Big Data Clustering And Its Applications Examination

Md Zaheer Ahmed, C Mahesh

ABSTRACT:Clustering is a type of mining process where the data set is categorized into various sub classes. Clustering process is very much essential in classification, grouping, and exploratory pattern of analysis, image segmentation and decision making. And we can explain about the big data as very large data sets which are examined computationally to show techniques and associations and also which is associated to the human behavior and their interactions. Big data is very essential for several organisations but in few cases very complex to store and it is also time saving. Hence one of the ways of overcoming these issues is to develop the many clustering methods, moreover it suffers from the large complexity. Data mining is a type of technique where the useful information is extracted, but the data mining models cannot utilized for the big data because of inherent complexity. The main scope here is to introducing a overview of data clustering divisions for the big data And also explains here few of the related work for it. This survey concentrates on the research of several clustering algorithms which are working basically on the elements of big data. And also the short overview of clustering algorithms which are grouped under partitioning, hierarchical, grid based and model based are seenClustering is major data mining and it is used for analyzing the big data.the problems for applying clustering patterns to big data and also we phase new issues come up with big data

I INTRODUCTION

Clustering is included as the main problems in the data mining and also in machine learning. Clustering is a technique of finding the same groups of the objects.several researchers has curiosity in the maintaining clustering algorithms. Very important issue in clustering is that we don't have specific knowledge about the data which is considered and input parameters as number nearest neighbors, kn amount of clusters in clusters is the complicated process.One of the efficient ways of controlling with these information is to divide the data into a group of classes. The Big data can define the data sets which is having high measurements and also with heavy velocity, hence it is very complex to manage the data sets by applying conventional tools and their methods, due to the fast spreading of the information, here we need solutions which efficiently controls and extracting useful information from these data sets.. To overcome these issues, Big Data is clustered in a compact format that is still an informative version of the whole data. The clustering techniques are highly essential for the data mining. There are many approaches to mine the data like neural algorithms, support vector machines, association algorithms, genetic algorithms Among all these mining techniques clustering techniques producing high quality of clusters with combining the unlabeled data. Clustering is the grouping of data which depends on their similar properties. The aim of this paper is to provide many clustering algorithms for Big Data.

Revised Version Manuscript Received on 16 September, 2019.

Md Zaheer Ahmed, Assistant Professor, Computer Science & Engineering, Vidya Jyothi Institute of Technology,Hyderabad.

C Mahesh, Associate Professor, Computer Science & Engineering, Veltech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai.

II PROBLEM STATEMENT

Clustering is an unsupervised method. Which could not have particular column. at the time we did not know anything about the data we can adopt clustering method for a better understanding .. As length improves Distance measure like Euclidean will became a serious problem. Finding optimal Cluster is very challenging .its iterative process

According to the author [1] the main issue is that huge data set is limited to only numeric values because of the algorithm will minimize the cost, by calculating the clusters mean. Hence to sort this issue, here presenting a fast clustering algorithm which is used to group the categorical data.so for this reason, here we are making use of k nodes algorithm which is nothing but is an extension of k-nodes algorithm when we compare to other clustering techniques k-means clustering is suitable for data mining. The main power of the k-means algorithm in data mining applications is its efficiency in clustering large data sets.

According to the author[2] major problem statement called K-means clustering which has to be sensitive to the outliers, moreover it is quite efficient in the calculation time. For this reason, K-medoids clustering are sometimes used, where representative objects called medoids are taken into consideration apart from centroids. hence it is depending on the most centrally located object in a cluster, it is low sensitive to outliers when it is compared with the K-means clustering. Among the present algorithms for K-medoids clustering, partitioning around medoids (PAM) proposed by Kaufman and Rousseeuw (1990) is known as powerful one.

