An Effective Technique on Clustering in Perspective of Huge Data Set

Muhammad Kalamuddin Ahamad, Ajay Kumar Bharti

Abstract*: Analysis of data plays a crucial job considering the different phenomenon. It explores the prior knowledge, consisting of development across the extensively different communities. Cluster technique is the collecting of data object placed into groups. Therefore objects are the same nature or similar place within a cluster different nature (i.e. dissimilar) put in other cluster. Differences and likeness are refereed on the attribute values say that the objects involved in measuring distance. We have reviewed a few clustering techniques for data sets in data mining of various field of computer science and engineering, statistical, machine learning and a novel attracting field of demanding efforts. Several closely related concepts of neural network, fuzzy and genetic algorithm are also discussed. In this research paper to also discussed the facebook data set to mining the attributes from the cluster set to changing the mean square error 0-10^{-6} and also discuss measuring the performance.*

Keywords: *Method of Partitioning, Method of Hierarchical, Method of Density-Based, Method of Grid-Based, Method of Model-Based, Method of Constraint-Based, Kernel Principle Component Analysis (KPCA), Self-Organizing Map (SOM), Fuzzy and Genetic Algorithm (GA).*

I. INTRODUCTION

The aim of this study of various research papers is to provide a complete review of different techniques of clustering in data mining. The Clustering technique is a partition of data into groups of similar or dissimilar objects. Each and every group is called the cluster and consists of objects that are similar among themselves in one group otherwise dissimilar in other group. This review paper presents different clustering algorithms by taking into consideration the attributes of huge data uniqueness such as size, noise, dimensionality of data set, computations of algorithms, outlier detection and shape of the cluster. Data mining deals with huge databases that need on analysis of clustering and several computations [1][6]. These challenges led to the coming out of powerful broadly related in data mining, the clustering technique reviewed as Scalability to huge datasets, ability to work with high dimensional data, discovery of clusters with irregular shape, handling of outliers, time complexity and ability to deal with the noisy data, data order dependency, labeling (hard or soft of fuzzy) and Interpretability of consequences or usability. This survey emphasizes on concepts of clustering in the field of data mining. Therefore, cluster technique is characterized by huge data-sets with many kinds of attributes [19][16]. Further the facebook data set is taken from the UCI Irvine MLR Archive because this is the business project for useful in real life. In the research paper comprises discussed in section-i Introduction, section-ii Method of clustering survey, section-iii Research papers related by cluster, section-iv Methodology, section-v Experimental result and Discussion and last section –vi Conclusion.

II. Method of Clustering

There are various methods of clustering can be characterized as follows:

- Method of Partitioning
- Method of Hierarchical
- Method of Density-based
- Method of Grid-Based
- Method of Model-Based
- Method of Constraint-based

A. Method of Partitioning

Every object is considered in the beginning as only one cluster. The collections of objects are divided into the number of partitions by iteratively manner and assigning the object between the partitions. Partition based algorithms as follows:

- *K-Means*

This algorithm is the well known tool of clustering used in scientific research and industrial applications. There are two versions of k-means. In the version first comprises of two iterations that (a) re-allocate the every objects to their nearby centroids, and (b) re-compute the centroids of recently group object to assemble. This version is recognized as Forgy's algorithm. This algorithm has many benefits like as work without difficulties, parallelization with respect to ordering of data [27][29].

In version second, the algorithm of K-Means re-assigns objects based on in depth analysis of effects on the goal of function due to moving an object from recent cluster to new one. If a move has a positive effect, the point is relocated and recomputed the two centroids.

Manuscript received on March 15, 2020.
Revised Manuscript received on March 24, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author

Muhammad Kalamuddin Ahamad *,Computer Application, Integral University, Lucknow U.P. ,India.Email: ahamad_kalam@rediffmail.com
Ajay Kumar Bharti, Computer Science, Maharishi University of Information Technology,`1

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Therefore, this is not more understandable that version possible of computation. Further for the reason that analysis of outlier are needs interior loop over every member of objects including the affected of cluster by shifts of centroids[26].

