THE SPECIFICS AND THE IMPACT OF DEFINING THE INITIAL SET OF CENTROIDS ON INFORMATION ANALYSIS

Ilie COANDĂ
Academy of Economic Studies of Moldova,
Str. Mitropolit Gavriil Bănulescu-Bodoni ,61, MD-2005, Chişinău,Republic of Moldova, +373402988,
ildirosv1@gmail.com
***Corresponding author:** ildirosv1@gmail.com*

Abstract

The purpose of this paper is to highlight the importance and influence of the way of defining the initial data for the realization of algorithms for classifying information. At first glance, the problem of choosing the initial set of data processing centers for information analysis purposes could be considered as simple and obvious. In fact, things are not so clear in the initial stage of data grouping, given that the choice of clustering centers, in most cases, is strongly influenced by the specificity of each formulated problem. Even if in the absolute majority of the fieldworkers this is highlighted, but there is no suggestion of a possible way, if such a one exists, to bypass this phenomenon. The very simple clustering algorithm provides us with a fairly fast convergence, which is why it is more important that the initial parameters of the process of involving this algorithm solve our problem at a respectable efficiency level, because an optimal solution, in terms of the mathematical strictures regarding the notion "optimum", in real life, does not exist. Then, for example, when we use the word like "good", it is necessary to specify in concrete, what we mean. In conclusion: for each concrete problem it is necessary to initiate a deeper study of the impact level of the phenomenon of dependence on the final result of the clustering, and, thus defining the initial data set of the algorithm, to lead us to a solution, being on as close as possible, which we could consider the most acceptable.

***Key words**: centroids, grouping, data, information, analysis.*

***Jel Classification*: C63, I21, I23, I25, I29**

INTRODUCTION

The K-Means algorithm is a very simple one to use, at the same time with a very high convergence speed. This is mentioned in most of the papers on this topic. On the other hand, it is also emphasized that, in the last period of time, the attention paid to this algorithm increases due to the extension of their implementation areas both horizontally and vertically. Nor is the phenomenon of very rapid growth of data volume - information, left in the shadow. It is true, not to mention that the performance of the HARD components of the tools for implementing this algorithm are also in a sharp rise, which diminishes the impact of the rapid increase of the volume of data.

The object clustering process is one of the many tools for studying, analyzing data, which is quite popular, that we have nowhere to determine, that it is the most of all the methodologies of data analysis. Of course, that some statistical study can be done, but that one will be with a margin of error. The special attention to this technology is very dictated by the simplicity of its implementation. But, rather quickly, some shortcomings also came to light, and one of those with a rather significant impact, being the very strong dependence of the final solution on the values of the initial parameters of the algorithm, in particular, of the initial set of centers (centroids).

The work [1] uses the expressions "the right number of clusters", "optimal gain" as a purpose of a technology called "divide-and-conquer" after the K-means algorithm has been used. Also here [1] the proposed algorithm is described in detail and some results are presented which: "This algorithm ensures the clustering of data in less time without affecting the accuracy of clusters" without reaching the problem of the uniqueness of the solution, also, without to reveal the essence of the expression "better results". It is also assumed that "Clusters (K) and data set is provided by the user", without specifying any information regarding the number of centers, as well as the definition of them. In conclusion, the focus is on efficiency in terms of the time required to solve the clasterization problem, the gain is more than 50%. As for another feature "accuracy" being considered important, the gain is not significant, it is less than 10%.

In another paper [2], there is also the problem of selecting the initial centers for the K Means algorithm. There are several ways to solve the problems - the shortcomings of the algorithm in question. As the basis of the experiments, the studies are taken into account data in 2D space, so that the study is more clearly performed and more clearly explained. The notion of density is also taken into account. However, the greatest attention is paid to the problem of defining the initial set of centers. He mentions, in particular, "Some of the studies reviewed focused on discovering the proper number of clusters prior to running the clustering process ... noted the importance of the number of centroids with respect to the number of clusters". Also there is the importance of the initial choice of the classification centers, thus emphasizing "weights based on probability, choosing a centroid location based upon density, subdividing clusters into smaller subsections prior to choosing a centroid, using a graph based method, and combinations of these techniques".

