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 Preserving Framework Using Wavecluster and K-Means Algorithm

Manjot Kiran Kaur Bedi[1*], Rekha Bhatia[2]
1*Department of computer science, PURCITM Mohali, Punjab, INDIA
2. Department of computer science, PURCITM Mohali, Punjab ,INDIA

Available online at: www.ijcseonline.org
Received: 17/Feb//2018, Revised: 23/Feb2018, Accepted: 18/Mar/2018, Published: 30/Mar/2018

ABSTRACT- A process of partitioning a set of data (or objects) into a set of significant sub-classes, called clusters. Also can be said as unsupervised classification which has no predefined classes. K- Mean is a type of unsupervised learning, which is used when we have any data without defined class or groups. The goal of this algorithm is to find the number of groups in the data, and represented by the variables K1, K2, up till KN. A wavelet based clustering approach for spatial statistics on very huge data. This is a grid based approach which applies wavelet transform in the quantized trait space and then senses the dense section in the transformed space. This paper discusses about the review of the clustering techniques which are observed and used by other numerous researchers for data mining. Further, this paper discusses about the advantages and limitations of the clustering techniques. As based on previous researchers' contribution that k-mean alone can't be efficient enough for the increased data set. So with time improved versions were introduced and combining two or more techniques for data mining clustering was practiced. This paper overcame the limitations and found more efficient way for clustering.

KEYWORDS: Data Mining, Clustring, K-Means Clustering, Wave Clustering

I. INTRODUCTION

The drawing out hidden information from large databases is an influential new technology with great potential to help companies focus on the most significant information in their data warehouse. Data mining is prepared for use in the companies because it is supported by three technologies that are now adequately grown-up:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Clustering - A process of partitioning a set of data (or objects) into a set of significant sub-classes, called clusters. Also can be said as unsupervised classification which has no predefined classes.

Application of clustering- Numerous of applications can be seen in this area of data mining namely:

- Market Research
- Document Classification
- Pattern Recognition
- Spatial Data Analysis
- Image Processing

Categories of clustering Methods-

- Partitioning algorithms: Constructing numerous partitions and then assess them by some decisive factor.
- Hierarchy algorithms: Creation of hierarchical breakdown of the set of data using a number of measures.
- Density-Based: Categorized on the basis of connectivity and density functions.
- Grid-Based: Categorized on the basis of levels granularity structure.
- Model-Based: A model is hypothesized for each of the cluster and the thought is to find the best fit of that is the representation to each other.

WaveCluster- A wavelet based clustering approach for spatial statistics on very huge data. According to Gholamhosein Sheikholeslami, Surojit Chatterjee, Aidong Zhang: this is a grid based approach which applies wavelet transform in the quantized trait space and then senses the dense section in the transformed space. The multi-resolution property of wavelet transform makes it happen that WaveCluster can distinguish the clusters at unlike scales and stages of aspect, which results extremely helpful in the user's applications.

K-Mean Algorithm - A type of unsupervised learning, which is used when we have any data without defined class

or groups. The goal of this algorithm is to find the number of groups in the data, and represented by the variables K1, K2, up till KN.

Steps for the K-Mean Algorithm:-

1. Decide on a value for k.
2. Initialize the k-cluster centers (aimlessly, if essential).
3. Run the k-Means algorithm on the level i illustration of the data
4. Use last centers from level i as first centers for leveli+1. This is attained by projecting the k centers returned by k-Means algorithm for the 2i space in the 2i+1 space.
5. If not any of the N objects altered membership in the last iteration, exit. Otherwise go to 3.

ADVANTAGES:

WAVECLUSTER ALGORITHM:

Noises get detached from the original feature space. In n number objects the complexity is O (n). Without difficulty handle numeric type data. Perceive the clusters at different scales and stages of aspect.

K-MEANS ALGORITHM:

The numeric type clusters it can handle short datasets and neighbor clusters. It also handles spherical clusters. Simple as compared to other techniques.

APPLICATIONS:

Land Use: Recognition of the region with same pattern of usage of land in a terrain study database.