In this work [3] author focusing on the problem called as high quantity of spatial data (usually, terabytes) which is getting from satellite images, medical equipment, video cameras etc. is very expensive and few times it is unrealistic for users to enhance the data in an efficient way. Hence, a Spatial data mining is used which is to analyze a knowledge discovery process. Hence it performs a major role in these things a) extracting interesting spatial patterns and features; b) capturing relationships between spatial and non-spatial data; c) presenting data regularity concisely and at higher conceptual levels; and d) helping to rearrange spatial databases to accommodate data semantics, as well as to achieve best results according to the author[4] araised problem statement called as limitations which has formed in the region is the finding of clusters but the past work di not performs on this problem and also the reducing the i/o costs of data sets.to overcome this drawback, in this paper proposing a data clustering which is called as BIRCH ((Balanced Iterative Reducing and Clustering using Hierarchies), and it also explains that which will suit for the large databases. BIRCH incrementally clusters incoming multi-dimensional metric data points and it is trying produce the best quality clustering with the present sources (i. e., available memory and time constraints).

According to the author[5]Main issues over right here are all of the recognized algorithms which needs input parameters that are very hard to discover, an in destiny for many data sets there might not exist a global parameter for many datasetsfor many data sets there might not exist a global parameter putting as end result of the clustering set of rules and is the reason about the shape correctly. For this reason on this paper we show the manner how automatically and correctly extract not best 'conventional' clustering records (e. G. Consultant points, arbitrary fashioned clusters), however also the intrinsic clustering structure. Larger and large amounts of facts are collected and saved in databases growing the need for green strategies for using the greater information in the data..main advantage of our approach, when compared to the clustering algorithms proposed in the literature, is that we do not limit ourselves to one global parameter setting..

According to the author [6]working on the issue of finding clusters of points leads to the spatial point which comes in several applications.. therefore, here we propose the new clustering algorithm called as DBCLASD which is designed to satisfy the grouping of their needs. Basically DBCLASD is An incremental approach. A point should be given to a cluster which is processed incrementally one by one.the concept of a cluster is depends up on the distance of the cluster points to their nearest neighbors and then enhance the required output of these distances for a cluster. The analyzation in this paper is depends up on the statement the points inside of a cluster are uniformly distributed.

In this paper author[7] is working on the problem of fast technological progress, as the amount of data which is stored in databases increases very fast. The types of data which are stored in the computer become increasingly complex. and other problem is that The improvement of the existing algorithm is somewhat limited.. hence to overcome these issues, In this paper, seen a new algorithm to clustering in huge multimedia databases called DENCLUE (Density based Clustering). The motivation of our new models are (1) it has a firm mathematical basis, (2) it has good clustering elements in data sets with huge amount of noise.

In this research author[8] working on the problem of large spatial databases which is trying to identify populated regions in data mining. Hence to overcome this problem, a clustering approach should be introduced which is very efficient. so here proposing a Wave Cluster method, a novel clustering method which based on wavelet transforms,it will satisfies all the requirements as want. Using multi resolution of wavelet transforms, it can easily find arbitrary clusters at several groups. here also explains about that Wave Cluster is largely efficient in time complexity. This research is supported by Xerox organisation.

In this paper author[9] is working on the problem called as Clustering of large high-dimensional databases is with complex performance and system resource needs. Most of the current clustering algorithms phases serious issues like scalability and/or accuracy related problems when used on databases with more number of records and/or attributes. But Only the few methods can easily control o manage numeric, nominal, and mixed data. Hence to overcome these problems, here introducing clustering method called as the orthogonal partitioning clustering which has the capacity of clustering efficiently , high dimensional databases with numeric and nominal value have other algorithms also that are applicable to very huge number of databases. O-Cluster relies on an active sampling approach to achieve scalability with large volumes of data and requires at most a single pass for the database.

