A Review of Various Clustering Techniques

Ejaz Ul Haq
School of Electrical and Computer Engineering
Xiamen University of Technology China.

Xu Huarong
School of Electrical and Computer Engineering
Xiamen University of Technology China.

Muhammad Irfan Khattak
University of Engineering and Technology (Kohat Campus)
Peshawar, Pakistan.

***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. Cluster analysis is a relegating task that is utilized to identify same group of object and it is additionally one of the most widely used method for many practical application in data mining. It is a method of grouping objects, where objects can be physical, such as a student or may be a summary such as customer comportment, handwriting. It has been proposed many clustering algorithms that it falls into the different clustering methods. The intention of this paper is to provide a relegation of some prominent clustering algorithms.**

***Keywords*— cluster analysis; comportment; relegation; algorithms; natural; distribution; hypothesis**

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.

Clustering is to divide data into group of kindred objects. Objects in each cluster are homogeneous amongst themselves and dissimilar to the objects in other clusters. Fewer clusters which are representing data achieve simplification but on other side it additionally loses certain fine information and detail. It represents many fine objects by few clusters, and hence, it model data by its clusters. Cluster analysis divides data into paramount or utilizable groups (clusters). If consequential clusters are our objective, then the resulting clusters should capture the "natural" structure of the data. Cluster analysis is only a subsidiary starting point for other purposes, e.g., data compression or efficiently finding the most proximate neighbors of points. Whether for understanding or utility, cluster analysis has long been utilized in a wide variety of fields: psychology and other convivial sciences, biology, statistics, pattern apperception, information retrieval, machine learning, and data mining. In this chapter we provide a short exordium to cluster analysis. We present a brief view recent technique, which utilizes a concept-predicated approach. In this case, the approach to clustering high dimensional data must deal with the "curse of dimensionality".

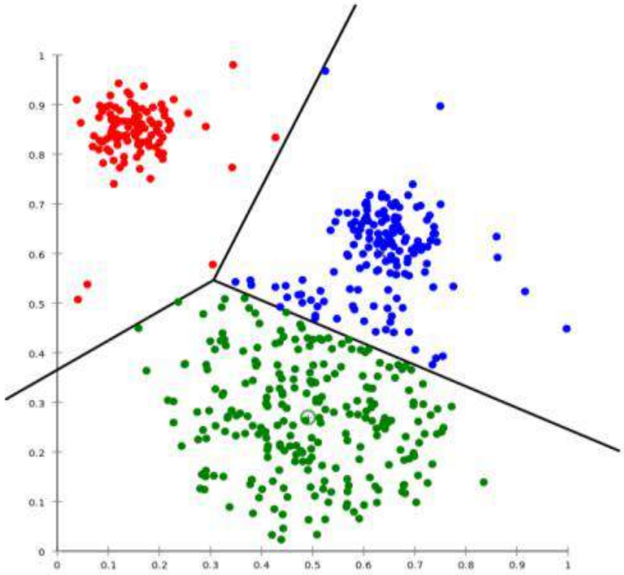


Figure 1. An illustration of making clusters.

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

Process of Clustering:

The overall process of cluster analysis involves four rudimentary steps as explicated below.

A. Feature Selection or Extraction

Feature selection is the process of identifying the most efficacious subset of the pristine features to utilize in clustering, whereas the feature extraction is the process of transforming one or more input features to engender incipient salient feature. Clustering process is highly dependent on this step. Infelicitous cull of features increases the involution and may result into impertinent clusters, additionally.

B. Clustering Algorithm Design or Selection

The impossibility theorem states that, "no single clustering algorithm simultaneously gratifies the three rudimentary axioms of data clustering, i.e., scale-invariance, consistency and richness". Thus it infeasible to develop a generalized framework of clustering methods for the application in the different scientific, gregarious, medical and other fields. It is consequently very consequential to cull the algorithm punctiliously by applying domain cognizance. Generally all algorithms are predicated on the different input parameters, like number of clusters, optimization/construction criterion, termination condition, proximity measure etc. This different parameters and criteria are additionally designed or culled as a prerequisite of this step.

