International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Dec 2025 www.irjet.net p-ISSN: 2395-0072

# Robust K-Means Clustering: A Unified Framework for Outlier Removal and Adaptive Distance Metrics

**Md. Mayn Uddin**

*Dept. of Electrical and Electronic Engineering, Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh-2224 , Bangladesh*

---

**Abstract -** *Outliers and inappropriate distance metrics remain major challenges in the successful improvement of the accuracy of K-Means clustering. Outliers distort centroid estimation, while conventional distance measurement methods often fail to capture true similarity among data points. This paper proposes a unified modification to the K-Means algorithm that combines systematic outlier detection and elimination before centroid calculation, with a novel adaptive distance metric for cluster assignment. By reducing the impression of anomalous data and refining similarity measurement, the modified algorithm achieves faster convergence and significantly higher clustering accuracy. An exploratory evaluation on nine benchmark multivariate datasets demonstrates up to 81% improvement in clustering performance compared to traditional K-Means. The proposed study using the python coding emphasizes the significance of robust preprocessing and metric design in unsupervised learning, and its practical implementation is illustrated using Python.*

***Key Words***: **K-Means Clustering, Outliers, Outliers removal, Adaptive Distance Function.**

## 1. INTRODUCTION

K means is an unsupervised clustering algorithm designed to divide unlabeled data into individual groups of selected numbers (that is, "K"). In other words, k means finds observations that share important functions and classifies them into clusters. An honest clustering solution is a solution that finds clusters so that the observations in each cluster are more similar than the cluster itself. In K Means, each cluster is at the average center of gravity (called the "center of gravity"). Re-presented. The dem cluster has assigned an information point. Centroids are also information that represents the center (mean) of the cluster and do not necessarily have to be members of the dataset. In this way, the algorithm runs an iterative process until each piece of information is closer to the center of gravity of its own cluster than the center of gravity of another cluster, minimizing the distance between the clusters at each step. For example, setting "k" to 2 group's records into two clusters, and setting "k" to 4 groups of knowledge into four clusters. K Means begins off evolved the system with a randomly decided on facts factor because the proposed centroid of the group, iteratively recalculates the brand new centroid, and converges at the very last clustering of facts factors. Specifically, the procedures are as follows: 1. The set of rules randomly selects the middle of gravity for every cluster. For example, in case you pick out 3 "k" s, the set of rules will randomly pick out 3 centroids. 2. K means maps all of the facts with inside the dataset to the nearest centroid. That is, an expertise factor is taken into consideration to be with inside the decided on cluster if it's far toward the middle of gravity of the cluster than the exchange middle of gravity. 3. For every cluster, the set of rules recalculates the centroid via means of getting the not unusual place cost of all factors with inside the cluster, decreasing the whole variance with inside the cluster with recognition to the relevance of the preceding step. When the centroid changes, the set of rules reassigns factors to the closest centroid. 4. The set of rules repeats the centroid calculation and factor allocation till the whole distance among the expertise factor and the corresponding centroid is minimised, the most variety of iterations is reached, or the centroid cost does now no longer change.

### 1.1 Background

Within the lion's share of information sets there are exceptions. This predominance of exceptions is indeed more noticeable in huge datasets since these are regularly assembled through a few computerized frameworks. This infers that there's likely no one physically checking for irregularities. And modern-day detecting frameworks ordinarily support simple gathering information over exactness. Confirmation is commonly the first costly portion, and other people assume that we are going to handle them later. In common exceptions are information designs that go astray from the quality or assumed conduct from the leftover portion of the data [1]. Given this definition, outliers are not one or the other neither loathsome nor extraordinary things; they're fair outliers, discovery (fund), illicit chasing or deforestation (normal sciences), alter in society's conduct (social sciences), among other exercises. In any case, those reasons have something in common: they're all intrigued. The interestingness or real-life significance of exceptions may well be a key highlight of peculiarity [1].