In this research author[10] concentrating on the problem called as which is fail to

identify the meaningful clusters because the real-world data sets are featured by a high dimensional, inherently sparse. Nevertheless, the data set sometimes containing the interesting clusters and those clusters are hidden in many sub spaces.. In this paper, introducing a SUBCLU (density-connected Subspace Clustering), it is an efficient way for solving the subspace clustering problem. as not analyzing any un required spaces, SUBCLU delivers for each subspace the related clusters DBSCAN would have found, when applied to this subspace separately.

According to the author[11] most difficult problem called as the diagnosis of fatal disease. As know that big data playing very important role in the health care .as the data mining techniques are used largely to for the problem of data analysis for identifying useful data from large data.. In this a large amount of data set is collected to form useful system. And one rough theory is used to discover the data and also can reduce the feature set which is presenting in the dataset. aim of the paper is at first the normal set theory is applied to the data set for explaining the data dependencies.

According to the author[12]the main problem definition of large data sets as all know the increase in the amount of data in the areas of biology, environmental, research , banking as many areas which manage huge data sets but is is becoming an issue of controlling huge data.

Hence to solve this problem in this The paper shows for many methods which is associated to several algorithms which is used to control huge data bases and it can also explains about the privacy concerns. different methodologies associated with different algorithms used to handle such large data sets. The main advantage of enhancing the data is to set the brand in market for satisfying the customer needs and demands also in each every way..

In this research author[13] is working on the problem called as in the GPS tracking field spreading the data which is related to their behaviors the data that relate to asset behaviors very hard to explain the asset usage distribution, operation, and the other major elements in the supply chain.. Mainly one essential thing here is the trip management where the every motion of the assets are can be tracked easily.

In this work author[14] focusing on the issue of clustering algorithms which are handling complex accuracy, reliability issues when it applies databases with huge number of records. But only some methods can control the numeric data and also the mixed data. Hence here to overcome this problem here proposing o-cluster algorithm. It has the capacity of clustering effectively and efficiently huge and dimensional data base with both numerical and the nominal values. The topdown partitioning technique shows tremendous scalability and the energy of the clustering solution

III RESEARCH ANALYSIS

here we have discussed so many clustering algorithms for the analysation of big data as it contains the huge amount of information, depends up on the particular requirement we are using specific algorithm. Here we have algorithms like o-clustering algorithm, grid based clustering algorithm, hybrid particle genetic optimisation, density connected sub space clustering and few more clustering algorithms. Each and every algorithm hastheir own priority in every aspects. And we seen merits and demerits associated to every paper.

AUTHOR	TITLE	RESEARCH METHODOLOGY	RESEARCH GAP
1.JHan, H Cheng, D Xin, X Yan -	Proceedings SIGMOD Workshop Res Issues Data Mining KnowledegeDiscovery; 1997	According to the author presenting a fast clustering algorithm which is used to combine the categorical data. Hence for this purpose here making use of knodes algorithm which is nothing but is an extension of k-nodes algorithm when compare to other clustering techniques k-means clustering is suitable for data mining. The major advantage of the k-means algorithm in data mining applications is its efficiency in clustering large data sets.	the main issue is the large data set is controlled to only numeric values because of the algorithm will minimise the cost, by calculating the clusters mean.
2.Hae-Sang Park, Chi-Hyuck Jun.	A simple and fast algorithm for K-medoids clustering	Here proposed K-medoids clustering are sometimes used, where representative objects called medoids are taken into consideration apart from centroids. Because it is based on the most centrally located object in a cluster, it is less sensitive to outliers when it is compared with the K-means clustering.	Here working on the problem statement called K-means clustering which is known to be sensitive to the outliers although it is quite efficient in terms of the calculational time.
3. Raymond T. Ng Department of Computer Science	Efficient and Effective Clustering Methods for Spatial Data Mining.	According to the author presented a clustering aIgorithm called as CLARANS which is depending up on randomized search and also developed two spatial data miningaIgorithms SD(CLARANS) and NSD(CLARANS). results shows that both algorithms are effective, and results to the discoveries which are complex to get with existing spatial data mining algorithms.	It has the problem called as high quantity of spatial data (usually, terabytes) which is getting from satellite images, medical equipment, video cameras etc. is very expensive and fewtimes it is unrealistic for users to enhance the data in an efficient way.
4. Tianza Computer science and engineering, Madison.	An efficient data clustering method for very large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data. 1996 Jun; 25(2):103–14.	in this paper presenting a data clustering method which is named as BIRCH ((Balanced Iterative Reducing and Clustering using Hierarchies), and it also explains that whcih will suit for the large databases. BIRCH incrementally clusters incoming multi-dimensional metric data points and it is trying to produce the best quality clustering with the present sources.	the problem statement here called as limitations which has occurred in the area is the discovery of clusters but the past work did not efficiently performs on this issue and also decreasing the i/o costs of datasets