Insurance- Identifying the typically high cost claimers which are entitled into various insurance policy holders.

City-Mapping- Grouping the houses as per their house build, cost, and size.

Marketing- Helps market researchers to find out distinctive groups in their customer bases, and then use this information to build up targeted promotion programs.

II. REVIEW OF LITERATURE

Various researchers in the past have done studies related to data mining, clustering, k-means and WaveCluster Algorithm. Few studies reviewed are as follows:

Jitendra Kumar and Binit Kumar Sinha [1] concluded that: "large amount of detailed personal data was regularly collected and shared, but this proved to be beneficial for data mining application. Information includes shopping habits, criminal records, medical history, credit records, etc. Such data have been always vital plus point to companies and governments for decision making. Alternatively, privacy policy and further privacy concerns may prevent data owners from sharing their personal data for data analysis. In order to share information while protecting the privacy of the data one need to find the appropriate solutions which attain the dual goal of privacy protection in addition to precise clustering result."

Michail Vlachos, Jessica Lin, Eamonn Keogh and Dimitrios Gunopulos [2] performed incremental clustering of time-series at various resolutions using wavelet transform. This study found that this approach yields faster execution time and clustering quality.

Shruti Dalmiya, Avijit Dasgupta and Soumya Kanti Datta [3] team used Wavelet Based technique along with K-mean algorithm for the large set of MRI images of mammograms. Which resulted in the wavelet transformation made the algorithm noise free as wavelets provided frequency information as well as time-space localization.

Rafal Ladysz [4] Clustering of evolving time series data- an effort to optimize the figure and early positions of cluster centroids for k-means clustering with consider to righteousness and exactness using the Euclidean and Dynamic Time Warping distance measures.

Kondra, Janardhan Reddy [5] Density-based clustering algorithm which employs DBSCAN theme exclusive of producing a clustering the data set openly, but as a stand-in, it generates an amplified ordering of that exacting database which stand for its density-based clustering arrangement.

Bikash Sharma and Aman Jain [6] Data mining services require precise input data for their results to be noteworthy, but privacy concerns may stress users to make available fictitious data. To protect customer privacy in the data mining procedure, techniques supports on a casual collection of data records are used. The randomization algorithm is the best option so that the collective properties of the data can be recovered with adequate correctness, while entity entries get considerably distorted.

Kaur, S., Chaudhary S., Bishnoi N. [7] Studied about the clustering methods in data mining. The paper purposed the different techniques of clustering as per the requirements and choosing an exact algorithm making desirable improvements. They concluded that k-means were better in results than k-mediods.

Vaidya J., Clifton C. [8] stated that privacy preserving data mining was designed for all kinds of research industry. They did use k-means algorithm when numerous of sites

contained various attributes for a common set of entities. Where the sites learned the clustering of each entity, but did not state anything about the attributes of the rest of the site.

Sachin Shinde and Bharat Tidke [9] stated that improved K-mean is more accurate and has more efficiency than the original k-mean. They proposed better initial centroids.

Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu [10] stated that the algorithm they used show how to compute the nearest center. Algorithm scales up in high dimensions.

Christopher and Divya [11] spotlighted the order to detect the outliers. They concluded that k-mean alone can't improve the efficiency it needs to be combined and they did combine it with CURE and CLARANS.

Shi Na et al. [12] found that the clustering is best analytical way in clustering. There paper reviewed the shortcoming of k-means algorithm and gave a future scope that it can be improved. They concluded that improved method will affect the speed of computing distance.

Kedar B. Sawant[13] analyzed that selection of random centers is very difficult. K-mean is the solution for this problem. Modified k-mean is used to reduce iterative problems and further enhances the time complexity.

Priti Maheshwari and Namita Srivastava [14] stated that a signal into different frequency sub bands. The method is used to compress data or use signals to extract features.

Kalaivani. R* Dr. R. Manicka Chezhian [15] stated that clustering is also useful in web mining. No crisp boundaries are required for mining. Clustering help in efficient delivery on the web.

