IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-6, Issue-3 E-ISSN: 2347-2693

A Comprehensive Review of Privacy Preservation Framework using Birch and K-Means Algorithm

Prabhjeet Kaur[1*], Rekha Bhatia[2]

[1*]Department of computer science, PURCITM Mohali, Punjab, INDIA
[2]Department of computer science, PURCITM Mohali, Punjab ,INDIA

Email: Prabhgill59@gmail.com , r.bhatia71@gmail.com

Available online at: www.ijcseonline.org

Received: 17/Feb//2018, Revised: 22/Feb2018, Accepted: 19/Mar/2018, Published: 30/Mar/2018

ABSTRACT: Clustering is important part in Data mining. Clustering is a technique, in which data is using in the form of clusters. A set of objects divided into groups these groups called clusters. K-MEANS is a basic type of clustering technique. It is an unsupervised learning. K-means clustering is a simple technique, which is use to group items into k clusters. BIRCH is one of the famous methods, which used with the k-means to improve the quality of data, which are present in clusters. BIRCH is an (Balanced Iterative Reducing and Clustering using Hierarchies). Birch is a scalable clustering method, which mainly designed for very large data sets. In this paper we discussed about review of other clustering technique which are used by others researchers for data mining. We also discussed the limitations and applications of clustering techniques, which are most popular for data mining. This paper also represents a current review about the K-MEANS and BIRCH algorithm.

***Keywords*:** Data Mining, Clustring, K-Means Clustering, Birch Clustering

I. INTRODUCTION

Researchers have estimated that amount of information in today's world double for every 20 month. This information increased because of the increment in the use of computer hardware and computer software and the rapid computization of business, large amount of data has been collected and stored in database. Data stored in database give us large amount of information about what type of data is used. Text mining is used to separate the high quality information from different text.[1] And it is also extract the information from hidden patterns in large data collections.

However, raw data cannot use directly for taking any type of information. Clustering algorithms used on the raw data and then form clusters of data, which gives needed information. Search Engines used to form a group of similar and dissimilar type of objects in different clusters. Clustering plays an important role in search engines to collect similar type of data easily.

Data mining is one of the youngest used activities, which used to extract only Interesting factors. Clustering algorithm can be identifying the cancerous data records. It used to find the data of people that they are included in which data the person is either cancerous or non-cancerous [2].

Handling the records of student is very difficult process. By clustering, it is easy to form clusters of different information like class, scores, name all these are from different datasets. These datasets grouped in different clusters by using K-MEANS and other algorithms.

Clustering, classification, regression and association are different patterns used for data mining process.

Stages of clustering in data mining:

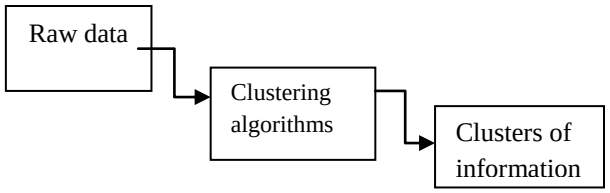

FIG.1

Rest of the paper is organized as follows, Section I contains the introduction of the current issues of clustering in data mining. Section II contains the algorithm of clustering which used in research. Section III contains the research review of clustering techniques, which already used; section IV contains the research gap between the studies of clustering algorithms. Last section V Conclude the result of research.

II. CLUSTERING

Clustering is a sub area of mining the data streams since the algorithms use to grouping the data into classes or groups so that objects within a class or a group have same high type of data. Which connected to one another but these objects are

very dissimilar to the objects that present in other groups or cluster of object, according to few relationship metrics.

A. Categories of Clustering Algorithms:
Any specific algorithm cannot solve clustering problems but it requires other algorithms to solve a particular problem. Every algorithm is use to solve a some problem like time complexity, small data clusters, different type of data sets in one clustering. All these problems are solving with the help of clustering algorithms.

