# Initial Seeds Selection for K-means Clustering Algorithm Using Elbow and Heron Mean Method

Ms. P. J. Patel[1],Krupa A. Patel[2], Mr. M. B. Chaudhari[3]
[1] *Assistant Professor, Department of CSE, G.E.C., Gandhinagar, India*
[2]*M.E Student, Department of CSE, G.E.C., Gandhinagar, India*
[3]*Professor, Department of CSE, G.E.C., Gandhinagar, India*

***Abstract-*** Cluster analysis is decades' old concept of data mining which performs division of data into groups of similar objects. It is used in various applications in the real world such as data/text mining, voice mining, image processing, web mining, medical data mining and many others. Overlapping K-means is improved K-means to find overlapping clusters which also inherits its advantages and disadvantages. Therefore, to improve the performance of OKM and to enhance its scope in other areas more improvements are required. In this proposed work optimum K value is generated using Elbow method and through proposed heron means method the position of initial centroids is calculated. This approach not only ensures that outlier data points doesn't get randomly selected as cluster centers but also helps in reducing the number of iterations the algorithm needs to make to effectively allocate clusters. This proposed method may increase the accuracy of the algorithm for medical data analysis.

***Keyword-*** Overlapping K-Means, Elbow Method, Heron Mean, FBCubed Metric

## I. INTRODUCTION

Clustering analysis is used to discover the natural grouping(s) of a set of points, or data objects by using statistical classification technique. It clusters the datasets into different groups by using quantitative comparisons of different attributes. Clustering should ensure that groups are homogenous within and heterogeneous outside.

Clustering methods can be categorized according to the following criteria [5]:

- Type of input data: To deal with different types of input data such as numerical, categorical and mixed, different clustering methods are used.
- Type of proximity measures: Different types of similarity measures are defined to deal with different type of input data, some of them are Euclidean distance, Manhattan distance etc.
- Type of generated cluster: In this category two types of clustering methods are defined one is Exclusive (Non-Overlapping) another one is Overlapping.
- Types of membership function: in terms of hard(crisp) of soft(fuzzy) clustering method.
- Type of clustering strategy used: In term of cluster strategy, clustering methods are divided into five groups like Partitioning, Hierarchical, Density-based, Model – based, Grid-based.

The K-Means algorithm is quite popular partitioning method for clustering, where the user inputs the number of clusters (k) in which he/she desires to partition the data. This divides the dataset of n objects into k clusters so that there is high homogeneity within each cluster and high heterogeneity among different clusters. The similarity of clusters is measured by mean value of the objects in a cluster, which can also be viewed as the cluster's centroid or center of gravity.

Since K-Means has some limitations as below[4]:
1) K-means algorithm assumes that value of *k(number of clusters)* is known in advance which is notnecessarily true in real-world applications.
2) The K-means algorithm is sensitive to initial centres selection.
3) K-means algorithm may converge to local minima.
This challenges makes it one of the wide research area.

There are some solutions and improvements have been developed so far. In paper [4] they have used Silhouette method to find number of cluster k and then apply it to simple K-means method and used SVM method for classification of result, in paper [5] combines K-Harmonic mean method with OKM and result of KHM has used as initial centroid value and apply it to OKM method for clustering, in paper [12]the Sorted K-Means which determines initial centroids after sorting the data points. Sorting procedure based on the type of data like merge sort or quick sort has been used.

In this paper we will overcome the limitation of K-Means method, the remainder of the paper is organized as follows. Section 2 provides a background of clustering algorithms. In Section 3, a full description of the proposed integrated overlapping clustering algorithm is presented. Section 4 includes the details of the experimental results, and the paper is concluded in Section 5.

## II. BACKGROUND KNOWLEDGE

**K-means Algorithm**

The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster , the aim of K-means is to minimize the objective function or the square-error criterion, defined as:

$$E = \sum_{j=1}^{k} \sum_{x_i \in \pi_j} \|x_i - z_j\|^2$$

Where E is the sum of the square error for all objects in the data set; $x_i$ is the point in space representing a given object; and $z_j$ is the mean of cluster $\pi_j$ (both $x_i$ and $z_j$ are multidimensional).

**Input:**

k: the number of clusters,

D: a data set containing n objects.

**Output:** A set of k clusters.

**Method:**

1. arbitrarily choose k objects from D as the initial cluster centers;
2. repeat
3. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. update the cluster means, i.e., calculate the mean value of the objects for each cluster;
5. until no change;

## III. PROPOSED METHOD

In this paper, an integrated OKM method is proposed for improving the overall performance of the algorithm. The integrated OKM algorithm is applied for dimensionality reduction to remove outliers and noisy data. The proposed system consists of following steps:

**Elbow Method:** The idea of the elbow method is to run k-means clustering on the dataset for a range of values of *k* (say, *k* from 1 to 10), and for each value of *k* calculate the sum of squared errors (SSE) and plot a line chart. Line chart looks like "elbow" on the arm is the best value of k. Choose small of k that still has a low SSE[3].

