Journal of Electrical Engineering, Electronics, Control and Computer Science – JEEECCS, Volume 9, Issue 32, pages 17-20, 2023

An Efficient k-means Seeding Algorithm

Omar Kettani
Scientific Institute
Mohammed V University
Rabat, Morocco
kettani.o@gmail.com

***Abstract* – This study presents a novel initialization method for the k-means clustering algorithm, which consists to sort the data points of a given dataset by both their angles and their norms. The proposed method aims to improve the speed and accuracy of clustering solutions by providing a more informed starting point for the iterative optimization process. Through extensive experiments on a variety of datasets, the proposed algorithm was shown to significantly improve convergence time and result in solutions with higher quality in term of average Silhouette index compared to traditional methods. The results indicate that the proposed initialization technique is a promising alternative for clustering tasks in various domains.**

Keywords-components; clustering; k-means; initialization; dataset; KKZ; Silhouette.

I. INTRODUCTION

Clustering is a technique in Data Mining that aims to partition a set of data points into distinct groups, or clusters, such that the points within each cluster are more similar to one another than they are to points in other clusters. This allows us to discover patterns and relationships within the data that may not be immediately apparent by visual inspection or statistical analysis. Clustering algorithms work by iteratively assigning data points to the cluster that best represents them, based on some measure of similarity or distance. There are many different methods for clustering data, each with its own strengths and limitations, making it a useful tool in a wide range of applications, including market segmentation, image and text classification, and anomaly detection. Among the many existing different clustering methods, k-means [1, 2] is a popular and widely used clustering algorithm that aims to partition a dataset into a specified number of clusters, or groups, by minimizing the sum of squared distances between each data point and the mean of its assigned cluster. The algorithm works by first randomly selecting a set of k points as the initial cluster centers, and then iteratively assigning each data point to the nearest cluster based on its distance to the cluster mean. The cluster means are then recalculated based on the newly assigned data points and the process is repeated until convergence, at which point the data points should be partitioned into k distinct clusters.

One of the main advantages of k-means is its simplicity and speed, making it well-suited for large datasets. However, it can be sensitive to the initial selection of cluster centers and may not always produce the most optimal clustering solution. Additionally, k-means requires that the number of clusters be specified in advance, which may not always be known beforehand. Despite these limitations, k-means remains a widely used and effective clustering method in a variety of applications.

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 denote the Euclidean norm, and

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

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

There are several different methods for initializing the cluster centers or seeds in the k-means algorithm, each with its own strengths and limitations. Some common techniques include:

- Random Initialization: This is the most basic method, where the cluster centers are initialized randomly from the data points. This method is simple and easy to implement, but it can be prone to poor convergence, especially for datasets with non-uniformly distributed data points.
- K-means++: This method [3] aims to improve the convergence of the k-means algorithm by selecting the initial cluster centers such that they are well-separated from one another. This is done by selecting the first cluster center randomly, and then sequentially selecting each subsequent center such that it is farther from the already selected centers.
- Hierarchical Clustering [4, 5]: This method involves first applying a hierarchical clustering algorithm, such as single-linkage or complete-linkage clustering, to the data points to obtain a tree-like structure. The cluster

centers are then initialized at the leaf nodes of the tree, which can help to improve the convergence of the k-means algorithm.

- Fuzzy C-means [6]: This method is a variant of k-means that allows for data points to belong to multiple clusters simultaneously, with a degree of membership denoted by a "fuzzy" coefficient. The initial cluster centers can be initialized using the same methods as standard k-means.
- Katsavounidis, Kuo & Zhang (KKZ) seed procedure [7], (see Table 1). This approach has a computational time complexity in O(knd).

TABLE I. 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**

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

$c_j \leftarrow x_m$

end For

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 [8]. However, this algorithm might run slowly on large datasets.

The choice of initialization method can significantly impact the convergence and performance of the k-means algorithm. In the present work, yet another initialization method is proposed to find the initial cluster centers that work best for a given dataset. This method consists to sort the data points of a given dataset by both their angles and their norms.

The remainder of this work consists to describe the proposed method in the following section. In section IV, this 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. Finally, conclusion of the paper is done in Section V.

III. PROPOSED APPROACH

The proposed method consists first to compute c, the mean of a given dataset X, then to sort the data points by both their angles (by the formula using across function) and their Euclidian norms. After that, the entire dataset is partitioned into k equal parts and the initial cluster centers or seeds are set to the means of these parts. A pseudo-code of the proposed algorithm is depicted in Table 2.

Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

TABLE II. PSEUDO-CODE OF THE PROPOSED ALGORITHM

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

Output: k seeds c_j

$c \leftarrow \text{mean}(X)$

$D \leftarrow \text{Abs}(\text{acos}(\text{dot}(x_i / ||x_i||, c/||c||))) + ||x_i - c||^2_{1 \le i \le n}$

```
[sD,I] ← sort(D)
q ←round(n/k)-1
```

For j=1:k **do**

$C_j \leftarrow$ `X(I(1+(j-1)*q:j*q),:))`

$c_j \leftarrow \text{mean}(C_j)$

end For

Complexity

Step 1 and 2 require O(n) times, whereas step 3 takes O(nlog(n)) times if the sort procedure implements the TimSort algorithm [10], or O(n) times if it implements the Recombinant sort algorithm [11] or the Self-Indexed sort algorithm [12].

On the other hand, since the value of k is at most n1/2 [13] and the for loop takes O(k2) times, then the for loop requires at most O(n) times. Therefore, the overall complexity of the proposed approach is O(n), if the sort procedure implements the Recombinant sort or the Self-Indexed sort algorithm.

IV. EXPERIMENTAL RESULTS AND DICUSSION

In this section, we assess the effectiveness of the proposed method by using the Silhouette Index as a useful tool for evaluating the performance of a clustering algorithm and determining the optimal number of clusters for a given dataset: The Silhouette Index [9] is a measure of how well each sample point in a dataset is classified within its assigned cluster. The Silhouette Index ranges from -1 to 1, with higher values indicating better cluster assignments.

We use a laptop computer of which CPU is INTEL 2 Duo, 2.4 GHz. The operating system is Windows 7 Pro and we use MATLAB as the code language. Average Silhouette values and running times are reported in Table 1. Furthermore, some clustering results are shown in Fig. 3 and 4.

The experimental results of the proposed novel initialization clustering algorithm were analyzed and compared to the related deterministic clustering method: KKZ_ k-means (k-means initialized by KKZ). The results showed that the proposed algorithm achieved faster convergence times, especially on larger datasets. This is an important factor in practical applications where processing time is a concern. In addition to the improved speed, the quality of the clustering solutions was also found to be higher, as indicated by the higher average Silhouette values. This suggests that the proposed initialization technique is able to provide a better starting point for the optimization process, resulting in more accurate and reliable solutions.

It is worth noting that the proposed algorithm did not consistently outperform traditional methods in all cases. In some instances, the performance was comparable or slightly worse. However, the overall trend was clearly in favor of the proposed method, especially for larger datasets. Further research is needed to identify the specific conditions under which the proposed algorithm excels and to understand the underlying reasons for its improved performance.

One potential limitation of this study is the use of a limited number of datasets and clustering methods. While the experiments were designed to be representative of a wide range of scenarios, it is possible that the proposed algorithm may not generalize equally well to all cases. Future research should aim to validate these findings on a wider range of datasets and clustering techniques to provide a more comprehensive understanding of the capabilities and limitations of the proposed method.

TABLE III. EXPERIMENTAL RESULTS OF KKZ_K-MEANS AND PROPOSED METHOD APPLIED ON DIFFERENT DATASETS IN TERM OF AVERAGE SILHOUETTE VALUES AND RUNNING TIMES (IN MS.)