In article [3], in order to define an initial set of N clusters to obtain a classification more appropriate to the real situation, a method of starting the initial centers based on an empirical study is proposed, using time series, thus obtaining some clarity, suggestions on classification results. In this paper [3] the study of the K - means methodology is carried out on the basis of a concrete problem, namely the solution of the problems arising regarding the conservation of electricity, which are required by law by the Government of Japan. As the main conclusion: "Initial centroids chosen using the percentile method from empirical cumulative distribution were found to be more accurate than the random initialization method and empty clusters were removed ...", "using the clustering technique, it is necessary to select the proper number of clusters… ". In particular, as an acceptable result, it is mentioned that "The calendar plot for three to five clusters using K-means clustering did not match the university schedule. For six clusters, the clustering result was similar to the university schedule with an accuracy of 89.3%. So, we found that six clusters were appropriate for Chubu University. "

The same problem of initiation of the Peruvian centrum K - Means methodology is also discussed in the paper [4]. And here again it is specified that "the original k-means algorithm is computationally expensive and the resulting set of clusters strongly depends on the selection of initial centroids", which led to a decision to study in the basis of a new method "a heuristic method to find better initial centroids as well as more accurate clusters with less computational time". As a main conclusion: "a modified k-means algorithm that finds the approximate centroids that reduces the number of iteration to assign the data into a cluster. One limitation of this algorithm is that we still need to provide the desired cluster number as input data".

The work on improving the processes of the K - means methodology is also dedicated [5]. In this paper a study is performed in which "a novel hybrid evolutionary model for k-means clustering (HE-kmeans) is proposed. This model uses meta-heuristic methods to identify the "good candidates" for initial centroid selection in k-means clustering method. As a main conclusion, the method used allows the K - Means technology to be enriched "... by using two separate datasets with different clustering and model parameters and our model outperformed the original and PSO-based k-means in clustering quality by approximately 30%."

METHOD, RESULTS AND DISCUSSIONS

In accordance with the purpose formulated above, the study was conducted on the basis of numerical experiments. The set of points was defined in: a) random distribution and b) uniform distribution. An application was created on the computer, which in a range of 400x400 pixels, defines N points (so we have 2D attribute space). With these N objects (points) defined, several calculations were performed (for different values of N, for both modes of object distribution on the surface 400x400. For each set of parameters (N, a), b)) it was performed the procedure of launching the K - Means methodology for different sets consisting of 3,4 5 and 6 initial centers of centroid. The results of the experiments we can visually analyze the images (photos) below.

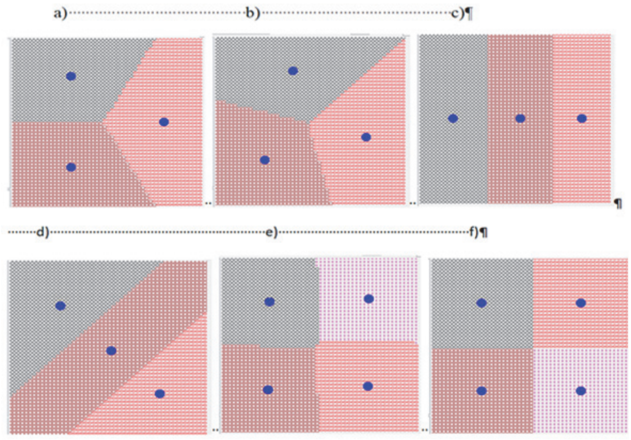

Photo 1 (a,b,c,d,e,f). Uniform distribution for 3 and 4 centers.

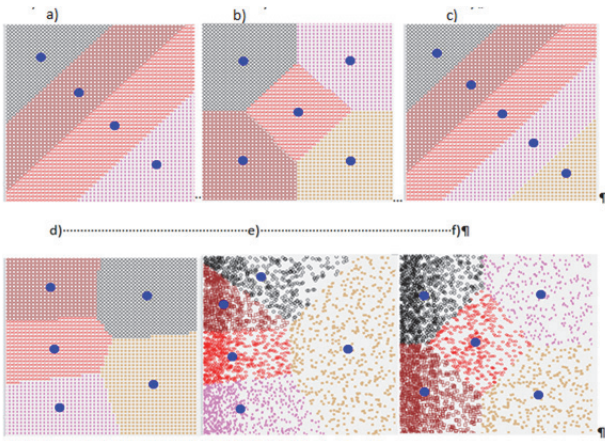

Photo 2.(a,b,c,d,e,f) Uniform and random distribution.

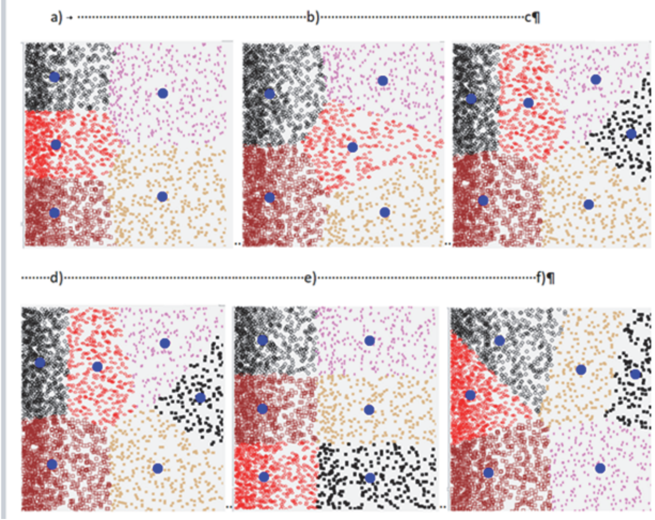

Photo 3 (a,b,c,d,e,f). Random distribution

Comments: For all segments a), b), c), d), e), f) from the above photos, the X coordinate is measured from 0 to 400 pixels, from the left - top right corner, the Y coordinate - it is measured from 0 to 400 pixels, from top to bottom.

