An Efficient K-Means++ Algorithm using Nearest- Neighbor Search

Syed Ahtesham Ullah[1] Praveen Shende[2]
[1]M.Tech Scholar [2]Assistant professor
[1,2]Department of Computer Science and Engineering
[1,2]Chhatrapati Shivaji Institute of Technology Durg, Chhattisgarh

Abstract— K-means algorithm is most popular partition based algorithm that is widely used in data clustering. A lot of algorithms have been proposed for data clustering using K-means algorithm due to its simplicity, efficiency and ease convergence. In spite this K-means algorithm has some drawbacks like it scales poorly computationally, initial cluster centers is supplied by the user and stuck in local optima etc. Determining the number of clusters is very complex and is usually done by an expert. Thus, this paper intends to overcome this problem by proposing a parameter-free algorithm for automatic clustering. It is based on successive adequate restarting of K-means algorithm based on Nearest Neighbor search. This proposed approach is more effective and gives a tough competition to the well-known algorithms, X-means and G-means in terms of clustering accuracy and estimation of the correct number of clusters.

Key words: K-Means++, G-Means, X-means, automatic clustering

I. INTRODUCTION

One of the most popular techniques for data analysis is data clustering, also known as unsupervised learning. According to Jain's definition, "The goal of data clustering, also known as cluster analysis, is to discover the natural grouping(s) of a set of patterns, points, or objects."[1]. Informally, we can define clustering as grouping of unlabeled objects into a set of groups, given a similarity metric. Data clustering has many applications. In Computer Vision, Image Segmentation can be defined as a clustering problem [2]. In Information Retrieval, document clustering is a very important method that can provide hierarchical retrieval and improvements in flat retrieval performance [3]. In Bioinformatics, clustering is used for improving multiple sequence alignment [4]. Many other applications also exist in other fields like: Online Shopping, Medicine, Online Social Networks, Recommender Systems, and etc. A valid clustering should have two characteristics:

1) Cohesion: the objects in one cluster should be as similar to each other as possible, and
2) Separation: clusters should be well separated i.e. the distance among the cluster centers must be large enough.

Many different approaches have been proposed for clustering problem, such as Multi Variant Analysis, Graph Theory, Expectation Maximization, and Evolutionary Computing. Amongst available clustering algorithms, maybe the most well-known one is the K-means algorithm. Although different clustering algorithms have shown good performance while applied to specific problems, but k-means has proven to be efficient and fast if applied to various domains [1]. Despite of simplicity and effectiveness of k-means, it has several disadvantages, too. The quality of clustering highly depends on the initial seeds. Choosing bad seeds can result very bad clusters. Another challenge of the k-means algorithm is the requirement of giving the number of clusters as an input parameter. However, determining the correct number of clusters is very complex and usually needs an expert. Some mechanisms have been proposed for automatic selection of number of clusters, like the X-means algorithm [6]. However, they are not accurate enough and usually result inaccurate cluster numbers.

An enhanced K-means Algorithm based on Nearest Neighbor Search [9] is proposed by Omar Kettani, Benaissa Tadili, and Faycal Ramdani which overcomes the drawback of accuracy and performance by using nearest Neighbor Search for choosing the initial cluster instead of initial clusters which are randomly chosen.

In this paper, an alternative parameter free method for automatic clustering for K- means is proposed. It is based on successive adequate restarting of K-means based on Nearest Neighbor search [9]. Algorithm validation and comparative study with X-means [6] and G-means [5], a related well known algorithm, are conducted using several real-worlds and artificial clustering data sets from the UCI Machine Learning Repository.

