IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# IMPLEMENTATION OF HEAT DIFFUSION FOR DATA CLUSTER APPROXIMATION FOR FAST SEARCH AND DENSITY PEAK FINDING

Dashrath P.D, B.M. Patil
Computer science and information technology,
M.B.E.S C.O.E.A, Ambejogai , India

***Abstract:*** In the era of knowledge, it is important to look at education; huge data is getting accumulated for educational purposes. It is the need of time to look at the data figure out patterns and check if the system needs any improvisation. To see useful trends in data, data will have to be clustered. To achieve this, various clustering techniques are used all over the world. This paper presents a comparative study of two such techniques; k-mean and Clustering by fast search and finding density peaks - heat diffusion (CFSFDP-HD). K-Mean, which has been used for a long time; is a robust algorithm and useful when the data is small. This is an iterative algorithm, by changing centroids in each iteration, this will give us a proper and optimum number of clusters. CFSFDP-HD algorithm follows a different and adaptive approach to cluster data where centroids are selected based on the density of data around them. We observed CFSFDP-HD be more efficient and robust as without too many iterations it gives proper clusters.

*Index Terms* - **Data Mining, Clustering, K-mean, Clustering by fast search, and finding density peaks - heat diffusion.**

## I. INTRODUCTION

K-mean is one of the clustering algorithms. K-mean is easy to understand. K-mean is iterative. K-mean selects centroids randomly. Based on the centroids selected whole data set is calculated by using the Euclidian distance formula. K-mean is a very popular clustering technique used for partitioning data objects [1] According to the result of first centroid selection Centroids are recalculated. Concerning the recalculate centroids once again whole data is calculated the process continues till the optimal cluster is formed. CFSFDP-HD is one of the clustering algorithms. It uses density, distance, decision graph for the clustering. With the help of the density and decision graph is plotted and we can find out the centroids. Density peaks assume that the cluster center is the highest dense point [2]. With the help of centroids by calculating distance and density it is possible to form a cluster. Input data is in the distance matrix.

CFSFDP-HD uses distance and density as the key features. The decision graph is also one of the important parts of the CFSFDP-HD. K-mean needs the centroid recalculations. But the CFSFDP-HD calculates the centroids only once the decision graph is used for the centroid calculations. The process of CFSFDP-HD is the clustering algorithm invented after the K-mean clustering algorithm. As Invented after K-mean it tries to overcome some of the drawbacks observed in k-mean. Due to adaptive nature CFSFDP-HD seems different than that of the iterative natured k-mean.

These clustering algorithms are needed in priority by various fields in software for mining data. Data mining comes into picture due to the increase in the use of software in a very huge amount. Due to the gathering of data, we have to consider the patterns which can be extracted from the data we have. K-mean thus comes in use by many people for the extraction of data patterns. But gradually Different clustering algorithms are invented. So as we are having different algorithms we will be in a position to figure out which one is better and why. Performance maintains CFSFDP-HD is different than that of the K-mean clustering algorithm. The result of both the algorithm is cluster formation. But how both works are different than the other. By using Both the algorithms for cluster formation it is possible to know the actual performance of both clustering algorithms.

K-mean is a simple, easy technique for data clustering. Cluster makes a group of similar data. K-mean is a technique used in very initial days for data mining. There is a need for random data centroid selection in the k-mean clustering algorithm. K-mean is easy to understand but seems difficult to maintain. K in K-mean is the number of clusters. And the K-mean thus is the integration of clusters and their means. The key feature in K-mean is the number of clusters and the mean. Clustering by fast search and finding density peaks - heat diffusion (CFSFDP-HD) uses a different algorithm than that of K-mean. Though both of them are used for the clustering purpose.

CFSFDP-HD is adaptive . K -mean is an iterative one. Centroids comparison is Frequently required in K mean. CFSFDP-HD centroids are needed to be selected only once. One time centroid selection is the advantage of CFSFDP-HD. Multiple times centroid comparison is the drawback of K mean. CFSFDP-HD overcomes the drawback of K -mean.

## II. LITERATURE REVIEW

For the distance calculation Euclidean distance is one of the methods. Euclidean distance method is easy to understand and implement. It represents the shortest distance between two points. Once we have the shortest distance then we can easily and accurately determine which data point lies in which cluster. The Euclidean distance thus plays a vital role in cluster formation. With the help of Euclidean distance, we can easily measure the distance of a particular point from the centroid selected. K-mean can form the clusters of the spherical shape. Clusters with other shapes are not possible by using the k-mean clustering algorithm. Which will have a problem when we need the clustering of different shapes [5].

