ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

Jerzy Korzeniewski[*]

A PROPOSAL OF A NEW METHOD OF CHOOSING STARTING POINTS FOR *k*-MEANS GROUPING

ABSTRACT. When one groups set elements with the help of *k*-means it is crucial to choose starting points properly. If they are chosen incorrectly one may arrive at badly grouped elements. In the paper a new method of choosing starting points is proposed. It is based on the distance matrix only. Starting points are chosen so as to improve the classical method of choosing points which are as far from one another as possible. The quality of grouping is assessed by means of silhouette indices – it is compared with the quality of grouping done with randomly chosen starting points and with maximum distance interval method. Sets from Euclidean spaces are generated with the help of CLUSTGEN software written by J. Milligana.

Key words: cluster analysis, *k*-means method, starting points, silhouette indices.

I. IDEA OF NEW ALGORITHM

There is a number of method of choosing starting points for *k*-means clustering. This choice influences heavily the outcome of grouping therefore it is very important to use most effective methods. Unfortunately there seems to be no universally good method i.e. a method that would perform well for all kinds of data sets. For example, the classical Hartigan-Wong's (Hartigan-Wong 1979) method (which will be later called the maximum distance interval method) works well for sets with clearly cut clusters but for slightly fuzzy sets it is actually on a par with the random choice method (see table 2). The search for a new, better method of choosing starting points was performed in a couple of directions.

In the first direction we applied the idea of comparing the distributions of pairwise distances between a fixed data point and all other points. The shape of this distribution is closely connected with the number of clusters that one should distinguish and even with the way of assigning points to clusters. This distribution (for all pairwise distances) for two two-dimensional data sets

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

depicted in Fig. 1, is presented in Fig. 2 and Fig. 3 . As it can be seen, e.g. the number of clusters in a data set is limited from below by the number of local maximums of the distribution. The way of defining a local maximum would, however, be a big problem. The investigation of similar distributions but for fixed single points may lead to interesting observations. For example, the points close to the centroids of clusters (good candidates for starting points for any grouping method) have similar shape of this distribution and the shape is rather different from the shape of the distribution for points lying far from the centroids. We tried to identify this shape by computing some measures of shape like asymmetry and curtosis. Points close to centroids usually have small asymmetry. However, this feature is not sufficient for picking up good starting points, probably, due to the fact that a point lying far from cluster centres, e.g. in between two clusters, may also have small asymmetry coefficient caused by the two small clusters relatively (in comparison with other clusters) close to this point. The method based on smallest asymmetry with a side condition preventing the choice of too close starting points, gave roughly twice smaller number of wrongly assigned (the criterion is given later) points than the random choice method.

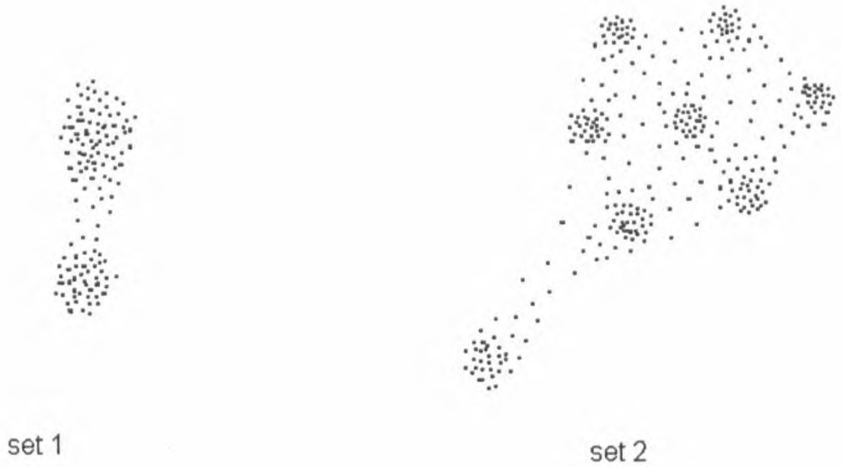

Fig. 1 Two two-dimensional data sets. The first set consists of two clusters and its diameter is about 100 units, the second set consists of eight clusters and its diameter is about 220 units.

