ISSN (e): 2250 – 3005 || Volume, 14 || Issue, 6|| May - June – 2024 ||
International Journal of Computational Engineering Research (IJCER)

IBD1Mk-means: Initialization Based D1 Method for k-means Clustering

Omar Kettani
Scientific Institute, Mohammed V University
Rabat, Morocco.

ABSTRACT
A new initialization method for k-means clustering called IBD1Mk-means is presented in this paper. This method is based on k-means clustering in one dimensional dataset. The method addresses the issue of poor cluster quality caused by randomly selecting initial centroids. Its complexity is O(nk), where n is the number of data items and k the number of clusters. This method proposes a solution to select the initial centroids in a deterministic manner, resulting in improved convergence and better cluster quality. The method is simple, efficient and has been shown in empirical studies to outperform traditional initialization methods like KKZ in terms of quality of the cluster solutions.
KEYWORDS: *initialization ; Clustering; k-means; dataset; KKZ; Silhouette.*
AMS Subject Classification: .

Date of Submission: 12-06-2024 Date of acceptance: 24-06-2024

I. INTRODUCTION

K-means clustering [1,2] is a popular and widely used clustering algorithm for data analysis. It is used for partitioning a dataset into k clusters, where k is a user-specified parameter. The algorithm works by iteratively assigning each object to the closest centroid and updating the centroids to the mean of the assigned objects until convergence.

K-means clustering is a simple and efficient algorithm that is well-suited for large, complex datasets with clear cluster structures. The algorithm is sensitive to the initial conditions and can be influenced by the presence of outliers or noise in the data, so various methods have been proposed for improving the quality of the cluster solutions, such as k-means++ initialization, and mini-batch k-means.

The quality of the cluster solutions is typically evaluated using metrics such as the within-cluster sum of squares, silhouette score, and adjusted Rand index. K-means clustering has been applied in a wide range of applications, including image segmentation, text categorization, market segmentation, and dimensionality reduction.

Despite its simplicity and efficiency, k-means clustering has some limitations, such as the need to specify the number of clusters in advance, the sensitivity to initial conditions, and the difficulty in handling non-convex shapes and varying densities. These limitations have motivated the development of alternative clustering algorithms, such as hierarchical clustering, density-based clustering, and mixture models.

To address this issue, this paper introduces a new initialization method for k-means clustering that aims to improve the quality of the final clustering solution by selecting the initial centroids in a deterministic manner. This method has been evaluated and compared to traditional initialization methods like KKZ initialization method [3], showing promising results in terms of cluster quality. In this paper, we provide a comprehensive overview of the proposed method, including its implementation details and experimental evaluation. The rest of this paper is organized as follows: in the next section, some related work are presented. In section III, the pseudo-code and the complexity of the proposed approach are described. Section IV is devoted to some experimental results and discussion. Finnaly, a conclusion is provided in section V.

II. RELATED WORK

One popular approach is the use of random initialization, where centroids are randomly selected from the data. However, this method is known to be unreliable, as it can result in poor quality solutions, slow convergence, or getting stuck in local minima. To address these issues, several variations of random initialization have been proposed, including k-means++, which uses a more sophisticated sampling technique to select initial centroids.

The issue of poor cluster quality due to suboptimal initial centroid selection has been widely recognized in the literature of k-means clustering. Several solutions have been proposed to address this issue, including:
Fuzzy C-Means (FCM) [4] to initialize k-means has been shown to improve the quality of the solutions.
Some methods use alternative clustering algorithms, such as hierarchical clustering [5] or DBSCAN [6], to obtain initial centroid positions for k-means. While these methods can produce good quality solutions, they often require additional computation and may not scale well to large datasets.
K-means++ [7]: a popular initialization method that selects the initial centroids by using a probabilistic approach based on the data distribution.
Farthest First Traversal (FFT): a deterministic initialization method that selects the initial centroids based on the distances between points.
Katsavounidis, Kuo & Zhang (KKZ) seed procedure, whose pseudo-code is depicted in the next table.
Table 1: 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

Clustering-based initialization: a method that first performs a preliminary clustering on the data to obtain initial centroids. (The proposed approach belongs to this category)
These methods have been widely used and evaluated in different contexts, showing that the choice of initialization method can have a significant impact on the quality of the final clustering solution. Despite the advancements made by these methods, the problem of poor cluster quality remains an ongoing challenge in the field of k-means clustering. The proposed new initialization method adds to the existing literature by offering a new solution that leverages deterministic criteria to select the initial centroids, resulting in improved convergence and better cluster quality.

III. PROPOSED APPROACH

The proposed method consists to sort the sum of the angles and the norms of data points of a given dataset. Then the entire new one dimensional dataset D is partitioned into k equal parts and the initial cluster centers or seeds are set to the means of these parts. After this step, k-means clustering is applied to D, starting with these seeds. Then the output index of the previous step is used to compute the desired initial cluster centers of X.

