ISSN : 3107 - 4308

EMERGING TRENDS IN DIGITAL TRANSFORMATION

Published By :
D3 Publishers

VOLUME-1 ISSUE-1,
INAUGURAL EDITION

ETDT

ETDT

Emerging Trends in Digital Transformation

Volume-1, Issue-1, April 2025 ISSN: **3107-4308** *Paper ID:* ***ETDT-V1-I1-02***

GENERALIZED K-MEANS CLUSTERING WITH CENTROID ENHANCEMENT: A STEP TOWARD ADAPTIVE DATA SEGMENTATION

[#1]K.Ramya,
B.Tech Student,
Department of CSE,

#2A.Anusha,
B.Tech Student,
Department of CSE,

#3Dr. SK Yakoob,
Associate Professor& HOD,
Department of CSE,

[#1,2,3]sai Spurthi Institute Of Technology-Autonomus - Sathupally.TG.
Corressponding author Mail: ramyak2002@gmail.com

ABSTRACT: One important unsupervised learning technique in data mining is the K-means clustering method. This method efficiently organizes big datasets by splitting objects into k separate clusters. It is possible to provide clearer data classification by ensuring that items in the same cluster are more similar than things in other clusters. The first step in creating clusters is to pick data points at random, ensuring that each one has an equal number of items. Improving K-means clustering's ability to handle a wide variety of data types and guarantee fair weight distribution, this research presents a new method for choosing the best cluster from uniform and non-uniform datasets.

Keywords:*Clustering, Initial Centroids, k-means Algorithm.*

1. INTRODUCTION

Clustering is essential to our data-driven operations as it enables the categorization of objects with analogous attributes and facilitates data exploration techniques to enhance processing efficiency. Clustering is primarily utilized in unsupervised learning, but it may also be applied in supervised methods to categorize entities according to a predetermined similarity metric. It is extensively utilized across various fields for identifying distributions and patterns in datasets, including statistics, computer science, psychology, economics, engineering, and medicine. The primary objective of clustering is to identify densely and sparsely inhabited regions within a dataset. This can be achieved primarily through two methods: exclusive clustering, which employs fuzzy sets to assign data points to a single cluster (similar to K-means), and overlapping clustering, which utilizes fuzzy sets to distribute data points across many clusters. Clustering, a fundamental data mining technique, provides valuable insights that improve decision-making and data management across several fields.

2. RELATEDWORKS

The limitations of k-means clustering, such as its reliance on arbitrary beginning centers that could reduce accuracy, are discussed in this work along with its extensive examination of its uses. To improve the algorithm's performance, it shows different ways to choose beginning nodes. Researchers compare the original k-means method with its updated variants in order to enhance centroid selection and attain a local optimum in clustering results. They do this by evaluating a varied array of datasets. Complex grouping patterns and high-dimensional

DOI : https://doi.org/10.5281/zenodo.16900105

datasets are among the complicated circumstances tackled in the work. In order to make the k-means algorithm more efficient and effective, especially when dealing with big and varied datasets, it suggests using new approaches.

Basickmeansclustering

Centroids are used as reference points by the famous k-means clustering technique, which uses partitions to divide data into k groups. The random initialization of these centroids, however, causes the clusters to vary greatly and frequently leads to convergence to a local optimum. Improving clustering outcomes for uniform and non-uniform datasets is the goal of this work, which presents a new method for picking initial centroids. The method assigns data points to the nearest cluster center iteratively based on computed distances (e.g., Manhattan or Euclidean distances) until the termination criteria are satisfied. By conceptualizing clusters as spheres or centers of gravity, the k-means method efficiently divides an n-item collection into k clusters. But it only works if the centroids are chosen exactly right to guarantee global optimality. By iteratively recalculating centroids, the suggested method improves the procedure and achieves clustering convergence. This happens when there is no longer any movement of data points across clusters and centroids have stabilized. This improved method makes the k-means algorithm more reliable and effective, especially when working with different types of datasets.

An algorithm can be defined as a set of rules or a systematic approach for solving a problem. The k-means clustering algorithm is a widely used unsupervised machine learning method. It creates k separate clusters from a dataset. There are n objects in the input dataset D.

