# Review of Existing Clustering Techniques

Amandeep Kaur, Aanshi Bhardwaj

***Abstract*— Data mining is an integrated field, depicted technologies in combination to the areas having database, learning by machine, statistical study, and recognition in patterns of same type, information regeneration, A I networks, knowledge-based portfolios, artificial intelligence, neural network, and data determination. In real terms, mining of data is the investigation of provisional data sets for finding hidden connections and to gather the information in peculiar form which are justifiable and understandable to the owner of gather or mined data. An unsupervised formula which differentiate data components into collections by which the components in similar group are more allied to one other and items in rest of cluster seems to be non-allied, by the criteria of measurement of equality or predictability is called process of clustering. In this paper, we representing a review of cluster types, its differential models and algorithms based on this models. Also a new approach is defined here to enhance the functionality of kmeans by introducing the formula of probability distribution for selection of initial seeds.***

***Keywords*— Clustering, Kmeans, Similarity Measures.**

## I. INTRODUCTION

Data mining is an integrated field, depicted technologies in combination to the areas having database, learning by machine, statistical study, and recognition in patterns of same type, information regeneration, A I networks, knowledge-based portfolios, artificial intelligence, neural network, and data determination. In real terms, mining of data is the investigation of provisional data sets for finding hidden connections and to gather the information in peculiar form which are justifiable and understandable to the owner of gather or mined data. The connections and hidden information gathered by data mining are represented as layouts or arrangements.

## II. CLUSTERING

An unsupervised formula which differentiate data components into count of collections by which the components in similar group are more allied to one other and items in rest of cluster seems to be non-allied, by the criteria of measurement of equality or predictability is called process of clustering.

Revised Version Manuscript Received on May 04, 2016.

**Amandeep Kaur**, Department of Computer Science, Lovely Professional University, Phagwara, India.

**Prof. Aanshi Bhardwaj**, Department of Computer Science, Lovely Professional University, Phagwara, India.

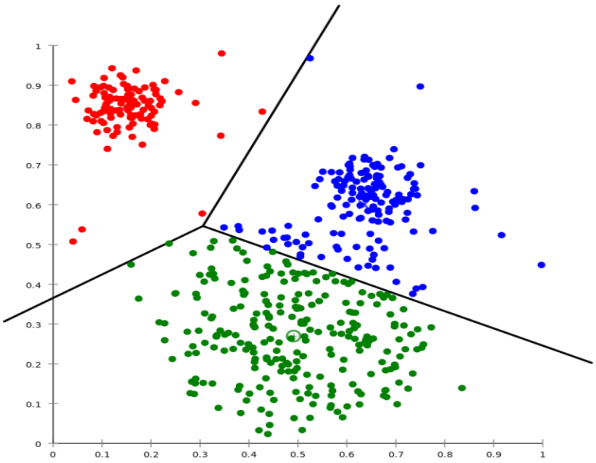


**Figure 1.1: An illustration of making clusters [15]**

The main achievement of clustering is allocate objects to the groups which are having similar behavior or attributes and nature, and non-likeness to rest of the instances.

**Components of Process of Clustering:**

Standard clustering methodology includes the specified components:

(i) Pattern presentation.
(ii) Foundation of common pattern occurrence.
(iii) Collective data patterns based on likeness.
(iv) Data hiding
(v) Estimate of outcome.

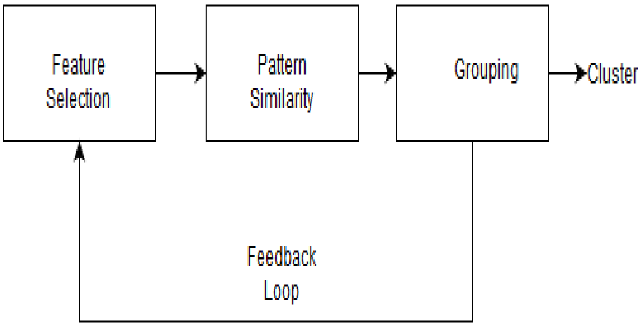


**Figure 1.2 Component of Clustering [13]**

## III. LITERATURE REVIEW

*Crime detection and criminal identification in India using data mining techniques*[1]: In this paper the applied approach is split up in six steps which are-1) data elicitation (DE), 2) data pre processing (DP), 3) clustering, 4) Google map re-presentation, 5) classification, and 6) WE KA_ implementation. Investigation of criminals is done by adopting k-means clustering algorithms, monotously produce 2 clusters of crimes which having same criminal record type. Google map is used to produce better graphic representation of classic algorithm. Crime authentication is done by K NN methodology. Peccant authentication of getting outcomes is working through WE KA. This approach devote goodness in the improvement in civilization by providing the procedure to inspecting firms in criminal detection and knowing of corrupt mans, and draining the peccant estimate.