C. Cluster Validation

As there is no macrocosmic algorithm for clustering, different clustering algorithm applied to same dataset engender different results. Even identically tantamount algorithm, with the different values of parameter engenders different clusters. Consequently it becomes compulsory to validate or evaluate the result engender by the clustering method. The evaluation criteria are categorized as:

1) Internal indices: The internal indices generally evaluate the clusters engenders by the clustering algorithm by comparing it with the data only.

2) External indices: The external indices evaluate the clustering results by utilizing the prior erudition, e.g. class labels.

3) Relative indices: As the designation suggest, this criteria compares the results against sundry other results engendered by the different algorithms.

D. Results Interpretation

The last step of clustering process deals with the representation of the clusters. The ultimate goal of clustering is to provide users with paramount insights from the pristine data, so that they can efficaciously analyze and solve the quandaries. This is still an untouched area of research.

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.

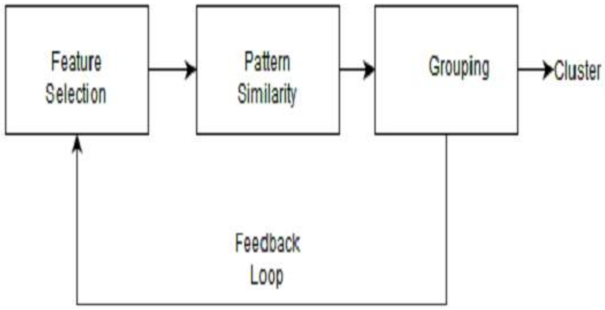


Figure 2. Component of Clustering

II. Literature Review

A. K-means versus K-means ++ Clustering Technique

This paper provides a path of computerizing k-means by selecting fluky starting midpoints with advanced efficient predictabilities. With merging k-means to a basic, flukier seeding terminology, a new articulated method that is (logk)-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.

B. Consensus Clustering Based on Particle Swarm Optimization Algorithm

In presented paper, the intended accession is the PSO 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 labeling is done on the result of grouped clustered data.

C. Automatic Identification of Replicated Criminal Websites Using Combined Clustering

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 fakeescrow. 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, and 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.

D. *Fast K-Means Clustering for Very Large Datasets Based on Map Reduce Combined with a New Cutting Method*

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 distributive by dividing it in to several parts and executing concurrently. By using it with K-Mean it provides facility to handle large data efficiently but the obstacle there is the increasing number of iterations which affects 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 is reduced up to 30% with increasing throughput.

E. *K-Means Based Clustering Approach for Data Aggregation in Periodic Sensor Networks*

In presented paper, an optimized criteria of old PFF method is proposed named as K-PFF. In the proposed methodology 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.

F. *Extensions of Kmeans-Type algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation*

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 criteria's are rely upon the classical k-means, W-k means, and AWA; 2) rest resembling updated axioms are made and the concurrence is proven and 3) empirical practical's are performed to analyzed the working efficiency.

G. *Map-Reduce Processing of K-means Algorithm With FPGA-accelerated Computer Cluster*

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.

H. *Asymmetric k-Means clustering of the Asymmetric Self-Organizing Map*

In the presented paper 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 divided 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 MM-based methodology, and DBSCAN. This approach improves the count of objects of the clusters.

I. *Data Clustering through Particle Swarm Optimization Driven Self-Organizing Maps*

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.

J. *A fuzzy clustering algorithm to detect criminals without prior information*

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.

K. *Applying K-Means Clustering Algorithm Using Oracle Data Mining to Banking Data*

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 is produced. By obtaining the cluster, comparisons between the banks is done on the basis of defined attributes in this paper.