In the image **Photo 1** (segments a), b), c), d)) are presented the results for the sets of three initial centers, for e), f) - four initial centers, the initial centers of Photo 1, a): X (50,50,200), Y (50,350,200) (so we have the centers (X1, Y1) = (50.50), (X2, Y1) = (50,350), (X2, Y1) = (200,350). this way of explaining the coordinates for each set of initial centers in each image, for each segment.) For: **Photo 1**, b): X (50,50,350), Y (50,350,350); **Photo 1**, c): X (50,250,350), Y (50,50.50);) Photo 1, d): X (50,250,350), Y (50,250,350); Photo 1, e): X (50,50,200,300), Y (50,350,200.50); Photo 1, f): X (50,50,350,350), Y (50,350,50,350); In **Photo 2** (segmental a), four initial centers (a): X (50,100,250,350), Y (50,100,250,350) ;, the other - b), c), d), e), f) - five initial centers: b): X (50,50,200,350,350),Y(50,350,200,50,350);, c): X(50,100,200,300,350), Y(50,100,200,350,350); d): X(50,50,50,50,350), Y(50,100,200,300,350); e): X(50,50,50,50,350), Y(50,100,200,300,350); f): X(50,50,200350,350), Y(50,350,200,50,350); In the image **Photo 3** (segments a), b) - five initial centers, segments c), d), e), f) - six initial centers. Thus for a): X (50,50,350,350,350), Y (50,350,200,50,350);, for b): X (50,50,350,350,350), Y (50,350,200,50,350) ;,for c): X (50,100,150,250,300,350), Y (350,350,350,350,350,350);, for e): X (50,50,50,350,350,350), Y (50,200,350,50,200,350);,) ;,for d): X (50,50,50,350,350,350), Y (50,200,350,50,200,350) for f): X (250,250,150,250,300,350), Y (350,350,350,350,350,350);

We mention that the random distributions were also made such that in the left regions of each segment of the Photo 3 image, they are with the highest density, and those on the right - with the lowest density. Thus we wanted to have the elements with different densities in order to have the possibility to conduct a study, including the effect of the uneven distribution, but also with different densities. There are presented the variants of diminishing density from left to right, making sure that the same effect can be obtained on the diagonal.

We emphasize that for the solutions (uniform distribution of points) predicted in Photo 1 (segments a), c), d), f)), Photo 2 (segments a), b), c)) are the results for which the centers Initial, practical, have not been modified. At the same time, the solutions (uniform distribution of points) of Photo 1 (segments b) and e)), Photo 2 segment d) correspond to the case when the initial centers are modified during the application of the K means procedure.

For the case of five initial centers, random distribution, increasing deity from left to right, we note that, for the variants, where the initial centers were placed with a small deviation of the locations from those obtained as a solution, the manual definition of the initial centers is reasonable - recommended. For segment e) of Photo 3, the initial centers, as a result of the K - Means procedure, were slightly shifted to the left (higher density region).

CONCLUSIONS

The presented results clearly indicate that it is reasonable that the definition of the initial centers for the application of K - means methodologies is reasonable, recommended. In our opinion, in many situations, the user, the beneficiary is the one who can substantially influence the results of the clustering operation.

REFERENCES

[1]Initialization of optimized K- Means centroids using Divide-and-conquer method, Web address:http://www.arpnjournals.org/jeas/research_papers/rp_2016/jeas_0116_3459.pdf

[2] On Initial Effects of the k-Means Clustering. Web address:http://worldcomp-proceedings.com/proc/p2015/CSC2667.pdf

[3] Analysis of Building Electricity Use Pattern Using K-Means Clustering Algorithm. Web address: https://ideas.repec.org/a/gam/jeners/v12y2019i12p2451-d242911.html

[4] Improvement of K-means clustering algorithm with better initial centroids based on weighted average. Web address: https://www.researchgate.net/publication/261233398_Improvement_of_K-means_clustering_algorithm_with_better_initial_centroids_based_on_weighted_average

[5] Clustering Quality Improvement of k-means Using a Hybrid Evolutionary Model. Web address:https://www.sciencedirect.com/science/article/pii/S1877050915029737