Therefore, in this case both versions have the same computational complexity O (n k d). The k-means has large data size, low dimensional, numerical type of data set, and also non convex shape of cluster [2].

K-Means Algorithm: Input: *k*: represent the cluster num, *D*: represent consisting of n object in a data set, Output: Indicate k a set of cluster.

Step1. At randomly select objects from n objects of data set as the prototype center of the cluster

Step2. Above steps (a) repeat

Step3. Re-allocate of every object in cluster for which is the similar. This similarity is based on find out the mean of object in cluster

Step4. Means value update of cluster

Step5. Until no change

- *K-Medoid*

In concept of the K-Medoid every cluster is represented by single object to found close proximity of cluster. The K-Medoid has complexity O (n2dt), non-convex shape of the cluster, small size of data, high dimensionality, and also categorical type of data set. PAM (Partitioning Around Medoid) was specific of the leading K-Medoid has complexity O (k (n-k) ^{2}), non-convex shape of the cluster, small size of data, low dimensionality, and also numerical type of data set [28] [18].

Algorithm of K-Medoid: Inputs: k: represent the total num. of clusters, D: represent a set of objects, Output: A stable of k number of clusters

Step1. Initially selects the object from the finite set D

Step2. Repeat the step (1)

Step3. Assign for all leftover the objects with nearby representative object.

Step4. Randomly select a none representative object

Step5. Work out total cost, exchanging illustrative object with non_rep object.

Step6. Calculate the cost < zero, then after required to swap rep-object with nonrep_object to make a new_reep objects.

Step7. Until no change

K-medoid techniques are more robust than k-mean. K-Mean is still the noise and outliers but medoid concept is less influence by outlier. Medoid is the measure center of cluster and other measure are also normally used in partitioning methods. The median can use, resulting a k-median, and the most frequent value for every attributes is used, resulting k-mode has complexity O (n), non-convex shape of the cluster, large size of data, high dimensionality, and also categorical type of data set [6][16].

- *Clustering Large Applications (CLARA)*

CLARA uses several (five) handle with huge data sets; this can used sampling based method. The concept follow from CLARA; in its place of tacking entire set of data object into consideration, a small segment of the real data is representative of database. Medoid are select for sampling using PAM. The object select likely be similar to which have been select from the entire data set CLARA draw the multiple sample of the data set, apply PAM on every sample and return its best outcome. CLARA has the many attribute such as complexity O (k (40+k) ^{2} +k (n-k)), non-convex shape of the cluster, large size of data, low dimensionality, numerical type of data set, and also no avoid the outlier [3][18].

- *Clustering Large Applications based upon Randomized (CLARANS)*

The CLARANS technique to enhance the scalability as well as quality of CLARA, combining the sampling technique with PAM. Draw a sample with randomly in every step of search process. The CLARANS is drawing a random sample with dynamically. The random sample of neighbors in each step of explore this not restricted the local area. If a better neighbors nod e found having more less error then approach to the neighborhood node and further again resume the process otherwise recent cluster to develop a local-minima. If the local- minima is found then after CLARANS start new local minima has found as output. CLARANS has complexity O (k n ^{2}), non-convex shape of the cluster, large size of data, low dimensionality, numerical type of data set, and also no avoid the outlier [6][18]. CLARANS has more successful and effective than duo PAM and CLARA The coefficient of silhouette a property of a data object how many object lying in the cluster [3].

B. Method of Hierarchical

Hierarchical clustering technique prepared by grouping of objects into a tree of clusters. It is classified into two categories one agglomerative and secondly divisive. Firstly agglomerative depending on decomposition is construct in merging (bottom to up) and secondly splitting fashion (top to down).

- *Agglomerative* (AGNES: Agglomerative Nesting)

This agglomerative concept depending on decomposition is constructing from merging. The iteratively nature is merging of clusters into make larger clusters, unless until every objects are place in one cluster and satisfied bottom up approach [1][3].

Algorithm:

Step1. Every object is place in its own cluster.

Step2. From all given recent clusters, select the two among them basis of smallest distance.