For data to be portioned we can use K-mean, CFSFDP-HD, and many more algorithms. K-mean is iterative. CFSFDP-HD is adaptive. Due to the adaptive nature using the decision graph the centroid selection becomes easy than that of the k-mean. Distance and density help in data maintenance and cluster formation too [6]. The decision graph is used for the cluster formation in CFSFDP-HD. Cluster selection is possible in CFSFDP-HD with minimum human interference [7]. CFSFDP-HD uses a decision graph with the use of which centroid recalculation is avoided. Using the decision graph for the centroid selection makes the CFSFDP-HD more robust than that of K-mean [8]. Arbitrary shaped clusters can be formed with the help of CFSFDP-HD [9].

candidates of cluster centers from the non-center data [10]. K-mean gradually grows and then able to put data in according to cluster [11]. it is necessary to manually select cluster centers through the decision graph and the selection of these centers is usually subjective [12]. the density of the high-class center is always surrounded by the neighbor points of low density and has a relatively large distance with the point which has a higher density than it [13]. clustering focuses on revealing underlying patterns embedded in data [14]. The CFSFDP-HD algorithm has the advantages of low computational complexity and high accuracy [15].K-mean clustering is used for the cluster analysis. K-mean classifies the given dataset through a certain number of clusters. K-mean makes the inner points of the cluster. The K-means is a state-of-the-art partition-based clustering Algorithm [6]. K mean is an unsupervised clustering technique. Data is partitioned based on the similarities between them. K-mean is easy to understand. and simple to implement.

Clustering by fast search and finding density peak is the nonparametric method for deciding the groups of datasets. Clusters are formed without iteration. The cluster center is a highly dense region. Non-parametric means data we are analyzing does not require to meet certain conditions. Non-parametric methods are applied to different types of data. The decision graph will help in deciding the centroids which will further help in cluster centers. Clustering groups objects in groups in such a way that they are more similar to each other. The density-based clustering algorithm is called Density Peaks. Cluster centers have a higher density than there neighbors.

K-mean partitions observations into clusters. Density and distance are used for cluster formation in (CFSFDP-HD). Due to higher technological use, a huge amount of data is generated. Generated data can be used for various purposes. For that Framing is needed on that data. The needed framing is accomplished using data mining techniques. Clustering is one of the data mining techniques. Due to cluster, We have data which is gathered according to similarities between them. Data according to one cluster set might be similar to other clusters. Once we can identify the hidden patterns from the collected data we can easily use them whenever needed.

The CFSFDP has characteristics to discover intrinsic hidden signals of interest from ambiguous data [6]. For data to be partitioned various algorithms can be used. K-mean is one of them. K-mean is iterative. So many iterations have to be performed to get the desired result. Instead of that (CFSFDP-HD) is adaptive. No iterations are needed for data to be partitioned. Data partitioning is done by considering two factors one is density and the other is distance. Density peak is the highest Dense point assumed as the center of the cluster. [6]. The minimum distance of the rest of the points from the cluster center which is calculated using the Euclidean distance formula is considered for a particular point for inside or outside the cluster.

K-mean was a robust and vary much widely used clustering technique. Due to K-means iterative nature (CFSFDP-HD) comes into the picture. (CFSFDP-HD) is a clustering approach that can be easily used for the clustering of data as it considers the density and distance for measuring is seems quite easy and robust nowadays. (CFSFDP-HD) is quite efficient as compared with the K-means [6]. The data mining approaches can further be improved to generate knowledge and provide intelligent assistance to the students. users can analyze data without any prior domain knowledge. In CFSFDP - HD, an adaptive method was used to estimate the density of the underlying dataset [6].

Less human interaction makes CFSFDP-HD more effective. In cluster analysis, the key challenge is to discover correct cluster centers in the datasets. The CFSFDP via heat diffusion (CFSFDP - HD) was proposed as a variant of CFSFDP, where limitations of CFSFDP are improved and users can analyze data without any prior domain knowledge [6].