Clustering categorized on the basics of what type of data is used. It is dividing in two types:

- Partitioning Based Clustering
- Hierarchical Based Clustering
- Density based Clustering
- Grid based Clustering
- Model based Clustering

K-MEANS CLUSTERING:

This algorithm is one of the simplest and high-speed execution algorithms. K-MEANS clustering used in many areas like retrieval of information; recognize the patterns of datasets, which used in clusters. An algorithm of unsupervised learning that is called K-means algorithm. Handling the records of student is very difficult process. By clustering, it is easy to form clusters of different information like class, scores, name all these are form different datasets. These datasets grouped in different clusters by using K-MEANS and other algorithms.[3]

This algorithm used to solve the clustering. K-means clustering is a simple technique, which is use to group items into k clusters. An iterative method assigns points to those clusters whose centers are nearest. In globular clusters, K-MEANS produce tighter clusters.
1) Randomly select *'c'* cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum from all the cluster centers.
4) Recalculate the new cluster center using.

BIRCH ALGORITHM:

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

Birch is a scalable clustering method, which designed for very large data sets. Only a single scan of data is required [4]. It is based on notation of CF (clustering features) a CF tree. It can handle spiral type clustering. In n number objects, the complexity is O (n).

Clustering features tree (CF) depends on the:

- Number of data cluster(N)
- Linear sum of data cluster(LS)
- Square sun of data clusters(SS)

Basic Algorithm:

- STEP 1: Load data into memory
- STEP 2: Condense data
- STEP 3: Global clustering
- STEP 4: Cluster refining

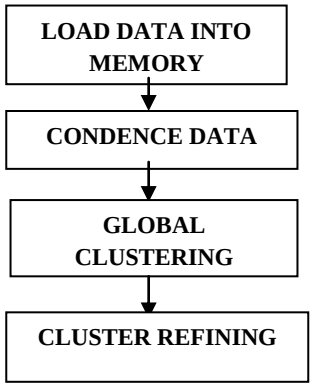

FIG 2.1 BIRCH ALGO

III. REVIEW OF LITERATURE

Various authors in the past have done studies related to data mining, clustering, k-means and BIRCH Algorithm. Few studies reviewed by me about my research proposal are as follows:

Christopher and Divya[5] focused on detection of outliers in data streams. According to them, it is a fundamental issue in data mining. They tried to analyze the performance of CURE with CLARANS and CURE with K-Means in order to detect outlier .they underwent through various performance measures for detecting outliers. The results signified that the efficiency of detection of outliers was more in case of CURE when combined with CLARANS then in case of CURE with K-Means.

Kaur et al. [6] studied clustering algorithms in data mining. The paper concluded that K-Means is a better method than K-Mediods type it is sensitive to noise whereas K-Mediods is not. BIRCH shows good reuse in case of large and small databases whereas CLARANS is better for small databases

only. In case of non-spherical clusters, CURE is best in handling large databases whereas ROCK is better in case of Boolean and categorical data variables. The author concluded that sCHAMELEON is best for all types of data and it forms non-spherical clusters.

Dr. Vijayarani et al. [7] explained that in data stream environment many challenges were faced when using clustering in data streams and detection of outlier. This research performed on detecting the outlier by using combination of different algorithms with K-means. Both CURE with K-means gave a better result.

Hendrik Fichtenberger et al.[8]explained that data stream algorithm for the k-means problem called BICO, that combines the data structure of the SIGMOD Test of Time award winning algorithm BIRCH with the theoretical concept of co-sets for clustering problems. The k-means problem asks for a set C of k centers minimizing the sum of the squared distances from every point in a set P to its nearest center in C. We compare BICO experimentally with popular and very fast heuristics (BIRCH, Mac Queen) and with approximation algorithms (Stream-KM++ [2], Stream LS) with the best-known quality guarantees. We achieve the same quality as the approximation algorithms mentioned with a much shorter running time, and we get much better solutions than the heuristics at the cost of only a moderate increase in running time.