Step3. Select the two clusters from mentioned the condition in step (2) to make new ones

Step4. Steps (1) and (2) repeat are until there is only one cluster remains in the group.

- *Divisive Analysis (DINA)*

Divisive analysis of hierarchical cluster technique works a top to down approach. It commence by putting every object in single cluster, which cluster is known as the root. Then it is divides the root clusters into few smaller sub-clusters and recursively separation those smaller clusters into only single object. It has the advantage of being more efficient if we do not generate a complete order all the way down to individual document leaves [1][16].

Algorithm

Step1. Place every object in single cluster

Step2. Repeat until all clusters are singletons

- Select a cluster for splitting
- Replace the select cluster with sub-cluster

- *BRICH (Balanced Iterative Reducing & Clustering Using Hierarchies*)

BRICH is intended for clustering a big quantity of numeric data by integrating of this clustering and other clustering method such as iterative partitioning. It is reduce the difficulty of agglomerative cluster is first one scalability and secondly the disaster to undo what was completed in the prior step. BIRCH uses the designs feature of cluster, and clustering feature of a tree (CF-tree) to denote a cluster hierarchy [18]. Given the n-dimensional objects of data set in cluster, further we are defined the following as centroids x_0, radius R, and diameter D, then after duo radius R and diameter D reveal the rigidity of cluster in the region of centroid. A feature of cluster (CF) has three dimensional information in the cluster, and then CF cluster defined as CF=< n, LS, SS>, n= represent the number of point in cluster, LS= represent linear sum of n points, SS= represent Square sum of data point [16]. BRICH has complexity O (n), non-convex shape of the cluster, and work on large size of data; low dimensionality, numerical type of data set, and also no avoid the outlier [6].

Algorithm: Input: N= finite set of object; T= represent of threshold value for construction CF tree; Output: C = indicate clusters set

Method:

Step1. Constructed the feature of cluster (CF) Tree is reading of a data, Following phases develop into fast, correct and fewer order of sensitive

Step2**.** Re-construct the feature of cluster tree with larger T

Step3**.** At the CF leaves require to need the presented of clustering algorithm

Step4**.** Work the additional passes on the data set, the re-allocating points of data nearest to centroids from step (1) to step (3)

Step5**.** Continue to till form k cluster

- *Chameleon Method*

It is an agglomerative hierarchical clustering established of rules that uses active modeling. It is a hierarchical method that methods the connection of two clusters based on dynamic model. This integration technique is using the dynamic model to facilitate discovery of expected and reliable clusters. The set of rules normally consist of two phases: first of all separating of data facts are completed to system sub clusters, use a partitioning of graph, after that have to do frequently integration of sub clusters that come from prior step to acquire final clusters [18]. Chameleon has complexity O (n^2), arbitrary shape of the cluster, and work on large size of data, high dimensionality, Numerical and categorical type of data set, and also no avoid the outlier [6].

- *ROCK Method*

Robust clustering via associations is a robust agglomerative hierarchical clustering set of rules created on the view of links. It is also related for managing the huge data sets. ROCK has complexity O $(n^2+nm_m m_a+n^2 logn)$, arbitrary shape of the cluster, and work on large size of data, low dimensionality, Numerical and categorical type of data set, and also no avoid the outlier [6][16]. The Concepts of ROCK has followed the few steps as:

Step1**.** Obtain the data sample randomly

Step2**.** Get hold of the goodness extent by carrying out connect to AGNES strategy on data. Further obtain the fact which can be compute at every step

Step3**.** Leftover data placing on the disk by required these data object to make clusters

C. Method of Density-Based

To find out the clusters with arbitrary shape, density based clustering technique have been developed. These typically regard clusters as dense regions of objects in data space that are divided by region of low density. One of the most well-known density based clustering algorithm is DBSCAN (Density Based Spatial Clustering of Applications with Noise) [16].