5.MihaelAnkerst, Markus M. Breunig, Hans-Peter Kriegel.	Optics: Ordering points for finding the clustering structure. Proceedings of the ACM SIGMOD International Conference on Management of Data.	Here, proposed a cluster analysis method based on the OPTICS algorithm. OPTICS computes an augmented cluster-ordering of the database objects. The main benefit of our approach, when compared to the clustering algorithms proposed in the literature, is that do not limit ourselves to one global parameter setting. Apart from the augmented cluster which contains data which is same as thedensity based clusteringassociated to a high range of parameters and thus is a versatile basis for both automatic and interactive cluster analysis.	working on the issues like all of the known algorithms which needsinput parameters which are very hard to find, and in future for many datasets there may not exist a global parameter setting as result of the clustering algorithm which explains about the structure accurately
6Xiaowei Xu, Martin EsterHans-Peter Kriegel, Jörg Sander.	A distribution based clustering algorithm for mining in large spatial databases.	According to the author he propose the new clustering algorithm called as DBCLASD which is designed to satisfy the grouping of their needs. (Application Based Clustering Algorithms for Mining in Large Spatial Databases) Basically DBCLASD is An incremental approach. A point should be givento a cluster which is processed incrementaly one by one.	the problem of identifyting clusters of points belongs to the spatial point which arises in several applications.. therefore, here proposed the new clustering algorithm called as DBCLASD which is designed to satisfy the grouping of their needs
7. Hinneburg A, Keim DA	An efficient approach to clustering in large multimedia databases with noise. Proceedings ACM SIGKDD ConfKnowl Discovery Ad Data Mining (KDD);	Here author introduce a new algorithm to clustering in huge multimedia databases called DENCLUE (Density based CLUstEring). The basic idea of our new approach is to model the overall point density analytically as the sum of influence functions of the data points..	Here the problem of fast technological progress, as the amount of data which is stored in databases increases very fast. The types of data which are stored in the computer become increasingly complex. and issue is that The effectiveness of the existing algorithm is somewhat limited,
8. Sheikholeslami G, Chatterjee S, Zhang A.	Wave cluster: A multi resolution clustering approach for very large spatial databases. Proceedings Int Conf Very Large Data Bases (VLDB); 1998. p. 428–39.	In this paperproposing a Wave Cluster method, a novel clustering approach which is depending up on the wavelet transforms, which will satisfies all the requirements as want. Using multiresolution property of wavelet transforms, it can effectively identify arbitrary shape clusters at different degrees of accuracy. this also explains about that Wave Cluster is largely efficient in time complexity.	working on the problem of large spatial databases which is trying to identify populated regions in data mining
9.Boriana L. Van de Graff Drive Burlington,	.Clustering Large Databases with Numeric and Nominal Values Using Orthogonal Projections.	here introducing clustering method called as the orthogonal partitioning clustering which enables the clustering particulerly , high dimensional databases with both numeric and nominal value have other algorithms also that are applies to the huge databases, and a few that address high-dimensional data. Moreover huge part of the data in data warehouses is nominal, few of these algorithms have addressed clustering nominal data and even fewer mixed data.	the problem called as Clustering of large high-dimensional databases is with complex performance and system resource needs. Most of the current clustering algorithms phases serious issues like scalability and/or accuracy related problems when used on databases with more number of records and/or attributes.