### A. *K-means Clustering Approach for Segmentation of Corpus Callosum from Brain Magnetic Resonance Images* [2]:

In presented paper, K-means clustering algorithm is applied for being dividing the locality of Callosum Corpus from brain's MR images. The outcomes of this distribution is being used for characteristics digging and segregation of curative predictions in future. The matured tool citing the CC region and its horizon to recognize diverse disorders. With apropos excerpt of mid points, the algo provide truthful evocation of Callosum Corpus and provide mechanization of departmentalizing and segregation of visual MR's by adopting distinct organized and Analytical Callosum Corpus characteristics in the future use.

### B. *An Enhanced K-Means Clustering Technique with Hopfield Artificial Neural Network based On Reactive clustering Protocol [3]:*

The proposing technique in presented paper providing a new criteria to improve the traditional K-Means clustering, whose performance is efficiently increased. The tentative result showing this proposed methodology provides the efficient outcome in comparison to the existing methodologies. This exploration work embedding a fresh assured Reactive protocol based on the sensing technology, to enhance the working of the classical K-means in W S N. One characteristic of existing algorithm is its ability to perform easily and its impressiveness in clustering's area or space. In this paper, neural network of Hopfield artificialness methodology is being implemented with K-means to dig the accurate count of clusters.

### C. *An Optimized Version of the K-Means Clustering Algorithm [4]:*

The presented paper introduced an upgraded adaptation of the traditional K-Means scheme. The main focus in this paper is on the optimization of running time and that concept realized by observing the relocation of data elements that occurred at a small rate after a few iterations. So, there was no need to rejuvenate data components. The work intended here in paper establish limb on those components that are not changing their positions in relocation process and which are changing their positions.

### D. *Applying K-Means Clustering Algorithm Using Oracle Data Mining to Banking Data [5]:*

Data clustering implies the scheme of merging data into distinct collections based on the inter class features. By the collaboration data is in the structure and consequently another preparation of the data is manufactured quite simpler. The paper purposed classical k-means algorithm investigated through Data Mining with oracle. Standard scheme of clustering is to apply to the eighteen attributes of 4 0 banks and 1 0 of the collective instances are produced. By obtaining the cluster, comparisons between the banks is done on the basis of defined attributes in this paper.

### E. *A fuzzy clustering algorithm to detect criminals without prior information [6]:*

The problem of recognizing criminals via communication network is resolved in this paper by proposing a technique named as a fuzzy clustering algorithm. By this algorithm, the hidden conspirators are analyzed which are not used any prior credentials. Fuzzy k means is applied on the global information. A weighted network is formed. Based on priority list, each node in the network that have link with local conjecture are mapped in to the global information cluster. This technique is applicable to large data sets as well as small data sets also. For e.g., TF-IDF method, Disease in biological network.

### F. *Data Clustering through Particle Swarm Optimization Driven Self-Organizing Maps [7]:*

In the presented paper Two techniques PSO (Particle Swarm Optimization) and SOM (Self Organizing Maps) are combined to perform clustering task. SOM is used here for unsupervised learning which maps data patterns with high dimension into reduced mapping of low level dimensions. This reduction makes that data more efficient and better visualization is done by that tool. PSO is the intelligent technique or the optimized algorithm which work on the population which called swarm. In proposed approach, the Lbest also known as input size and Pbest are randomly chosen for each neuron particle.

### G. *Asymmetric k-Means clustering of the Asymmetric Self-Organizing Map [8]:*

In the presented paper, the approach of scrutiny of data is being represented which have two steps. The first step contains visualization of data which is done through asymmetric S O M, whereas the second step of approach is the data visualization through disorganized data that was being divide in allied collections by applying the asymmetric K-means. The outcomes of the performed work proved the effectiveness of the intended scheme upon the traditional algorithms of clustering that are the classical K-means algo the G M M-based methodology, and DB SCAN. This approach improves the count of objects of the clusters.