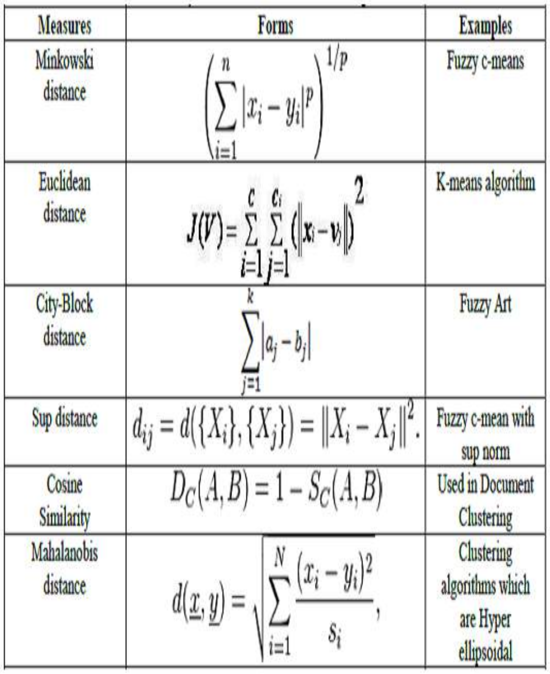
L. An Optimized Version of the K-Means Clustering Algorithm

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.

III. Similarity Measures In Clustering

The hierarchal clustering method which is in the form of trees makes 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 1: 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

IV. 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 Classification of clusters is explained in the following section.

A. Partitional Clustering

Partitional techniques engender a one-level (unnested) partitioning of the data points. If K is the desired number of clusters, then partitional approaches typically find all K clusters at once. Contrast this with traditional hierarchical schemes, which bisect a cluster to get two clusters or merge two clusters to get one. Of course, a hierarchical approach can be used to engender a flat partition of K clusters, and likewise, the reiterated application of a partitional scheme can provide a hierarchical clustering. The cluster must have two properties. They are each group must contain at least one object and each object must belongs to precisely one group.

In this type of clustering, the familiar algorithms are K-Means, K-Medoids, CLARANS, Fuzzy K-Means, and K-Modes.

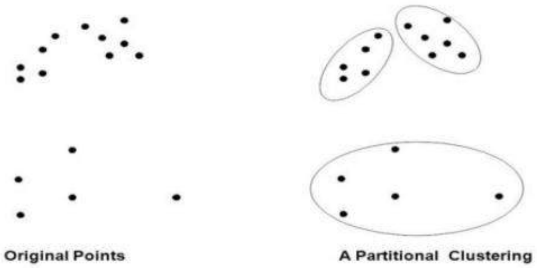


Figure 3. Partitional Clustering

B. Hierarchical Clustering

In Hierarchical type of clustering, more diminutive clusters are merged into more astronomically immense ones, or more sizably voluminous clusters are splitted into more minuscule clusters. The result of the algorithm is a tree of clusters, called dendrogram, which shows how the clusters are cognate. By cutting the dendrogram at a desired level, a clustering of the data items into disjoint groups is obtained. A hierarchy of clusters is built by hierarchical clustering. Its representation is a tree, with individual elements at one end and a single cluster with every element at the other .A hierarchical algorithm yields a dendrogram representing the nested grouping of patterns and kindred attribute levels at which groupings change.

Cutting the tree at a given height will give a clustering at a culled precision. In the above example, cutting after the second row will yield clusters {a} {b c} {d e} {f}. Cutting after the third row will yield clusters {a} {b c} {d e f}, which is a coarser clustering with fewer clusters. The merging or splitting ceases once the desired number of clusters has been composed. In general, each iteration involves merging or splitting a dyad of clusters predicated on a certain criterion, often quantifying the proximity between clusters. Hierarchical techniques suffer from the fact that interiorly taken steps (merge or split), possibly erroneous, are irreversible.

In Hierarchical Clustering, the familiar algorithms are AGNES, DIANA, CURE, CHAMELEON, BIRCH, and ROCK.