- *Introduction to Density-Based Connectivity*

DBSCAN algorithm is target the low-dimensional data for the major indicating in these categories. DBSCAN has more attribute as complexity O(n log n), arbitrary shape of the cluster, and work on large size of data, low dimensionality, Numerical type of data set, and also no avoid the outlier [24].There are two input argument first one epsilon and second Min-points. These arguments are used here and defined as follows:

Step1. An epsilon neighborhood $N_{epsilon}$ ((a)) = {b belongs in set A| d (a, b) ≤ epsilon} of the point a

Step2**.** Collect number of points including greater than Min-points (called core points)

Step3. In defined step (1), the point b density-reachable from core point a

Step4**.** Here point a and b define in step (1) represent the density-connectivity

- *Method of OPTICS (Ordering point to identify the clustering structure)*

It creates the linear sequence of objects lying in database. In DBSCAN technique where use the two argument one is epsilon indicate Max distance and second Min-point to number objects required to form a cluster. Core as well as Reach-ability distance are required classify to sequence of objects in database. OPTICS has complexity O(n log n), arbitrary shape of the cluster, and work on large size of data, low dimensionality, Numerical type of data set, and also avoid the outlier[16][18].OPTICS algorithm finds the clusters as the following steps:

Step1. Generate a gathering of the objects in a database, storing the distance of every object duo core and exact reach ability. The high density of Clusters will be carried out first.

Step2. OPTIC produced the sequencing of information that required to extraction of clusters

Step3. Measuring the distance, when the condition is satisfy epsilon '< epsilon, then after extract clusters from Density based

- *DENCLUE(Density based Custering)*

This algorithm is based on the density function.Good qualty of cluster is develop the arbitrary shape for large data set with high dimensionalty, avoid the outliers, numerical data types, and it'scomplexity O(log|D|) [6].

Algorithm:
Step 1**.** Let the data set in grid structure and determine high density cells on highest mean.
Step 2**.** If density(mean of cluster(c1) , mean of cluster(c2))less than four times of the mean from step (a) then interconnect c1 and c2.
Step 3**.** Determine the density attractors using hill climbing technique, local maxima is the density function.
Step 4**.** Attractors are merge and further identied as clusters.

D. Method of Grid Based

In this technique the space is arranging into structure of grid. It quantizes the object space into a restricted number of cells that system a grid structure on which every operations for clustering are well performed. It has the improvement of processing time significantly.

- *STING(Statistical Information Grid based)*

This method is like a based on grid and multi-resolution clustering skill in which the drive in spatial area of the input objects is distributed into rectangular cells [16]. This is same as to BIRCH algorithm to develop a cluster with relating to data bases. Statistical Information Grid based Method has complexity O(k) , spatial shape of the cluster, and work on large size of data, low dimensionality, arbitrary type of data set, and also avoid the outlier[6][18].
Algorithm:
Step1**.** At beginning the spatial data are place into rectangular cells for requiring hierarchical grid.
Step2. Divide the all cell into four child cells and further approaches to the next level matching the all child to parent cell (quadrant).
Step3**.** Find the probability of all cell, which related or not. If the cell is related then apply same calculations on every cell one by one.
Step4. Determine the regions of related cells in which order to cluster are produce.

- *CLIQUE Method*

It is an easy grid-based method for discovery density based clusters in subspaces. CLIQUE dividers of every dimension into non-covering the intervals, thus segregate the entire set in space of objects into cells. This subspace clustering for numerical attributes to bottom to up strategy is applied to develop the clusters. CLIQUE has complexity O(C k + m k) ,arbitrary shape of the cluster, and work on large size of data, low dimensionality, Numerical type of data set, and also avoid the outlier [18].
Algorithm:
Step1**.** Select the set of data points at one pass relate to same width of grid cells.
Step2**.** Rectangular cells into subspace, whose thickness go beyond, are sited into equal grids.
Step 3**.**Repeated recursively to form (d-1) dimensional units to d dimensional units.
Steps 4.Subspaces are linked to every other and develop the cluster with equal width.

E. Method of Model-Based

Model-based clustering is based on the assuming that data are produced by sum up of probability distributions, and they optimize the feasibility. When facing an unidentified data allocation, selecting a proper one from the model based candidates is still have major challenges. On clustering based, this is suffers from huge computational cost, in particular when the scale of data is very huge.