Shi Na et al.[9]found that clustering analysis method is one of the main analytical methods in data mining, the method of clustering algorithm will influence the clustering results directly. This paper discusses the standard k-means clustering algorithm and analyzes the shortcomings of standard k-means algorithm, such as the k-means clustering algorithm has to calculate the distance between each data object and not all cluster centers in iteration make the efficiency of clustering is high. This paper proposes an improved k-means algorithm in order to solve this question, requiring a simple data structure to store some information in iteration, which used in the next iteration. The improved method avoids computing the distance of each data object to the cluster centers repeat, saving the running time. Experimental results show that the improved method can effectively improve the speed of clustering and accuracy, reducing the computational complexity of the k-means.

Kumar J. and Sinha B.K. [10]found out that privacy regulations and other privacy concerns used for Data analysis. Data analysis prevents data owners, which used for sharing the information. These data owner always gives good results for privacy preservation and accurate clustering. After removing the data owner a new approach Vector quantization used for, solve the problem. This approach used for piecewise on datasets, which are divided datasets of each row into segments. Some results presented which tries to find the optimum value and these quantization parameters performed between data privacy and clustering utility to take good results.

Vadiya and Clifton[11]presented the method for k-means clustering where the existence of different sites, attributes but common entities. They implied that cluster of each entity is known by each site but not about the attribute. The authors focused on reducing communication cost with ensured reasonable privacy. The paper tried to combine data mining algorithms and provided a direction to implement them with privacy and other information based disclosure properties. The authors hoped for a standard method of development in data mining to ease the task of privacy preserving.

Tian Zhang et al[12] analyzed that for processing of large data sets data clustering do not address the problem with limited amount of sources which are used in clustering to address the problem. To solve the problem BIRCH (balanced iterative reduction and clustering hierarchies) is used which deal with the real problems. For these problems, following solutions are used:

- Building a pixel classification tool (iterative and interactive)
- Generating the initial codebook for image compression

Raghu Ramakrishnam et al.[13] analyzed that clustering not handle the large data sets. Many algorithms are used to cure this problem then it faces many other problems with high dimensional like noise in data. To solve this problem BIRCH algorithm is used which deal with the multi-dimensional metric data points to produce the best quality clustering with the available resources. BIRCH deals with the following problems:

- It can improve the quality of data in one scan.
- More improve the quality in additional scans.

Kedar B. Sawant[14] analyzed that in Partitioned clustering selection of random centers is very difficult and this clustering technique is only select single portion of data instead of full clustering structure. K-MEANS algorithm used to overcome this problem but not all problems are handles by K-MEANS so this paper gives a modified K-MEAN algorithm to reduce this find this iterative problem and improves the time complexity.

Sachin Shinde et.al[15] examined that to search and read a single research paper consumed two to three hours per paper. To overcome the time complexity problem traditional K-MEANS used but this algorithm not efficient .This paper gives improved K-MEANS algorithm which is used to making algorithm very efficient and it will reduce the time complexity to find a research paper.

Rajesh N.et.al[16]analyzed that privacy preservation play an important role in data mining. Mined data should be secure. Many techniques like Randomization. For privacy purpose K-anonymity, technique is used. In this p researchers give privacy preservation with fuzzy logics, neural network learning which overcome the problems.

Raj bala[17]analyzed that every algorithm in clustering is not efficient and accurate for mining the data from data set clusters. It is comparative analysis of K-MEANS and density based algorithm. In this research researcher, conclude that K-means is best for efficiency and accuracy.

Maneesh Upmanyu et.al[18] analyzed for privacy preservation protocol for K-MEANS algorithm. For sharing personal information of any organization, privacy is major part. In this, we use the paradigm of secret sharing in which data divided into multiple shares and processed separately at different servers. This paradigm gives solution of negligible communication in cloud computing. This process is million times faster than other used protocols for privacy preservation.

Ipsa De[19]analyzed that data mining is very difficult when datasets are distributed between two different parties and more difficult when both parties do not want to share their data sets with each other. This researcher gives a solution of this problem by using privacy preservation two party hierarchical clustering algorithms over vertically partitioned datasets.

