

# A Survey on Various K-Means algorithms for Clustering

Malwinder singh<sup>#1</sup>, Meenakshi bansal<sup>#2</sup>

Department of Computer Engineering, Punjabi University

Yadavindra College of Engineering, Talwandi Sabo Punjab, India

## Abstract

Data Mining is the process to find out the data from large data sets and transform into the valuable information. In this paper we are presenting about the clustering techniques and the impact noise on clustering techniques. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). Clustering can be done by the different no. of algorithms such as hierarchical, partitioning, grid and density based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centroid based clustering. K-Mean is widely used clustering algorithm.

## Keywords

Data mining, Clustering, K-Mean, Noise

## 1. Introduction

Data mining is the process of collecting, searching through, and analyzing a large amount of data in a database, as to discover patterns or relationships. Data mining is the process of analysing data from different perspectives and summarizing it into useful information. Data mining consists of extract, transform, and load transaction data onto the data warehouse system, Store and manage the data in a multidimensional database system, Provide data access to business analysts and information technology professionals, Analyse the data by application software, Present the data in a useful format, such as a graph or table. Data mining involves the anomaly detection, association rule learning, classification, regression, summarization and clustering.

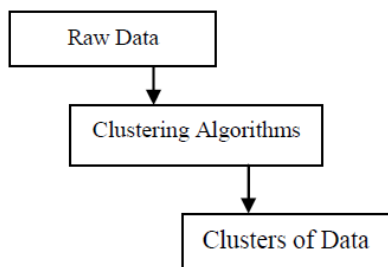


Fig 1. Stages of Clustering

Cluster is the process of partition or grouping of a given set of data into disjoint cluster. This is done in such a pattern that the same cluster are alike and data belong to

different set is different. The accessed data can be stored in one or more operational databases, a data warehouse or a flat file. In data mining the data is mined using two learning approaches i.e. supervised learning or unsupervised clustering.

**Supervised Learning** In this training data includes both the input and the desired results. These methods are fast and accurate. The correct results are known and are given in inputs to the model during the learning process. Supervised models are neural network, Multilayer Perceptron, Decision trees. **Unsupervised Learning** The model is not provided with the correct results during the training. It can be used to cluster the input data in classes on the basis of their statistical properties only. Unsupervised models are different types of clustering, distances and normalization, k-means, self organizing maps.

Data Mining is a four step: Assemble data, Apply data mining tools on datasets, Interpretation and evaluation of result, Result application.

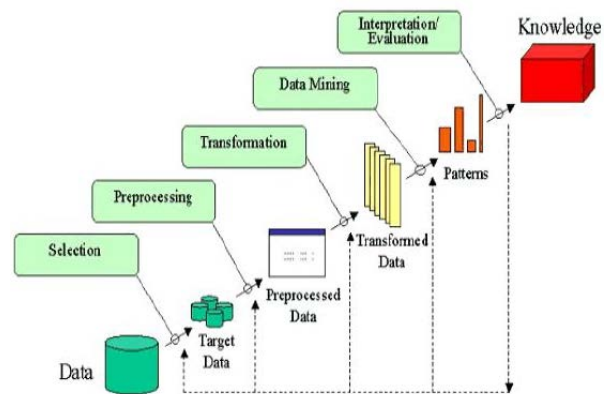


Fig 2. Steps of Data Mining Process

## 2. Clustering Algorithm

### 2.1 The k-Means Algorithm

K-mean is the most popular partitioning method of clustering. It was firstly proposed by MacQueen in 1967, though the idea goes back to Hugo Steinhaus in 1957. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation.