$D=\{d1, d2, \ldots\ldots\ldots\ldots, dn\}$

K= The number of desired clusters Output:

A set of k clusters containing data from dataset D. Method:

- The clustering process begins by picking k items at random from dataset D to serve as initial centroids.
- Depending on how similar an object is to the centroid of dataset D, we will place it in the cluster that holds that centroid.
- Find the average placement of all items assigned to each cluster or the new mean for each cluster. When the cluster centroids' coordinates stop changing, we say that the convergence condition is satisfied, and we must continue with steps 1 and 2 until then.
- When processing massive datasets, the k-means algorithm provides a straightforward and easy-to-implement option.
- Because of its ease of use and robust features, the k-means approach has become standard practice for handling massive datasets. However, it must be noted that the methodology has a number of major flaws:

The method's O(nkl) temporal complexity—where n is the total number of elements in the dataset, k is the number of clusters required, and l is the number of repetitions—makes it computationally expensive. There is a strong correlation between the precision of the first centroid determination and the subsequent clusters' accuracy. So, even when using the same input data again, the results could be different each time.

DOI : https://doi.org/10.5281/zenodo.16900105

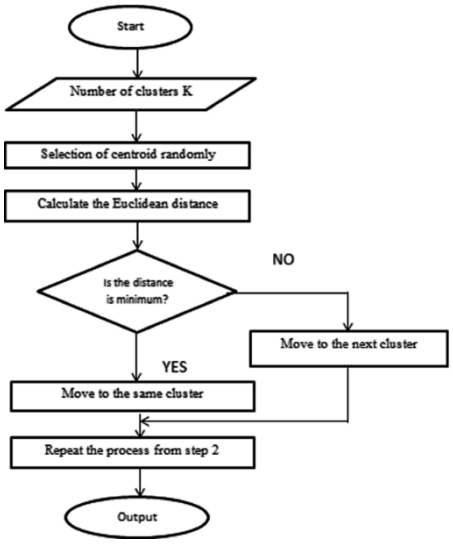


Fig.1:The following process flow diagram shows the main steps of the k-means clustering approach.

3. PROPOSED ALGORITHM FOR IMPROVEMENT OF INITIALCENTROIDS

This research lays forth a thorough method for finding the first centroid. In order to get reliable and consistent results from several algorithm iterations on the same dataset, a strict technique is used to generate centroids. After that, a two-dimensional coordinate system is used to plot the given data points. There ought to be some good quality in every data point. Features that have negative values need to be transformed into their positive corresponding ones. To achieve this, take the supplied dataset and remove the attribute with the lowest value from each data point. Because the investigated approach needs to find the distance from the origin to each data point, this adjustment is crucial. When unaltered, it is possible for data points to reach comparable distances from the origin as measured by Euclidean distance. The choice of the initial centroids may be erroneous as a result.

Here is the algorithm:

There are n distinct pieces of data in the X dataset.

$$X=\{x1, x2, \ldots\ldots\ldots\ldots, xn\}$$

The quantity of clusters that were searched is represented by K. The initial centroids, k, are what make up the output.

- At the outset, we must determine how far each data point (d) is from the starting point.
- Organizing the distances that were accumulated in the previous phase is crucial. Using these distances, the initial dataset is sorted.

It is recommended to divide the sorted data components into k equal pieces, and the user has previous expertise with academic writing.

The following is the formula for calculating the distance from the origin to each data point using the Euclidean distance metric:

The origin of the coordinate system is at (0,0). Each set of coordinates (x, y) represents a single piece of data.

DOI : https://doi.org/10.5281/zenodo.16900105

Using this formula, we can find the distance in geometric terms between points O and P:

$$\sqrt{\left((x-0)^2+(y-0)^2\right)}$$

After that, the previously stated distances are shown in either a descending or an ascending order. According to the identified distances, the initial data points show a similar pattern of sorting. Sorting this data compilation into k equal parts is the next step. The next step is to calculate the average for each subdivision. Averaging the values of each partition is used to construct the first centroids. Centroids produced using this technique can be used with many different types of datasets. The data points show two possible outcomes when the values are biased toward one border rather than being distributed uniformly across the split.