### H. *Map-Reduce Processing of K-means Algorithm With FPGA-accelerated Computer Cluster [9]:*

This paper proposed an approach in which the k-means clustering algorithm is designed and implemented on an FPGA-accelerated computer cluster. The map-reduce models used with the map and reduce procedures executed paralleled by the CPU on concurrent FPGAs. In this technique two types of communication channel is used that are in first type is used for retrieval of intended instances of primary storage method which are refined through surveyors, second is the transfer of intermediate values in the mappers and reducers. By implementing k-means, system's computation and I/O functioning of FPGA era is analyzed. As compared to the Hadoop environment this approach's performance is improved.

### I. *Extensions of Kmeans-Type algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation [10]:*

*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

In presented paper, a chain of algorithms of clustering by expanding the current traditional k-means is suggested by merging the intra cluster likeness and inter cluster division. The features and effectiveness of proposed algorithms are experimented on different real-life data sets. The presented paper includes the under defined phenomenon: 1) the 3 proposed new judicious criterias are rely upon the classical k-means, W-kmeans, and A W A; 2) rest resembling updated axioms are made and the concurrence is proven and 3) empirical practical's are performed to analyzed the working efficiency.

### *J. K-Means Based Clustering Approach for Data Aggregation in Periodic Sensor Networks [11]:*

In presented paper, an optimized criteria of old P F F method is proposed named as K-P FF. In the proposed methodology, K mean algorithm of clustering is embedding and it is applied before the older PFF technique is applied on the generated clusters. By using K-mean the iteration of comparisons are reduced for finding the similar data. Hence it resulted in the reduced overhead of network and also reduced data latency.

### *K. Fast K-Means Clustering for Very Large Datasets Based on MapReduce Combined with a New Cutting Method [12]:*

This paper proposing a new technique in the clustering environment based on Map reducing method. A new feature is also embedding in it that is called a new cutting method. Map Reduce method helped in executing the job distributively by dividing it in to several parts and executing concurrently. By using it with K-Mean it provide facility to handle large data efficiently but the obstacle there is the increasing number of iterations which effects the overall performance. The proposed method providing solution for this obstacle by introducing a new characteristic called cutting method. By using this property, the iteration are reduced up to 30% with increasing throughput.

### *L. Automatic Identification of Replicated Criminal Websites Using Combined Clustering [13]:*

In presented paper a combined clustering method is presented which is used to link the replicated extortion websites even the criminals' use techniques to hide details. The proposed technique is used for semi-automated extortions or frauds. For this data is taken from databases of two websites that are: high yield investment programs (HYIPs) and fake-escrow. After getting the data attributes of input data are extracted. Then in clustering's first stage computation of clustering is done for each input attributes by hierarchical clustering algorithm. A combined matrix is obtained on attribute basis, then in the next stage of clustering is done with that matrix and clusters with criminal data are produced. The result implies that this technique worked efficiently as compared to general purpose methods.

### *M. Consensus Clustering Based on Particle Swarm Optimization Algorithm [14]:*

In presented paper, the intended accession is the P S O which used to illuminate the problem of allied collection of consensus. It is conclude the Particle Swarm Algorithm is working efficiently regarding present problem. In this paper firstly the algorithms is described which is used to create cluster as a group and consensus functions in implementation. For building the group of clusters five distinct clustering algorithms are being used, that are- K-means using the Euclidean equality schema, K-means using Manhattan equality schema, Expectation–maximization algorithm (E M), Hierarchical schemes and P S O clustering. Presented algorithms generated the individual allied collections using similar data sets. By previous using the consensus method on the obtained clusters using algorithms, the labelling is done on the result of grouped clustered data.