10. Kailing K, Kriegel HP, Kroger P.	Density-connected subspace clustering for high-dimensionality data. Proceedings of the 2004 SIAM International Conference on Data Mining; 2010. p. 246–57.	In this paper making use of concept of subspace clustering which ahs seen at recent times, this aimsa t automatically discovering the subspaces of the feature space in where the clusters will exist. In this paper, introducing a SUBCLU (density-connected Subspace Clustering), it si an efficient way for solving the subspace clusteing problem.	The problem here is which is fail to find the meaningful clusters because temoe real-world data sets are featured by a high dimensional, inherently sparse. Nevertheless, the data set sometimes containing the interesting clusters and those clusters are hidden in various subspaces.
11.YasodhaP, Ananathanarayanan NR.	Analyzing Big Data to build knowledge based system for early detection of ovarian cancer. Indian Journal of Science and Technology. 2015 Jul; 8(14):1–7.	In this work author using The Hybrid Particle Genetic Swarm Optimization (PGSO) for minimizing the particular characteristics to classify the cancer, at normal or various stages of ovarian cancer. The main goal of the paper is at first the normal set theory is applied to the data set for explaining the data dependencies and then the hybrid particle is used to minimize the set feature reduction to classify the cancer tumours	Here trying to solve the problem called as the diagnosis of fatal disease.. as the data mining methods used largely to for the problem of data analysis for identifying useful data from large volumes of data.
12.Yadav C, Wang S, Kumar M	Algorithms and approaches to handle large data sets - A survey. International Journal of Computer Science and Network. 2013; 2(3):1–5	in this research work author explains about many methodologies which is related to the various algorithms which is used to manage huge data bases .The main advantage of enhancing the pdata is to set the brand in market for satisfying the customer needs and demands also in each every way	problem definition of large datasets as knows the increase in the amount of data in the areas of biology, environmental, research , banking like these had many areas.
13Qing Cao, BouchraBouqata, Patricia D. Mackenzie, Daniel Messier, Josheph J. Salvo Computing and Decision Sciences GE	A Grid-Based Clustering Method For Mining Frequent Trips From Large-Scale, Event-Based Telematics Datasets	According to the author here , a grid-based hierarchical clustering algorithm which is used to identify the trip patterns in high scale GPS datasets. At first record all the trips which is depending up on the grid indexing method, after that during the hierarchical clustering method ,only the trips which is shared similar neighbourhood are compared.	Here working on the problem called as in the GPS tracking field spreading the data which is related to their behaviors the data that relate to asset behaviors very hard to explain the asset usage distribution, operation.
14.Milenova BL, Campos M	Clustering large databases with numeric and nominal values using orthogonal projections. O Cluster; 2006. p. 1–11.	Here author proposing o-cluster algorithm. It has the capacity of clustering effectively and efficiently huge and high dimensional data base with both numerical and the nominal values. O-Cluster depends up on the active sampling method to gain reliability with high volumes of data.	the problems of clustering algorithms which will phase critical accuracy ndreliability issues when it used on databases with huge number of records..

IV RESULTS

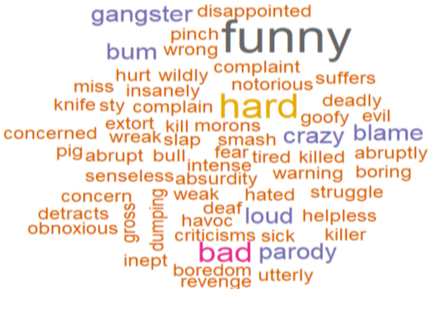

Figure: 1 Negative dictionary words

Fig.1 explain that negative words from proposing dictionary words these are many words from collecting negative reviews.