In 1965, E.W.Forgy published essentially the same method, which is why it is sometimes referred to as Lloyd-Forgy. K-mean is a unsupervised, non-deterministic, numerical, iterative method of clustering. In k-mean each cluster is represented by the mean value of objects in the cluster. Here we partition a set of n object into k cluster so that inter-cluster similarity is low and intra-cluster similarity is high. Similarity is measured in term of mean value of objects in a cluster. K-means algorithm uses an iterative procedure in order to cluster database. It takes the number of desired clusters and the initial means as inputs and produces final means as output. Mentioned initial and final means are the means of clusters. If the algorithm is required to produce K clusters then there will be K initial means and K final means. In completion, K-means algorithm produces K final means which answers why the name of algorithm is K-means. After termination of K-means clustering, each object in dataset becomes a member of one cluster. This cluster is determined by searching throughout the means in order to find the cluster with nearest mean to the object. Shortest distanced mean is considered to be the mean of cluster to which examined object belongs. K-means algorithm tries to group the items in dataset into desired number of clusters. To perform this task it makes some iteration until it converges. After each iteration, calculated means are updated such that they become closer to final means. And finally, the algorithm converges and stops performing iterations. Different techniques can be used in K-means clustering in order to measure the distance between objects and means. Most popular two distant metrics are Manhattan Distance and Euclidean Distance. Selecting of initial means is up to the developer of clustering system. This selection is independent of K-means clustering, because these means are inputs of K-means algorithm. Some developers prefer to select initial means randomly from dataset while some others prefer to produce initial points randomly.

In general, we have n data points  $X_j$ ,  $i=1...n$  that have to be partitioned in k clusters. The goal is to assign a cluster to each data point. K-means is a clustering method that aims to find the positions  $\mu_i$ ,  $i=1...k$  of the clusters that minimize the distance from the data points to the cluster. K-means clustering solves

$$\arg \min_c \sum_{i=1}^k \sum_{x \in c_i} d(x, \mu_i) = \arg \min_c \sum_{i=1}^k \sum_{x \in c_i} \|x - \mu_i\|_2^2 \tag{1}$$

where  $c_i$  is the set of points that belong to cluster i. The K-means clustering uses the square of the Euclidean distance. This problem is not trivial, so the K-means algorithm only hopes to find the global minimum, possibly getting stuck in a different solution.

The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. Therefore

$$w_j = i_l, j \in \{1, \dots, k\}, l \in \{1, \dots, n\} \tag{2}$$

Figure 3 shows a high level description of the direct k-means clustering algorithm.  $C_j$  is the  $j^{\text{th}}$  cluster whose value is a disjoint subset of input patterns. The quality of the clustering is determined by the following error function:

$$E = \sum_{j=1}^k \sum_{i_l \in C_j} |i_l - w_j|^2 \tag{3}$$

The appropriate choice of k is problem and domain dependent and generally a user tries several values of k. Assuming that there are n patterns, each of dimension d, iterations I, the time Complexity is  $O(n * K * I * d)$ .

The centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point it is necessary to re-calculate k new centroids as bar centres of the clusters resulting from the previous step. After obtaining these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop, one may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function.

Where a chosen distance measure between a data point and the cluster centre is an indicator of the distance of the n data points from their respective cluster centers.

1. Place k points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the k centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move.

This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The algorithm is significantly sensitive to the initial randomly selected cluster centres. The k- Means algorithm can be run multiple times to reduce this effect.

K-Means is a simple algorithm that has been adapted to many problem domains and it is a good candidate to work for a randomly generated data points. One of the most popular heuristics for solving the k-Means problem is based on a simple iterative scheme for finding a locally minimal solution. The two key features of k-means which make it efficient are often regarded as its biggest drawbacks:

- Euclidean distance is used as a metric and variance is used as a measure of cluster scatter.
- The number of clusters  $k$  is an input parameter: an inappropriate choice of  $k$  may yield poor results. That is why, when performing k-means, it is important to run diagnostic checks for determining the number of clusters in the data set.
- Convergence to a local minimum may produce counterintuitive results.

## 2.2 Variations

In data mining,  $k$ -medians clustering is a cluster analysis algorithm. It is a variation of  $k$ -means clustering where instead of calculating the mean for each cluster to determine its centroid, one instead calculates the median. This has the effect of minimizing error over all clusters with respect to the 1-norm distance metric, as opposed to the square of the 2-norm distance metric. This relates directly to the  $k$ -median problem which is the problem of finding  $k$  centres such that the clusters formed by them are the most compact. Formally, given a set of data points  $x$ , the  $k$  centres  $c_i$  are to be chosen so as to minimize the sum of the distances from each  $x$  to the nearest  $c_i$ . The criterion function formulated in this way is sometimes a better criterion than that used in the  $k$ -means clustering algorithm, in which the sum of the squared distances is used. The sum of distances is widely used in applications such as facility location. *Hartigan et.al 1979* [23] given method to escape local optima by swapping points between clusters. *Dhillon et.al 2001*[5] proposed spherical  $k$ -means algorithm for clustering unlabeled document collections. The algorithm outputs  $k$  disjoint clusters each with a concept vector that is the centroid of the cluster normalized to have unit Euclidean norm. Fuzzy C-Means Clustering is a soft version of K-means, where each data point has a fuzzy degree of belonging to each cluster. In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that

