International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Extended Advanced Method of Clustering Big data to achieve high dimensionality

N.SreeRam, M.H.M.Krishna Prasad, K.Satya Prasad

Abstract: ***Clustering is one of the relevant knowledge engineering methods of data analysis. The clustering method will automatically directly affect the result dataset. The proposed work aims at developing an Extended Advanced Method of Clustering (EAMC) to address numerous types of issues associated to large and high dimensional dataset. The proposed Extended Advance Method of clustering will repetitively avoid computational time between each data cluster object contained by the cluster that saves execution time in term. For each iteration EAMC needs a data structure to store data that can be utilized for the next iteration. We have gained outcomes from the proposed method, which demonstrates that there is an improvement in effectiveness and pace of clustering and precision generation, which will decrease the convolution of computing over the old algorithms like SOM, HAC, and K-means. This paper includes EAMC and the investigational outcomes done using academic datasets.***

Keywords : convolution of computing, Extended Advanced Method, high dimensional dataset, HAC, pace of clustering, precision generation, and SOM.

I. INTRODUCTION

Clustering is a practice that reorganizes data objects into their names of disconnect classes. It's an unsupervised learning. The exact answers are not known to the user in the unsupervised technique, whereas classification technique will assign classes to set of data objects. Initially for a given dimensional points in a given set, an operation performed will give the points difference in a set, we are needed to evaluate the centers of clusters in a way that, each of the points should fall within the same cluster, which are similar in points related to different cluster of dissimilar type. Majority of the techniques are developed using pattern matching or statically methods, the goal is to cluster the number of objects of data. Whereas in Data mining, the focus is to cluster larger datasets [1,2]. Developing various types of grouping method to the effect of generation of efficient cluster dynamically and rapidly for the growing various types of datasets has been an identical problem and a challenge. Different types of clustering methods were developed in solving clustering problems. Some of the mostly employed clustering algorithms are SOM, HCA and K-Means.

The basic K-mean Method will calculate each data object distance with all the centers of k- clusters for each time on every iteration which is executed, rather it takes large time for execution especially for high dimensional datasets. in this method cluster centers are generated arbitrary; it may not generate the required results. Many efficient of the K-Mean method relies on the centroid of the cluster. Initial cluster centroid will mainly rely on number of iterations and the running time of the K-Mean method. The execution complexity of the K-Means Method will be larger and won't provide quality clusters, when high dimensional datasets are used in clustering [3]. Kohonen SOM method is used to reduce or represent high dimensional datasets into lower dimensional space. This process is of reducing the dimensional vector a dynamic way of compression technique used in vector quantization. In added to Kohonen method has created a network information storage based on topological relationship with the training data set which is maintained. The main and interesting aspect in SOMs is to classify the data automatically without supervision. SOM consist of map units of neurons, which are having a position which is variant in dimensional space as well using discrete collection of data repeatedly and is presented to SOM topology for mapping from very high dimensional space to two-dimensional data output vector space. The reduced dimensional vector provides the property related to SOM, which is made easy for visualization of data. Each of the SOM are different in individual related to visualization of data, we should be careful in concluding the results which are drawn out from the datasets [4-7].

Network based clustering is either bottom up or top down. Bottom up methods assumes each of the documents as a single point of cluster and will merge all the pairs of clusters, such that all the clusters are merged into a single cluster which will be contained in single document. Bottom up network clustering is also called as HAC. In top down clustering, each of the clusters are spilling into small clusters. The process of splitting the clusters is done recursively into smaller clusters of each individually into various documents.

Manuscript published on November 30, 2019.
* Correspondence Author
N.Sreeram*, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India Research Scholar JNTUK, Kakinada A.P, India Email:sriramnimmagadda@gmail.com
Dr.M.H.M. Krishna Prasad, Professor of CSE, Vice-Principal & Coordinator-TEQIP-III,, University College of Engineering Kakinada(A) J.N.T.U. KAKINADA - 533003. Email: krishnaprasad_mhm@yahoo.co.in
Dr.kK. Satya Prasad, Professor of ECE (Rtd), ,, University College of Engineering Kakinada(A) J.N.T.U. KAKINADA - 533003.
Email: prasad_kodati@yahoo.co.in

This method is very much sensitive in generating outliers and at times difficult in identifying the accurate number of clusters from the Dendrogram. Numerous methods have been proposed from various literatures by various authors have analyzed SOM, K-Means & HCA fails when it comes to optimize high dimensional dataset clustering, because of its complexity and trends which makes it difficult, when larger dimensions are

added to the cluster. In the area of research, this type of problem is called "Dimensionality of Curse". Our work mainly deals with problem related to high dimensional datasets [10,11]. Various type of clustering algorithms was developed until now, each of the method will address various requirements and specification. We don't have a single method which can address all the concepts of solution the problems. They lack with some feature of adoptability, which is a greater challenge by the user in selecting an appropriate algorithm or method for solving a task. So, cope up with multiple problem, we propose a new advanced algorithm name Extended Advanced method of clustering.