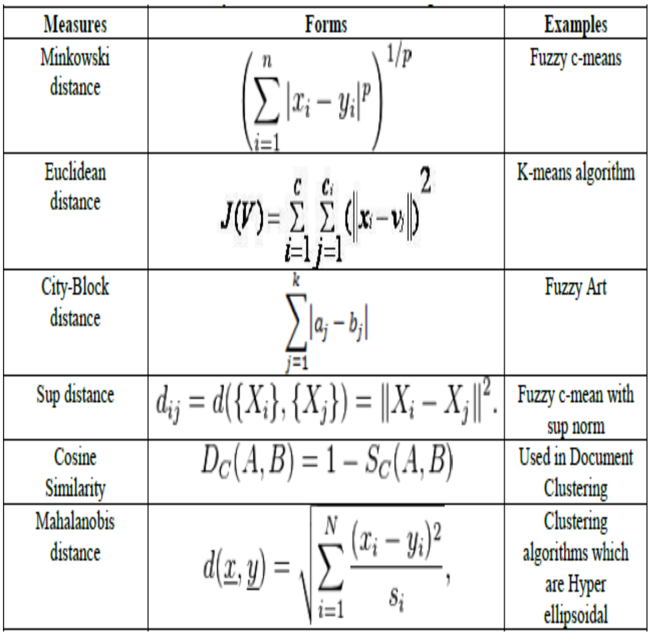
### *N. K-means versus K-means ++ Clustering Technique [15]:*

This paper provide a path of computerizing k-m eans by selecting fluky starting midpoints with advanced efficient predictabilities. With merging k-m eans to a basic, flukier seeding terminology, a new articulated method that is (log k) -competitive having the optimal efficiency can be produced. This terminology guarantees an approximated ratio O (log k) in which k is count of allied collection.

## IV. Similarity Measures In Clustering

The hierarchal clustering method which is in the form of trees make use of the equality and gap in the production of instance's clusters. For collaboration and dividing the components some specific criteria are used named as similarity. For e.g., clustering of fast food is done on the basis of calories contained, price and taste, type. Multi dimensions areas are the most significant method for evaluating the distances of objects. Researcher's main concern is with the measurement of gap rather it is obtained through the pure method or technique, or it is imitated through simulated terminology.

**Table I [2] Similarity Measures used in different Algorithms**



| Measures | Forms | Examples |
|---|---|---|
| Minkowski distance | $\left(\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p\right)^{1/p}$ | Fuzzy c-means |
| Euclidean distance | $J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\lVert x_i - v_j \rVert\right)^2$ | K-means algorithm |
| City-Block distance | $\sum_{j=1}^{k} \lvert a_j - b_j \rvert$ | Fuzzy Art |
| Sup distance | $d_{ij} = d(\{X_i\}, \{X_j\}) = \lVert X_i - X_j \rVert^2.$ | Fuzzy c-mean with sup norm |
| Cosine Similarity | $D_C(A,B) = 1 - S_C(A,B)$ | Used in Document Clustering |
| Mahalanobis distance | $d(\underline{x}, \underline{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i}},$ | Clustering algorithms which are Hyper ellipsoidal |

## V. CLUSTERING ALGORITHMS

The clustering algorithms are classified on the basis of clustering models. The algorithms are many in numbers but not all the algorithms are correct. The algorithms are chosen for a specific problem on the basis of experimental study. The partition approach uses a traditional algorithm called k-mean algorithm which partition the data space into k-clusters. Hierarchal clustering approach are using linkage clustering which is represented by the dendrogram. DBSCAN and optics are the most popular algorithms which are used in the density based clustering model. The overview of this algorithm is as follows:-

The clustering algorithms are classified on the basis of clustering models. The algorithms are many in numbers but not all the algorithms are correct. The algorithms are chosen for a specific problem on the basis of experimental study. The partition approach uses a traditional algorithm called k-mean algorithm which partition the data space into k-clusters. Hierarchal clustering approach are using linkage clustering which is represented by the dendrogram. DBSCAN and optics are the most popular algorithms which are used in the density based clustering model. The overview of these algorithm is as follows:-

### K-Mean Algorithm

- Divide a data space into predefined no of clusters.
- The key objective is to define k centers as one for each cluster. The centers should be placed very cleverly as there will be different results for different locations. So, it is good option to position them as far as possible from each other.
- Now, pick each point of the given data and place it among the nearest center.
- When there is no point left, this completes step one and also an early group age is completed. At this point again calculate the k new cluster centroids from the output of the previous step.
- When the new k centroids have been created, there is a need to bind again the new points between the same data set points and the nearest new center.
- It generates a loop. Due to this loop it is observed that k centers alters their location in step by step till no more alterations are done or we can say that centers do not move any more.
- This algorithm focuses on minimizing an objective function known as squared error function given by:

$$\mathcal{J}(v) = \sum_{i=1}^{c}\sum_{j=1}^{ci}(\|xi - vj\|)^\wedge 2$$

where,

*‘||x_i - v_j||’* is the Euclidean distance between $x_i$ and $v_j$.