data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. *Elkan et al. 2003*[6] proposed some methods to speed up each k-means step using corsets or the triangle inequality. It shows how to accelerate it, while still always computing exactly the same result as the standard algorithm. The accelerated algorithm avoids unnecessary distance calculations by applying the triangle inequality in two different ways, and by keeping track of lower and upper bounds for distances between points and centres. *Frahling et al 2006* [7] developed an efficient implementation for a k-means clustering algorithm. The novel feature of algorithm is that it uses corsets to speed up the algorithm. A corset is a small weighted set of points that approximates the original point set with respect to the considered problem. The main strength of the algorithm is that it can quickly determine clustering of the same point set for many values of  $k$ . This is necessary in many applications, since, typically, one does not know a good value for  $k$  in advance. *David et al 2006*[21] explains that k-means offers no accuracy guarantees; its simplicity and speed are very appealing in practice. By augmenting k-means with a simple, randomized seeding technique, he obtain an algorithm named k-mean++ that is  $O(\log k)$ -competitive with the optimal clustering. *Vattani et al*[22] tells that the recently proposed k-means++ initialization algorithm achieves proper initialization, obtaining an initial set of centres that is provably close to the optimum solution. A major downside of the k-means++ is its inherent sequential nature, which limits its applicability to massive data: one must make  $k$  passes over the data to find a good initial set of centres. New algorithm k-means|| is presented, parallel version for initializing the centres. This algorithm reduces the number of passes needed to obtain, in parallel, a good initialization.

## 2.3 Applications

K-means clustering in particular when using heuristics such as Lloyd's algorithm is rather easy to implement and apply even on large data sets. As such, it has been successfully used in various topics, ranging from market segmentation, computer vision, geostatistics, and astronomy to agriculture. It often is used as a pre-processing step for other algorithms, for example to find a starting configuration.

*Vector quantization*:- k-means originates from signal processing, and still finds use in this domain. For example in computer graphics, colour quantization is the task of reducing the colour palette of an image to a fixed number of colours  $k$ . The k-means algorithm can easily be used for this task and produces competitive results. Other uses

of vector quantization include non-random sampling, as k-means can easily be used to choose k different but prototypical objects from a large data set for further analysis.

*Cluster analysis:-* In cluster analysis, the k-means algorithm can be used to partition the input data set into k partitions (clusters). However, the pure k-means algorithm is not very flexible, and as such of limited use (except for when vector quantization as above is actually the desired use case!). In particular, the parameter k is known to be hard to choose (as discussed below) when not given by external constraints. In contrast to other algorithms, k-means can also not be used with arbitrary distance functions or be use on non-numerical data. For these use cases, many other algorithms have been developed since.

*Feature learning:-* k-means clustering has been used as a feature learning (or dictionary learning) step, which can be used in the for supervised learning or unsupervised learning.<sup>1</sup> The basic approach is first to train a k-means clustering representation, using the input training data. Then, to project any input datum into the new feature space, threshold matrix-product of the datum can be used with the centroid locations, the distance from the datum to each centroid, or simply an indicator function for the nearest centroid, or some smooth transformation of the distance. Alternatively, by transforming the sample-cluster distance through a Gaussian RBF, one effectively obtains the hidden layer of a radial basis function network.

### 3. RELATED STUDY

**Amrinder et al.[9]** discusses about the performance of clustering technique in the presence of noise. Noise can appear in many real world datasets and heavily corrupt the data structure. The performance of many existing algorithms is degraded by presence of noise in terms of space and time. Noise is the data which is not relevant in data mining and it may affect the performance of clustering. Noise is major problem in clustering analysis.