## III. RESEARCH METHODOLOGY

### 3.1 K-mean Flowchart :

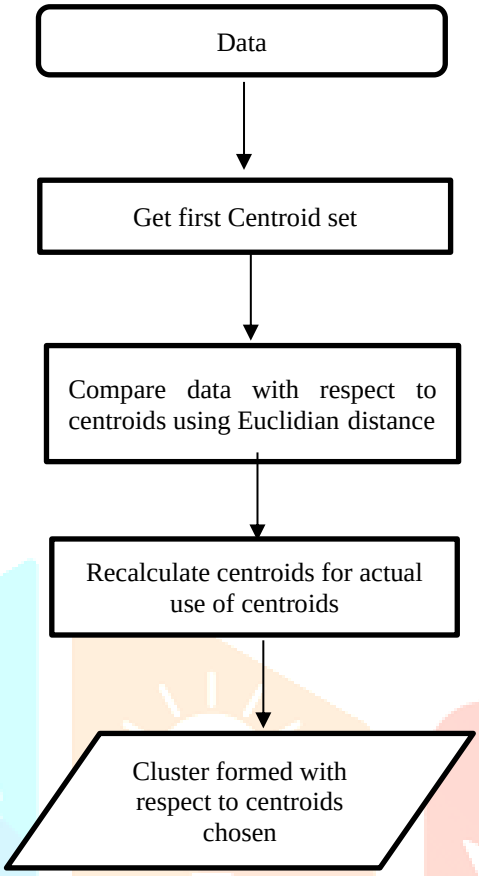


Fig 3.1: K-mean flowchart

### 3.2 K -Mean Algorithm

1. Pick random points called Centroids.
2. Distance calculation concerning centroids using Euclidian distance.
3. Recalculate the Centroids for finding actual centroids.
4. Reposition the centroids till we get the final cluster.

Centroids are randomly chosen from the dataset we have. Then the whole dataset is compared concerning the centroids chosen and Euclidian distance is used for the distance calculation. Then based on the distances we got again we recalculate the centroids and then the whole dataset is calculated concerning the new centroids we got. The process of centroids recalculation was repeated several times until we got the optimized clusters. This is the whole process called K-mean clustering algorithms. K-mean clustering algorithm gives us the centroids in the spherical shapes only.

K-mean calculates the centroids several times. The First time centroids are chosen randomly. Cluster formation with the help of chosen centroids needs the repositioning of centroids. Frequent centroids recalculation and repositioning is the task of K-mean clustering algorithm. K mean is simple to understand.

### 3.3 CFSFDP-HD

1. A distance matrix is given as an input.
2. All the data points are plotted on a decision graph based on density and distance from the data points.
3. Users will be able to choose the centroids from the decision graph.
4. Clusters are formed according to the centroids chosen..

### 3.4 CFSFDP-HD

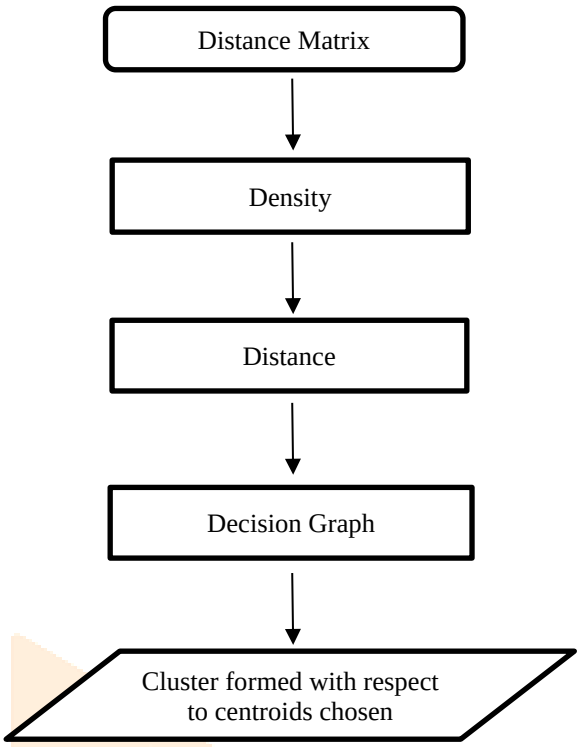


Fig 3.2: CFSFDP-HD flowchart

Density $\rho_i$ is calculated with the help of equation

$$\hat{f}(x,t) = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=-\infty}^{\infty} e^{-k^2\pi^2\frac{t}{2}}\cos(k\pi x)\cos\left(k\Pi x_j\right) \quad (3.1)$$

Where x represents the data-points and preparatory probability is distributed through the data-points {x1, x2, x3, . . . , xn}. The method evolves for a time t. The function Of in Equation (3.1) can be expressed as below

$$X(x) = \begin{cases} 1 & x<0 \\ 0 & \text{otherwise} \end{cases} \quad (3.2)$$

Equation (3.2) can be used for distance $\partial_i$ calculation With the help of equation (3.1) and (3.2) we can calculate the density and distance both then we will be in a position to chose centroids for the particular clusters. After choosing centroids. Clustering by using CFSFDP-HD uses the pairwise Distance matrix for the input. The distance matrix maintains a pairwise distance. Density and distance are calculated. Plotting the density and distance on the graph we will be able to figure out the centroids. The decision graph gives us the correct cluster centroids. Due to the decision graph centroids selected without recalculations. K-mean requires centroid recalculation. Clustering by fast search and finding density peaks - heat diffusion (CFSFDP-HD) selects the centroids once's and then proceed for further process. This is the biggest advantage of CFSFDP-HD over the K-mean clustering algorithm.