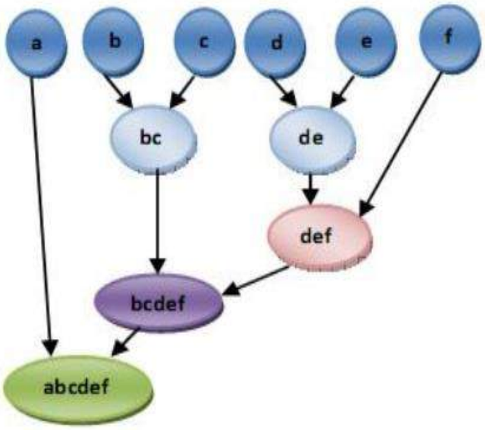


Figure 4. Hierarchical Clustering

C. Density-Based Clustering

Density-based Clusters are defined as areas of higher density than the remnant of the data set. Objects in these sparse areas that are required to separate clusters are customarily considered to be noise and border points. It requires just two parameters and is mostly in sensitive to the injuctively authorizing of the database. The quality of density-predicated clustering depends on the distance measure utilized in the function. It does not require one to designate the number of clusters in the data a priori. This method has been developed predicated on the notion of density that is the no of objects in the given cluster, in this context. The general conception is to perpetuate growing the given cluster as long as the density in the neighborhood exceeds some threshold; that is for each data point within a given cluster; the neighborhood of a given radius has to contain at least a minimum number of points. The density bases algorithms can further relegated as: density predicated on connectivity of points and predicated on density function.

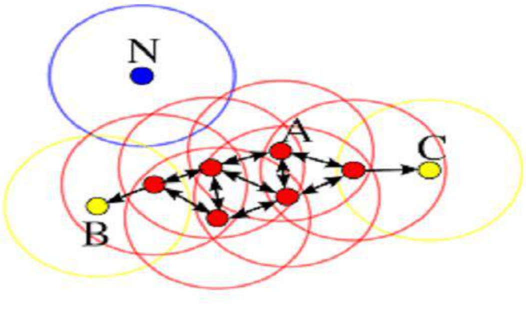


Figure 5. Density based Clustering

The algorithms in this method include DBSCAN, DENCLUE and OPTICS.

D. Grid-Based Clustering

The Grid-based type of clustering approach utilizes a multi resolution grid data structure. It quantizes the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The grid-based clustering approach differs from the conventional clustering algorithms in that it is concerned not with the data points but with the value space that circumvents the data points. In general, a typical grid-predicated clustering algorithm consists of the following five rudimental steps:

- Creating the grid structure, i.e., partitioning the data space into a finite number of cells.
- Calculating the cell density for each cell.
- Sorting of the cells according to their densities.
- Identifying cluster centers.
- Traversal of neighbor cells.

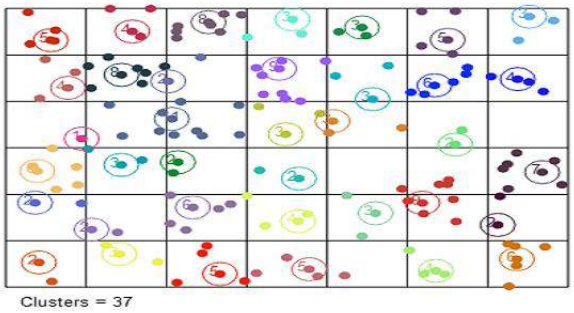


Figure 6. Grid-Based Clustering

The important algorithms in this method include STING, Wavelet and CLIQUE.

V. Problem Formulation

As we all know kmeans algorithm has some short comings which are firstly it chooses the initial seeds for center of clusters randomly which leads to wrong formation of clusters. In the presented approach a new technique is appended in kmeans algorithm to overcome these shortcomings and to reduce the iterations of algorithm. In the presented approach we implemented two new formulas by which the initial seeds for centers are selected on probability distribution basis and for calculating the distance respectively. The data points which have highest probability must be the initial center 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 center of predefined clusters. Then on the basis of result of these 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 iteration 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 algorithms and graphically confront the difficulties and effectiveness of the algorithm.

VI. Conclution

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] Jiawei Han and M Kamber, Data Mining: Concepts and Techniques, Second Edition.

[2] 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.

[3] 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.

[4] Drew Jake, Moore Tyler (2014) "Automatic Identification of Replicated Criminal Websites Using Combined Clustering", Security and Privacy Workshops (SPW), 2014 IEEE, pp. 116-123.

[5] 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.

[6] 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.

[7] Rui Xu, and Donald Wunsch II , ìSurvey of Clustering Algorithms, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005.

[8] Pooja Batra Nagpal and Priyanka Ahlawat Mann, comparative Study of Density based Clustering Algorithms, International Journal of Computer Applications (0975 ñ 8887) Volume 27ñ No.11.

[9] 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.

[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.