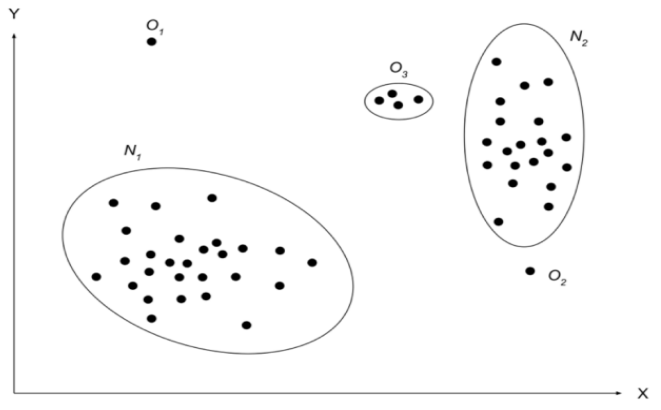


**Figure 1**: Outlines a collection of focuses during a two-dimensional space. There are two fundamental clusters and some outliers.

## 1.2 Causes of Outliers

The taking after are a number of the common causes of the presence of exceptions amid a given information set:

**Measurement Blunder** - this can be regularly caused when the estimation instrument utilised appears to be faulty.

**Data section Blunder** - Human mistakes like blunders caused amid information collection, recording, or passage can cause exceptions in data.

**Experimental Mistake** - These blunders are caused amid information extraction or test arranging or whereas executing an experiment.

**Data Preparing Blunder** - These are caused when control or extraction of the data set is performed.

**Sampling Blunder** - This happens when one extricates or blends information from the erroneous or different sources.

**Intentional Exception** - These are sham exceptions made to check discovery methods.

**Natural Exception** - When an exception isn't counterfeit i.e. caused by a slip-up, it's a common exception. Inside the method of fabricating, collecting, handling and analyzing information, exceptions can come from numerous sources and conceal in numerous measurements.

### 1.3 Effect of outliers on a data set

Exceptions have an expansive effect on the coming of information examination and different measurable measures. Some of the first common impacts are as follows:

- ✓ If the exceptions are no randomly dispersed, they'll diminish normality.
- ✓ It increments the blunder fluctuation and diminishes the office of measurable tests.
- ✓ They can cause inclination and/or impact estimates.
- ✓ They can moreover affect the elemental suspicion of relapse besides as other measurable models.

### 1.4 Mechanism of Outliers Removal:

In terms of the exceptions expulsion component, Fig. 2 appears as a case of the centroid estimation at each cycle amid the exceptions expulsion preparation. Fig. 2(a) outlines a diagram of the boxplot parameters. During this case, the data of the Sepal length trait of the Iris dataset is checked in each single cluster amid the exceptions expulsion process. Altogether three clusters, the remaining information at the extreme emphasis is completely outliers-free as appeared in Fig. 2(b), (c) and (d). It is famous that at the essential cycle there are a number of the exceptions that were recognized and reserved from the Sepal length quality in each cluster. Moreover, the proposed procedure for centroid estimation is distinctive from the centroid of the standard k-means at the extreme cycle of expelling the exceptions. The proposed centroid tends to be close to the cruelty of the outliers-free information. The proposed centroid is adjusted at the middle of the blue box where the information is focused on. During this cluster, the information median is shifted at the lower boundary as shown in Fig. 2(d).