II. LITERATURE REVIEW

Despite the actual fact that getting an optimum range of clusters k for a given knowledge set is an NP-hard downside [6], several methods are developed to seek out k mechanically Pelleg and Moore [7] introduced the X-means algorithm, which proceed by learning k with k-means using the Bayesian Information Criterion (BIC) to score each model, and choose the model with the highest BIC score. However, this method tends to over fit when it deals with data that arise from non-spherical clusters. Tibshirani et al. [7] proposed the Gap statistic, which compares the likelihood of a learned model with the distribution of the likelihood of models trained on data drawn from a null distribution. This method is suitable for finding a small number of clusters, but has difficulty when k increases. Hamerly and Elkan [3] proposed the G-means algorithm, based on K-means algorithm, which uses projection and a statistical test for the hypothesis that the data in a cluster come from a Gaussian distribution. Fast Approximate K-means via Cluster Closures [12] uses cluster closure and active points for the fast approximation of K-means. Scalable K-MEANS ++ [13] uses Lloyd's algorithm for iteration and defines the reduction in the number of passes needed the obtain massive data .A Deterministic K-means Algorithm based on Nearest Neighbor Search [9] uses simple deterministic method based on Nearest Neighbor Search as a preprocessing step is used to overcome the drawback of accuracy and performance due to the initial choice of cluster center which are randomly generated. A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm [14] improves the performance of K-means by generating initial cluster centers with the help of binary search method and after that K-means algorithm is applied.

III. METHODOLOGY

The methodology starts with taking the data set as input; the data sets are taken from the UCI Machine Learning Repository. These data sets include Iris, Wine and Glass. Now the second step is enhancing the K-means++ algorithm, after that the clusters are obtained as output and at the last the obtained output clusters are compared with the output of the G-means algorithm.

This enhanced K-means++ algorithm finds the correct number of clusters in the dataset using a deterministic K-means algorithm based on Nearest Neighbor search[8].The algorithm starts by initializing k= mod(square root(n)) , where n is the number of objects in the given data set. This can be taken by the fact that the number lies in the range from 2 to square root of n, as reported by Pal and Bezdek [9]. Now K-means is applied with these initial k centroids, and centroid of the smallest cluster is removed, then K-means restarts with the remaining centroids.

In each iteration, the maximum of the cluster validity index [10] of the current position is stored. We use this index because it is relatively inexpensive to compute, and it generally outperforms other cluster validity indices as reported by Milligan and Cooper in [11]. This process is repeated until k=2. Finally, the algorithm outputs the optimal k and partition corresponding to the maximum value of CH stored.

Algorithm

Input: Dataset D ={ x_1 , x_2 ,…. x_n }
Output: An integer k as the number of clusters
Step1. Initialize k=mod (sqrt (n)), X← D
Step2. For j=1 to k
Do
C_j←KNNsearch(x1, X, mod (n/k))
c_j←Σx_i/mod (n/k)
X-C_j
End for
Step3. Applying K-means (D, c, k).
Step4. Storing the values of k and cluster index.
k_o← k
CH_o ←CH (I)
Step5. Removing the smallest cluster and calculating the highest cluster index
While k>2
Do
j←argMin (modC_i)
i<=k
cj = []
k←k-1
Applying K-means (D, c, k)
If CH_o<CH (I) then
k_o←k
CH_o←CH (I)
End if
End While
Step6. Optimal value of k is stored in k_o .

IV. RESULT AND CONCLUSION

We use two measures to evaluate enhanced K-means++ algorithm. The first measure is accuracy which defined as the average ratio of the number of cluster determined by the algorithm to the number of actual clusters. The second criterion is speed, which is the average amount of time that an algorithm requires to solve a clustering problem regardless of the accuracy.

Algorithm validation is conducted using real-world clustering data sets, namely iris, wine, glass and cmc data sets from the UCI Machine Learning Repository. This approach can obtain the correct number of clusters in almost all tested data sets.