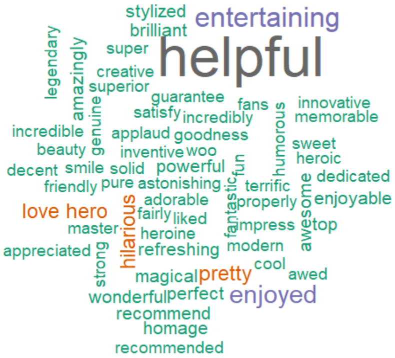

Figure: 2 positive words

Fig.2 explains that negative words from review from different websites like Face book, wattapp, and amazon, book my show etc.

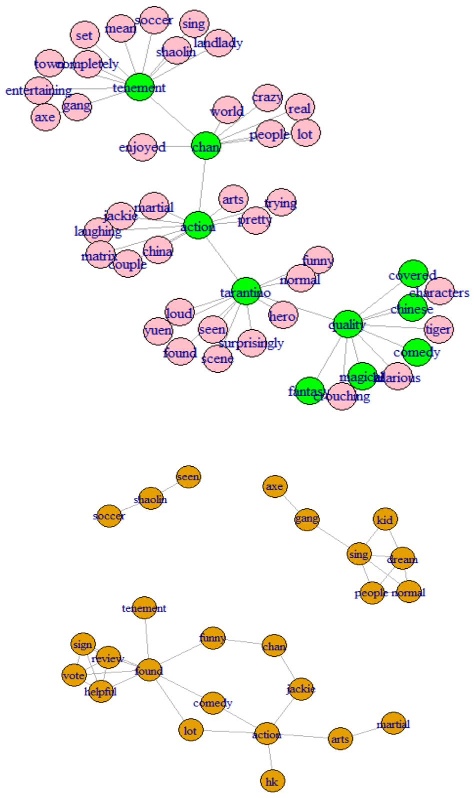

Fig: 3 negative and positive nodes

Fig: 4 frequently used words

Fig: 5 Sentiment Degree Dictionary

Consider an instance in real time in figure 5 below. The phrases in the blue font apply to phrases of feeling. The phrases in the yellow font apply to phrases in the degree of feeling. The light blue font phrases apply to expressions of negation. The white font phrases apply to the phrases of the combination. if (rating<2.0 and rating >=0)

{ Movie rating = 1 star }

if (rating>=2 and rating<3)

{ Movie rating = 2 star }

if (rating>=3 and rating<4)

{ Movie rating = 2.5 star }

if (rating>=4 and rating<5)

{ Movie rating = 3 star }

If (rating>=5 and rating<6)

{ Movie rating = 3.5 star}

Figure 5. Real time example

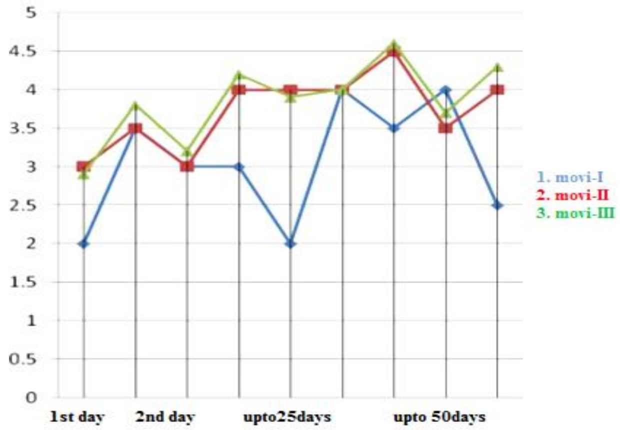

Fig: 6 output from experiment

Fig.5 and 6 explains that this is a rate prediction optimization which is obtain from R-studio data base.Fig.8 explains that rate prediction based on review from different reviewers, these are collected for rating prediction purpose.

parameter	Existing method	Proposed method	% change
Efficiency	65.54%	91.78%	26.24% increase

V CONCLUSION

By seeing all the papers as above , we have studied so many clustering algorithms which are presently used for analysing the big data cluster quality and their advantages and disadvantages are shown. After studying the above methods it is observed that some new techniques are also required for analysing the big data. as these methods are not efficient for the analyzation of data. in future enhancement, we can work on our research for the specific clustering domain and presenting new techniques for the exact clusters.