Dr. S. Vijayarani, Ms.P.Jothi [16] performed data outlier detection using clustering. The analysis required clustering accuracy and outlier detection accuracy.

Gholamhosein Sheikholeslami, Surojit Chatterjee, Aidong Zhang [17] stated that much different management application requires spatial data and it should be insensitive towards noise. This makes it highly efficient for time complexity.

Ling Chen, Ting Yu and Rada Chirkova [18] WaveCluster with Differential Privacy they showed technique on synthetic data generation. Preserving data is required with less random noise.

Ahmet Artu Yıldırım and Cem Özdoğan [19]] adopted master-slave model and replicated approach to increase the performance. This method did improved the execution input and output time. The model approach did helped to find the limitations and improve them in future.

K. Chitra and Dr. D.Maheswari [20] paper helped to see clear comparison between different clustering technique and which can be used for the particular application. Hence using wave cluster and k-mean is opted for privacy preservation in this paper.

III. TABLE

List of research paper discussed

S.No	Author Name	Extracted features	Clustering technique
1.	Jitendra Kumar and Binit Kumar Sinha	Privacy preservation	Quantization approach with K-MEANS
2.	Michail Vlachos, Jessica Lin, Eamonn Keogh and Dimitrios Gunopulos	Faster execution time	Incremental clustering of time-series at various
3.	Shruti Dalmiya, Avijit Dasgupta and Soumya Kanti Datta	Noise free	Wavelet Based technique along with K-mean
4.	Rafal Ladysz	Effort to optimize the figure and early positions	K-means clustering
5.	Kondra, Janardhan Reddy	Amplify the order exacting database	DBSCAN
6.	Bikash Sharma and Aman Jain	Preserve customer privacy	K-Means
7.	Kaur, S., Chaudhary S., Bishnoi N.	Faster Execution	K-Means
8.	Vadiya and Clifton	Privacy preservation	K-MEANS
9.	Sachin Shinde and Bharat Tidke	Enhanced K-Mean	Improved K-Mean

10.	Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu	Llyod's K-Mean	K-Mean
11.	Christopher and Divya	Detection of outliers	K-Mean, CURE and CLARANS
12.	Shi Na et al	Centers repeat	Analysis of K-MEANS
13.	Kedar B. Sawant	Random centers	K-MEANS
14.	Priti Maheshwari and Namita Srivastava	Feature Extract	WaveCluster
15.	Kalaivani. R and Dr. R. Manicka Chezhian	Web Mining	Clustering
16.	Dr. S. Vijayarani, Ms.P.Jothi	Outlier Detection	K-mean
17.	Gholamhosein Sheikholeslami, Surojit Chatterjee, Aidong Zhang	Spatial Approach	WaveCluster
18.	Ling Chen, Ting Yu and Rada Chirkova	Data Preserving	WaveCluster
19.	Ahmet Artu Yıldırım and Cem Özdoğan	Parallel Clustering	WaveCluster
20.	K. Chitra and Dr. D.Maheswari	Comparison between clustering methods	k-mean, wavecluster, cure, clarans, etc.

IV. RESEARCH GAP

In past years various techniques being used by scientist and researchers to get efficient outcomes for the clustering algorithms.

In 2002 many researchers studied the improved or enhanced k-mean algorithm. In which they improved the selection of center in K-means. But still there was need for more efficiency in results as database increased with the time.

Further in the year of 2003 researchers found the algorithm application in privacy preservation. And they found the way to secure the large data.

Like this researchers improved the techniques in the past years but faced many limitations. K-Means and WaveCluster deal with some detailed limitations like they both are number dependent on wrong value can cause the issues of wrong clustering and may outlier the useful data. WaveCluster need to be accurate which makes the process bit complex as noise need to be eliminated depending on the K-means outlier results. Empty cluster is a day to day problem in K-means. It creates empty clusters which consist of no data and software may pick random values. Outlier represents data set but it is not useful in the cluster results. It is important to remove them for better results.