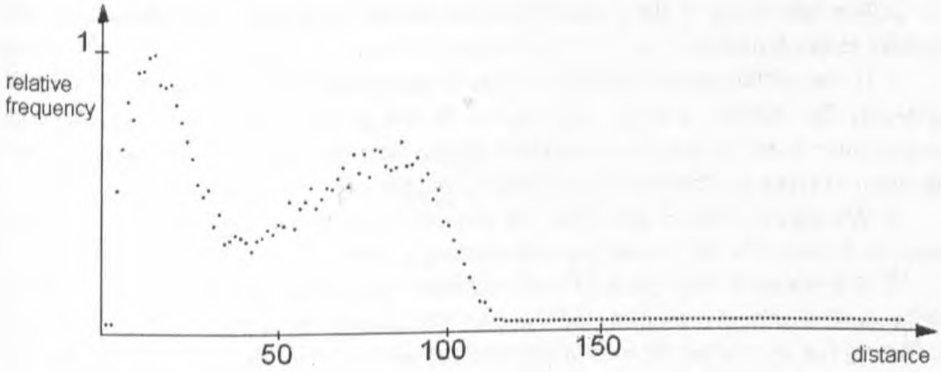

Fig. 2 The shape of the distribution of relative frequency of pairwise distances for all pairs of points for set 1. The frequency is presented in classes of width 2 and is related to the frequency of the most frequent class

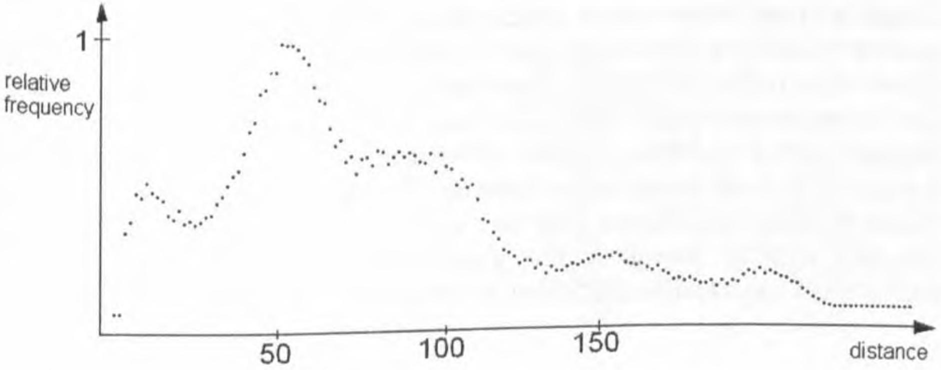

Fig. 3 The shape of the distribution of relative frequency of pairwise distances for all pairs of points for set 2. The frequency is presented in classes of width 2 and is related to the frequency of the most frequent class

In the second direction, and this approach turned out to be more successful, we tried to start in the first stage in a similar way as in the maximal distance interval method and to refine the points in the next stage. Thus, if k represents the number of clusters (and starting points) the new proposal consists of the following steps:

1. We take first k data points from the list of the data set points and call them current starting points.

2. We take the k+1 data point from the list and compute the distances to all current starting points.

3. If the distances computed in step 2 are greater than all of the distances between the current starting points we exchange one of the current starting points (one from the pair with smallest single distance out of all the distances to all other current starting point) for the k+1 data point.

4. We repeat steps 2 and 3 for the rest of the data set points arriving, in this way, at the set of k far spaced current starting points.

5. We consider the pair of two current starting points with the smallest distance d. Each of the points of this pair we change for a point whose distance to this point is smaller than ½*d and whose sum of distances to all other points with the same property is smallest.

6. We repeat step 5 for all other pairs of current starting points respectively to growing distances between pairs. Thus, we get the final set of k starting points.

