Optimizing Big Data Analysis based on Hybrid K-Means Clustering Approach and Proposed Genetic Approach

Shobhit Jain [1]

M. Tech. Scholar, Department of CSE

SATI, Vidisha, M.P. India.

ABSTRACT: Clustering of large data analysis has received much attention recently. Our research aim healthcare data analysis, Health data is any data related to health conditions, reproductive outcomes, causes of quality of life and health issue for an individual or population. A plurality of health data are collected and used when individuals interact with improve health care systems. The purpose of Health data analysis is to provide better care for patients and help achieve health equity. Health data analysis supports recording of patient data to improve healthcare system. Present a new proposed genetic approach and compare it with popular data hybrid k-means clustering approach. Different clustering comparison methods are based on optimizing of the large data and also the well known, classical group hybrid k-means clustering model. Specifically, in use hybrid k-means clustering approach, in set random values in dataset and hybrid k-means clustering approach using representatives for different numerical values based on comparisons. Previous work has established that hybrid k-means clustering produces more error in data analysis in clusters, produces suboptimal solution. Performed experiments to show that hybrid k-means clustering initialization issues that cause failures. On the other hand proposed genetic approach almost always finds partitions that accurately labeled of data, it is optimizing data analysis, improvement of clusters size, minimization error and additional iteration.

Keywords: **processing, massive Datasets, agglomeration Technique, modified K-Mean agglomeration, basic K-Means agglomeration, Machine Learning, unattended Learning, and PGA.**

I. INTRODUCTION

Data mining is method of extraction, transformation and loading of data to/from information or warehouse system. Storing and managing information give access to information analyst and information person to analyses the information for advantage of their business the wide-spread use of distributed data systems ends up in the development of huge information collections in business, science and on the net. These information collections contain a wealth of data, that but has to be discovered. Medical dataset will learn from their group action information a lot of regarding the behavior of their patient and thus will improve their medical health care by exploiting this data. Science will acquire from experimental information (e.g. satellite data) new insights on analysis queries. Internet usage data are often analyzed and exploited to optimize data access [1]. Data processing provides strategies that enable extracting from massive information collections unknown relationships among the information things that are helpful for higher cognitive process. So data processing generates novel, unexpected interpretations of information [2].

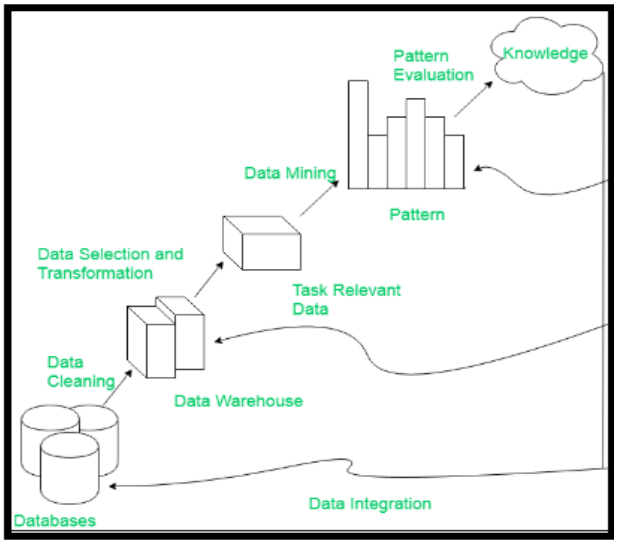

Fig1 Knowledge Discovery Process

Machine Learning

There are 2 learning methodology presents to mine helpful information from data. 1. Supervised Learning: during this form of learning, dataset is given as input and find output as desired, however in presence of trainer. Trainer typically trains the input dataset and classifies it. Example of supervised learning techniques are: Neural network, Multilayer perception, call tree. 2. Unsupervised Learning: the specified result's not provided to the unsupervised model throughout learning procedure. This methodology is wont to

cluster the computer file in categories on the idea of their applied math properties only. These models are for numerous forms of cluster, k-means, distances and standardization, self-organizing maps. This paper reviews numerous strategies and techniques utilized in literature and its benefits and limitations, to research the any want of improvement of k-means rule [3].