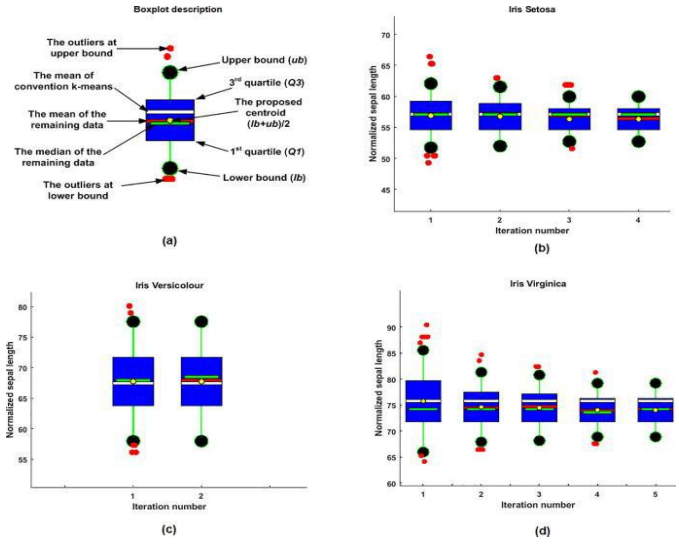


**Figure 2:** k-means algorithm to detect and remove outliers from the dataset.

### 1.5 Purpose of the Work

Numerous ponders moving forward the clustering exactness of the k-means calculation upheld different methods of exceptions evacuation. In terms of exceptions detection-based separate metrics, a few ponders have recognized the exceptions supporting the hole between the information point and its closest centroid (Sarvani et al., 2019, Barai (Deb) and Dey, 2017). In these strategies, the information point with an indeed greater separation to the closest centroid is recognized as an exception. Also, the information focuses with both tenuity and gigantic remove to their centroids are considered as exceptions as displayed in He et al. (2020). In an awfully exceptionally distinctive approach, nearby look strategies (Gupta et al., 2017, Friggstad et al., 2019) are usual to help the k-means for exceptions location. The nearby lookpoints to initiate deter a few information focuses from the information inside the cluster for minimising the target work.

On the off chance that the evacuated information focuses have minimised the target work at that point those information focuses are considered as exceptions and gathered in an exceedingly isolated cluster. In terms of preprocessing methods, kmeans++ is utilized as an additional sifting step in Im et al. (2020) to initiate end of information focuses as exceptions some time recently applying the standard kmeans. In spite of the fact that empowering clustering comes about from these methods, the clustering preparation was as it were performed on the remaining information which is outlier-free. The exception information is totally evacuated and not classified to any known cluster as collected initially.

In other things, exception discovery is utilised as a reward to partition a question from its foundation like in picture handling (Tu et al., 2020, Tu et al., 2019). Be that as it may, few ponders tended to relieve outliers' impacts from the cruel estimation and classifying all information into known clusters as collected at first. In Olukanmi et al. (2017), a k-means# is proposed to dispense with the outliers' impacts from the clusters' centroid. The recognized exceptions are totally avoided from the cruel estimation as it were but they're included afterward inside the clustering preparation. In this way, the impact of the exceptions is relieved from the centroid estimation and improves the clustering exactness. In spite of the fact that the proposed procedure beated the standard k-means, the data point with qualities was disposed of totally from centroid estimation. Amid this case, the calculation cannot recognize an outlier's nearness in each property freely.

Advancement of the clustering exactness from the point of remove metric is illustrated in different ponders. Probabilistic removal for ICA blend models (PDI) is proposed in Safont et al. (2018). The hole measures the harshness between the likelihood thickness of the information, particularly to the parameters of each ICAMM show. The source partition of ICAMM is progressed by supporting PDI separately particularly after altering a limit esteem. In spite of the fact that, the great execution for identifying the failings and hence the variety in electroencephalography (EEG). In Meng et al. (2018), a few orders of subordinate data are measured between the compared vectors and included to the hole metric. The included data of the subsidiaries is valuable for capturing the contrasts between the compared utilitarian information. Be that as it may, this method is computational complex due to the calculation of a few subsidiary orders of the useful information.

In terms of half breed separate measurements which is ordinarily utilized for moving forward clustering exactness, a most recent remove metric named "direction-aware" is created in Gu et al. (2017) to zest up the clustering exactness of k-means. The proposed remove combines the quality Euclidean separately to handle the spatial similitude whereas the cosine metric calculates the shape similitude. Compared to the beginning measurements, the hybridization of both measurements amid one one has moved forward the clustering virtue. Besides, a weighted entirety of the Euclidean and Pearson separate is presented in Immink and Weber (2015) utilizing a weight for summation of both the Euclidean and Pearson coefficients.

The hybrid distance has improved the similarity between the compared signals once the noise is added significantly. The clustering precision may moreover be moved forward as long as the exception is evacuated some time recently measuring the cluster's centroid as talked about amid this segment. Taking the preferences of commonly utilized remove measurements like Euclidean, cosine and relationship into a most recent single likeness metric can prepare the information from distinctive viewpoints. The potential of crossover separate approaches for progressing the clustering exactness is high particularly after killing the impact of the exceptions from the centroid of information clusters [3].