**Chen et al.[1]** proposed a particle swarm optimization algorithm-based technique, called PSO-clustering. Particle swarm optimization is used to search the cluster centre in the arbitrary data set automatically. PSO can search the best solution from the probability options. This method is simple and valid. PSO based on the minimum object function J to search automatically the data cluster centres of n-dimension Euclidean space. Traditional cluster algorithm such as K-means may get stuck in local optimal solution. PSO needs the less parameter to decide and it has better performance than the traditional clustering analysis algorithms.

**Chunfei et al. [25]** depicts that the traditional K-means algorithm is a widely used clustering algorithm, with a

wide range of applications. This paper introduces the idea of the K-means clustering algorithm analysis the advantages and disadvantages of the traditional K-means clustering algorithm elaborates the method of improving the K-means clustering algorithm based on improve the initial focal point and determine the K value. Simulation experiments prove that the improved clustering algorithm is not only more stable in clustering process, at the same time, improved clustering algorithm to reduce ore even avoid the impact of the noise data in the dataset object to ensure that the final clustering result is more accurate and effective.

**Ghosh et al.** analysed the K-means and Fuzzy C Means algorithms. These algorithms are applied and performance is evaluated on the basis of the efficiency of clustering output. K-Means or Hard C-Means clustering is basically a partitioning method applied to analyse data and treats observations of the data as objects based on locations and distance between various input data points. This algorithm is used for analysis based on distance between various input data points. Results conclude that K-Means algorithm is better than FCM algorithm. FCM produces close results to K-Means clustering but it still requires more computation time than K-Means because of the fuzzy measures calculations involvement in the algorithm

**Merwe et al.[11]** describes that PSO can be used to find the centroids of a user specified number of clusters. The algorithm is then extended to use K-means clustering to seed the initial swarm. This paper investigated the application of the PSO to cluster data vectors. Two algorithms were tested, namely a standard gbest PSO and a Hybrid approach where the individuals of the swarm are seeded by the result of the K-means algorithm. The two PSO approaches were compared against K-means clustering, which showed that the PSO approaches have better convergence to lower quantization errors, and in general, larger inter-cluster distances and smaller intra-cluster distances.

**Parthipan et al.[16]** presented an improved Particle Swarm Optimization (IPSO) and K-means algorithm for solving clustering problems for document and avoid trapping in local optimal solution. Partition clustering algorithms are more suitable for clustering large datasets. K-means algorithm is mostly used algorithm due to easy implementation and efficiency in terms of execution time. IPSO+ K-means produce more accurate, robust and better clustering results. The IPSO+ K-means algorithm combines the ability of globalized searching of the PSO algorithm and the fast convergence of the algorithm and can avoid the drawback of both algorithms. The result from the IPSO algorithm is used as the initial seed of the K-means algorithm. Experimental results illustrate that using this IPSO+K-means algorithm can generate higher compact clustering than using either PSO or K-means alone.

**Sakthi et al [18]** presents that due to the increase in the quantity of data across the world, it turns out to be very complex task for analyzing those data. Categorize those data into remarkable collection is one of the common forms of understanding and learning. This leads to the requirement for better data mining technique.

**Shafeeq et al.[19]** presented a modified K-means algorithm with the intension of improving cluster quality and to fix the optimal number of cluster. The K-means algorithm takes number of clusters (K) as input from the user. But in the practical scenario, it is very difficult to fix the number of clusters in advance. The proposed method works for both the cases i.e. for known number of clusters in advance as well as unknown number of clusters. The user has the flexibility either to fix the number of clusters or input the minimum number of clusters required. The algorithm computes the new cluster centres by incrementing the cluster counter by one in each iteration until it satisfies the validity of cluster quality. This algorithm will overcome this problem by finding the

optimal number of clusters on the run. The main drawback of the proposed approach is that it takes more computational time than the K-means for larger data sets.