The basic modification of the well known classical maximum distance interval method is contained in step 5 of the new proposal. In this step we tried a number of new ideas – most of them being based on picking up points with smallest (possibly negative) asymmetry of the distribution of distances to some chosen other points. All of these ideas did not give satisfying results, probably, due to reasons mentioned while describing the first approach. The, seemingly, simplest method of picking up point with smallest mean (or summary) distance to some other points turned out to be better. The only artificial choice here is the choice of half of the distance between the pair of starting points. Such a way is definitely artificial (though at first glance seems natural), however, this fact creates some opportunities for further investigations and possible modifications.

III. PERFORMANCE ASSESSMENT

We used the Milligan's CLUSTGEN programme, (see. *Milligan* 1985, available at *http://www.pitt.edu/~csna/Milligan/readme.html*), to generate 216 data sets, each containing 100 elements. The sets were distributed equally with respect to the dimensions of the Euclidean spaces i.e. 72 sets in each of R^4, R^6 and R^8 spaces. The division with respect to the number of clusters was also equal i.e. 54 sets with 2 clusters, 54 with 3 clusters, and 54 with 4 clusters and 54 with 5 clusters. This experiment was done twice, first time sets with well separated clusters were generated, second time 40 uniformly distributed points were added to each set so as to make the clusters slightly fuzzy i.e. not so well separated. Then, the k-means method (for three different methods of choosing starting points) was applied to group each set in the form of the number of clusters equal to the number predetermined for the set's generation. 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 shuld be rather assigned to some other cluster. Thus, the percentage of points with negative value of the silhouette index was used as the measure of the quality of grouping. The results are presented in tables 1 and 2.

Table 1

Arithmetic mean percentages of wrongly classified points for sets with well separated clusters

Number of clusters	Method		
	Random choice	Maximum distance interval	New proposal
2 clusters	18,6%	1,3%	1,2%
3 clusters	23,8%	2,4%	2,2%
4 clusters	24,1%	3,7%	3,5%
5 clusters	28,0%	2,9%	2,9%

Source: own investigations.

Table 2

Arithmetic mean percentages of wrongly classified points for sets with fuzzy clusters

Number of clusters	Method		
	Random choice	Maximum distance interval	New proposal
2 clusters	21,6%	8,3%	6,7%
3 clusters	25,2%	21,4%	8,5%
4 clusters	26,3%	22,2%	11,8%
5 clusters	29,9%	16,0%	10,7%

Source: own investigations.

The new proposal turned out to be of the same quality (or even maybe fractionally better) for sets with well separated clusters and much better for fuzzy sets than the classical method of maximum distance interval. It seems that the method of the new proposal has its prospects because its idea is based on modifying the classical approach by means of analysing the distribution of pairwise distances. The very analysis of the distribution of pairwise distances so far did not give good results.

REFERENCES

Gordon A. D., *Classification*, Chapman & Hall, 1999.

Hartigan J. A., Wong M. A., *A K-means clustering algorithm*, Applied Statistics 28, 100–108 1979.

Milligan G. W., *An algorithm for generating artificial test clusters*, "Psychometrika", vol. 50, no. 1, 123–127, 1985.

Jerzy Korzeniewski

PROPOZYCJA NOWEJ METODY WYBORU PUNKTÓW STARTOWYCH DO GRUPOWANIA METODĄ *K*-ŚREDNICH

Gdy grupujemy punkty zbioru metodą *k*-średnich to zasadniczym problemem jest właściwy wybór punktów startowych. Jeśli są one źle wybrane to grupowanie może być złe. W artykule zaproponowana jest nowa metoda wyboru punktów startowych. Metoda ta jest oparta wyłącznie na znajomości macierzy odległości. Punkty startowe są wybierane tak, by poprawić wybór , który otrzymamy przy pomocy metody klasycznej polegającej na wyborze punktów możliwie jak najbardziej od siebie oddalonych. Jakość grupowania jest oceniana przy pomocy indeksów sylwetkowych – porównywana jest z jakością grupowania otrzymanego przy losowym wyborze punktów startowych oraz przy wyborze metodą klasyczną. Zbiory z przestrzeni euklidesowych są generowane przy pomocy programu CLUSTGEN autorstwa J. Milligana.