- *EM (Expectation Maximization)*

Expectation maximization has complexity O (k n p), non convex shape of the cluster, and work on large size of data; high dimensionality, spatial type of data set, and also no avoid the outlier. It is based on two parameters one Expectation (E), and second maximization (M) [6]16].
Step 1.Expectation (E):
In every cluster assign the object by fractionally based on this PD (Posterior Distribution) as

$$Q(\theta,\theta^{T}) = E(logp(x^{g},x^{m}|\theta)x^{g},\theta^{T}) \qquad (1)$$

Step2 .Maximization (M):

Reevaluating of assignment by fractionally, the parameters with the maximum likelihood rule as

$$\theta^{T+1} = MAX(\theta,\theta^{T} \qquad (2)$$

Step 3**.**Repeated until the convergence.

- *Self Organized Map (SOM)*

This model is based on incremental clustering algorithm by formation grid. The self organized map produced by well known Prof. Kohonen. It has been established and useful in more applications such as competitive learning networks, unsupervised learning, feature inheriting, preserving the mapping from high-dimensional space to map unit lying in two dimensional of lattice, and generalized to capability. SOM has non convex shape of the cluster, and work on small data size; high dimensionality, multi variant type of data set, no avoids the outlier, and complexity O (n^2 m) [23].
Algorithm:
Step1. Pick the last layer of network and set to initialize the positive value for each nearby distance.
Step2. Place weight w from input to output to randomly small, and let t= 1, until not exceed to computational bound

- Select l_i as an input sample.
- Determine square of Euclidian distance from $of \quad l_i$ from weight vector w_i related to every output node

$$\sum_{k=1}^{n}\left(i_{l,k} - w_{j,k}(t)\right)^2 \qquad (3)$$

- Select output node j^* with minimum value return from step (2) of equation (3).
- Update the weight of all nodes.

$$w_j(t+1) = w_j(t) + \propto(t)\left(l_i - w_j(t)\right) \qquad (4)$$

Where learning rate inequality
$0 < \propto(t) \leq \propto(t-1) \leq 1$

- t++

step3. End while.

III. RESEARCH PAPERS RELATED BY CLUSTER

Mine a linear projection of the data by principle component analysis .There are required to determine three compute such as matrix of covariance, eigenvectors and projection in PCA. But

Kernel trick is used to better performance in higher dimensional space. Here given a set of training samples (x_1 x_2 $x_{3,\ldots\ldots}x_l$) in R^N and a non linear mapping function$\emptyset$: $R^N \rightarrow F$, $x \rightarrow \emptyset(x)$. Get extracting a nonlinear projection of data by kernel with PCA are required to compute covariance matrix, eigenvalues and extracting the projection [17]. Authors introduce is selecting the various trick of kernel, it can manage the problem of nonlinear projection for huge data. The RBF function is applicable as the kernel trick. RBF has good performance and it can broadly need in neural network [23]. Introducing the hybridizing cluster technique of huge data set with PCA and WK-Means. Concluding the result cosisting by three steps. The step first normalize the data point uzing z- score, secondly apply the concept of principal component to rduced the data set , and third select the fittness of population from the data set but it fit on the origional k-means [7].Authors proposed to combine the two module,one PSO and second is k-mean. Module of PSO apply on the traditional K-Mean algorithm. It can to help for measuring the execution time, good initialization and applicable on big data.[10]-[12]. Min-Max K-Means cluster technique partitioning of a data objects to several subsets is called clustering. In this research paper author to proposed to more effective clster than k-means and also shows the more detection rate[11].Therefore, the K-means algorithm is more popular in mining. K-Mean algorithm is slightly modified with size k≤ n for many applications [13]. Author, proposed G-Means algorithm with help of a undirected graph *G (V, E)* and a constant *k* cluster, where, V is finite set of vertex (object), *E* is finite set of edges. It is generate the large number of clusters by supply the input data [8]. EKMG algorithm try to overcome the few problems occurred in K-Mean. It is motivated by genetic algorithm but preserve the concept of K-Means. The new population (i.e. data) is generated by fitness value and termination state can be achieved when getting the result is either exceeds the generation or small fitness value [14].

