International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

A Deterministic Seeding Approach for k-means Clustering

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

Article Info
Volume 7, Issue 2
Page Number: 192-195
Publication Issue :
March-April-2021
Article History
Accepted : 02April2021
Published :08April2021

ABSTRACT

In this work, a simple and efficient approach is proposed to initialize the k-means clustering algorithm. The complexity of this method is O(nk), where n is the number of data and k the number of clusters. Performance evaluation was done by applying this approach on various benchmark datasets and comparing with the related deterministic KKZ seed algorithm. Experimental results have demonstrated that this approach produces more consistent clustering results in term of average silhouette index.

Keywords: clustering, k-means, initialization, KKZ, silhouette.

I. INTRODUCTION

Cluster analysis is the most widely used technique in Data Mining. Clustering consists of grouping a given dataset into a predefined number of disjoint sets, called clusters, so that the elements in the same cluster are more comparable to each other and more different from the elements in the other cluster. This optimization problem is known to be NP-hard, even when the clustering process deals with only two clusters [1]. Therefore, many heuristics have been proposed, in order to find near optimal clustering solution in reasonable computational time. The most prominent clustering method k-means is a greedy algorithm which has two stages: Initialization, in which we set the seed set of centroids, and an iterative stage, called Lloyd's algorithm [6]. Additionally, Lloyd's algorithm has two steps: The assignment step, in which each object is assigned to its closest centroid, and the centroid's update step. The time required for the assignment step is O(nkd), while the centroid's update step and the computation of the error function is O(nd). The main advantage of k-means is its fast convergence to a local minimum. A major drawback of k-means is its sensitivity to the initial clustering centers (namely, seed). To achieve a better initialization, many techniques have been proposed. In this study, yet another k-means initialization technique is proposed.

In the next section, some related works are briefly discussed. Then the proposed method and its computational complexity are described in Section 3. Section 4 consists to apply this approach to some standard data sets and reports its performance. Finally, conclusion of the paper is summarized in Section 5.

II. RELATED WORK

Several initialization methods have been proposed in the literatures. Among them, Katsavounidis et al. [4]

utilize the sorted pairwise distances for initialization which has been termed as the KKZ algorithm. This algorithm chooses the vector with maximal norm as the first seed, then For j = 2, . . . , k, each centroid m_j is set in the following way: For any remaining data x i , its distance di to the existing centroids is computed. d $_i$ is calculated as the distance between xi to its closest existing centroid. Then, the point with the largest d $_i$ is selected as m_j. The computational complexity of KKZ is O(nk).

Another initialization method that is based on simple probabilistic seeding procedures. In particular, the k-means++ method, proposed by Arthur and Vassilvitskii in [3], consists of randomly selecting only the first centroid from the dataset. The greedy k-means++ method probabilistically selects k centers in each round and then greedily selects the center that most reduces the SSE. It chooses the first center randomly and the i-th (I ∈ {2, 3, . . . , k}) center is chosen to be x' ∈X with a probability of

$$md(x')^2/\sum_{j=1}^{n} md(xj)^2 \qquad (1)$$

where md(x) denotes the minimum-distance from a point x to the previously selected centers. The drawback of this approach is referred to its sequential nature, as well as to the fact that it requires k scans of the entire dataset, therefore it has a complexity of O(nkd).

III. PROPOSED APPROACH

Pseudo-code:

```
INPUT: a dataset X whose cardinality is n and an integer k
OUTPUT: k seeds $c_i$
c=mean(X)
for i=1 to k do
      j=ArgMax(d($X_h$,c))
          1<=h<=n
    c=c+ $X_j$
    output(c)
end;
```

Complexity:

Clearly, the complexity of this method is O(nk) .

IV. EXPERIMENTAL RESULTS

Algorithm validation is conducted using different data sets from the UCI Machine Learning Repository [2] . We evaluated its performance by applying on several benchmark datasets and compare with KKZ_ k-means. In preprocessing step, the data were normalized.