V. CONCLUSION:

This paper does discuss the approaches of K-means and WaveCluster algorithm for clustering. Numerous researchers concluded the various advantages, application and limitations of these clustering algorithms. A lot of improvement can be observed from the past work till now. Removing the noise from the data for better clustering and to have an efficient output. So, this paper does reach to conclusion that there is a lot to explore and research regarding the algorithm and their respective outcome patterns.

REFERENCES

[1]. Jitendra Kumar and Binit Kumar Sinha ID CODE-1789, Department of Computer Science (NIT ROURKELA, ODISHA) Privacy Preserving Clustering in Data Mining

[2]. Michail Vlachos, Jessica Lin, Eamonn Keogh and Dimitrios Gunopulos Computer Science & Engineering Department University of California - Riverside A Wavelet-Based Anytime Algorithm for K-Means Clustering of Time Series.

[3]. Shruti Dalmiya, Avijit Dasgupta and Soumya Kanti Datta International Journal of Computer Applications (0975-8887) Application of Wavelet based K-means Algorithm in Mammogram Segmentation.

[4]. Rafal Ladysz FINAL PROJECT PAPER for INFS 795 CLUSTERING OF EVOLVING TIME SERIES DATA

[5]. Kondra, Janardhan Reddy Privacy Preserving Optics Clustering ID CODE-8539, Department of Computer Science (NIT ROURKELA, ODISHA)

[6]. Bikash Sharma and Aman Jain Privacy preserving data mining ID CODE -4218, Department of Computer Science (NIT ROURKELA, ODISHA)

[7]. Kaur, S., Chaudhary S., Bishnoi N. (2015) a survey: Clustering Algorithm in Data Mining. International Journal of Computer Applications, (0975-8887), 12-14.

[8]. Vaidya J., Clifton C. (2003) Privacy Preserving K-Means Clustering Over Vertically Partitioned Data. SIGKDD. 206-215.

[9]. Sachin Shinde et al (2003) Improved K-means Algorithm for Searching Research Papers, International Journal of Computer Science & Communication Networks,Vol (6),197-202

[10]. Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu (2002) An Efficient k-Means Clustering Algorithm:Analysis and Implementation IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 7, july 2002.

[11]. Christopher and Divya (2005) 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.

[12]. Na S., XuminL., Yong G.(2010), Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm, IITSI '10 Proceedings of the 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 63-67.

[13]. Kedar B. Sawant Shree Rayeshwar (2015) Institute of Engineering and Information Technology IT Department, Shiroda-Goa Volume 3, Issue 1, 2015, ISSN 2349-4395 (PRINT) & ISSN 2349-4409 (ONLINE).

[14]. Priti Maheshwary et al. (2011) International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397 Vol. 3 No. 2

[15]. Kalaivani. R and Dr. R. Manicka Chezhian (2013) A Competent Data Set Grouping in Clustering Algorithms Volume 3, Issue 8, August 2013 ISSN: 2277 128X

[16]. Dr. S. Vijayarani, Ms.P.Jothi (2014) Partitioning Clustering Algorithms for Data Stream Outlier Detection. International Journal of Innovative Research in Computer and Communication Engineering. ISSN (Online): 2320-9801

[17]. Gholamhosein Sheikholeslami, Surojit Chatterjee, Aidong Zhang WaveCluster: a wavelet-based clustering approach for spatial data in very large databases. The VLDB Journal (2000) 8: 289–304

[18]. Ling Chen, Ting Yu and Rada Chirkova (2015) WaveCluster with Differential Privacy Department of Computer Science, North Carolina State University, Raleigh, USA.

[19]. Ahmet Artu Yıldırım and Cem Özdoğan Parallel WaveCluster: A linear scaling parallel clustering algorithm implementation with application to very large datasets J. Parallel Distrib. Comput. 71 (2011) 955–962

[19]. K. Chitra and Dr. D.Maheswari (2017) International Journal of Computer Science and Mobile Computing ISSN 2320–088X

Authors profile:

Miss Manjot Kiran Kaur Bedi, completed B.tech (ECE) 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.