The author proposed a framework a task of detecting the communities by clustering message from large stream of social data. Eliminate the difficulty of K-Means for selecting the centroids using GA, and get the more accurate result by OCD [5].

The building of new models as well as data analyze by cluster of fuzzy concept, which learning concept is more influential unsupervised. There are two specific type of fuzz first one is soft and second hard clustering. In fuzzy c-mean is usually applicable where the object not enforced fit in any single classes but degree of membership to assign between 0 and 1. In 1974 J.Dunn report, this cluster concept in our literature paper. This fuzzy clustering concept can be enhanced by Bezdek in 1981. In fuzzy portioning is to employ the FCM, and the data points lying every cluster with different membership lying between 0 and 1 [15] [25].

IV. METHODOLGY

The Facebook metric data set is taken from online archive of UCI Irvine Machine Learning Repository (MLR). This data set is created in 2016 for business project. The path of data source is https:/archive.ics.uci.edu/ml/datasets/Facebook+metrics. The data set of FBM (Face Book Metric) is consisting of multivariate data characteristics, integer characteristic attribute, 501 instances,19 attributes and in addition of 4 cluster. The outcome of performance is business oriented field where applicable it project. In this research paper go through the surveys for various cluster techniques like partitioning, hierarchical, density, grid based, model based and some other related papers. The K-Mans algorithm is more efficient technique to optimize the facebook data set with KPCA (kernel trick) because the huge data set is reduced the higher dimension to lower dimension. For our side surveys and we have applied the MATLAB R21013a tool find the MSE performance of facebook metric data reduction and also used the MS-Excel to representation of graph for facebook data set.

V.EXPERIMENTAL RESULT AND DICUSSION

The data set of facebook like four clusters is to post (i.e. uploading) links, status, video and photo for business purpose. The data set of facebook to illustrate in blow table 1, there four cluster types and few most attributes are used in daily life of business field for posting the few attributes comprises like as post_month, post_weekend, post_hour, categories, paid e.tc. , and hit count on facebook 110, 1200, 3946 and 1455 respectively. Find out the evaluation type of clusters accuracy.

Table 1: Field of Attributes with Cluster Set

S.N.	Cluster Set	PM	PW	PH	Ct	P
1	LINK	14	8	1	1	0
2	STATUS	10	4	4	1	1
3	VIDEO	8	10	2	2	1
4	PHOTO	6	2	10	1	1

Abbreviations

Where PM= Post_Month, PW= Post_Weekend, PH= Post Hour, Ct= Categories, P= Paid

Retrieval Number: F9185038620/2020©BEIESP
DOI: 10.35940/ijrte.F9185.038620
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

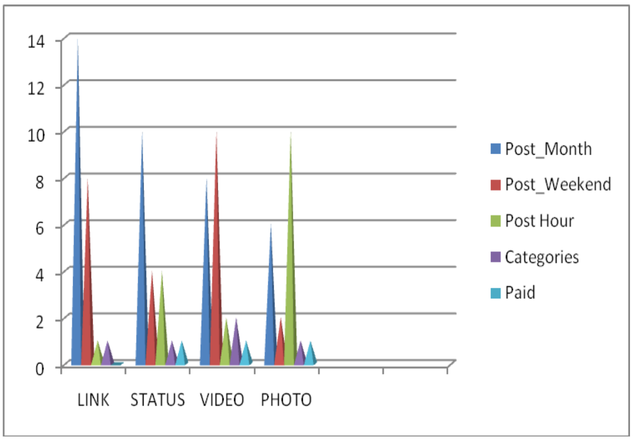

Figure v-a, Cluster Types versus various Attributes

From the above figure v-a, it is observed that when the various ranges apply from 0- 14 for posting the various attributes online to make four cluster types link, status, video and photo. Illustrate in the figure to which attribute hits ranges.