Partitioning Clustering

It is the method of discovering teams and totally different structures within the information. Role of cluster in data processing is that it provides the user with the benefit to know that during a information set however we will do natural grouping of a structure. Uses of cluster are often done either as a tool that may stand alone to travel well within the distribution of information or as a pre-processing step for the other formula. Cluster is that the most essential a part of data processing. Cluster could be a technique during which we have a tendency to cluster totally different information supported their similarities and dissimilarities from the opposite. Information in one information set can posses similar properties as compare to the opposite. Clusters will take issue from one another in terms of form size and density. The similarity between the objects is calculated by the employment of a similarity perform. This can be primarily helpful for organizing documents, to enhance recovery and support browsing. Centriod –based technique: - during this variety of grouping methodology a vector of values is employed for referring each cluster. Each object becomes a part of that cluster whose worth distinction is smallest, as compared with another cluster. Disadvantage with this methodology is that no. of clusters ought to be predefined. Distributed –based: - this kind of methodology combines those objects whose worth belongs to identical distribution. this kind of technique wants a model that is advanced however well outlined to act during a higher means with real information. Connectivity-based: - each object in information set is said to its neighbors, which is able to depend upon the degree of the connection of the space between information set and its neighbor. Clusters are created with the objects that are near , and might be delineate as a most distance limit. Density- based: - during this technique clusters are created on the idea of the high density of members of an information set, during a determined location [4].

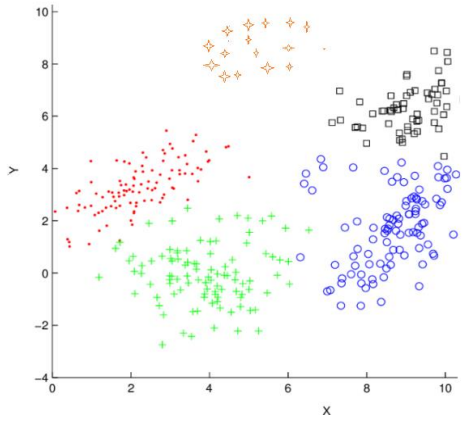

Fig 2 number of clustering

Among all the data cluster algorithmic programs K-means is thus standard as a result of it's the only and therefore the wide used algorithm for clustering of the information sets. it's thus as a result of it uses unattended learning technique to resolve the accepted problems with the cluster. it's conjointly similar temperament for the big knowledge sets. K-mean algorithmic program is that the most generally used algorithms for cluster analysis of enormous no. of information sets. K-mean algorithmic program could be a sort of Centriod-based agglomeration technique. K-mean agglomeration algorithmic program is efficient in generating clusters for several applications. In K-mean algorithmic program centroids are wont to represent a cluster that is essentially the mean of the points of the cluster. These figures show the operating of K-mean algorithmic program on the given information set. during this algorithmic program the complete method is allotted in a pair of phases, within the initial part the Centriod is chosen at random and assignment information objects to the closest cluster then recomputed the cluster Centriod, and within the second part the on top of method is sustained till the clusters become stable. Information objects are partitioned off into non overlapping clusters in order that every and each object is in barely in one set. The explanation of division of the information into many sets is that checking of the all doable subset systems is computationally not feasible; there are sure greedy heuristics schemes that are utilized in the shape of reiterative improvement. this implies completely different relocation schemes that iteratively assign points between the k clusters [5].

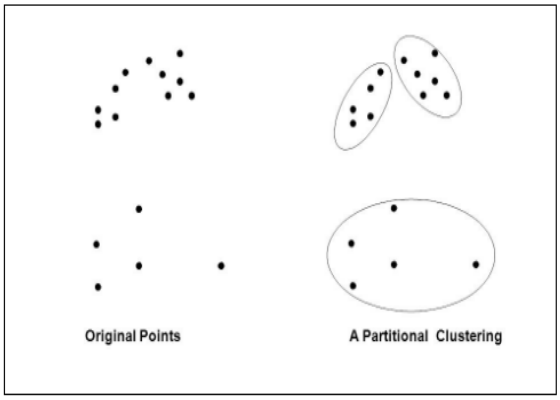

Fig3 before and after partitioning Clustering

II. LITERATURE WORK

Junatao Wang et al. [21] propose an improved kmeans algorithm using noise data filter in this paper. The shortcomings of the traditional k-means clustering algorithm are overcome by this proposed algorithm. density based clustering detection and characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By pre-processing the data to exclude these noise data before clustering data sets the cluster cohesion of the clustering results is improved significantly and the impact of noise data on kmeans algorithm is decreased effectively and the clustering results are more accurate. In present method gets suboptimal result so future work focus on improved accuracy.