II. PROPOSED EXTENDED ADVANCED METHOD OF CLUSTERING

Self-Organizing Maps is superlative clustering algorithm among K-means, Hierarchical Agglomerative clustering [12] as shown by experimental results. SOM project multidimensional data in reduced dimensional spaces, and SOM is non-deterministic and can produce different results in different run. For the shortcomings of the above SOM algorithm, this paper provides an innovative k-means clustering algorithm known as Extended Advanced d K-Mean Clustering Algorithm. The most important concept of the algorithm is to set two simple records to retain the labels of cluster and the distance of all the data objects to the nearest cluster during each iteration that can be used in next iteration. We calculate the gap between this object and also new cluster center, if the computed distance is smaller than or adequate to the gap to the recent center, the object stays in its cluster that was appointed to in prior iteration. Hence, there is no ought to calculate the distance from this data object to the other k-1 clustering centers, saving the accessing time to the k-1 cluster centers [13]. Otherwise, we must calculate the distance to all k cluster centers and find the closest cluster center. This point is assigned to the nearest cluster center and then distance to its center is recorded separately. Because some data points remain in the original cluster in each iteration, it means that some parts of the data points will not be calculated, saving a total distance calculation time, there by increasing the algorithm's efficiency.

A. Algorithm

The process of the Extended Advanced Method of Clustering is described as follows'

Input: The number of desired clusters K.
Dataset S.
D= {d1, d2,…dn} containing n data objects.
di={ x1, x2,…xm }// Set of attributes of one data point.
Output: A set of K clusters.
1.Draw multiple sub-samples {SI, S2, . . . ,Sj} from the original dataset.
2.Repeat step 3 for m=1 to n.
3.Apply combined approach for sub sample.
4.In each set, take the middle point as the initial centroid.
5.Compute the distance between each data point to all the initial centroids
6. For each data point find the closest centroid and assign to nearest cluster.
7. Choose minimum of minimum distance from cluster center criteria.
8. Now apply new calculation again on dataset S for K clusters.
9. Combine two nearest clusters into one cluster.
10.Recalculate the new cluster center for the combined cluster until the number of clusters reduces into k.

Fig. 1.EAMC algorithm

B. Computational Time

The proposed clustering method will initial the cluster first, then algorithms time complexity is calculated and stated as O(nk). Few data points in the original cluster are kept moving to other nearing clusters. As and if, points remain in the original cluster are not moved the complexity is O(k) else O(1). Based on the convergence based on the clustering methods, the movements of data points within the clusters are reduced. Based on the constraints, if one half of the points are moved from the cluster, its time complexity is O(n/2) [14], Therefore the whole complexity is O(nk). The time complexity of SOM method will not be known because, it gives different results on different type of system when it is run. Hence the proposed method will improve the clustering speed and will reduce the computational complexity.

C. Experimental Results

We have chosen academic dataset which are taken from the repository database of machine language for testing the efficiency of Extended Clustering Method and with the standardized available methods such as (HAC, SOM and K-Means). Two experiments of simulation are demonstrating in showing the performance of the extended Clustering method in our paper [15,16]. Clustering process is done on real datasets using WEKA DM tool. We have performed two experiments and their times of generation are calculated. The academic datasets are applied as input to clustering algorithms of standard and the proposed clustering method.

Comparisons are performed with the standard clustering methods to that of extended clustering method in measuring the accuracy and time execution on experimental basis. We have validated using Windows 8 operating system and Java Programming Language [17]. The proposed paper uses test datasets of academic activity and which provides a brief output from the gained experimental results. The characteristics of the datasets are shown in table 1 and the results of experiments are shown in figure 2-6.