## 2. Methodology

Each machine learning build needs their calculations to make exact expectations. These sorts of learning calculations are frequently classified as administered or unsupervised. K-means clustering is an unsupervised method that needs no labeled reaction for the given computer file. K-means clustering can be broadly utilized. Approach for clustering. By and large, professionals start by learning almost the design of the dataset. K-means clusters information focuses into interesting, no overlapping groupings. It works o.k. when the clusters have a round frame. Be that as it may, it endures from the genuine reality that clusters' geometric shapes leave from round shapes. Moreover, it doesn't learn the number of clusters from the information and wants

## 3. Background

Inconsistencies or outliers are described as observations that deviate sufficiently from the remaining observations to suggest that they were caused by a unique process (Hawkins, 1980). These signs can result from a series of processes such as measurement errors, or specific events such as stack events and climatic events (Chandola et al., 2009). One of the key areas of research within the data processing community is the identification of outsiders.

Many applications have been enhanced, including alien identification fraud detection, transportation networks, and military surveillance. For example, field yield data (the subject of my paper) has shown several times, even to a limited extent, how outsiders affect the standards of the entire dataset (Griffin et al), 2008; Taylor et al., 2007). Interestingly, depending on the scope, you may want to remove the underdog from the dataset (because it affects quality and relevance), while finding and identifying the underdog that is focused on a particular opportunity. You may be particularly interested in doing it, (For example, a sharp rise in temperature may indicate a fire outbreak.) [4] Cluster analysis is a basic task of data processing and machine learning, with different groups of information points.

The purpose is to divide it into two so that similar points can be assigned to the same cluster. Cluster analysis has been studied for a very long time, but thanks to its wide range of applications, from customer segmentation [5] to information retrieval [6], and from recommender systems [7] to resource allocation [8]. Therefore, cluster analysis has also been widely studied in science. Kmeans is one of the leading clustering techniques for finding K-prototypes because it represents information points with the closest centre of gravity. Developed for graph partitions, spectral clustering minimizes the weights of intersecting edges to get individual subgraphs of nearly uniform size. The Gaussian mixture model estimates the K normal distribution using means and variances to fit the information.

Although much effort has been put into cluster analysis, most current methods assume that each information point needs to be assigned a cluster label. That is, there are no anomalous data points in the clustering process. Unfortunately, this is not always the case, especially for unattended tasks. Potential anomalies or outliers inevitably degrade clustering performance. For example, there are few outliers that can easily destroy a cluster structure derived from Kmeans and produce a strange distribution in a Gaussian mixture model. Several robust clustering techniques have been proposed to recover clean data to handle outliers and noisy data. Distance function learning aims to find a powerful distance function that resists outliers [9]. The L1 norm is used to mitigate the negative effects of outliers on the cluster structure [10]. In addition, some methods aim to find more practical expressions with some restrictions. The low-ranked representation assumes that unique or clean data is present in the lowdimensional manifold [11].

Subspace sparse clustering uses sparse coefficients for expression learning to examine selfexpression properties. Nowadays, consensus clustering first creates a basic partition and then uses the required partition based on a robust partition representation. Note that these methods assign a cluster label to all information instead of explicitly removing the anomaly point. Several unsupervised outlier detection methods have been implied in many ways to address the negative effects of outliers during the clustering process. Scores are usually calculated on a perinformation basis, detecting the extent of outliers and returning the best candidate for K outliers. The outlier factor is one of the common density-based methods of identifying outliers by comparing the local densities of information points and their neighbors [12].

Similarly, local distancebased outlier detection uses the object's relative position to its neighbors to see how far the item deviates from its neighbors [13]. Angle-based outlier detection focuses on the variance within an angle between a one-degree difference vector with respect to the opposite point. Here, there are some deviations between the outliers and the angles of the two randomly selected points [13], [14]. Other typical methods include ensemble-based iForest [15], eigenvector-based OPCA [16], and clusterbased TONMF [17].

Outlier detection and cluster analysis are typically performed as two separate tasks, although outlier detection methods are often considered pre-processing for cluster analysis. In fact, they are tightly bound. Cluster structures are often easily destroyed by some outliers [18]. Outliers, in contrast, are defined by the concept of clusters. The cluster is detected because the point does not belong to any cluster [11].