K. A. Abdul Nazeer et al. [22] proposes k-means algorithm, for different sets of values of initial centroids, produces different clusters. Final cluster quality in algorithm depends on the selection of initial centroids. Here two stage first original k-means algorithm determining primary centroids and second for assigning data points to the nearest clusters and then recalculating the clustering mean.

G.P and MARTY et.al [23] examines within the paper, how cluster technique is helpful to identify totally different data by considering varied examples and one will see wherever the similarities and ranges agree. By examining one or additional attributes or categories, you'll cluster individual items of information along to create a structure opinion. At an easy level, cluster is victimization one or additional attributes as your basis for distinguishing a cluster of correlating results. Cluster will work each ways in which. You'll assume that there's a cluster at sure purpose then use our identification criteria to visualize if you're correct. In present method gets More error then low accuracy so future work focus on improved accuracy.

P Purohit et al. [24] proposed an improved approach for original K-means clustering algorithm due to its certain limitations. The main reason for poor performance of K-means algorithm is selection of initial centroids randomly. The proposed algorithm deals with this problem and improves the performance and cluster quality of original k-means algorithm. The new algorithm selects the initial centroid in a systematic manner rather than randomly selecting. It first find out the closest data points by calculating Euclidian distance between each data point and then these points are deleted from population and forms a new set. This step is repeated on new set by finding data points that are closest to each other. Performance comparison is done using Mat lab tool. The proposed algorithm gives more accurate results and also decreases the mean square distance**,** In present method gets Error data more then original pattern lose so future work focus on improved accuracy.

Alin Fan et al. [25] Elobrates k-means clustering algorithm based on coefficient of variation. The coefficient of variation is defined as ratio of standard deviation to the mean value. Existing k-means algorithm uses Euclidean distance as the similarity metric which gives inaccurate results due to the effect of useless data. To overcome with this problem, proposed algorithm uses coefficient of weight factor to elicit the effect of ouliers.Weight values. are assigned to all the features in clustering to remove irrelevant, noisy data so as to increase cluster quality. The results are evaluated using popular data sets i.e. Iris, Wine and Balance scale. The results prove that the modified algorithm presents more clustering accuracy and the number of iterations required for clustering is less than original k-means. The problem faced by proposed algorithm is that the number of clusters required as output is needed to be initially defined. In present method gets Number of iteration more and copy data more so future work focus on improved accuracy.

Md Sohrab Mahmud et al. [26] have proposed a heuristic method that takes less computational time than traditional k-means. It is discover average values of dataset and determine normal of data points and then the method sort data points. The nearest possible data point to the mean is chosen as initial centroid. . In present method gets inaccurate results due to the effect of useless data so future work focus on improved accuracy.

D. Arthur et al. [27] k-means++ is one of the most popular centroid initialization technique because it makes k-means converge faster reducing number of iterations with better sum of squared distances error. It is a method to initialize the k centroids where k (required number of clusters) is given as an input to the k-means algorithm. Probabilities depending on the minimum-distance from a point x to the previously selected centers are used to select next centre. The greedy version of this method probabilistically selects log (K) centers in each round and then greedily selects the centroid that most reduces the sum of squared error. that are close to each other. In present method gets less computational time and calculating average of data points so future work focus on improved accuracy.

Kiri Wagstaff et al. [28] proposed constraint based k-means. Author has suggested that in most of the cases the experimenter has some background knowledge about which instances should be and should not be grouped together. This information is expressed in terms of instance-level constraints and is given as input. The modified algorithm assigns an object to a cluster only when none of the constraints are violated. Cluster is not found object values then returns types empty cluster. Beginning the answer of the experimentation one can say that incorporation of background knowledge has improved accuracy of traditional k-means significantly. In present method cluster is not found for an object then method returns empty cluster so low accuracy so future work focus on improved accuracy.