**Heron Mean Method:** Heron mean is one of the families that interpolate between arithmetic mean and geometric mean[2]. It is defined as a convex combination of the arithmetic and geometric mean. Heron mean family $HeM_\alpha(a,b)$is defined as[2]

$$HeM_\alpha(a,b) = (1-\alpha)GM(a,b) + \alpha AM(a,b)$$

Or

$$HeM_\alpha(a,b) = (1-\alpha)\sqrt{ab} + \alpha\frac{a+b}{2}$$

Where $0 \ll \alpha \ll 1$. This family is the linear interpolate between the geometric and the arithmetic mean. Perhaps because of its naivete, it has received less attention than other families of interpolating means. This is a convex function, its minimum value is $\alpha(1/2) = 0$, and its maximum value is $\alpha(0) = \alpha(1) = 1$.

The quantity below is called Heronian mean or Heron mean [2]

$$HeM_{\frac{2}{3}}(a,b) = \frac{1}{3}(a + \sqrt{ab} + b)$$

**Input**: Dataset D containing n objects

**Output**: A set of k clusters

**Algorithm steps:**

1. Preprocessing of data including cleaning and normalization
2. Find appropriate number of clusters (k) by using Elbow method.
3. Proposed algorithm to find the centroids of clusters using Heron Mean (HeM) method
4. Initialize KM method to find final overlapping clusters
5. Analyze result

Fig.1: Proposed method steps

## IV. RESULT ANALYSIS

Proposed model has been applied on Lung cancer, Diabetes and Dermatology datasets collected from UCI repository. Same dataset has been applied on simple KM and compared its result with proposed integrated method. Below table shows comparisons with different evaluation matrics.

| User | K | Initial Centroid Position | Precision | Recall | F-measure | Rand Index | Overlap | BCubed Precision | # Iterations |
|---|---|---|---|---|---|---|---|---|---|
| 1_R | 2 | NA | 0.143 | 0.904 | 0.247 | 0.244 | 1.475 | 0.107 | 11 |
| 1_U | 2 | 56,789 | 0.141 | 0.911 | 0.243 | 0.223 | 1.456 | 0.114 | 8 |
| 2_R | 3 | NA | 0.159 | 0.802 | 0.266 | 0.338 | 1.695 | 0.104 | 7 |
| 2_U | 3 | 45,578, 899 | 0.142 | 0.798 | 0.242 | 0.253 | 1.786 | 0.116 | 9 |
| 3_R | 4 | NA | 0.169 | 0.851 | 0.282 | 0.294 | 2.084 | 0.1 | 9 |
| 3_U | 4 | 23,456, 678,890 | 0.169 | 0.801 | 0.28 | 0.326 | 1.755 | 0.113 | 10 |
| 4_R | 5 | NA | 0.188 | 0.861 | 0.301 | 0.319 | 2.333 | 0.096 | 10 |
| 4_U | 5 | 27,105, 480,733, 940 | 0.183 | 0.738 | 0.293 | 0.372 | 2.085 | 0.144 | 11 |
| Average | | | 0.162 | 0.833 | 0.27 | 0.296 | 1.833 | 0.112 | 10 |
| Proposed OKM algorithm | 3 | 174,461, 961 | 0.193 | 0.71 | 0.304 | 0.514 | 1.281 | 0.185 | 5 |

Table 1 comparison of different results for lung cancer dataset

| User | K | Initial Centroid Position | Precision | Recall | F-measure | Rand Index | Overlap | BCubed Precision | # Iterations |
|---|---|---|---|---|---|---|---|---|---|
| 1_R | 2 | NA | 0.118 | 0.88 | 0.208 | 0.185 | 1.259 | 0.11 | 10 |
| 1_U | 2 | 24,375 | 0.118 | 0.879 | 0.208 | 0.185 | 1.258 | 0.11 | 18 |
| 2_R | 3 | NA | 0.102 | 0.636 | 0.175 | 0.274 | 1.34 | 0.11 | 18 |
| 2_U | 3 | 39,284, 641 | 0.115 | 0.828 | 0.203 | 0.208 | 1.704 | 0.089 | 20 |
| 3_R | 4 | NA | 0.112 | 0.637 | 0.191 | 0.144 | 1.668 | 0.11 | 19 |
| 3_U | 4 | 28,186, 379,760 | 0.107 | 0.65 | 0.183 | 0.205 | 1.703 | 0.099 | 17 |
| 4_R | 5 | NA | 0.115 | 0.607 | 0.193 | 0.183 | 1.783 | 0.102 | 20 |
| 4_U | 5 | 18,164, 452,530, | 0.112 | 0.608 | 0.189 | 0.265 | 1.896 | 0.111 | 20 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 701 | | | | | | | |
| Average | | | 0.112 | 0.715 | 0.193 | 0.218 | 1.576 | 0.105 | 17 |
| Proposed OKM algorithm | 3 | 266,748, 364 | 0.101 | 0.626 | 0.174 | 0.279 | 1.318 | 0.111 | 7 |