*‘c_i’* is the number of data points in $i^{th}$ cluster.

*‘c’* is the number of cluster centers.

### DBSCAN

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is an algorithm which is based on density. The clusters are made on the basis of high density parts in region of data space. The DBSCAN's working is based on two main parameters which are The Eps and the Minpts. The working of DBSCAN is having each object of cluster, a minimum number of objects (*MinPts* ) must be contained in the neighborhood of a given radius ( *Eps* ). The quality of clusters made in DBSCAN algorithm heartily depend on Eps and minpts, are not on the database. The value of parameters are chosen by users which help in the computation of clusters. The parameters are tobe selected properly for the better results; with mistake of chosen parameters the algorithm can provide results with intrusions.

**Table II [2] Complexities of different Clustering Algorithms**

| Cluster Algorithms | Complexities |
|---|---|
| K-mean | *O(NKD)*(time)<br>*O(N+K)* (space) |
| Fuzzy K-mean | Near *O(N)* |
| Hierarchal Clustering | $O(n^2)$(time)<br>$O(n^2)$(space) |
| CLARA | O(ks)^2+ k(n-k)) |
| CLARANS | Quadratic in total performance |
| DBSCAN | *O(N)* (time) |
| BRICH | *O(N Log N)* (time) |

## VI. PROBLEM FORMULATION

As we all know kmeans algorithm has some short comings which are firstly it choose the initial seeds for centre of clusters randomly which leads to wrong formation of clusters. In the presented approach a new technique is appended in kmeans algorithm to overcome this shortcomings and to reduce the iterations of algorithm.

In the presented approach we implemented two new formulas by which the initial seeds for centres are selected on probability distribution basis and for calculating the distance respectively. The data points which have highest probability must be the initial centre of cluster.

1. New cluster centroid using formula of average

$$V_i = \left(\frac{1}{C_i}\right)\sum_{i=1}^{c_i} x_i$$

2. Improved K Means ---------Distance

$$bn = \sum_{j=1}^{n} \max(d_{k-1}^{j} \quad - \|x - xj\|^\wedge 2{,}0)$$

According to these formulas, firstly we apply the cluster centroid formula to calculate the initial centre of predefined clusters. Then on the basis of result of this formulas the data is distributed into clusters. Now the distance formula is applied to calculate the new distance of clusters according to new introduced formula.

These enhanced approaches provide the less iterations as compared to the classical kmeans. The error rate is also reduced to a great extent.

An emerging technology which is implemented in number of fields, the basic moto of this emerging scheme is to distillate enlightenment by applying KDD to coarse data and then do the makeover into an easily accessible, ordered and understandable conformation for another use is often named as mining of data.

Clustering is one of the main aspects used in mining of data. An unsupervised attainments formula which differentiate data components into count of collections by which the components in similar group are more allied to one other and items in rest of cluster seems to be non-allied, by the criteria of measurement of equality or predictability is called process of clustering.

K-means is traditional clustering algorithms, but its usage with the bulk computations, make its performance quite low. The proposed schema can upgrade or boost the execution process of classical programmability of K-Means by enhancing it introducing seed selection criteria and new distance matrix method.

By enhanced collaboration of these two features in algorithms, this can implement in large scale application with reduced amount of calculation and reduced iterations. The scope of the implementing terminologies in a pace originality point of view and execution span for the specific employment would be propagandize as the performance measurement criterion. This scheme's intentions are to contrap these algorithm and graphically confront the difficulties and effectiveness of the algorithm.