Jinfei Liu[20]analyzed DBSCAN algorithm is used for that data-sets which are distributed between two parties which gives a results of using data without inform the parties which is not private. In this paper researcher, address the problem of privacy preservation DBSCAN. In this, two protocols for privacy preservation of DBSCAN over horizontally or vertically partitioned data. In this paper researcher, also provide the analysis of performance of protocols and give the privacy proof of the solution.

R. Sasikala[21] analyzed that privacy preservation has become most popular in data mining. In this paper researcher, introduce a new K-Anonymity algorithm, which used to transform a K-Anonymity datasets into K-Anonymity datasets. This algorithm used to transform a table so that no one can make high probability association between records in the table. In this, each entity is different from at least k-1 records.

Dr. T. Christopher[22]analyzed that data streams clustering is helpful to cluster the similar type of data items in data streams. Outlier detection is major issue in data mining. Clustering algorithms used for outlier detection. However, data streams in clustering are new researching area in data mining. This refers to the process of extracting knowledge from fast growing data.

Tapas Kanungo[23]This researcher present a new algorithm of K-MEANS clustering is Lloyd's algorithm which is call filtering algorithm. This algorithm is using for distortion in BIRCH algorithm and find the nearest centre.

B.S.Sangeetha[24]proposed a research based on searching latest research papers easily by using K-MEANS clustering algorithm. This researcher gave an arranged research papers. This is less time consuming and gives update information of latest research paper.

T.Vijaya Kumars et.al[25] proposed a graph based clustering based on web usage pattern. In which it first session uses time oriented approach, which based on sessions and page requests. Then its data points are generated which is generated by CHAMELEON ALGORITHM.

P.Prabhu et al[26]proposed a improved complexity in high dimensional data-sets which are used in K-MEANS clustering algorithm. It improve the efficiency and accuracy in high dimensional data by using dimensional reduction method such as Principal component Analysis (PCA)

TABLE 3.1

List of research paper discussed

S. No.	Author Name	Extracted features	Clustering technique
1.	Christopher and Divya	Detection of outliers	• CURE with K-MEANS • CURE with CLARAS
2.	Kaur et al	Noise reduction	• K-MEANS • K-MEDIODS • BIRCH
3.	Dr. S. Vijayarani et al	Detection of outliers with data streams	CURE with K-MEANS
4.	Hendrik Fichtenberger et al	Time complexity	BIRCH

5.	Shi Na et al	Centers repeat	Analysis of K-MEANS
6.	Vadiya and Clifton	Privacy preservation	K-MEANS
7.	Tian Zhang et al.	Large datasets	BIRCH
8.	Kumar J. and Sinha B.K	Privacy preservation	Quantization approach with K-MEANS
9.	Raghu Ramakrishnam	Large datasets	BIRCH
10.	Kedar B. Sawant	select random centers	K-MEANS
11.	Sachin Shinde	Time complexity	K-MEANS
12.	Rajesh N.et.al	Privacy preservation	• FUZZY LOGIC • NEURAL NETWORK
13.	Raj Bala	Efficiency Accuracy	• K-MEANS • DENSITY BASED
14.	Maneesh Upmanyu et.al	Privacy preservation	• K-MEANS • SECRET SHARING PROTOCOL
15.	Ipsa De	Privacy between two hierarchical parties	HIERARCHICAL CLUSTERING
16.	Jinfei Liu	Privacy preservation	DBSCAN
17.	R. Sasikala	Privacy in tables	New K-Anonymity algorithm
18.	Dr. T. Christopher	Knowledge extraction	Data streams
19.	Tapas Kanungo	Exact centre distortion	Lloyd algorithm
20.	B.S.Sangeetha et.al	Time complexity	K-MEANS
21.	T.Vijaya Kumars et.al	Clustering of web uses	CHAMELEON ALGORITHM
22	P.Prabhu et al	High dimensional data-sets	K-MEANS