K-mean selects centroids for the first time randomly there is no fixed process for that. But the CFSFDP-HD has the decision graph for the centroid selection. Once we select the centroids in CFSFDP-HD using the decision graph we need not calculate them frequently. One fixed process for centroid selection avoids the centroid recalculation. The pairwise distance matrix is used for the data set input which maintains the pairwise distance between dataset. For using the decision graph we have to calculate density, distance, and also their plotting is needed.

CFSFDP-HD forms arbitrary clusters that are not possible with the help of K-mean. K-mean can form the clusters in spherical shapes only. No other type of cluster can form with the help of the k-mean.

Only spherical cluster formation is one of the drawbacks of K-mean. CFSFDP-HD can form an arbitrary cluster which is useful for the dataset. The cluster formation process is easy and more helpful than that of the K-mean clustering algorithm. Once a cluster is formed it can be used for any purpose.

Centroid selection, arbitrary Cluster formation process of CFSFDP-HD makes it more robust than that of the K-mean. K-mean is suitable for the small dataset as it needs frequent recalculations of centroids and only spherical cluster formation is possible with the help of K-mean.

CFSFDP-HD is suitable for large datasets too due to arbitrary cluster formation and decision graph. CFSFDP-HD overcomes some of the drawbacks of K-mean. K-mean was invented before CFSFDP-HD. Though K-mean is also the clustering algorithm just like the CFSFDP-HD decision graph and arbitrary cluster becomes more powerful.

As the decision graph is used for the cluster centroid selection decision graph is a key feature of CFSFDP-HD. Centroid recalculation is the key feature of the K-mean clustering algorithm. CFSFDP-HD forms the cluster in two steps first is a decision graph and the second is the cluster formation. CFSFDP-HD when compared with the K-mean clustering algorithm sounds better than that of the K-mean due to the process and cluster shapes too. Structured data clustering is very easy and powerful too using CFSFDP-HD rather than that the K-mean clustering algorithm. CFSFDP-HD works in a more organized manner than that of the K-mean clustering algorithm. CFSFDP-HD can be implemented in industry for the meaningful data extraction from the available dataset and cluster formation.

## IV. RESULT

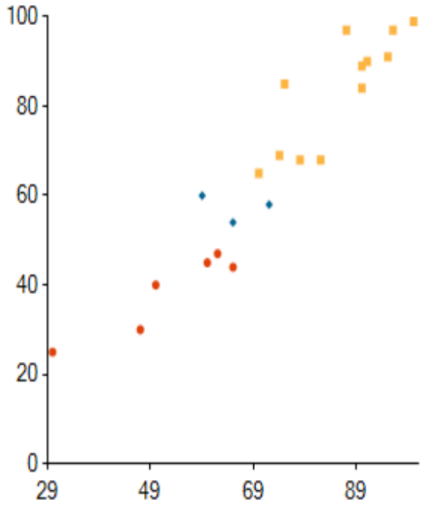


Fig 4.1 The cluster formed using K-mean

According to data points, we are having k clusters in an iterative nature. And the clusters we obtain from the k-mean algorithm are spherical shapes. Cluster formation depends on the initial values of the centroids we have selected.

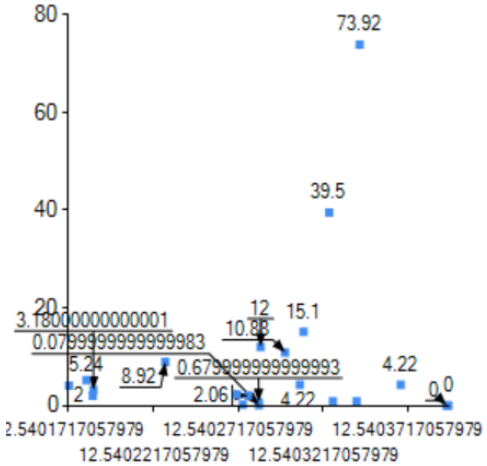


Fig 4.2 Decision graph using CFSFDP-HD

A decision graph is used to select the centroids. Upper points are considered as center points of expected clusters. The decision graph helps with centroids selection. And afterword clusters will be formed according to data points we are having. Centroids are chosen once. No need to recalculate the centroids, unlike k-mean. Frequent recalculation of centroids is not required.