Umutoni Nadine et al. [29] presented a new hybrid competitive recommendation approach to improving the effectiveness through the competition process among a series of algorithms. Author situation that grouping of unusual algorithms determination make available more accurate and effective recommendations than a single algorithm. The drawbacks of one algorithm are smoothened by some other algorithm. The collective come within reach of of multiple proposal techniques can decrease the weaknesses of an individual technique in a combined model. Keeping these facts in mind, a new hybrid method is suggested. Existing method achieves more stable performance through the competition among a series of algorithms. The ranking is computed using individual algorithms and finally, the combined rank for each product is calculated using a variety of ranking functions. In present method gets new hybrid method based on multiple recommendation techniques so suboptimal solution so future work focus on improved accuracy.

Bruno Fernandez et al. [30] has compared various data mining algorithm such as Bayesian techniques and LMT for classification and association on UCI repository breast cancer dataset, dermatology dataset and vertebral column dataset. After analysis Bayes Net was found to be the best classification algorithm for Dermatology and breast cancer dataset with 97% of classification accuracy. For vertebral column the best algorithm was logistic model tree with 85% of accuracy. In present method gets Here only used cancer dataset and low accuracy so future work focus on improved accuracy.

III. Overview Proposed Methodology

In this section they present our algorithm, which produces good initialization centers for the K-Means instead of random ones. They introduce the intuition behind the algorithm at first, and describe related notations and background. They will demonstrate our algorithm at last. It is designed to overcome the dependence of split data on selection an ordering. This model also divides the big data split into small data that can be loaded in the system memory then performs proposed method on each break apart to obtain (n sets) n centroids. The n centroids obtained from the n split data are then clustered using proposed method, where the numbers of values are the cardinalities of the clusters, to obtain the final n centroids. Then, as in proposed method, the big data are labeled using the proposed method .

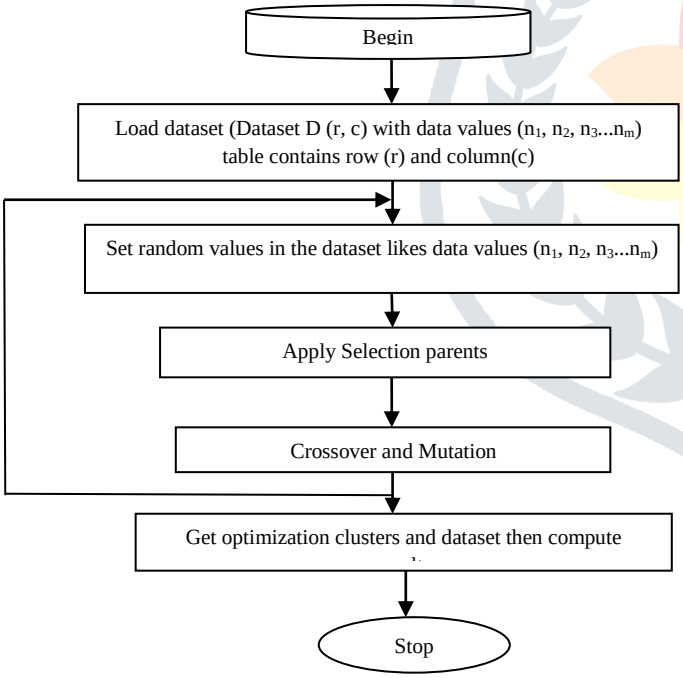

Fig4 Working Architecture

It is a proposed clustering algorithm that randomly samples a fixed number of points from the large dataset so that the representative points (hopefully) retain the geometry of the entire dataset. Let the sampled dataset be *S*, and assume that it contains *n* clusters. Each cluster is represented by a fixed number *m* of well-scattered points called representative points, which are shrunken toward the mean of the cluster by a fraction *x* to have a compact representation of the cluster as well as to minimize the effect of outliers. Each cluster is a unique entry in the fitness values and has the information regarding the number of points, representative points, and mean of all the points in the cluster and is arranged in increasing order of the distance from its nearest cluster. The proposed clustering stores the representative points of all the clusters and can find the nearest neighbor of each point in constant time. Each representative point forms a separate cluster in the beginning. The nearest clusters are merged to form a big cluster until the prespecified number of clusters is obtained. The distance between two clusters is the minimum distance between the representative points of the two clusters.