Calinski-Harabasz (CH), this index obtained the best results in the work of Milligan and Cooper. It is a ratio-type index where the cohesion is estimated based on the distances from the points in a cluster to its centroid is used in the algorithm for finding the correct number of the clusters

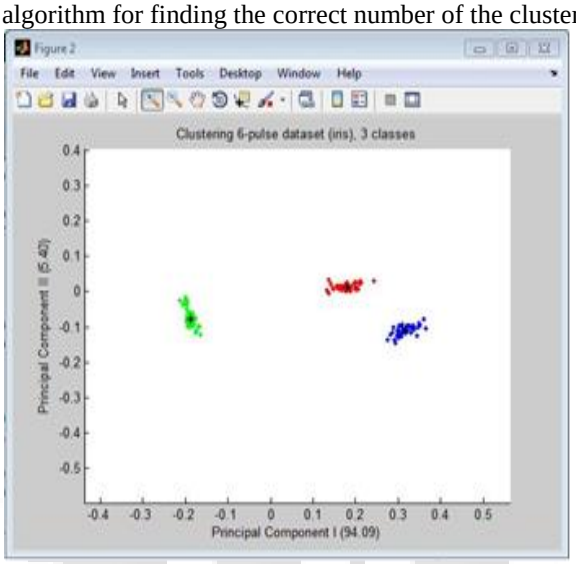

Fig. 1: Figure shows the principal component analysis of the iris data set having 3 clusters with their centroid

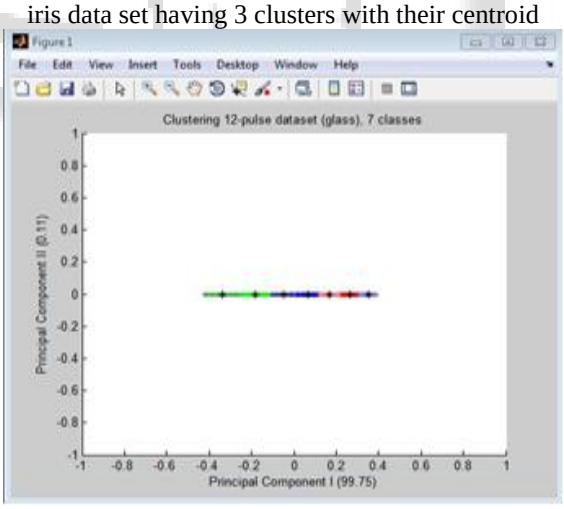

Fig. 2: Figure shows the principal component analysis of the glass data set having 7 clusters with their centroid

This method was compared with the related well known algorithm, X-means and G-means, which is improved for finding the correct number of clusters. The comparisons also showed that the proposed approach is better than G-means in terms of clustering accuracy.

Data sets	Iris(150 instances)	Glass(214 instances)	Wine(178 instances)	Cmc(1473 instances)

Actual value of K	3	7	3	2
X-means (K)	17	85	71	384
Time taken by X-mean	0.767 sec.	3.953 sec.	4.825 sec.	66.779 sec.
G-mean (K)	4	9	5	1
Time taken by G-mean	0.023 sec.	0.127 sec.	0.091 sec.	0.0093 sec.
Enhanced K-mean++ (K)	3	7	4	3
Time taken by Enhanced K-mean++	0.393 sec.	1.284 sec.	0.731 sec.	4.786 sec.

Table 1: Shows the comparison between X-means, G-means And Enhanced K-means++ in terms of actual number of cluster (K) And the response time for calculating the actual number of cluster (K).

The above comparison table shows that the enhanced K-means++ algorithm determines the accurate number of clusters of the datasets in minimum response time as compared to the X-means where the number of cluster is very large and in G-means the response time is less but the number of cluster determined is inaccurate.

The enhanced K-means++ algorithm is the best algorithm as compared to the X-means and the G-means algorithm

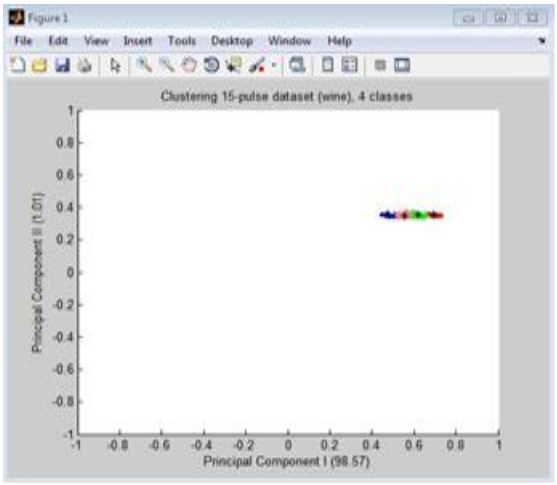

Fig. 3: Figure shows the principal component analysis of the wine data set having 4 clusters with their centroid