Table 2 comparison of different results for diabetes dataset

| User | K | Initial Centroid Position | Precision | Recall | F-measure | Rand Index | Overlap | BCubed Precision | # Iterations |
|---|---|---|---|---|---|---|---|---|---|
| 1_R | 2 | NA | 0.0 | 1.0 | 0.0 | 0.218 | 1.341 | 0.0 | 9 |
| 1_U | 2 | 35,175 | 0.0 | 1.0 | 0.0 | 0.218 | 1.341 | 0.0 | 12 |
| 2_R | 3 | NA | 0.061 | 0.765 | 0.114 | 0.333 | 1.626 | 0.051 | 12 |
| 2_U | 3 | 1,112, 287 | 0.058 | 0.728 | 0.108 | 0.326 | 1.626 | 0.049 | 20 |
| 3_R | 4 | NA | 0.064 | 0.669 | 0.116 | 0.469 | 1.598 | 0.054 | 9 |
| 3_U | 4 | 18.98, 157,236 | 0.06 | 0.696 | 0.109 | 0.456 | 1.612 | 0.054 | 18 |
| Average | | | 0.0405 | 0.784 | 0.074 | 0.337 | 1.524 | 0.035 | 13 |
| Proposed OKM algorithm | 3 | 45,117, 348 | 0.058 | 0.658 | 0.107 | 0.385 | 1.48 | 0.055 | 7 |

Table 3 comparison of different results for dermatology dataset

| Ratio of between_SS/total_SS | | | |
|---|---|---|---|
| | Lung Cancer | Diabetes | Dermatology |
| K=3_ Elbow Method | 88.60% | 68.00% | 86.80% |
| K=2_Silhoutte Method | 74.80% | 57.40% | 73.00% |

Table 4 Comparison of different k value

## V. CONCLUSION AND FUTURE SCOPE

Initial seeds selection is proposed in this paper to improve the performance of the existing algorithm, instead of selecting random value of cluster k and centroid of that cluster using Elbow method we will find optimal value of k and using Heron Mean method the sensitivity of outlier data point selection as centroid has been removed. Result analyzed in RStudio and shows that the proposed method improves the accuracy of than simple K-means method. In Future work we will analyze result for another medical dataset.

## VI. REFERENCES


[1]. Jiewei Han MichelineKamber, data mining Concepts and techniques.

[2]. R.Bhatia , "Interpolating the arithmetic-geometric mean inequality and its operator version" 2005 Elsevier

[3]. T. M. Kodinariya,Dr. P. R. Makwana "Review on determining number of Cluster in K-Means Clustering", IJARCSMS 2013

[4]. Sandeep Kaur and Dr. SheetalKalra, "Disease Prediction using Hybrid K-means and Support Vector Machine" 2016 IEEE

[5]. SinaKhanmohammadi, NaiierAdibeig, SamanehShanehbandy, "An Improved Overlapping k-Means Clustering Method for Medical Applications", Expert Systems With Applications 2016 Elsevier

[6]. ArgenisAroche, José Francisco Martínez-Trinidad, José Arturo Olvera-López, Airel Perez-Suarez "Study of Overlapping Clustering Algorithms Based on Kmeans through FBcubed Metric", Springer 2014.

[7]. ManishaGoyal, Mr. M.B. Chaudhary, Ms. Pinal Patel, A Survey: Different Improvements and Integrated Approaches of K-means. IJIRT 2018

[8]. S. Baadel, F. Thabtah, and J. Lu. Multi-Cluster Overlapping K-means Extension Algorithm. In proceedings of the XIII International Conference on Machine Learning and Computing, ICMLC'2015. 2015.

[9]. TanawatLimungkura, PeeraponVateekul. Enhance Accuracy of Partition-based Overlapping Clustering by Exploiting Benefit of Distances between Clusters, 2016 International Conference on Knowledge and Systems Engineering

[10]. Said Baadel, FadiThabtah, Joan Lu. Overlapping Clustering: A Review, SAI Computing Conference 2016

[11]. Syed Zishan Ali , Nikhil Tiwari and SushmitaSen "A Novel method for clustering using K-means and apriorialgorithm",IEEE 2016

[12]. Preeti Arora, DeepaliVirmani, Himanshu Jindal and Mritunjaya Sharma, "Sorted K-Means Towards the Enhancement of K-Means to Form Stable Clusters", Proceedings of International Conference on Communication and Networks, Springer 2017