PGA cluster is an unsupervised hard partitioning cluster technique. The target is to search out k clusters and minimum size of clusters of information points supported the space obtained, the data points are appointed to the cluster with minimum distance from the centroid. Once the points are clustered the mean of all points belonging to the cluster is found. The aim of hybrid K-Means algorithm rule is to satisfy the target performs however suboptimal resolution but PGA as improvement approaches and search out an optimum solution. GA starts with a group of potential solutions (chromosomes). Next, genetic operators (selection, mutation and crossover) are applied one once another to get a brand new generation of chromosomes. This method is continuing developed the termination condition in mat lab.

The fundamental steps of PGA are following below.

Step I begin load given data set like Osmotolerant Yeasts,Hyper_thyroidism_dataset,Escherichia_coili,DCIS_breast_cancerdataset. (Dataset D (r, c) with N instance values (n_1, n_2, n_3..n_m), here a set of samples X, number of feature F, a label function l, number of iteration.

Step II Set random values (n_1, n_2, n_3..n_m), in the dataset, Here dataset hold number of samples X and also set valid types of samples.

Step III Set samples X after that training dataset D (r, c) and an analysis or evaluated to dataset D (r, c) is created by repeating the following steps:

i) Selection: Select two parents (chromosomes) from a population according to their fitness value. The possibility for every chromosome to be selected, as a parent, is determined according to its fitness.

ii) Crossover: According to the crossover probability (Pc), new offspring (children) is generated from parents.

iii) Mutation: According to mutation probability (Pm), new offspring at each locus (position in chromosome) is mutated.

Step IV Accepting new offspring based on vector space model (V1, V2…Vn) is placed in optimal values (the new population). Go to stepII.

Step V Generated output based on minimization error values in dataset. To find clusters from the data based on the objective function E given in Eq. (1).

$$E=\sum i=1k\sum j=1Nd2\ (Ci-Xj)$$

Where d^2 (C_i-X_j) is the squared Euclidean distance between ith cluster centroid and jth data point N is the total number of data points Based on the distance obtained, the points are assigned to the cluster with minimum distance from the centroid. After the points are clustered the mean of all points belonging to the cluster is found.

Step VI If the end condition is satisfied, return valid values and also optimization values, the outputs in current dataset and go to step VII.

Step VII Stop.

IV.MATLAB Setup

It is simulating on MATLAB-2013 Now, let's introduce the experiment environments and the results obtained from the proposed algorithms and the neighborhood model method. The experiments are performed on a laptop with 1.83 GHz Intel core 2 processor and 2GB byte memory running Windows XP, window7 operating systems etc. MATLAB (R2013a) programming language is used to code the algorithms. It is a high-level technical calculate language and interactive setting for algorithmic program development, information image, records analysis, and numeric computation Mat research laboratory could be a code program that enables you to try and do information manipulation and image, calculations, science and programming. It may be accustomed do terribly easy also as terribly refined tasks. Database, analysis, image, and algorithmic program development. You can perform efficient data retrieve enhancement. Many functions in the toolbox are multithreaded to take benefit of multicore and multiprocessor computers. An additional package, Simulink, netural network.MATLAB (matrix laboratory) could be a multi paradigm numerical computing scenario and fourth generation artificial language It is developed by math work

V. EXPERIMENTAL RESULT ANALYSIS

In research in field of data mining based optimizing dataset in clusters based on data analysis using hybrid approach to clustering in big data analysis using k-means clustering algorithm. Find minimum error of medical dataset analysis and best possible solution.

(i)Experimentation1 Base on Osmotolerant Yeasts Dataset

Table 2 Results analysis for the osmotolerant yeasts dataset of clustering algorithms

Algorithm	No. of Random different Point Values	Time (s)	Error Rate (%)	Iteration
HKMA	0.6435	2.51162	4.00499	4
PGA	0.6435	3.51002	1.30723	5
HKMA	0.9	0.483603	2.6552	3
PGA	0.9	0.312002	0.0087105	4

Experimentation1 base on osmotolerant yeasts dataset analysis using HKMA and PGA. HKMA are error is more and time also more but iteration min and PGA error less and average time but iteration more.

