Journal of Electrical Engineering, Electronics, Control and Computer Science – JEEECCS, Volume 7, Issue 25, pages 47-50, 2021

A Fast Clustering Method for Large Data Sets

O. Kettani
Scientific Institute
Mohammed V University
Rabat, Morocco

Abstract **– In this paper a deterministic clustering method based on the Katsavounidis, Kuo & Zhang (KKZ) seed procedure, is proposed. The computational complexity of this approach is lower than the complexity of the prominent k-means algorithm. Comparison of our method with the related deterministic clustering method: KKZ_k-means (k-means initialized by KKZ) was done and performance evaluation demonstrates its effectiveness in term of average Silhouette index in various benchmark datasets.**

Keywords-component; Clustering; k-means; Dataset; KKZ; Silhouette

I. INTRODUCTION

The clustering problem has many applications in Machine Learning, Pattern Recognition and Statistics. Clustering consists of grouping similar data into groups called clusters, so that the objects in the same cluster are more comparable to each other and more distinct from the objects in the other clusters [1]. This is an NP-hard optimization problem, even when the clustering process deals with only two clusters [2]. Till now, many heuristics have been proposed to tackle this problem in reasonable computational time. In the present study, yet another clustering approach which has the benefit of low computational complexity is suggested.

This paper is organized as follows: after the introduction, we discuss briefly some related work. Then the proposed approach and its computational complexity are described in Section 3. In section 4, this clustering method is applied to some standard data sets and comparison with the related deterministic clustering method, KKZ_k-means (k-means initialized by KKZ) is done. Lastly, conclusion of the paper is done in Section 5.

II. RELATED WORK

Given a set of n data points (objects) X = {x1 , . . . , xn } in Rd and an integer k, the clustering problem consists to determine a partition (Cj)1 ≤ j ≤ k of X , in order to minimize the following Sum of Square Error (SSE) function:

$$SSE = \sum_{i=1}^{k} \sum_{x_i \in C_i} (x_j - c_i)^2 \quad (1)$$

Where || . || 2 denotes the Euclidean norm, and k

$$c_i = \frac{1}{|C_i|} \sum_{x_i \in C_i} x_j \quad (2)$$

denotes the centroid of cluster Ci whose cardinality is |Ci|.

Among the many existing clustering methods, the k-means algorithm [3][4] is the most commonly used clustering approach, because its simplicity. However, one of its major drawbacks is its sensitivity to initial seeds. Several methods have been proposed in the literature to overcome this issue, among them, the Katsavounidis, Kuo & Zhang (KKZ) seed procedure [5], (see Table 1 a)). This approach has a computational time complexity in O(knd). Another existing method is the FICA algorithm, proposed by Kettani & Ramdani in a previous work [6] (see Table I b)). Recently, Vo-Van et al. have suggested a new clustering algorithm based on the definition of epsilon radius neighbors, that can automatically determine the number of clusters and can find clusters with different sizes, shapes, and densities [7]. However, this algorithm might run slowly on large datasets. In the present paper, an alternative approach to the k-means algorithm is proposed: Its initialization process consists to run the KKZ seed procedure, then its assigns each data point to its "nearest cluster", using a criterion based on the following lemma which aims to minimize the SSE.

TABLE I A): PSEUDO-CODE OF THE KKZ SEED PROCEDURE

Input: A data set X with cardinality n and an integer k
Output: k center c_j

$$c_1 \leftarrow \text{Arg}(\text{Max}(|| x_h ||))_{1 \le h \le n}$$

For j=2:k **do**

$$m \leftarrow \text{Arg}(\text{Max}(\text{Min}(|| x_i - c_h ||)))_{1 \le i \le n, 1 \le h \le j-1}$$

$c_j \leftarrow x_m$

end For

TABLE I B): PSEUDO-CODE OF THE FICA ALGORITHM

Input: A data set X with cardinality n and an integer k
Output: k cluster C_j

$c_1 \leftarrow \text{Arg}(\text{Max}(|| x_h ||))_{1 \le h \le n}$

For j=2:k **do**
$m \leftarrow \text{Arg}(\text{Max}(\text{Min}(|| x_i - c_h ||)))_{1 \le i \le n, 1 \le h \le j-1}$
$c_j \leftarrow x_m$

end For

For i=1:n **do**
$j \leftarrow \text{Arg}(\text{Min}(|| x_i - c_h ||))_{1 \le h \le k}$