However, the integrated framework handles cluster analysis and outlier detection is just a small part of the work available. DBSCAN is one of the pioneering studies of densitybased cluster analysis, with outliers set as additional output [19]. Here, all information points are categorized according to density into three categories: Two connecting core points and their connected boundary points. Strictly speaking, DBSCAN does not belong to co-cluster analysis and outlier detection, which first identifies and removes outliers, that is, clustering. As far as we know, Kmeans [20] is the first work in this direction.

It aims to detect outliers and subdivide the remaining points into Kclusters. Instances isolated from the nearest centroid are considered outliers during the clustering process. Since this hassle can be a discrete optimization hassle in essence, it`s herbal that Lagrangian Relaxation

(LP) [21] formulates the clustering with outliers as an integer programming hassle with numerous constraints, which desires the cluster introduction charges due to the enter parameter. Although those pioneering works offer new instructions for joint clustering and outlier detection, the round shape assumption of K-approach and consequently the authentic characteristic area restricts its potential for complicated records evaluation, and additionally the setup of enter parameters and time complexity in LP make it infeasible for large-scale records.

During this paper, we deal with the joint cluster evaluation and outlier detection hassle, and recommend the Clustering with Outlier Removal (COR) algorithm. Since the outliers rely upon the idea of clusters, we rework the primary area into the partition area by jogging a few clustering algorithms (e.g. K Means) with specific parameters to get a collection of diverse fundamental walls. By this implies, the persistent records are mapped right into a binary area thru one warm encoding of fundamental walls. Inside the partition area, a goal characteristic is supposedly supported Holoentropy [20] to increase the compactness of each cluster after a few outliers are removed.

With similar analyses, we rework the partial hassle of the goal characteristic right into a K-approach optimizatimization. To offer an entire and neat solution, an auxiliary binary matrix derived from fundamental walls is introduced. Then COR is performed at the concatenated matrix, which absolutely and efficaciously solves the tough hassle thru a unified Kmeans with theoretical support.

To choose the overall performance of COR, we conduct widespread experiments on severa records units in diverse domains. Compared with K-approach and diverse outlier detection methods, COR outperforms competitors in phrases of cluster validity and outlier detection via means of 4 metrics. Moreover, we show the excessive performance of COR, which suggests it is appropriate for large-scale and excessive-dimensional records evaluation. Some key elements in COR are similarly analysed for sensible use. Finally, an utility on flight trajectory is furnished to illustrate the effectiveness of COR inside the real-international scenario. Here we summarise our main contributions as follows.

i. As far as we know, we are the first company to perform outlier clustering by removing outliers in the partition space, achieving consensus clustering and outlier detection at the same time.

ii. Supported holo entropy. Design the objective function from the perspective of detecting outliers. This is partially resolved by kmeans clustering. By introducing a binary auxiliary matrix, we completely transform non-K Means clustering problems into Kmeans optimizations with theoretical support.

iii. Extensive experimental results have shown that the proposed COR is far more effective and efficient than its cutting-edge competitors in terms of cluster validity and outlier detection.

## 4. Aims and Objectives:

The goal of information clustering is to spot homogeneous groups or clusters from a group of objects. In other words, data clustering aims to divide a collection of objects into groups or clusters specified objects within the same cluster are more kind of like one another than to things from other clusters. As an unsupervised learning process, data clustering is usually used as a preliminary step for data analytics. For instance, data clustering is employed to spot the patterns hidden in organic phenomenon data, to supply an honest quality of clusters or summaries for giant data to handle the associated storage and analytical issues, to pick representative insurance policies from an outsized portfolio so as to make met model models. Many clustering algorithms have been developed within the past sixty years. Among these algorithms, the k-means algorithm is one amongst the oldest and most typically used clustering algorithms. Despite being employed widely, the k-means algorithm has several drawbacks. One drawback is that it`s sensitive to noisy data and outliers. As an example, the k-means algorithm isn't ready to correctly recover the 2 clusters shown in Fig. 3(a) thanks to the outliers. As you can see from Figure 3 (b), the three points were incorrectly grouped [6].