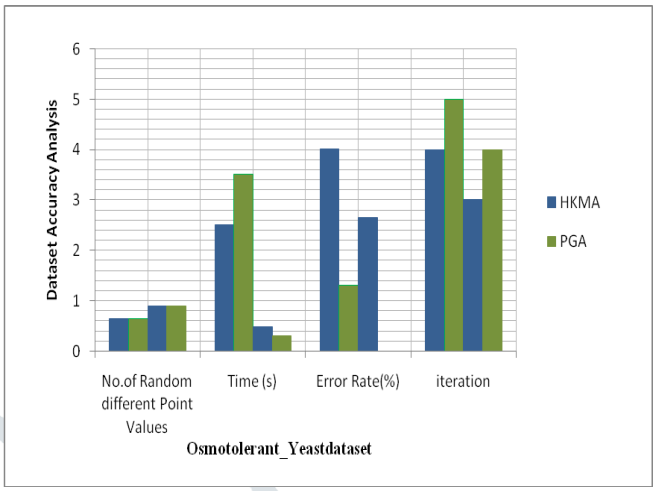

Fig5 osmotolerant yeasts dataset analysis between HKMA and PGA

Above graph results analysis between HKMA and PGA using osmotolerant yeasts dataset. HKMA are error is more and time also more but iteration minimum and PGA error less and average time but iteration more.

VI. CONCLUSION

In study of some well known algorithms concerned with data processing technique .Under the clustering techniques of knowledge mining varied algorithms specifically likes k-means, hierarchal clustering and k-mean algorithms are studied. The results are compared between HKMCA and PGA and analyzed in accordance to their efficiencies. For classification, the hybrid k-means clustering approach and PGA formula were implemented and compared. under the clustering techniques of information mining varied algorithms specifically k-mean data processing could also be a broad area that deals among the analysis of big volume of data by the mix of techniques from several fields like machine learning, pattern recognition, statics, technology and direction system. Medical mining is one major application space wherever accuracy is very important. They've got observed an outsized kind of algorithms to perform data analysis tasks. A hybrid approach is to partition the information, avoiding the necessity to run algorithms on very large datasets. Different types of healthcare information set analysis in processing supported bunch algorithms hybrid k-means clustering approach (HKMCA) and PGA with their much dataset analysis and remove copy data and overcome existing technique issues. Clustering algorithmic program could also be a developed cluster scheme, autonomous of any initial circumstances and grants outstanding outcomes in terms of the whole of the square error traditional. Our projected genetic approach executes gets higher accuracy as compare to the HKMCA and show in results. Planned approach is improved information optimization and clusters size minimize quick supported time and more iteration however error minis and

similar time and it is show in results then gets the best solutions , less error clusters.

REFERENCES

[1]Waikato ML Group. User Manual Weka: The Waikato Environment for Knowledge Analysis. Department of Computer Science, University of Waikato (New Zealand), June 1997.

[2] Malwindersingh, Meenakshibansal ,î A Survey on Various KMeans algorithms for Clustering, IJCSNS International Journal of Computer Science and Network Security, VOL.15 No.6, June 2015.

[3] Mitchell, Tom M. "Machine learning and data mining." Communications of the ACM 42.11 ,1999.

[4] Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features." In European conference on machine learning, pp. 137-142. Springer, Berlin, Heidelberg, 1998.

[5] Watkins, Andrew, Jon Timmis, and Lois Boggess. "Artificial immune recognition system (AIRS): An immune-inspired supervised learning algorithm." Genetic Programming and Evolvable Machines 5.3 (2004): 291-317.

[6] Namrata S Gupta, Bijendra S Agrawal, Rajkumar M. Chauhan, ìSurvey On Clustering Technique of Data Mining, American International Journal of Research in Science, Technology, Engineering & Mathematics, ISSN:2328-3491.

[7] Narender Kumar, Vishal Verma, Vipin Saxena "Cluster Analysis in Data Minning Using K-mean Method", International Journal of Computer Applications, volume 76, No.- 12, August,2013.

[8] Verma, Manish, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, and Nidhi Gupta. "A comparative study of various clustering algorithms in data mining." International Journal of Engineering Research and Applications (IJERA) 2, no. 3: 1379-1384, 2012.

[9] Omran, Mahamed GH, Andries P. Engelbrecht, and Ayed Salman. "An overview of clustering methods." Intelligent Data Analysis 11, no. 6 (2007): 583-605.