Silhouette index [5] which measures the cohesion based on the distance between all the points in the same cluster and the separation based on the nearest neighbor distance, was used in these experiments in order to evaluate clustering accuracy.

The proposed approach was compared with a related deterministic clustering method: KKZ_k-means (k-means initialized by KKZ).

Experimental results are reported in table 1 and figure 1.

Table 1. Experimental results of KKZ_k-means and proposed approach on different datasets in term of average Silhouette value.

Data set	**k**	**KKZ_k-means**	**Proposed**
Iris	3	0.7542	**0.8152**
Ruspini	4	0.9086	**0.9097**
Aggregation	7	0.6536	**0.7410**
Compound	6	0.6484	**0.7464**
Pathbased	3	**0.7316**	0.7253
Spiral	3	**0.5206**	0.4953
D31	31	0.5881	**0.7629**
R15	15	0.5966	**0.7475**

Jain	2	0.6719	**0.9078**
Flame	2	0.5347	**0.8760**
s1	15	**0.7333**	0.7262
s2	15	0.6024	**0.7375**
s3	15	0.6117	**0.6253**
s4	15	**0.6330**	0.6303
t4.8k	8	0.5841	**0.6789**
dim2	9	0.7816	**0.7896**
dim3	9	0.3966	**0.6190**
dim4	9	0.5849	**0.8716**
dim5	9	0.4490	**0.7565**
dim6	9	0.6308	**0.7346**
dim7	9	0.5652	**0.8858**
dim8	9	0.4604	**0.6553**
dim9	9	0.3778	**0.4865**
dim10	9	0.3738	**0.6424**
dim11	9	**0.5092**	0.4209
dim12	9	0.4329	**0.6813**
dim13	9	0.6241	**0.6637**
dim14	9	**0.6909**	0.4973
dim15	9	**0.5527**	0.3829
a1	20	0.5542	**0.7048**
a2	35	0.5970	**0.6811**
a3	50	0.5752	**0.6917**

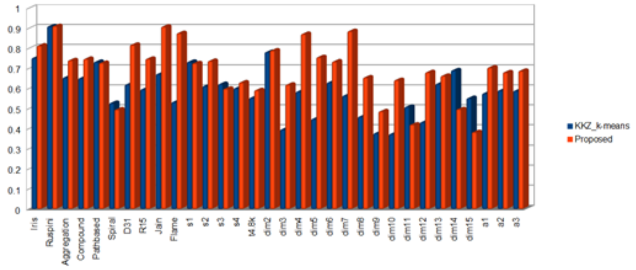

Fig 1: Chart of kkz_kmeans and proposed approach on different datasets in term of average Silhouette value

V. CONCLUSION

In this study, a deterministic initialization method for the k-means algorithm was suggested. Its time complexity is 0(nk) and experimental results have demonstrated that the proposed approach produces quickly consistent clustering results in term of average silhouette index.Future work will consist to consider a possible improvement of this method by avoiding dataset outliers, in order to obtain more accurate clustering results.

VI. REFERENCES

[1]. 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.

[2]. Arthur D., Vassilvitskii S.: k-means++: the advantages of careful seeding. In: Proceedings of the 18th annual ACM-SIAM Symp. on Disc. Alg, pp. 1027 – 1035 (2007).

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

[4]. Katsavounidis I., Jay Kuo C. C., and Zhang Z. 1994 . A new initialization technique for generalized lloyd iteration. IEEE Signal Processing Letters, vol. 1, pp. 144–146, Oct. 1994.

[5]. Kaufman L. and Rousseeuw P. , 2005 Finding groups in data: an introduction to cluster analysis. Wiley.

[6]. Lloyd., S. P. (1982). "Least squares quantization in PCM". IEEE Transactions on Information Theory 28 (2): 129–137. doi:10.1109/TIT.1982.1056489.

Cite this article as :

Omar Kettani, "A Deterministic Seeding Approach for k-means Clustering", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 2, pp. 192-195, March-April 2021. Available at doi : **https://doi.org/10.32628/CSEIT217246**