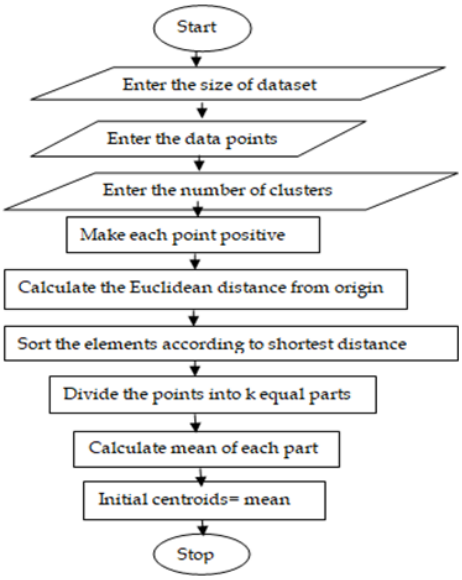


Fig2:The suggested k-means implementation's process flow diagram is shown here.

IMPLEMENTATION AND RESULT:

Table1:Diabetes-related data are of particular importance.

preg	plas	pres	skin	insu	mass	pedi	age
6	148	72	35	0	33.6	0.627	50
1	85	66	29	0	26.6	0.351	31
8	183	64	0	0	23.3	0.672	32
1	89	66	23	94	28.1	0.167	21
0	137	40	35	168	43.1	2.288	33
5	116	74	0	0	25.6	0.201	30
3	78	50	32	88	31	0.248	26
10	115	0	0	0	35.3	0.134	29
2	197	70	45	543	30.5	0.158	53
8	125	96	0	0	0	0.232	54
4	110	92	0	0	37.6	0.191	30
10	168	74	0	0	38	0.537	34
10	139	80	0	0	27.1	1.441	57
1	189	60	23	846	30.1	0.398	59
5	166	72	19	175	25.8	0.587	51
7	100	0	0	0	30	0.484	32

DOI : https://doi.org/10.5281/zenodo.16900105

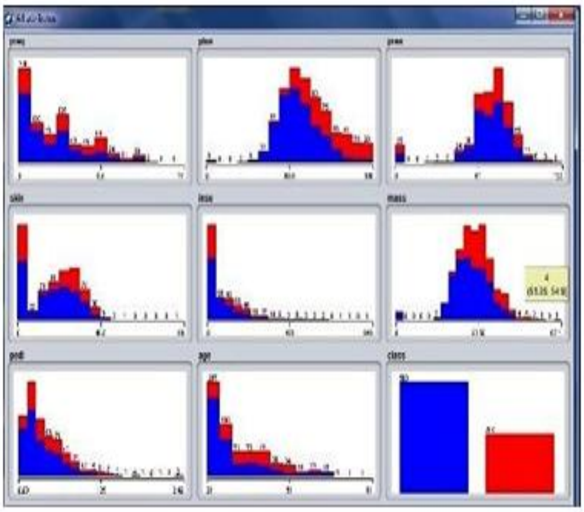


Fig3:There are positive and negative results for every attribute in the test.

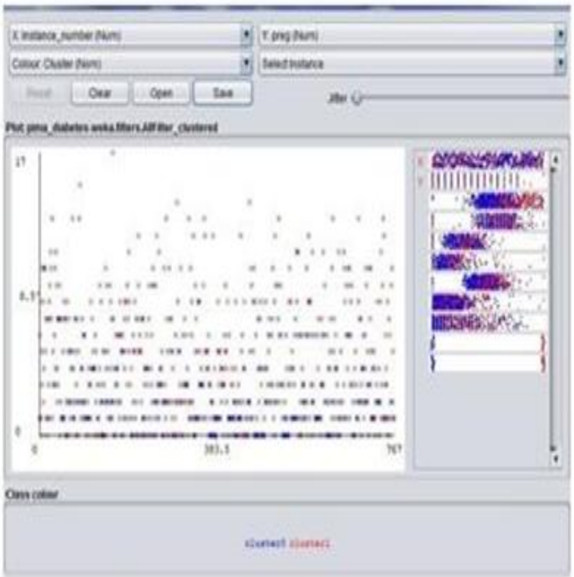
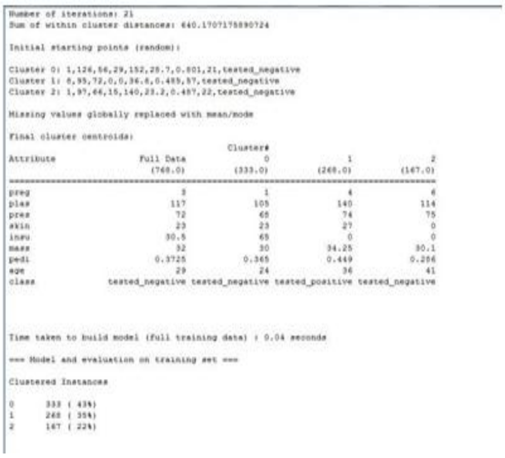