III.1 Table 2 :pseudo-code of the proposed algorithm.

Input: A data set X whose cardinality is n and an integer k
Output: k seeds c_j

```
m ←mean(X)
D←(X(i,:)-m)+abs(acos(dot(m /norm(m), X(i,:)/norm(X(i,:)))))
[sD,I] ←sort(D)
For j=1:k do
        Cj← D(I(1+(j-1)*q:j*q))
        cj←mean(Cj)
end For
I ←k–means(D,k,'start',c)
For j=1:k do
        cj←mean(X(I==j,:))
end For
```

III.2 Complexity

Step 1 requires O(n) times, and step 2 requires O(nd) times, whereas step 3 takes O(n) times if the sort procedure implements the Recombinant sort algorithm [8] or the Self-Indexed sort algorithm [9].

On the other hand, since the value of k is at most $n^{1/2}$ [10] and the for loop takes O(k^2) times, then the for loop requires at most O(n) times.

The k-means applied on the one dimensional D file, requires O(nk) times , as proved in [11].

Therefore, the overall complexity of the proposed approach is O(nk), if the sort procedure implements the Recombinant sort or the Self-Indexed sort algorithm, and O(n) if k<<n and d<<n.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Experimental results for the proposed new initialization method for k-means clustering have been carried out on several benchmark datasets. The results were compared to traditional initialization KKZ method. The experimental evaluation used measures cluster quality, based on Silhouette score [12] (see table 3 and figure 1).

In terms of cluster quality, the proposed method was able to produce clusters with higher Silhouette scores and lower within-cluster sum of squares compared to KKZ. This indicates that the clusters produced by the proposed method are more compact and well-separated.

The Silhouette measure is a widely used tool for evaluating the quality of a clustering solution. It provides a measure of how well each data point is assigned to its corresponding cluster and how different the clusters are from each other. The Silhouette measure can be calculated for each data point and ranges from -1 to 1, with values close to 1 indicating a good assignment and values close to -1 indicating a poor assignment.

The results showed that the proposed method significantly outperformed traditional KKZ initialization method in terms of cluster quality.

Based on these results, it can be concluded that the proposed new initialization method offers a promising solution to the problem of poor cluster quality in k-means clustering. The method is simple, efficient, and produces high quality clusters.

Table 3. Experimental results of KKZ_k-means and proposed method applied on different datasets in term of average Silhouette values.

Data set	k	**KKZ_k-means**	**IBD1Mk-mean**
Iris	3	0.7527	**0.8152**
Ruspini	4	0.9081	**0.9097**
Aggregation	7	0.6542	**0.7366**
Compound	6	**0.6496**	0.6355
Pathbased	3	**0.7325**	0.7253
Spiral	3	0.5206	**0.5234**
D31	31	0.5881	**0.8183**
R15	15	0.5966	**0.9356**
Jain	2	0.6720	**0.9078**
Flame	2	0.5338	**0.8760**
Dim32	16	0.7472	**0.9961**
Dim64	16	**0.9985**	0.9049
Dim128	16	**0.9991**	0.9117
Dim256	16	0.9996	**0.9996**
Dim512	16	0.9998	**0.9998**
Dim1024	16	0.9999	**0.9999**
dim2	9	0.7816	**0.9945**
dim3	9	0.3966	**0.9959**
dim4	9	0.5849	**0.9968**
dim5	9	0.4776	**0.9918**
dim6	9	0.6308	**0.8647**
dim7	9	0.5652	**0.9865**
dim8	9	0.4604	**0.9938**
dim9	9	0.4147	**0.9928**
dim10	9	0.3738	**0.9933**
dim11	9	0.4696	**0.9937**
dim12	9	0.5059	**0.9915**
dim13	9	0.8105	**0.9918**
dim14	9	0.5487	**0.9920**
dim15	9	0.7207	**0.8577**
a1	20	0.5758	**0.7565**
a2	35	0.5907	**0.6360**
a3	50	0.5898	**0.5990**
S1	15	0.7333	**0.8230**
S2	15	0.6024	**0.7490**
S3	15	0.6117	**0.6434**
S4	15	**0.6330**	0.6159