Table- I: Datasets of Academic Data

DATA SETS	NO OF ATTRIBUTES	RECORD COUNT
ACADEMIC	16	5508

Table- II: Analysis between traditional (K-Means, SOM, HAC) and Advanced Method of Clustering

TYPE OF PARAMETERS	SOM	K-MEANS	HAC	AMC
ERROR RATIO	0.8287	0.8357	0.8469	0.3762
TIME OF EXECUTION	298ms	1271ms	1431ms	999ms
Access time	Fast	Slow	Slow	Very fast
No of clusters	6	6	6	8

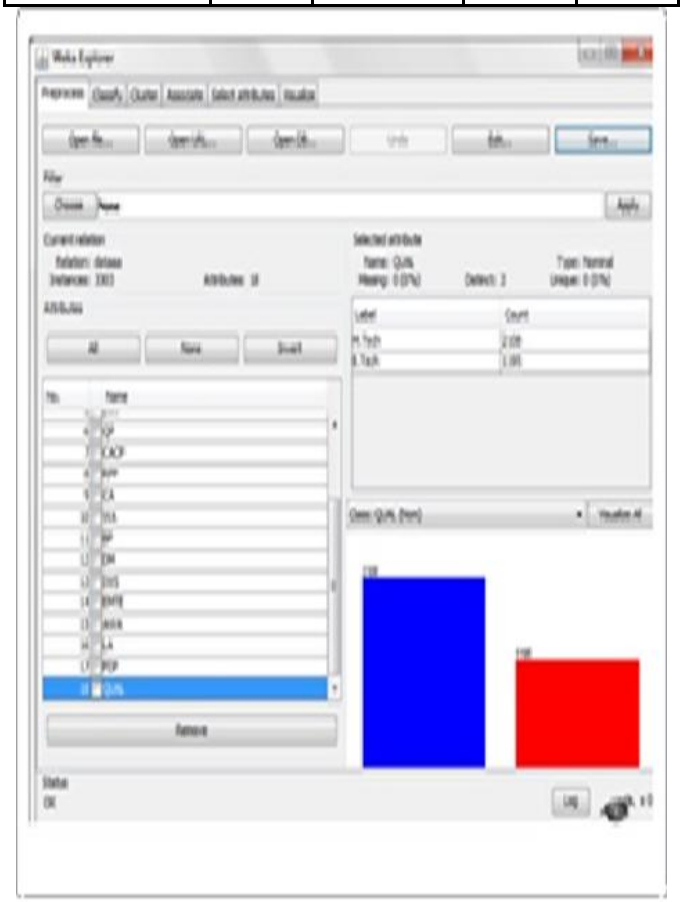

Fig. 2. Display Data Set According To Class Attributes

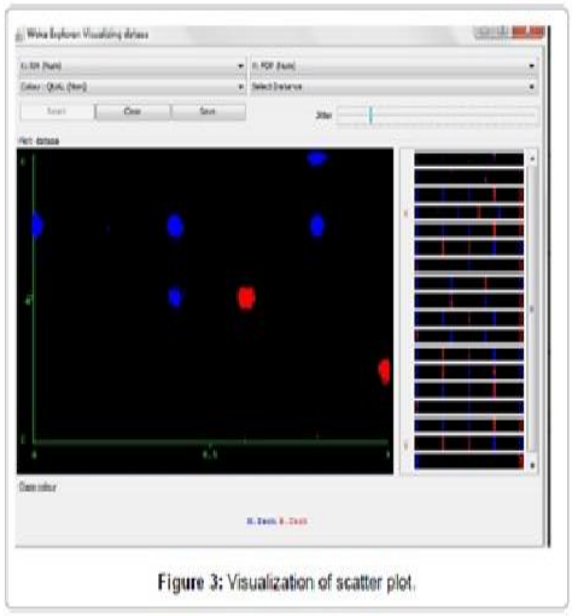

Fig. 3. visualization of scatter plots

Fig.4 . Performance comparison based on Error rate

The above figure shows the performance of proposed EAMC method in terms of error rate. By executing the proposed algorithm on academic data, it showed that the error rate of the proposed algorithm lesser than the traditional algorithms such as SOM, K-means and HAC. The experimental results also showed that the execution time and the number of clusters generated by the proposed algorithm are better than the other traditional algorithms which are showed in the following figures.

Fig. 5. Performance comparison based on Execution time

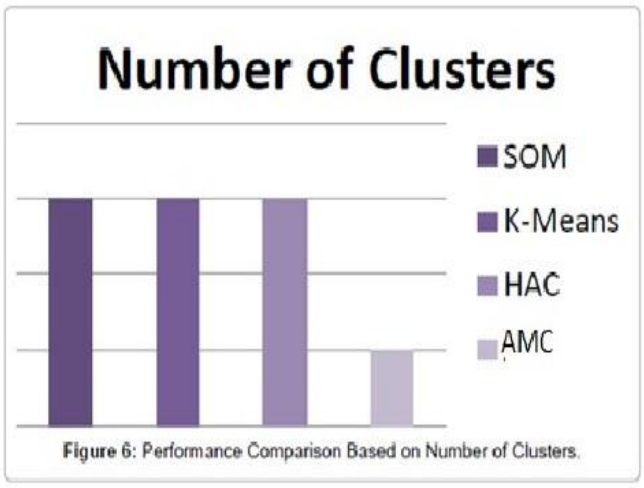

Fig. 6. Performance comparison based on number of clusters

III. CONCLUSION

The SOM method is a cluster algorithm which is used to cluster large datasets. The paper shows