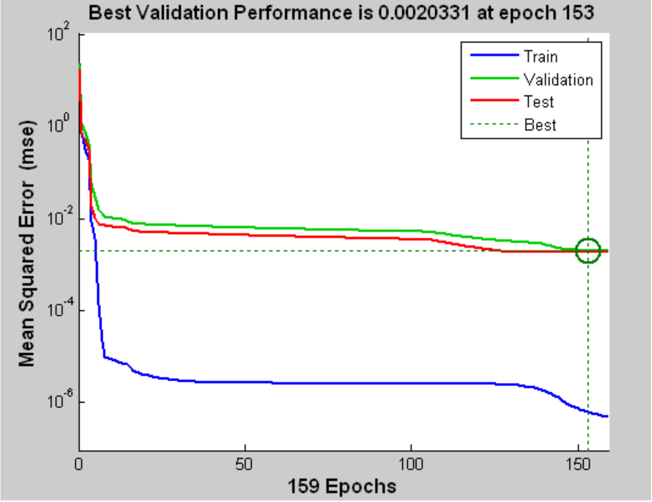

Figure v-b, MSE versus Iterations

From the above figure v-b, plot between MSE versus apply the number of iterations (Epochs) on data set. It is observed that when apply the MATLAB R21013a tool to measure the Mean Square Error (MSE). In the figure indicate increase the number of iteration then to minimize MSE from ranges 0- 10^{-6} and show the validation performance is 0.00203 at 153 iteration.

VI. CONCLUSION

In this research paper we analyzed the various cluster technique to require for processing of huge mine of data set. The cluster techniques are unsupervised learning methods which create the cluster of objects or documents based on their similarity or dissimilarity. An object which exhibits the same feature is placed in one cluster and those which are not similar are placed in other cluster. Clustering can be done by the various procedures such as the algorithm based on partitioning, hierarchical, grid, density, and model. The few machine learning concepts are useful to optimize the cluster quality, avoid the outlier and accuracy. Therefore we taken facebook data set from UCI machine learning repository which data are used in business oriented project and extracting few attributes from data set itself and to hit the site. The extracting of data is to minimize the MSE, reduce the dimension from kernel trick then after improves the efficiency of clusters.