IV. RESEARCH GAP

In 1997, Tian Zhang and his other researchers used a clustering BIRCH algorithm, which is only useful for large data sets. Then In 2003, Vadiya and Clifton named researcher used K-means algorithm in privacy preservation but this algorithm only handle small-scale type of data. Then many researchers used many clustering algorithm for privacy preservation in data mining, time complexity and for large datasets in data mining. Nevertheless, no algorithm used for handling privacy preservation in data mining for large data sets. BIRCH algorithm only handles the large data sets and numeric data and result in BIRCH algorithm is very difficult to modify again once the merging or splitting decision made while In K-MEANS algorithm only handles the privacy preservation in small segments of data and k-means cannot handle the data in different sizes and shapes. K-means algorithm shows outliers when the different execution on the same data is always different. In this research, we show a new result by using K-MEANS and BIRCH algorithm for privacy preservation large datasets. In which birch used for large data sets and K-means gives privacy.

CURE	Numeric	Arbitrary	O(n^2 log N)	Yes
BIRCH	Numeric	Convex	O(n)	Yes
CHAMELEON	Numeric	Arbitrary	O(n^2)	No

V. CONCLUSION

Table 4.1
Comparison of different clustering Algorithms:

Cluster algorithm	**Data type handling**	**Cluster shape**	**Complexity**	**Handling high dimensional data**
K-MEANS	Numeric	Convex	O(nkt)	No

In this paper, we have discussed different clustering approach of K-MEANS and BIRCH algorithm. Different researchers search the applications, limitations, advantages of these clusters algorithms in their work that done previously. In this review, paper we observed that lot of improvement has been making in K-MEANS and BIRCH algorithms in past years. Most work done improved accuracy, time complexity and reliability of clusters. Setting of initial numbers of clusters and different type of clusters is always a challenging task in K-MEANS and BIRCH Algorithm. Hence, this paper concludes that yet

there is still so much to explore and research regarding these algorithms.

REFERENCES

[1] Shraddha Shukla and Naganna (2014)S, "A Review ON K-means DATA Clustering APPROACH" International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1847-1860

[2] Wang, X. Y., & Garibaldi, J. M. (2005, June) "A comparison of fuzzy and non-fuzzy clustering techniques in cancer diagnosis" In Proceedings of the 2nd International Conference in Computational Intelligence in Medicine and Healthcare, BIOPATTERN Conference, Costa da Caparica, Lisbon, *Portugal* (Vol. 28).

[3] Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." IEEE transactions on pattern analysis and machine intelligence 24.7 (2002): 881-892.

[4]Rafsanjani, Marjan Kuchaki, Zahra Asghari Varzaneh, and Nasibeh Emami Chukanlo. "The Journal of Mathematics and Computer Science." TJMCS Vol .5 No.3 (2012) 229-240

[5] Christopher, T and Divya, T. (2015). "A Study of Clustering Based Algorithm for Outlier_Detection in Data streams". *International Journal of Advanced Networking and Applications*, Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications, 194-197.

[6]Kaur, S., Chaudhary S., Bishnoi N. (2015). "A Survey: Clustering Algorithms in Data Mining." *International Journal of Computer Applications*, (0975- 8887), 12-14.

[7]Vijayarani S., Jothi P.,(2014). "Hierarchical and Partitioning Clustering Algorithms for Detecting Outliers in Data Streams". *International Journal of Advanced Research in Computer and Communication Engineering*, 3(4), 6204-6207.