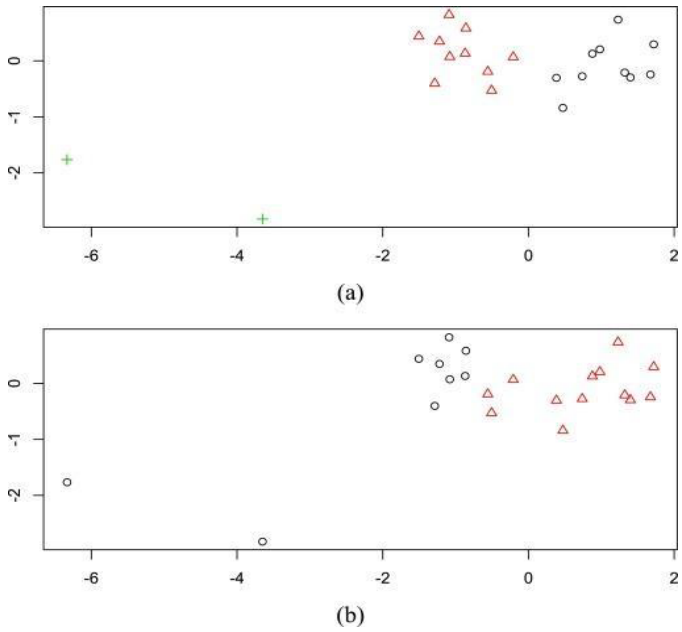


**Figure 3:** An illustration showing that the k-means algorithm is sensitive to outliers.

(a) A data set with two clusters and two outliers. The two clusters are plotted by triangles and circles, respectively. The two outliers are denoted by plus signs. (b) Two

clusters found by the k-means algorithm. The two found clusters are plotted by triangles and circles, respectively.

## 5. K-means Clustering

The KMeans algorithm is an iterative algorithm that attempts to divide a dataset into non-overlapping subgroups (clusters) of k Predefined, where each piece of information belongs to only one group. Try to make the data points in the cluster as similar as possible while keeping the cluster as distinct (wide) as possible. Assign data points to the cluster so that the sum of the squares of the distances between the information points and the center of gravity of the cluster (the arithmetic mean of all information points belonging to this cluster) is minimized. The less variation within a cluster, the more homogeneous (similar) the information points within the same cluster. The algorithm k means that:

i. Gives the number of clusters K.

ii. Initialize the centroid by first shuffling the dataset and then randomly selecting the K data points of the centroid without replacing them.

iii. Repeat until the change in the center of gravity stops. This means that the knowledge point assignments to the cluster will not change.

iv. Calculate the sum of the squares of the distance between the data points and each center of gravity.

v. Match each piece of information to the closest cluster (center of gravity).

vi. Calculate the centroid of the cluster using the standard values for all data points that belong to each cluster.

Approach k means that solving the problem is called expected value maximization. Estep assigns information points to the nearest cluster. Step calculates the centroid of each cluster. Below is a breakdown of how to solve it mathematically.

The objective function is:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{k} \omega ik \left|\left| x^i - \mu k \right|\right|^2 \text{.....(1)}$$

where wik=1 for information xi if it belongs to cluster k; otherwise, wik=0. Also, μk is the centroid of xi's cluster. It's a minimization problem of two parts. We first minimize J w.r.t. wik and treat μk fixed. Then we minimize J w.r.t. μk and treat wik fixed. Technically speaking, we differentiate J w.r.t. wik first and update cluster assignments (E-step). Then we differentiate J w.r.t. μk and recompute the centroids after the cluster assignments from the previous step (M-step). Therefore, E-step is:

$$\delta j \backslash \delta \omega ik = \sum_{i=1}^{m} \sum_{k=1}^{k} \omega ik \left|\left| x^i - \mu k \right|\right|^2$$

$$\Rightarrow \omega ik = \begin{cases} 1 & if\ k = argminj \left|\left| x^i - \mu j \right|\right|^2 \\ otherwise & 0 \end{cases} \text{ ..... [2]}$$