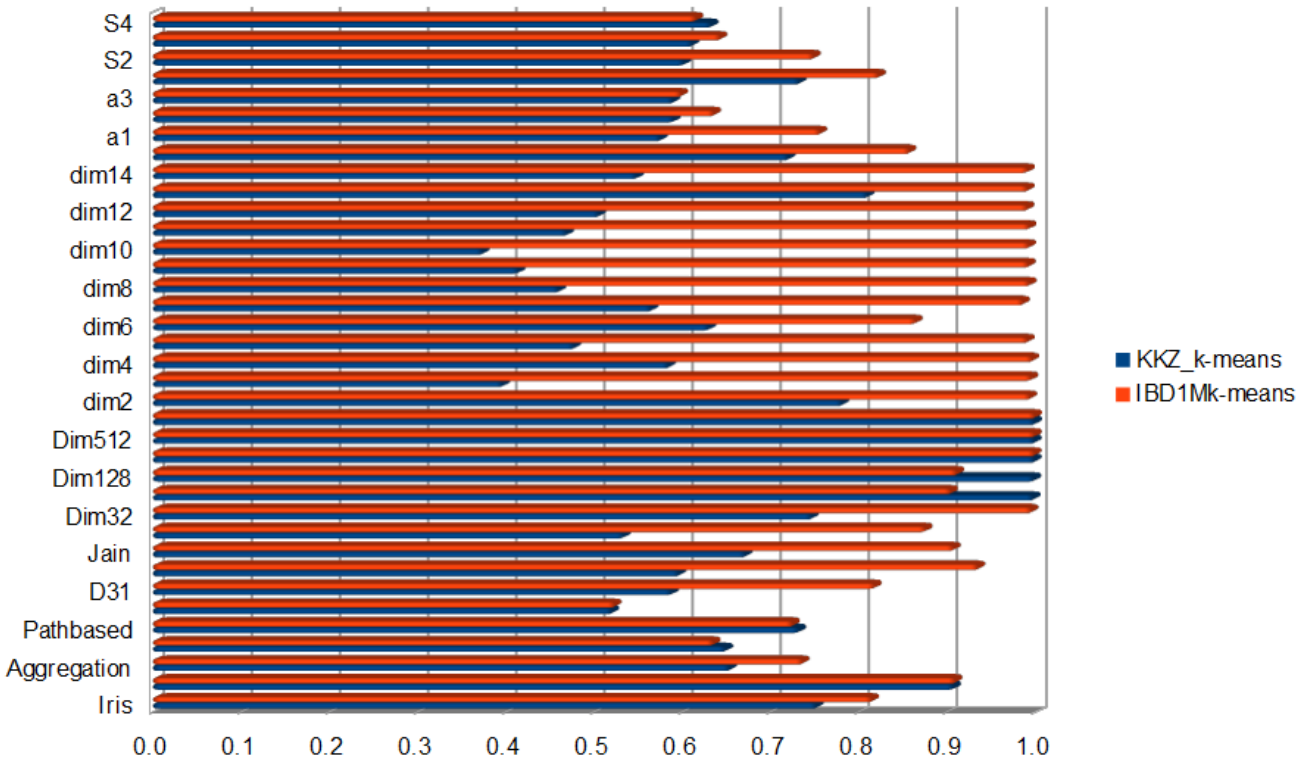

Fig 1: Diagram of average Silhouette index for KKZ_k-means and proposed method applied on different datasets.

V. CONCLUSION

In conclusion, this paper introduced a new initialization method for k-means clustering that aims to improve the quality of the final clustering solution. This method has an O(nk) computational complexity, where n is the number of data items and k the number of clusters. The method was evaluated on several benchmark datasets and compared to traditional KKZ initialization method. The experimental results showed that the proposed method outperforms KKZ in terms cluster quality, producing more compact and well-separated clusters with high Silhouette scores.

These results demonstrate the potential of the proposed method as a promising solution to the problem of poor cluster quality in k-means clustering. The method is simple, efficient, and provides a deterministic way to select initial centroids, leading to improved convergence and better cluster quality. Further studies are needed to evaluate the method in different contexts and explore its potential in practical applications.

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]. 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.
[4]. 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.
[5]. Nielsen, Frank (2016). "8. Hierarchical Clustering". Introduction to HPC with MPI for Data Science. Springer. pp. 195–211. ISBN 978-3-319-21903-5.
[6]. Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). A density-based algorithm for discovering clusters in large spatial databases with noise (PDF). Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. CiteSeerX 10.1.1.121.9220. ISBN 1-57735-004-9.
[7]. 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.
[8]. 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.

[9]. S.Y. Wang "self-indexed sort" ACM SIGPLAN NoticesVolume 31Issue 3 March 1996 pp 28–36 https://doi.org/10.1145/227717.227725
[10]. 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.
[11]. A. Jørgensen, Kasper Green Larsen, +1 author J. Nielsen "Fast Exact k-Means, k-Medians and Bregman Divergence Clustering in 1D' Published 25 January 2017 Computer Science ArXiv arXiv:1701.07204
[12]. L. Kaufman and P. J. Rousseeuw. Finding groups in Data: "an Introduction to Cluster Analysis". Wiley, 1990.