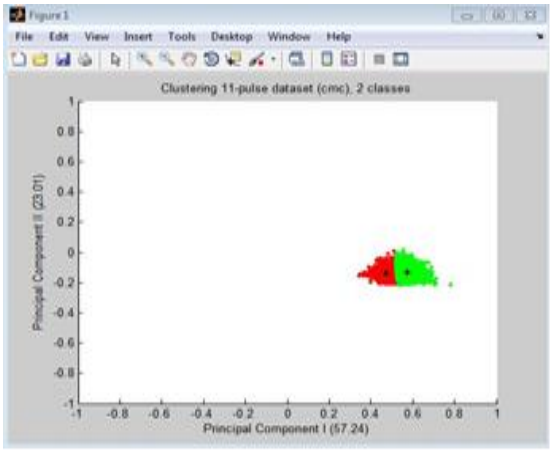

Fig. 4: Figure shows the principal component analysis of the Cmc data set having 2 clusters with their centroid.

V. FUTURE WORK

The algorithm finds the correct number of clusters in the dataset using a deterministic K-means algorithm based on Nearest Neighbor search.

In future work, it will be of interest to find a tighter upper bound on the number of clusters, instead of square root of n, in order to reduce the number of computation's steps of the proposed approach. Another is to fit the proposed algorithm into the Map Reduce programming model.

VI. REFERENCES

[1] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit.Lett., vol. 31, no. 8, pp. 651–666, Jun. 2010.
[2] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," Comput. Vs. Image Underst. vol. 110, no. 2, pp. 260–280, 2008.
[3] R. Baeza-Yates, B. Ribeiro-Neto, and others, Modern information retrieval, vol. 82. Addison-Wesley New York, 1999.
[4] D. J. Miller, Y. Wang, and G. Kesidis, "Emergent unsupervised clustering paradigms with potential application to bioinformatics.," Front. Biosci. vol. 13, pp. 677–690, 2008.
[5] H. Spath, Clustering Analysis Algorithms for Data Reduction and Classification of Objects, Ellis Harwood, Chichester, 1980.
[6] Dan Pelleg and Andrew Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In Proceedings of the 17th International Conf. on Machine Learning, pages 727–734. Morgan Kaufmann, 2000.
[7] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a dataset via the Gap statistic. Journal of the Royal Statistical Society B, 63:411–423, 2001. http://dx.doi.org/10.1111/1467-9868.00293.
[8] 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.
[9] Kettani, O.; Tadili, B. and Ramdani, F. - A deterministic k-means algorithm based on nearest neighbor search.

International Journal of Computer Applications (0975 – 8887), Vol. 63, No.15, February 2013. http://dx.doi.org/10.5120/10544-5541.

[10] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. Communications in Statistics, 3:1–27, 1974.

[11] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. Psychometrica, 50:159– 179, 1985.

[12] Jing Wang, Peking University, Jingdong Wang Microsoft, Qifa Ke Microsoft ,Gang Zeng, Peking University Fast Approximate K-means via Cluster Closures .

[13] Behman Behmani, Stanford University Stanford CA, Ravi Kumar- Yahoo! Research Sunnyvale, CA, Benjamin Moseley, University of Illinois, Urbana IL, Andrea Vattani, University of California San Diego, CA Scalable K-MEANS ++ .

[14] Omar Kettani, Benaissa Tadili, Faycal Ramdini LPG Lab, Scientific Institute Mohammed V University, Rabat A Deterministic K-means Algorithm based on Nearest Neighbor Search.

[15] Yugal Kumar, G.Sahoo Department of Information Technology, Birla Institute of Technology, Mesra, India A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm.

[16] Aggarwal, C. C., & Reddy, C. K. An introduction to cluster analysis. In C. C. Aggarwal, & C. K. Reddy, Data Clustering Algorithms and Applications (pp. 1-22). CRC Press