$C_j \leftarrow C_j \cup \{x_i\}$
$n_j \leftarrow |C_j|$
$c_j \leftarrow \text{mean}(C_j)$

end For

Lemma [8]:
Let a cluster C_j consist of p points x_i , i = 1, . . . , p
With a mean

$$x_p = \frac{1}{p}\sum_{i=1}^{p} x_i \qquad (3)$$

Let the objective function be

$$O_p = \sum_{i=1}^{p} (x_i - x_p)^2 \qquad (4)$$

Then O_p is increased by

$$\frac{p}{p+1}(x_p - x_{p+1})^2$$

when a point x_{p+1} is added to cluster C_j

Proof:
x_{p+1} , the center of $C_j \cup \{x_{p+1}\}$ is at

$$x_{p+1} = \frac{p x_p + x_{p+1}}{p+1} = x_p + \frac{x_{p+1} - x_p}{p+1}$$

then

$$O_{p+1} = O_p + x_{p+1}^2 + p x_p^2 - (p+1) x_{p+1}^2 = O_p + \frac{p}{p+1}(x_p - x_{p+1})^2$$

which proves the lemma.

III. PROPOSED APPROACH

The pseudo-code of the proposed approach FCM (Fast Clustering Method) is shown below. Notice that it differs at the assignment loop from the FICA [6] code by the factor (nh /(1+nh)) suggested by the previous lemma.

TABLE II: PSEUDO-CODE OF THE PROPOSED FCM APPROACH

Input: A data set X with cardinality n and an integer k
Output: k cluster C_j

$c_1 \leftarrow \text{Arg}(\text{Max}(|| x_h ||))_{1 \le h \le n}$

For j=2:k **do**

$m \leftarrow \text{Arg}(\text{Max}(\text{Min}(|| x_i - c_h ||)))_{1 \le i \le n, 1 \le h \le j-1}$

$c_j \leftarrow x_m$

end For

For i=1:n **do**

$j \leftarrow \text{Arg}(\text{Min}(\frac{n_h}{1+nh} || x_i - c_h ||^2))_{1 \le h \le k}$

$C_j \leftarrow C_j \cup \{x_i\}$
$n_j \leftarrow |C_j|$
$c_j \leftarrow \text{mean}(C_j)$

end For

Complexity

The running time of step 1 and 2 (which correspond to the KKZ procedure) is O(knd) [5], and the for loop in step 3 requires O(nkd) operations, so the overall running time complexity of FCM is O(nkd), which corresponds to the complexity of one iteration of the k-means algorithm.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

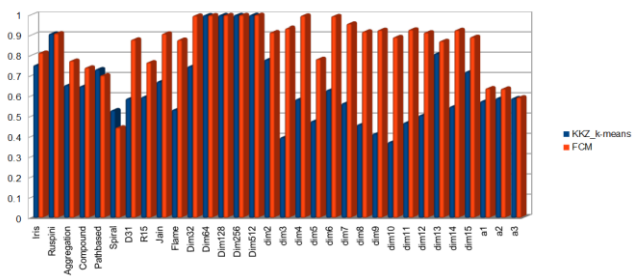
We evaluated algorithm performance by applying on several benchmark datasets from the UCI Machine Learning Repository [9] and compare with KKZ_ k-means. In a preprocessing step, the data were normalized. Silhouette index [10] was used in these experiments in order to evaluate clustering accuracy and experimental results are reported in Table III and Figure 1.

Advantages of this proposed method are: it is deterministic and more performing than KKZ in term of average silhouette value. An inconvenient of this method is that is slightly more complex than FICA [6].