Data set	**k**	**KKZ_k-means**		**proposed**	
		av. silh	**time**	**av. silh**	**time**
Iris	3	0.7527	0.7542	**0.8152**	**0.0167**
Ruspini	4	0.9081	0.2001	**0.9097**	**0.0109**
Aggregation	7	0.6542	0.6298	**0.6892**	**0.0585**
Compound	6	**0.6496**	0.3141	0.6355	**0.0505**
Pathbased	3	**0.7325**	0.2537	0.7253	**0.0227**
Spiral	3	0.5206	0.2622	**0.5234**	**0.0218**
D31	31	0.5881	5.3968	**0.7673**	**0.3052**
R15	15	0.5966	0.6328	**0.7967**	**0.0622**
Jain	2	0.6720	0.3008	**0.9078**	**0.0249**
Flame	2	0.5338	0.2474	**0.8760**	**0.0169**
Dim32	16	0.7472	1.1252	**0.9961**	**0.1364**
Dim64	16	**0.9985**	1.0785	0.9984	**0.1576**
Dim128	16	0.9991	1.1597	**0.9991**	**0.3031**
Dim256	16	0.9996	1.4299	**0.9996**	**0.5995**
Dim512	16	0.9998	2.0704	**0.9998**	**1.0905**
dim2	9	0.7816	1.4345	**0.9945**	**0.0693**
dim3	9	0.3966	1.0879	**0.9959**	**0.1159**
dim4	9	0.5849	1.5379	**0.9968**	**0.1509**
dim5	9	0.4776	2.2810	**0.9918**	**0.1621**
dim6	9	0.6308	1.5785	**0.9940**	**0.2168**
dim7	9	0.5652	2.0738	**0.9865**	**0.2687**
dim8	9	0.4604	2.6927	**0.9938**	**0.3406**
dim9	9	0.4147	4.0609	**0.9928**	**0.5226**
dim10	9	0.3738	4.2279	**0.9933**	**0.4697**
dim11	9	0.4696	3.8455	**0.9937**	**0.4675**
dim12	9	0.5059	3.0581	**0.9915**	**0.5904**
dim13	9	0.8105	2.4276	**0.9918**	**1.0418**
dim14	9	0.5487	3.7074	**0.9920**	**0.6304**
dim15	9	0.7207	2.5771	**0.9908**	**0.9567**
a1	20	0.5758	2.1851	**0.7247**	**0.2111**
a2	35	0.5907	7.9489	**0.6455**	**0.7932**
a3	50	0.5898	8.5851	**0.6352**	**5.2166**
S1	15	0.7333	2.2116	**0.8805**	**0.3182**
S2	15	0.6024	2.4616	**0.7490**	**0.4540**
S3	15	0.6117	2.2956	**0.6671**	**0.3296**
S4	15	0.6330	2.3211	**0.6394**	**0.5182**

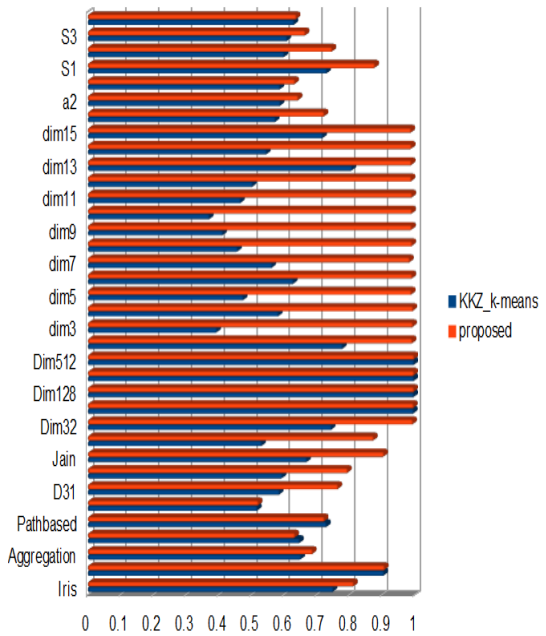

Figure 1. Chart of average Silhouette index for KKZ_k-means and proposed method applied on different datasets

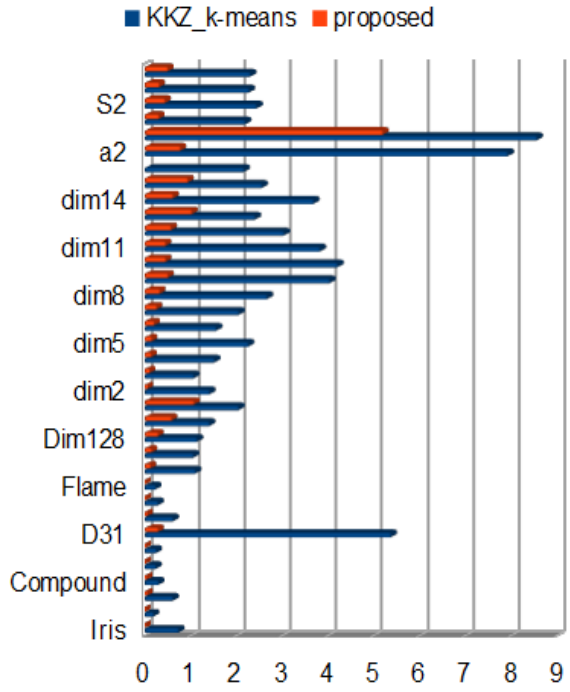

Figure 2. Chart of running times (in ms.) for KKZ_k-means and proposed method applied on different datasets