**Xiong et al.[24]** proposed new method named as a hyper clique-based data cleaner (HCleaner). HCleaner tends to have better noise removal capabilities than the outlier based approaches. Most existing data cleaning methods focus on removing noise that is the result of low-level data errors that result from an imperfect data collection process, but data objects that are irrelevant or only weakly relevant can also significantly hinder data analysis. HCleaner is based on the concept of hyper clique patterns, which consist of objects that are strongly similar to each other. In particular, every pair of objects within a hyper clique pattern is guaranteed to have a cosine similarity above a certain level. A measure of association that describes the strength or magnitude of a relationship between two objects. HCleaner filters out all objects that do not appear in any hyper clique pattern.

#### 4. Comparative study

S.No.	Paper	Technique used	Results of paper
1	Comparisons Between Data Clustering Algorithms[15]	k-means algorithm, hierarchical clustering algorithm, self-organizing maps algorithm, expectation maximization clustering algorithm	These algorithms are compared according to the following factors: size of dataset, number of clusters, type of dataset and type of software used. Some conclusions that are extracted belong to the performance, quality, and accuracy of the clustering algorithms.
2	A Comparative Study of Various Clustering Algorithms in Data Mining[12]	k-Means Clustering, Hierarchical Clustering, DB Scan clustering, Density Based Clustering, Optics, EM Algorithm.	These clustering techniques are implemented and analysed using a clustering tool WEKA. Performance of the 6 techniques are presented and compared.
3	Performance analysis of k-means with different initialization methods for high dimensional data[20]	Principal Component Analysis (PCA)	It used Principal Component Analysis (PCA) for dimension reduction and to find the initial centroid for k-means
4	A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Set [4]	PCA	Principal component analysis and linear transformation is used for dimensionality reduction and initial centroid is computed, then it is applied to K-Means clustering algorithm.
5	Evolving limitations in K-means algorithm in data mining and their removal [10]	k means algorithm	Limitations and methods to remove these limitation are discussed.
6	Comparative Analysis of Two Algorithms for Intrusion Attack Classification Using KDD CUP Data set	classification. Bayes net and J48 algorithm	Evaluated the performance to two well known classification algorithms for attack classification. Bayes net and J48 algorithm are analyzed The key ideas are to use data mining techniques efficiently for intrusion attack classification
7	Compression, Clustering, and Pattern Discovery in Very High - Dimensional Discrete-Attribute Data Sets[13]	PROXIMUS	PROXIMUS, provides a technique for reducing large data sets into a much smaller set of representative patterns, on which traditional (expensive) analysis algorithms can be applied with minimal loss of accuracy
8	A Modified K-Means Algorithm for Circular Invariant Clustering [3]	split and merge approach (SMCK means)	Introduces a distance measure and a K-means based algorithm, namely, Circular K-Means (CK-means) to cluster vectors containing directional information in a circular-shift invariant manner.

Table 1:- Comparison of different techniques implemented with K-Mean

## 5. CONCLUSION

It is shown in the paper that there is several methods to improve the clustering with different approaches. Noise can appear in many real world datasets and heavily corrupt the data structure. Noise can degrade the quality of data clustering. Large amount of data is also a complex task for analysis. The advantage of the K means algorithm is its favourable execution time. Its drawback is that the user has to know in advance how many clusters are searched for. It is observed that K means algorithm is efficient for smaller data sets.