[10] Xu, Rui, and Donald C. Wunsch. "Survey of clustering algorithms." ,2005.

[11] Kanungo, Tapas, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. "An efficient k-means clustering algorithm: Analysis and implementation." IEEE Transactions on Pattern Analysis & Machine Intelligence 7 (2002): 881-892.

[12] Moftah, Hossam M., Ahmad Taher Azar, Eiman Tamah Al-Shammari, Neveen I. Ghali, Aboul Ella Hassanien, and Mahmoud Shoman. "Adaptive k-means clustering algorithm for MR breast image segmentation." Neural Computing and Applications 24, no. 7-8 (2014): 1917-1928.

[13] Park, Hae-Sang, and Chi-Hyuck Jun. "A simple and fast algorithm for K-medoids clustering." Expert systems with applications 36, no. 2: 3336-3341, 2009.

[14] Zhang, Qiaoping, and Isabelle Couloigner. "A new and efficient k-medoid algorithm for spatial clustering." In International conference on computational science and its applications, pp. 181-189. Springer, Berlin, Heidelberg, 2005.

[15] Bandyopadhyay, Seema, and Edward J. Coyle. "An energy efficient hierarchical clustering algorithm for wireless sensor networks." In IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428), vol. 3, pp. 1713-1723. IEEE, 2003.

[16] Yuan, Yufei, and Michael J. Shaw. "Induction of fuzzy decision trees." Fuzzy Sets and systems 69, no. 2: 125-139, 1995.

[17] Zeng, Zhi-Qiang, Hong-Bin Yu, Hua-Rong Xu, Yan-Qi Xie, and Ji Gao. "Fast training support vector machines using parallel sequential minimal optimization." In 2008 3rd international conference on intelligent system and knowledge engineering, vol. 1, pp. 997-1001. IEEE, 2008.

[18] Srivastava, Jaideep, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. "Web usage mining: Discovery and applications of usage patterns from web data." Acm Sigkdd Explorations Newsletter 1, no. 2 (2000): 12-23.

[19] Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In Kdd, vol. 96, no. 34, pp. 226-231. 1996.

[20] Xie, Xuanli Lisa, and Gerardo Beni. "A validity measure for fuzzy clustering." IEEE Transactions on Pattern Analysis & Machine Intelligence 8 (1991): 841-847.

[21] Junatao Wang, XiaolongSu, "An Improved K-means Clustering Algorithm, Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on 27 may, (pp. 44-46), 2011.

[22] K. A. Abdul Nazeer, M. P. Sebastian,"Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

[23] Marty ,Babu, G.P. and M.N. 1994. Clustering with evolution strategies Pattern Recognition, 27, 2, 321-329.

[24] Pallavi Purohit "A new Efficient Approach towards k-means Clustering Algorithm" ,International journal of Computer Applications,Vol 65-no 11,march 2013.

[25] Shuhua Ren &Alin Fan "K-means Clustering Algorithm Based On Coefficient Of Variation" 2011 4th International Congress on Image and Signal Processing2011 IEEE

[26] Md Sohrab Mahmud, Md. Mostafizer Rahman, Md. Nasim Akhtar, " Improvement of K-means Clustering algorithm with better initial centroids based on weighted average", IEEE 7th International Conference on Electrical and Computer Engineering, 2012, pp. 647-650

[27] D. Arthur, S. Vassilvitskii, "k-means++: The advantages of careful seeding", Proceedings of the 18th annual ACM-SIAM symposium on discrete algorithms, pp. 1027–1035, 2007.

[28] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schroedl, "Constrained K-means Clustering with Background Knowledge", ICML Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577–584, 2001.

[29] Umutoni Nadine, Huiying Cao, Jiangzhou Deng, "Competitive Recommendation Algorithm for E-commerce", 2016 12th International Conference on Natural Computation, Fuzzy Systems, and Knowledge Discovery, (ICNC-FSKD) 978-1-5090-4093-3/16 , IEEE, 2016.

[30] Bruno Fernandes Chimieski1, Rubem Dutra RibeiroFagundes," Association and Classification Data Mining Algorithms Comparison over Medical Datasets", J. Health Inform. Abril-Junho; 5(2): 44-5, 2013.