REFERENCES

1. E.U. Haq, X. Huarong, and M.I. Khattak, “A review of various clustering techniques,” Empirical Research Press Ltd., 2017.
2. B. Patel,and C.Gondaliya,” Student performance analysis using data mining technique,” Journal of computer science and mobile computing, 6th ed. vol. 5, 2017, pp. 64-71
3. Y. Bae, Y.S. Kim, F.C.H. Rhee, Y.T. Kim, and C.W. Tao, “Editorial message: Special issue on fuzzy system in data mining and knowledge discovery: Modeling and Application,” International journal of fuzzy systems, 19 ed. Vol .4, 2017, pp. 1157-1157.
4. X.Wang, and Y. Bai,” A modified Min-Max K-Means algorithm based on PSO,” Computer intelligence and neuroscience, 2016.
5. A. Alsayat, and H. El-Sayed,” Social media analysis using optimized K-Means clustering,” In 2016 (14th) International conference on software engineering research, management and applications(SERA), IEEE , 2016, pp61-66.
6. T. Sajana, C.M. Sheela Rani, and K.V. Narayana,” A survey on clustering techniques for big data mining ,” Indian journal of science and technology, 9th ed. Vol. 3, 2016, pp. 1-12.
7. F. Boobord, Z.Othman, and A. Abubakar,” PCAWK: A hybridized clustering algorithm based on PCA and WK-Means for large size of data set,” International journal advanced soft computing applications, 7th ed. Vol. 3, 2015.
8. R.A. Haraty, R.A. Dimishkieh, and M. Masud,” An enhanced k-means clustering algorithm for pattern recovery in healthcare data,” International journal of distributed sensor networks,11th issue vol. 6, 2015, pp. 615740.
9. P. Rathore, and D. Shukla,” Analysis and performance improvement of k-means clustering in big data environment,” International conference on communication networks (ICCN) (ICCN) (pp. 43-46), IEEE, 2015.
10. N. Kamel, I. Ouchen, and K. Baali, “A sampling PSO k-means algorithm for document clustering,” In genetic and evolutionary computing (pp. 45-54), Springer, 2014.
11. M. Eslamnezhad, and A.Y. Varjani,”Intrusion detection based on min max k-means clustering,” In 7th International Symposium on Telecommunications (IST2014) (PP. 804-808), IEEE, 2014.
12. G. Saini, H. Kaur, “K-mean clustering and PSO: A review”, International journal of engineering and advanced technology, 3rd ed. Vol. 5, 2014, pp. 112-114.
13. N. Ganganath, C.T. Cheng and K.T. Chi, “ Data clustering with cluster size constraints using a modified k-means algorithm”, In 2014 International conference on cyber enabled distributed computing and knowledge discovery, (pp. 158-161), IEEE, 2014
14. M. Anusha, and J.G.R.Sathiaseelan,” Enhanced k-means genetic algorithm for optimal cluster,” IEEE International conference on computational intelligence and computing research (pp. 1-5), IEEE, 2014.
15. R. Suganya, and R. Shanthi,” Fuzzy c-means algorithm a review”, International journal of scientific and research publications, 2nd ed. Vol. 11, 2012.
16. J. Han, J. Pei, and M. Kamber,” Data mining: concepts and techniques,” Elsevier, 2011.
17. P. Hongxia, W. Xiuye, and H. Jinying,” Fault feature extraction based on KPCA optimized by PSO algorithm,” In 2010 8th IEEE International conference on industrial informatics (pp. 102-107), IEEE, 2010.
18. R. Xu, and D.C. Wunsch, “Survey of clustering algorithms,” In2005 Transaction on neural networks, IEEE, 2005.
19. P. Berkhin, and J. Becher, “Learning simple relations: Theory and applications,” In proceeding of 2nd SIAMICDM, pp. 420-436, Arlington, VA., 2002.

20. P. Berkhin," Survey of clustering data mining techniques," 2001.
21. A. Jain, M. Murthy, and P. Flynn," Data clustering: A review," ACM computer survey, vol. 31, 1999, pp. 264–323.
22. P. Hansen, and B. Jaumard, "Cluster analysis and mathematical programming," Mathematics program, vol. 79, 1997, pp. 191–215.
23. K. Mehotra, C. K. Mohan, and S. Ranka," Elements of artificial neural networks," MIT Press., 1997, pp. 187-202.
24. M.Ester, H. P. Kriegel, J. Sander, and X. Xu," A density based algorithm for discovering cluster in large spatial databases with noise," In proceeding of 2nd ACM SIGKDD , Portland ,Oregon,1996, pp. 226-231.
25. L. Kaufman, and P. Rousseau," Finding groups in data: An introduction to cluster analysis," Wiley, 1990,
26. R. Duda, and P. Hart," Pattern classification and scene analysis," John Wiley & Sons, New York (NY), 1973.
27. J. Hartingan," Clustering algorithms," John Wiley & Sons, New York, NY, 1975.
28. J. Hartingan, and M Wong," K-means clustering algorithm," Applied statistics, vol. 28, 1979, pp. 100-108.
29. E. Forgy," Cluster analysis of multivariate data: efficiency versus interpretability of classification," 1965.

AUTHORS PROFILE

Muhammad Kalamuddin Ahamad received the degree of Mater of Computer Application from recognized government engineering college then latter completed the M.Tech in information technology in year 2010. I am working in academic field and also pursuing the Ph.D. from Maharishi University of Information Technology Lucknow state Uttar Pradesh of department of computer science and engineering. I am interesting in the research field of data mining and soft computing.

Dr. Ajay Kumar Bharti received the degree of Mater of Computer Application from recognized government engineering college. He was Doctorate from Dr. Baba Bhim Rao Ambedkar University Lucknow U.P. a central status university. He is serving as Prof. and Dean of computer science in Maharishi University of Information Technology Lucknow Uttar Pradesh. He is supervising in the various research fields.