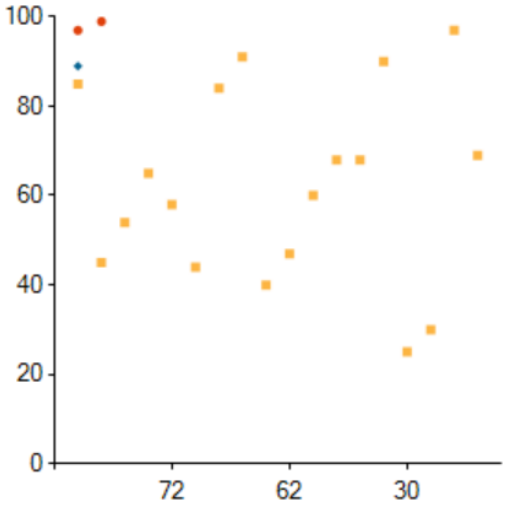


Fig 4.3 Result of CFSFDP-HD

Finally, Clusters are formed according to the data points we are having. Clusters formed with the help of the CFSFDP-HD give the arbitrary in shapes. Centroids needed to be calculated only once so it's easy.

## V. CONCLUSION AND FUTURE SCOPE

Clustering by fast search and finding density peaks - Heat Diffusion is more robust than K-means . K-means to obtain meaningful clusters users are required to repeat the clustering process multiple times. Clustering by fast search and finding density peaks via Heat Diffusion is less iterative than K-mean.

K-mean can form spherical clusters. CFSFDP-HD can form arbitrary clusters. K-mean does not have a fixed way of finding the cluster centroids for the first time. CFSFDP-HD overcomes it by using a decision graph for calculating the cluster Centroids. K-mean Sounds better with a small data set. For large and complex data sets Clustering by fast search and finding density peaks via Heat Diffusion sounds better.

In future work proposed system shall use cosine similarity instead of Euclidian distance for distance calculation. Even if two similar documents apart by Euclidian distance by its size they could still have a similar angle between them. Which will helps in mining data and forming clusters too.

## REFERENCES

**[1]** Performance Analysis of Student Learning Metric using K-mean Clustering Approach K-Mean Cluster 2016 IEEE.
**[2]** An Adaptive Clustering Algorithm Based on CFSFDP Feng Yang, Jinming Cao, Kuang Zhou, Pengyan Zhang,2018 IEEE.
**[3]** Efficient and Privacy-Preserving k-means clustering For Big Data Mining Zakaria Gheit.
**[4]** Improvement in the Accuracy of the Moving Object Position by Eliminating Erroneous Sensors with K-Means Clustering Approach 2020 IEEE.
**[5]** Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means.2020 IEEE.
**[6]** In the Personalized E-Learning System SAMINA KAUSAR1,2, XU HUAHU1, IFTIKHAR HUSSAIN 3,4, of Computer Engineering and Science.
**[7]** Adaptive fuzzy clustering by fast search and find of density peaks, Rongfang Bie1, Rashid Mehmood1, Shanshan Ruan1, 2016.
**[8]** GDCLU: a new Grid-Density based clustering algorithm, Gholamreza Esfandani.
**[9]** Clustering by fast search and finding density peaks DOI: 10.1126/science.1242072Science 344, 1492 (2014); Alex Rodriguez and Alessandro Laio.
**[10]** Evaluating the Density Parameter in Density Peak Based Clustering, Jian Hou, Weixue Liu, 2016 IEEE.
**[11]** Vegetable Disease Detection Using K-Means Clustering And Svm 2020 IEEE.
**[12]** Research on the Optimized Clustering Method Based on CFSFDP, Longlong Sun, Ming Jiang2019 IEEE.
**[13]** A New Measurement Partition for Extended Target Tracking Based on CFSFDP Algorithm, ChiLuo-Jia, Feng Xin-xi 2017 IEEE.
**[14]** Incremental CFS Clustering on Large Data Liang Zhao_y, Zhikui Chen_, Yi Yangyzx.
**[15]** Sparse learning based on clustering by fast search and find of density peaks Pengqing Li1Xuelian Deng1,2.