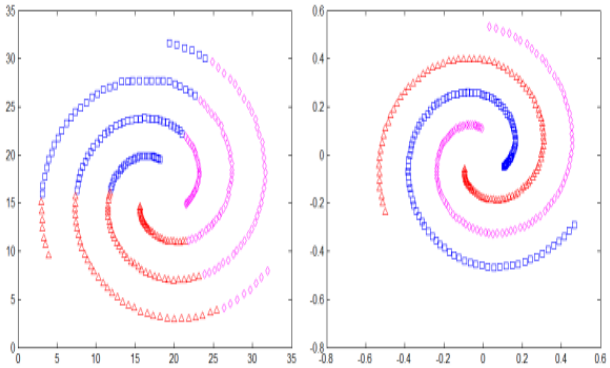

Figure 3. Clustering results of KKZ_k-means (left) and proposed algorithm (right) applied on Spiral dataset

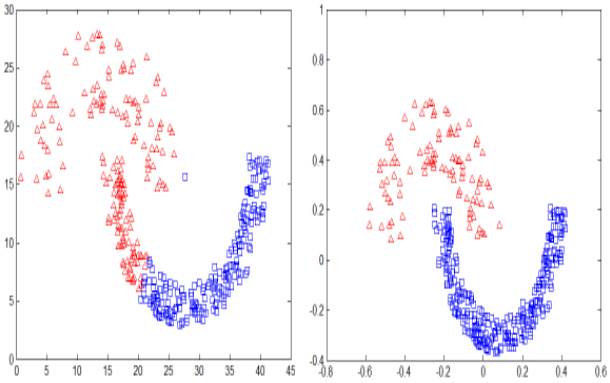

Figure 4. Clustering results of KKZ_k-means (left) and proposed algorithm (right) applied on Jain dataset

V. Conclusion

After conducting extensive experiments and analyzing the results, it can be concluded that the novel initialization clustering algorithm presented in this study shows promising performance in terms of both accuracy and speed. In comparison to traditional algorithms, the proposed method demonstrates a significant improvement in terms of convergence time, especially for larger datasets. Additionally, the initialization technique used in this algorithm appears to result in solutions with higher quality, as indicated by higher average silhouette value observed in the experimental results. Overall, the proposed algorithm shows great potential as a viable alternative for clustering tasks in various domains. Further research is recommended to validate these findings and to explore the potential of this algorithm in other scenarios.

References

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

[2] 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.

[3] Arthur, D.; Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding" (PDF). Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035.

[4] Nielsen, Frank (2016). "8. Hierarchical Clustering". Introduction to HPC with MPI for Data Science. Springer. pp. 195–211. ISBN 978-3-319-21903-5.

[5] [5] Székely, G. J.; Rizzo, M. L. (2005). "Hierarchical clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method". Journal of Classification. 22 (2): 151–183. doi:10.1007/s00357-005-0012-9. S2CID 206960007.

[6] Dunn, J. C. (1973-01-01). "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". Journal of Cybernetics. 3 (3): 32–57. doi:10.1080/01969727308546046. ISSN 0022-0280.

[7] 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.

[8] 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.

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

[10] Peters, Tim. "listsort.txt". CPython git repository. Retrieved 5 December 2019..

[11] P. Kumar et al. 'Recombinant Sort: N-Dimensional Cartesian Spaced Algorithm Designed from Synergetic Combination of Hashing, Bucket, Counting and Radix Sort' . Ingénierie des Systèmes d'Information. Vol. 25, No. 5, October, 2020, pp. 655-668.

[12] S.Y. Wang "self-indexed sort" ACM SIGPLAN NoticesVolume 31Issue 3 March 1996 pp 28–36 https://doi.org/10.1145/227717.227725

[13] Pal, N.R. and Bezdek, J.C. (1995) On Cluster Validity for the Fuzzy c-Means Model. IEEE Transactions on Fuzzy Systems, 3, 370-379. http://dx.doi.org/10.1109/91.413225.