONCLUSIONS

There was no significant difference with respect to the dimension of the Euclidean space and to the number of clusters, therefore, we present only arithmetic means for each method in two cases: well seperated clusters and fuzzy clusters. As it can be seen the modifications almost always significantly improve the performance of the traditional grouping methods. The only exception is the complete link method which is very hard to be upgraded in the way proposed. This is probably due to the fact that the complete link method is very promiscuous itself and the counting of points lying "in between" does not bring much new as far as the mean of the number of wrongly classified observations is concerned. It is also worth observing that, in this case, the dispersion of the results was higher than for other methods (i.e. for many sets the traditional method was better, for many others its modification).

The modifications are not recommendable for large data sets because they work much longer than the traditional methods (from 3 to 6 times longer).

REFERENCES

Gordon A. D. (199), *Classification*, Chapman & Hall.

Jerzy Korzeniewski

PROPOZYCJA MODYFIKACJI ALORYTMÓW AGLOMERACYNYCH KONSTRUOWANIA SKUPIEŃ

W pracy przedstawiono propozycję modyfikacji dowolnego algorytmu aglomeracyjnego łączenia obserwacji w skupienia. Ideą modyfikacji jest położenie większego nacisku na łączenie skupień w tych obszarach, w których lokalna gęstość rozkładu obserwacji jest większa. Modyfikację zastosowano do czterech klasycznych algorytmów: aglomeracji pojedynczego połączenia, całkowitego połączenia, środka ciężkości i średniej odległości klasowej. Jakość otrzymywanych grupowań była oceniana przy pomocy odsetka obserwacji o ujemnym indeksie sylwetkowym. Wyniki pokazują, że zaproponowane modyfikacje prawie zawsze poprawiają tradycyjne algorytmy.