## VII. CONCLUSION

The main idea here is to investigate a universal efficient segregation, quick response to improved schema, of defined officials into a peculiar count of allied collections. The methodology is designed here for same kind of obstacles. With the change of segmentation obstacle like an Optimized obstacle, an improved partitioning accession is intended. After that the improved approach merged with K-means algorithm to scale the algorithm. Simulations will be performed to obtain effective execution of the improved algorithm and matched with the rest of the programs. It will help in reducing the iterations and computational time of algorithm. Also overcome the problem of increased error rate.

## REFERENCES

1. Tayal Devendra K., Jain Arti, Arora Surbhi, Agarwal Surbhi, Gupta Tushar, Tyagi Nikhil (2015) "Crime detection and criminal identification in India using data mining techniques", AI & SOCIETY, 30(1), Springer-Verlag London 2014, pp. 117-127.
2. Bhalerao Gaurav Vivek, Dr. Sampathila Niranjana (2014) "K-means Clustering Approach for Segmentation of Corpus Callosum from Brain Magnetic Resonance Images'', Circuits, Communication, Control and Computing (I4C), 2014 International Conference. IEEE, pp. 434-437.
3. Jassi Kaur Navjot, Wraich Singh Sandeep (2014) "An Enhanced K-Means Clustering Technique with Hopfield Artificial Neural Network based On Reactive clustering Protocol", Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference, IEEE, pp. 821-825.
4. Poteras Cosmin Marian, Mihˇaescu Marian Cristian, Mocanu Mihai (2014) "An Optimized Version of the K-Means Clustering Algorithm", Computer Science and Information Systems (Fed CSIS), 2014 Federated Conference, IEEE, pp. 695-699.
5. Hilala Jafarova and Rovshan Aliyev (2015) "Applying K-Means Clustering Algorithm Using Oracle Data Mining to Banking Data" , Proceedings of the Ninth International Conference on Management Science and Engineering Management, Springer Berlin Heidelberg, pp. 809-816.
6. Fan Changjun, Xiao Kaiming, Xiu Baoxin, Lv Guodong(2014) "A fuzzy clustering algorithm to detect criminals without prior information", Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference ,IEEE, pp. 238-243.
7. Gonsalves Tad and Nishimoto Yasuaki (2015) "Data Clustering through Particle Swarm Optimization Driven Self-Organizing Maps", Intelligence in the Era of Big Data, Springer Berlin Heidelberg, pp. 212-219.
8. Olszewski Dominik (2015) "Asymmetric k-Means Clustering of the Asymmetric Self-Organizing Map", Artificial Intelligence and Soft Computing, Springer International Publishing, pp. 772-783.
9. Choi Yuk-Ming, Hayden Kwok-Hay So (2014) "Map-Reduce Processing of K-means Algorithm With FPGA-accelerated Computer Cluster", Application-specific Systems, Architectures and Processors (ASAP), 2014 IEEE 25th International Conference, IEEE, pp. 9-16.
10. Huang Xiaohui, Ye Yunming, and Zhang Haijun (2013) "Extensions of Kmeans-Type algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation", Neural Networks and Learning Systems, IEEE Transactions, 25(8), pp. 1433-1446.
11. Harb Hassan, Makhoul Abdallah and Laiymani David, Jaber Ali and Tawil Rami (2014) "K-Means Based Clustering Approach for Data Aggregation in Periodic Sensor Networks", Wireless and Mobile Computing, Networking and Communications (WiMob), 2014 IEEE 10th International Conference, IEEE, pp. 434-441.
12. Hieu Duong Van and Meesad Phayung (2015) "Fast K-Means Clustering for Very Large Datasets Based on MapReduce Combined with a New Cutting Method", Knowledge and Systems Engineering, Springer International Publishing, pp. 287-298.
13. Drew Jake, Moore Tyler (2014) "Automatic Identification of Replicated Criminal Websites Using Combined Clustering", Security and Privacy Workshops (SPW), 2014 IEEE, pp. 116-123.
14. Esmin Ahmed. A. A., Coelho Rodrigo A. (2013) "Consensus Clustering Based on Particle Swarm Optimization Algorithm", Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference, pp. 2280-2285.
15. Agarwal Shalove, Yadav Shashank and Singh Kanchan (2012) "K-means versus K-means ++ Clustering Technique", Engineering and Systems (SCES), 2012 Students Conference, IEEE, pp. 1-6.