In other words, assign the data point xi to the closest cluster judged by its sum of squared distance from cluster's centroid [22], and M-step is:

$$\delta j \backslash \delta \mu k = 2 \sum_{i=1}^{m} \omega ik (x^i - \mu k) = 0$$

$$\Rightarrow \mu k = \sum_{i=1}^{m} \omega ik\, x^i / \sum_{i=1}^{m} \omega ik \text{ .....[3]}$$

### 5.1 Few things to notice here:

vii. 1. Clustering algorithms, including k-means, use distance-based measurements to check for similarities between data points, so each dataset features most of the time with a mean of 0 and a standard deviation of 1. It is recommended to standardize the information so that it becomes. Measurement units such as age and income are different.

viii. 2. Given the repeatability of the k-means clustering and the random initialization of the centroid at the start of the algorithm, different initializations can result in different clusters. This is because k means that the algorithm stays at the overly local optimum point and does not converge. Therefore, it is advisable to run the algorithm with different centroid initializations and select the execution result with the smaller sum of squares of distances.

ix. 3. Assigning an example that does not change is the same as not changing the variation in the cluster.

### 5.2 Steps for solving the problem:

a. This article is considered to be derived from vector space.

b. Algorithm for clustering N data points into K disjoint subsets S containing data points to minimize the sum of squares criteria.

c. Simply put, k-clustering is an algorithm for grouping objects supported by k-number groups.

d. K can be a positive integer.

This paper proposes a modified k means algorithm to improve data clustering, but both online clustering and large-scale data clustering are outside the scope of this

work. The approach of k means relies on spherical clusters during which the info points converge surrounding the cluster`s centroid. The kmeans splits a group of knowledge points X=x1, x2, x3, ⋯, xN into k known number of clusters. Randomly, the k means selects k set of centroids C=c1, c2, c3, , ck where k≤N. Thereafter, each datum xi is assigned to the closest cluster Cj supported the littlest Euclidean distance. [24]

The Euclidean distance formula is employed to search out the space between two points on a plane. This formula says the gap between two points (x1 1 , y1 1 ) and (x2 2 , y2 2 ) is d = √[(x2 – x1)2 + (y2 – y1)2].

### 5.3 Demonstration of the standard algorithm

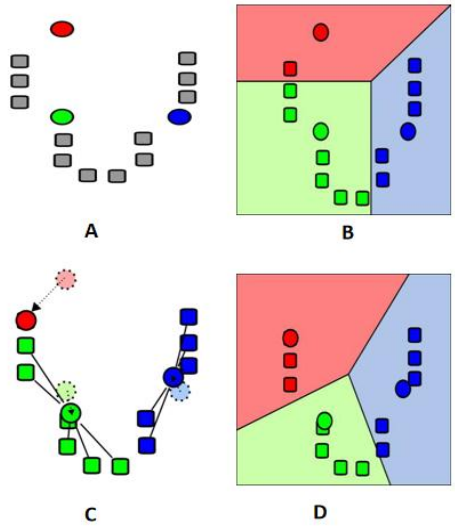


**Figure 4:** standard k-means clustering process

A. k initial "means" (in this case k=3) are randomly generated within the domain (shown in colour).

B. k clusters are created by associating every observation with the closest mean. The partitions here represent the Voronoi diagram generated by the means.

C. The centroid of every of the k clusters becomes the new mean.

D. Steps 2 and three are repeated until convergence has been reached [7].

### 5.4 The advantages of the algorithm are:

• Generally tall productivity with simple execution.

• High-quality clustering.

• Plausibility of parallelization.

• The presence of the numerous modifications.

### 5.5 The disadvantages of the algorithm are:

• The sum of clusters may be a parameter of the calculation

• Affectability to beginning conditions -initialization of cluster centres essentially influences the comes about of clustering.

• Affectability to emanations and clamour Emanations that are reserved from the centres of those clusters are still taken beneath thought when calculating their centres.

• The chance of joining an regional ideal - an iterative approach doesn't ensure meeting an ideal solution[8].