Fig4:The objective of this research is to create a graphical depiction of the clustering results, which are the k-means algorithm's categorization of sets of data components into separate clusters.



```
Number of iterations: 21
Sum of within cluster distances: 640.1707175890724

Initial starting points (random):

Cluster 0: 1,126,56,29,152,28.7,0.801,21,tested_negative
Cluster 1: 6,95,72,0,0,36.8,0.485,57,tested_negative
Cluster 2: 1,97,66,15,140,23.2,0.487,22,tested_negative

Missing values globally replaced with mean/mode

Final cluster centroids:
                                         Cluster#
Attribute              Full Data                0                1                2
                         (768.0)          (333.0)          (268.0)          (167.0)
===================================================================================
preg                           3                1                4                6
plas                         117              105              140              114
pres                          72               65               74               75
skin                          23               23               27                0
insu                        30.5               65                0                0
mass                          32               30            34.25             30.1
pedi                      0.3725            0.365            0.449            0.256
age                           29               24               36               41
class            tested_negative  tested_negative  tested_positive  tested_negative


Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       333 ( 43%)
1       268 ( 35%)
2       167 ( 22%)
```

Fig5:Clusters are created via the clustering process.

DOI : https://doi.org/10.5281/zenodo.16900105

4. CONCLUSION

The research's approach to finding the k-means clustering algorithm's initial centroid placements is easier to apply and more efficient. In addition to working with a wide variety of datasets, our suggested strategy solves the problem of duplicate findings. Problems with both homogeneous and heterogeneous distributions of data points are successfully addressed by the suggested methods. This method significantly shortens the time required to reach the convergence threshold compared to the traditional k-means algorithm, provided that the starting centroids are chosen correctly. Only numerical data can be processed using the k-means method. On the other hand, values of numerical and categorical data are ubiquitous in real life. It is possible that this method might make the k-means algorithm work better with mixed-type data.

REFERENCES:

[1]. Chen Zhang and Shixiong Xia, K-means Clustering Algorithmwith Improved Initial center, in Second International WorkshoponKnowledgeDiscoveryandData Mining(WKDD),pp. 790-792,2009.

[2]. F. Yuan, Z. H. Meng, H. X. Zhangz, C. R. Dong, A New Algo-rithmtoGettheInitialCentroids,proceedingsofthe3rdInterna-

[3]. tionalConference on Machine Learning and Cybernetics, pp. 26-29, August2004.

[4]. S. Deelers and S. Auwatanamongkol, Enhancing K-means Algo-rithm with Initial Cluster Centers Derived from Data Partitioningalong the DataAxiswiththe HighestVariance,InternationalJournalofComputerScience,Vol. 2, Number4.

[5]. Huang, Extensions to the k-meanss Algorithms for ClusteringLarge Data Sets with Categorical Values, Data Ming and Knowl-edgeDiscovery,vol. 2,no.3,pp.283-304,1998.

[6]. S.A. Rauf, S. mahfooz, S. Khusro, H. Javed, enchanted k-meansclustering algorithm to reduce number of iterations and time com-plexity, Middle–EastJ. Sci.Res.12(7),959-963(2012)

[7]. C.S. Li, Cluster center initialization method for K-means algo-rithm over dataset with two clusters, in Proceeding of interna-tionalconferenceon AdvancesinEngineering,pp.324-328,2011

[8]. A.K.JainandR.C.Dubes,AlgorithmsforClusteringData.EnglewoodCliffs,NJ:Prentice-Hall,1988.

[9]. S. Z. Selim and M. A. Ismail, K-means type algorithms: a gener-alized convergence theorem and characterization of local optimal-ity, in IEEE Transaction on Pattern Analysis and Machine Intelli-gence, vol.6,No.1,pp.81--87,1984.

[10]. Jieming Zhou, J.G. and X. Chen, An Enhancement of K-meansClustering Algorithm, in Business Intelligence and Financial En-gineering,BIFE'09.InternationalConferenceon,Beijing,2009.

[11]. M.P.S Bhatia, Deepika Khurana, Analysis of Initial Centers fork-meanss Clustering Algorithm , International Journal of Com-puterApplications(0975–8887)Volume71–No.5,May2013

DOI : https://doi.org/10.5281/zenodo.16900105