TABLE III. EXPERIMENTAL RESULTS OF KKZ_K-MEANS AND FCM APPLICATION ON DIFFERENT DATASETS IN TERM OF AVERAGE SILHOUETTE VALUE

Data set	k	KKZ_k-means	FCM
Iris	3	0.7527	**0.8142**
Ruspini	4	0.9081	**0.9093**
Aggregation	7	0.6542	**0.7745**
Compound	6	0.6496	**0.7410**
Pathbased	3	**0.7325**	0.7025
Spiral	3	**0.5299**	0.4449
D31	31	0.5881	**0.8786**
R15	15	0.5966	**0.7664**
Jain	2	0.6720	**0.9081**
Flame	2	0.5338	**0.8775**
Dim32	16	0.7472	**0.9961**
Dim64	16	**0.9985**	0.9984
Dim128	16	**0.9991**	**0.9991**
Dim256	16	**0.9996**	**0.9996**
Dim512	16	**0.9998**	**0.9998**
dim2	9	0.7816	**0.9148**
dim3	9	0.3966	**0.9340**
dim4	9	0.5849	**0.9968**
dim5	9	0.4776	**0.7830**
dim6	9	0.6308	**0.9938**
dim7	9	0.5652	**0.9553**
dim8	9	0.4604	**0.9184**
dim9	9	0.4147	**0.9260**
dim10	9	0.3738	**0.8909**
dim11	9	0.4696	**0.9281**
dim12	9	0.5059	**0.9140**
dim13	9	0.8105	**0.8717**
dim14	9	0.5487	**0.9258**
dim15	9	0.7207	**0.8918**
a1	20	0.5758	**0.6384**
a2	35	0.5907	**0.6366**
a3	50	0.5898	**0.5936**

Figure 1. Chart of average Silhouette index for FCM and KKZ_k-means applied on different datasets



CONCLUSION

In this study, a fast clustering method was suggested. This approach is an alternative to the k-means algorithm for producing better clustering with less computational time and experimental results have showed that it is effective in finding consistent clustering results, when it is applied on various datasets.

Future work will consist to compare this method to others existing clustering algorithms. A possible improvement will consist to consider a parallelization of this method, for faster clustering.

REFERENCES

[1] Ankerst, M., M. Breunig, H.P. Kriegel and J. Sander, 1999.OPTICS: Ordering points to identify the clustering structure. Proceeding of ACM SIGMOD International Conference Management of Data Mining, May 31-June 3, ACM Press, Philadelphia, United States, pp: 49-60.

[2] Aloise, D.; Deshpande, A.; Hansen, P.; Popat, P. (2009). "NP-hardness of Euclidean sum-of-squares clustering". Machine Learning 75: 245–249. doi:10.1007/s10994-009-5103-0.

[3] Lloyd, S.P., 1982. Least square quantization in PCM. IEEE Trans. Inform. Theor., 28: 129-136.

[4] MacQueen, J.B., 1967. Some Method for Classification and Analysis of Multivariate Observations, Proceeding of the Berkeley Symposium on Mathematical Statistics and Probability, (MSP'67), Berkeley, University of California Press, pp: 281-297.K. Elissa, "Title of paper if known," unpublished.

[5] Katsavounidis, I., C.C.J. Kuo and Z. Zhen, 1994. A new initialization technique for generalized Lloyd iteration. IEEE. Sig. Process. Lett., 1: 144-146.

[6] Kettani O. & Ramdani, F. FICA: Fast Incremental Clustering Algorithm.International Journal of Computer Applications 164(1):34-39 , April 2018. 179(33):35-38 DOI: 10.5120/ijca2018916747

[7] T. Vo-Van, A. Nguyen-Hai, M. V. Tat-Hong, T. Nguyen-Trang, "A New Clustering Algorithm and Its Application in Assessing the Quality of Underground Water", Scientific Programming, vol. 2020, Article ID 6458576, 12 pages, 2020. https://doi.org/10.1155/2020/6458576

[8] Pawel Kalczynski et al . « The Importance of Good Starting Solutions in the Minimum Sum of Squares Clustering Problem » https://arxiv.org/abs/2004.04593

[9] Asuncion, A. and Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html] Irvine, CA: University of California, School of Information and Computer Science.

[10] L. Kaufman and P. J. Rousseeuw. Finding groups in Data: "an Introduction to Cluster Analysis". Wiley, 1990.