[8] Fichtenberger H., Gillé M., Schmidt M., Schwiegelshohn C., Sohler C. (2013) "BICO: BIRCH Meets Coresets for *k*-Means Clustering. In: Bodlaender H.L., Italiano G.F. (eds) Algorithms". ESA 2013. Lecture Notes in Computer Science, vol 8125. Springer, Berlin, Heidelberg

[9]Na S., XuminL.__,Yong G.(2010), "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm". *Third International Symposium on Intelligent Information Technology and Security Informatics*, pp.63-67

[10] Kumar J. and Sinha B.K. (2010), "Privacy Preserving Clustering In Data Mining", *A thesis submitted for degree of bachelor in Technology in Computer science and engineering of* Department of Computer Science and Engineering ,National Institute of Technology Rourkela

[11] Vaidya J.,Clifton C. (2003). "*Privacy-Preserving K -Means Clustering over Vertically Partitioned Data".* KDD-2003 Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. SIGKDD. 206-215

[12] Zhang, T., Ramakrishnan, R. & Livny, M. "BIRCH: A New Data Clustering Algorithm and Its Applications" Data Mining and Knowledge Discovery(1997) , Volume 1, Issue 2, pp 141–182

[13]Raghu Ramakrishnam Zhang, "BIRCH: A New Data Clustering Algorithm and Its Applications" Data Mining and Knowledge Discovery(1997) , Volume 1, pp 141–182

[14] Kedar B. Sawant Shree Rayeshwar "international journals of advances in management and enginerring sciences, volume 4, issue 6(1) January 2015, PP 22-27 ISSN 2349-4395(Print) & ISSN2349-4409(Online)

[15] *Prateeksha Tomar, Amit Kumar Manjhvar, "*Clustering Classification for Diabetic Patients using K-Means and M-Tree prediction model*", International Journal of Scientific Research in Multidisciplinary Studies , Vol.3, Issue.6, pp.48-53, 2017.*

[16]Rajesh N. "Survey on Privacy Preserving Data Mining Techniques using Recent Algorithms". International Journal of Computer Applications (0975 – 8887) Volume 133 – No.7, January 2016

[17] Ashaq Hussain Bhat, Puniethaa Prabhu, "*OTU Clustering: A window to analyse uncultured microbial world*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.6, pp.62-68, 2017

[18]Maneesh Upmanyu, Anoop M. Namboodiri, Kannan Srinathan, and C.V. Jawahar, "Efficient Privacy Preserving K-Means Clustering". Pacific-Asia workshop on intelligence and security informatics ,PAISI 2010 : Intelligence and security informatics pp 154-166

[19]Animesh Tripathy1, Ipsa DE1, "Privacy Preserving Two-Party Hierarchical Clustering Over Vertically Partitioned Dataset" A Journal of Software Engineering and Applications, 2013, 6, 26-31

[20] Jinfei Liu, Li Xiong, Jun Luo, Joshua Zhexue "Privacy Preserving Distributed DBSCAN Clustering". Transactions of data privacy vol 6 issue 1, april 2013 page 69-85

[21] R. Sasikala T. Bhuvaneswari ,Assistant Professor Department of Computer Science and Engineering ,Sankara College of Commerce and Science

[22] Dr. T. Christopher, "A Study of Clustering Based Algorithm for Outlier Detection in Data streams" Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications.pp194-197

[23] Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." IEEE transactions on pattern analysis and machine intelligence 24.7 (2002): 881-892.

[24]B.S.Sangeetha,K.N.Nithya,N.Suganya Devi, A.Shyamala Gowri, "An Innovation approach in searching of research paper using text clustering with feature selection" IJMDRR E-ISSN-2395-1885 ISSN -2395-1877

[25] T.Vijaya Kumar, Dr. H.S.Guruprasad "Clustering of Web Usage Data using Hybrid K-means and PACT Algorithms" BIJIT - BVICAM's International Journal of Information Technology

[26] P.Prabhu "Improving the performance of K-means clustering for high dimensional data sets." International Journal on Computer Science and Engineering (IJCSE)

Authors profile:

Miss Prabhjeet kaur, completed B.tech (CSE) from Rayat Bahra group of college of enginerring for women kharar in 2016. She is currently pursuing M.tech in CSE from PURCITM Mohali. Her interested area is data mining and cloud computing.

Dr. Rekha Bhatia, is currently working as Associate Professor in department of computer science in PURCITM Mohali. She received her Ph.D degree from Punjabi university Patiala. Her Research area is artifical intelligence and information security. She has published many papers in national and international journals.