### REFERENCES

- [1] Chen,C., Ye,F., "Particle Swarm Optimization Algorithm and Its Application to Clustering Analysis" Proceedings of the 2004 IEEE international Conference on Networking, Sensing Control. Taiwan. Page(s):789 - 794 Vol.2, 2004
- [2] Christopher M. Bishop and Michael E. Tipping, "A Hierarchical Latent Variable Model For Data Visualization ," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, Page(s):.281 - 293, Mar. 1998.
- [3] Dimitrios CharalampidisI, " A Modified K - Means Algorithm for Circular Invariant Clustering ," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 12, Page(s): 1856 - 1865, Dec 2005.
- [4] D.Napoleon,S.Pavalakodi," A New Method for Dimensionality Reduction using K - Means Clustering Algorithm for High Dimensional Data Set ," International Journal of Computer Applications (0975-8887),vol. 13, no.7, Page(s): 41 - 46, Jan 2011.
- [5] Dhillon,I., Modha,M., "Concept decompositions for large sparse text data using clustering". Machine Learning 42 (1): Page(s): 143-175 (2001).
- [6] Elkan, C. "Using the triangle inequality to accelerate k-means". Proceedings of the Twentieth International Conference on Machine Learning (ICML) 2003.
- [7] Frahling,G., Sohler,C., "A fast k-means implementation using coresets". Proceedings of the twenty-second annual symposium on Computational geometry SoCG. 2006
- [8] Ghosh,S., dubey,S., " Comparative Analysis of K-Means and Fuzzy C-Means Algorithms" International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013
- [9] Kaur,A., Kumar,P.,Kumar,P., "Effect of Noise on the Performance of Clustering Techniques" International Conference on Networking and Information Technology IEEE Page(s): 504-506 ,2010
- [10] Kehar Singh, Dimple Malik and Naveen Sharma, "Evolving limitations in K-means algorithm in data mining and their removal,"IJCEM International Journal of Computational Engineering &Management, vol. 12, Page(s):105- 109, Apr. 2011.
- [11] Merwe,D., Engelbrecht,AP., "Data Clustering using Particle Swarm Optimization"
- [12] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining ," International Journal of Engineering Research and Applications (IJERA) ISSN: 2248 - 9622 www.ijera.com, vol. 2, Issue 3, Page(s):1379 - 1384,May - Jun. 2012.
- [13] Mehmet.K.,Ananth.G., and Naren. R., "Compression, Clustering and Pattern Discovery in Very High - Dimensional Discrete - Attribute Data Sets, "IEEE Transactions on Knowledge and A Data Engineering", vol. 17, no. 4, Page(s):447-461, Apr 2005.
- [14] N.S.Chandolika,V.D.Nandavadekar," Comparative Analysis of Two Algorithms for Intrusion Attack Classification Using KDD CUP Dataset ," International Journal of Computer Science and Engineering (IJCS),vol.1, Page(s):81 - 88,Aug 2012
- [15] Osama.A.,Abbas, Jordan, " Comparisons Between Data Clustering Algorithms ,"The International Arab Journal of Information Technology, vol. 5, no. 3, Page(s):320 - 326,Jul. 2008.
- [16] Parthipan,L., Rani., A., "Clustering Analysis by Improved Particle Swarm Optimization and K- Means Algorithm" third International Conference on Sustainable Energy and Intelligent System,VCTW, Tiruchengode, Tamilnadu, India, 2012.
- [17] Singh,R., and Bhatia,M., "Data Clustering with Modified K-means Algorithm", IEEE International Conference on Recent Trends in Information Technology, Page(s): 717-721, 2011,.
- [18] Sakthi,M., Thanamani,A., "An Enhanced K Means Clustering using Improved Hopfield Artificial Neural Network and Genetic Algorithm", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Vol-2, 2013
- [19] Shafeeq, A., Hareesha,K.,"Dynamic Clustering of Data with Modified K-Means Algorithm" International Conference on Information and Computer Networks, vol. 27, 2012
- [20] Tajunisha and Saravanan," Performance analysis of k - means with different initialization methods or high dimensional datasets, " International Journal of Artificial Intelligence & Applications (IJAA), vol. 1, no.4, Page(s):44 - 52,Oct. 2010.
- [21] Vassilvitskii. S., David. A., "k-means++: The Advantages of Careful Seeding" ProceedingSODA '07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms Page(s):1027-1035, 2007
- [22] Vattani. A., Moseley. B., Bahmani. B., "Scalable KMeans++" Proceedings of the VLDB Endowment, Vol. 5, No. 7 Page(s): 622-633, 2012
- [23] Wong,A., Hartigan,J., "A K-Means Clustering Algorithm". Journal of the Royal Statistical Society, Series C 28 (1): Page(s): 100-108. 1979.
- [24] Xiong,H., Pandey,G., Steinbach,M., Kumar,V., "Enhancing Data Analysis with Noise Removal" IEEE transactions on knowledge and data engineering, vol. 18, Page(s): 304-319,2006
- [25] Zhang,C.,Fang,Z., "An Improved K-means Clustering Algorithm", Journal of Information & Computational Science, Page(s): 193-199 ,2013