Advanced method of clustering and analysis will overcome the standard clustering algorithms like SOM, HAC and K-Mean. Since the complexity of the standard clustering methods is very constraining provided that the data points are assigned to each iteration again which makes the efficiency of standard clustering methods very poor. Our work has proposed a method which will simply assigns the data points to the cluster in an efficient way. Our paper ensures that the clustering process takes O(nk) time without changes the clusters accuracy. The results gained from experiments show that EAMC will improve the computational time on high dimensional dataset. So proposed work is much feasible than other standard clustering methods.

REFERENCES

1. Yuan F, Meng ZH, Zhang HX, Dong CR (2004) A New Algorithm to Get the Initial Centroids. Proceedings of 2004 International Conference on Machine Learning and Cybernetics 2: 1191-1193.
2. Sun J, Liu J, Zhao L (2008) Clustering algorithms research. Journal of Software 19: 48-61.
3. Sun S, Qin K (2007) Research on Modified k-means Data Cluster Algorithm. Fine particles, thin films and exchange anisotropy. Computer Engineering, Jacobs IS, Bean CP editors, 33: 200-201.
4. ftp://ftp.ics.uci.edu/pub/machine-learning-databases.
5. Fahim AM, Salem AM, Turkey FA (2006) An efficient enhanced k-means clustering algorithm. Journal of Zhejiang University Science A 10: 1626-1633.
6. Zhao YC, Song J (2001) GDILC: A grid-based density isoline clustering algorithm. 2001 International Conferences on Info-tech and Info-net, 2001. Proceedings. ICII 2001-Beijing. 3: 140-145
7. Toor AK, Amarpreet S (2013) A Survey paper on recent clustering approaches in data mining. International Journal of Advanced Research in Computer Science and Software Engineering 3.
8. Toor AK, Amarpreet S (2013) Analysis of Clustering Algorithm based on Number of Clusters, error rate, Computation Time and Map Topology on large Data Set. International Journal of Emerging Trends & Technology in Computer Science 2: 94-98
9. Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery 2: 283-304.
10. AbdulNazeer KA, Sebastian MP (2009) Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. Proceeding of the World Congress on Engineering 1: 1-5
11. Fred ALN, Leitão JMN (2000) Partitional vs Hierarchical clustering using a minimum grammar complexity approach. Lecture Notes in Computer Science 1876: 193-202
12. Gelbard R, Spiegler I (2000) Hempel's Raven paradox: a positive approach to cluster analysis. Computers & Operations Research 27(4): 305-320.
13. Huang Z (1997) A fast clustering algorithm to cluster very large categorical data sets in data mining. Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Tucson, 146-151.
14. Ding C, He X (2004) K-Nearest-Neighbor in data clustering: Incorporating local information into global optimization. Proceedings of the 2004 ACM symposium on applied computing, 584-589
15. Hinneburg A, Keim D (1998) An efficient approach to clustering in large multimedia databases with noise. American Association for Artificial Intelligence 58-65
16. Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: An efficient data clustering method for very large databases. SIGMOD International Conference on Management of Data. Montreal, ACM Press, Canada, 103-114
17. Birant D, Kut A (2007) ST-DBSCAN: An algorithm for clustering spatial-temporal data. Data & Knowledge Engineering 60: 208-221.

AUTHORS PROFILE

N.SreeRam, working as Assistant Professor Department of CSE of KLEF Vijayawada, A.P, India , and a research scholar of JNTUK Kakinada A.P, India. Published 8 research papers in various international journals. Life member of Computer Society of India.

Dr MHM Krishna Prasad B.E., M.Tech, MIUR Fellow(U. of Udine, Italy), Ph.D., Professor of CSE, Vice-Principal and Coordinator-TEQIP-III, University College of Engineering Kakinada, J.N.T.U. KAKINADA – 533003 Andhra Pradesh, INDIA State Teacher Awardee from the Govt of Andhra Pradesh

Dr.K.Satya Prasad, worked as a professor of ECE University College of Engineering Kakinada, J.N.T.U. KAKINADA – 533003 Andhra Pradesh, INDIA