REFERENCES

1. Proceedings SIGMOD Workshop Res Issues Data Mining Knowledege Discovery; 1997JHan, H Cheng, D Xin, X Yan - Data mining and knowledge discovery, 2007
2. A simple and fast algorithm for K-medoids clustering. Expert Systems Applications.2009Mar
3. Efficient and Effective Clustering Methods for Spatial Data MiningRaymond T. Ng Department of Computer Science University of British Columbia Vancouver,B.C., V6T 124, Canada rng@cs.ubc.ca
4. An efficient data clustering method for very large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data. 1996 Jun; 25(2):103–14.Karypis G, Han EH, Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. IEEE Computer. 1999 Aug; 32(8): 68–75.
5. Optics: Ordering points to identify the clustering structure. Proceedings of the ACM SIGMOD International Conference on Management of Data. 1999 Jun; 28(2):49–60.
6. A distribution-based clustering algorithm for mining in large spatial databases. Proceedings 14th IEEE InternationalConference on Data Engineering (ICDE);Orlando, FL. 1998 Feb 23-27. p. 324–31.
7. Hinneburg A, Keim DA. An efficient approach to clustering in large multimedia databases with noise.Proceedings ACM SIGKDD ConfKnowl Discovery Ad Data Mining (KDD); 1998. p. 58–65
8. Sheikholeslami G, Chatterjee S, Zhang A. Wave cluster: A multi resolution clustering approach for very large spatial databases. Proceedings Int Conf Very Large Data Bases (VLDB); 1998. p. 428–39.
9. Clustering Large Databases with Numeric and Nominal Values Using Orthogonal Projections.
10. 10. Kailing K, Kriegel HP, Kroger P. Density-connected subspace clustering for high- dimensionalitydata. Proceedings of the 2004 SIAM International Conference on Data Mining; 2010. p. 246–57.
11. Yasodha P, Ananathanarayanan NR. Analyzing Big Data to build knowledge based system for early detection of ovarian cancer. Indian Journal of Science and Technology. 2015 Jul; 8(14):1–7.
12. Yadav C, Wang S, Kumar M. Algorithms and approaches to handle large data sets A survey. International Journal of Computer Science and Network. 2013; 2(3):1–5
13. A Grid-Based Clustering Method For Mining Frequent Trips From Large-Scale, Event-Based Telematics Datasets Qing Cao, BouchraBouqata, Patricia D. Mackenzie, Daniel Messier, JoshephJ. Salvo Computing and Decision Sciences GE Global Research Center One Research Circle, Niskayuna, NY 12309
14. Milenova BL, Campos M. Clustering large databases with numeric and nominal values using orthogonal projections. O Cluster; 2006. p. 1–11.

AUTHORS PROFILE

Mr Md Zaheer Ahmed has more than 10+ years of experience in the field of teaching. He was awarded M.Tech in Computer Science & Engineering from Jawaharlal Nehru Technological University, Hyderabad. Presently working as an Assistant Professor in Vidya Jyothi Institute of Technology,Hyderabad. He is a Research Scholar at Veltech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai. His area of interest includes Data Mining and Big Data.

Dr C Mahesh has more than 20 years of experience in the field of teaching. He was awarded B.E in Electrical and Electronics Engineering from Madras University, Chennai. He was awarded M. E in Computer Science, and Engineering, Anna University Chennai. He was awarded Doctorate in Computer Science and Engineering in the year 2016. Currently he is working as an Associate Professor in Veltech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai. His area of interest includes Neural Networks and Natural Language Processing.