## Conclusion

In this paper we presented outlier detection methods in both K-Means and Hierarchical Clustering. to get rid of outliers is a very important task. We proposed algorithms by which we will remove outliers. We work on a benchmark dataset and after implementing our proposed algorithm it's proved that our proposed algorithm is more efficient than the previous one. After removing the outliers' accuracy is increased. The approach must be implemented on more complex datasets. Future work requires an approach applicable for varying datasets.

## REFERENCES

[1] Chandola , V, Banerjee, A. & Kumar, V. (2009), "Anomaly Detection: A survey"

[2] W. T . Williams, "Principles of Clustering", Annual Review of Ecology and Systematics Vol. 2 (1971), pp. 303-326 (24)

[3] J. Parienti, O. Kuss, " Cluster-crossover design: A method for limiting clusters level effect in community-intervention studies " Contemporary Clinical Trials, Volume 28, Issue 3, May 2007, Pages 316-323.

[4] Nawaf H.M.M Shrifan, Muhammad F. Akbar, Noor Ashidi MatIsa, "An Adaptive Outlier Removal Aided K-Means Clustering Algorithm" Computer and Information Sciences, Available online 13 July 2021.

[5] C. Tsai, Y. Hu and Y. Lu, "Customer Segmentation Issues and Strategies for An Automobile Dealership With Two Clustering Techniques," Expert Systems, vol. 32, no. 1, pp. 65–76, 2015.

[6] R. Campos, G. Dias, A. M. Jorge and A. Jatowt, "Survey of Temporal Information Retrieval and Related Applications," ACM Computing Surveys, vol. 47, no. 2, p. 15, 2015.

[7] A. Shepitsen, J. Gemmell, B. Mobasher and R. Burke, "Personalised Recommendation in Social Tagging Systems Using Hierarchical Clustering," in Proceedings of ACM Conference on Recommender Systems, 2008.

[8] R. Grandl, G. Ananthanarayanan, S. Kandula, S. Rao, and A. Akella, "Multi-Resource Packing for Cluster Schedulers," ACM SIGCOMM Computer Communication Review, vol. 44no4,pp.455–466,2015.

[9] J. V. Davis, B. Kulis, P. Jain, S. Sra and I. S. Dhillon, "Information Theoretic Metric Learning," in Proceedings of International Conference on Machine Learning,.J. V

[10] C. Ding, D. Zhou, X. He, and H. Zha, "R 1-pca: Rotational Invariant L 1-norm Principal Component Analysis for Robust Subspace Factorization," in Proceedings of International Conference on Machine Learning,2006.

[11] G. Liu, Z. Lin and Y. Yu, "Robust Subspace Segmentation by Low-Rank Representation," in Proceedings of International Conference on Machine Learning, 2010.

[12] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying Density-Based Local Outliers," in ACM Sigmod Record, vol. 29, no. 2, 2000, pp. 93–104.

[13] K. Zhang, M. Hutter, and H. Jin, "A New Local Distance Based Outlier Detection Approach for Scattered Real-World Data," in Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2009. IEEE TRANSACTIONS ON KNOWLEDGE.

[14] H. Kriegel and A. Zimek, "Angle-Based Outlier Detection in High Dimensional Data," in Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008.

[15] N. Pham and R. Pagh, "A Near-Linear Time Approximation Algorithm for Angle-Based Outlier Detection in High-Dimensional Data," in Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and DataMining, 2012.

[16] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in Proceedings of IEEE International Conference on Data Mining, 2008.

[17] Y. Lee, Y. Yeh, and Y. Wang, "Anomaly Detection Via Online Oversampling Principal Component Analysis," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no.7,pp.1460–1470,2013.

[18] R. Kannan, H. Woo, C. C. Aggarwal, and H. Park, "Outlier Detection for Text Data," in Proceedings of SIAM International Conference on Data Mining, 2017.

[19] A. Georgogiannis, "Lof: Identifying Density-Based Local Outlier," in Proceedings of Advances in Neural Information Processing Systems, 2016.

[20] M. Ester, H.-P. Kriegel, J. Sander, X. Xu et al., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases With Noise." in Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1996.

[21] S. Chawla and A. Gionis, "K-Means-: A Unified Approach to Clustering and Outlier Detection," in Proceedings of SIAM International Conference on Data Mining, 2013.