International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 9, September-2018

Improved Privacy Preservation with Clustering and Cryptographic Technique in Data Mining

Jeetendra Mittal, Dr. Akash Saxena (Prof.)

Department of Computer Science
Compucom Institute of Technology & Management, Jaipur, India

Abstract*—Data Mining is the way toward separating knowledge escaped substantial volumes of raw data .The knowledge must be new, not obvious, and one must be able to use it. The primary data is altered by the disinfection procedure to hide sensitive knowledge before discharge so the issue can be addressed. Privacy preservation of delicate knowledge is tended to by a few specialists in the form of association rules via suppressing the frequent item sets. Clustering is technique which makes cluster of useful objects which have resemble characteristics. Anonymization is to protect the identity of the individual this encrypt identifiers like unique number and the name whereas the data which is not encrypted provides less or no guarantee. Advanced Encryption Standard (AES) is an algorithm to provide the security to the data and it is very difficult to apply attacks. By the proposed work, privacy preservation of the data increased and it can be shown with the help of the results. AES provide the result in minimum time which show that propose produce result faster than existing approaches.

Keywords*—PPDM, Anonymization, Hierarchical Clustering, Data Encryption Standard, Advanced Encryption Standard.

I. INTRODUCTION

Privacy preserving data mining (PPDM) is categorize into different classes. We will survey the essential ideas of PPDM & distinctive studies performed in the region of PPDM under different classes. We will focus on measurement that are utilized to calculate the side effects came about because of privacy preserving procedure [1]. We will talk about heuristic based algo. Though several diverse methodologies are utilized to secure critical information in the today's organized environment, these techniques often fail [2]. One approach to make information less susceptible is to organize intrusion detection system (IDS) in basic PC framework. In case of PC framework is organized an early recognition is the key for recuperating lost or harmed information without much difficulty. In current years, analysts have proposed a assortment of methodologies for expanding the intrusion detection efficiency & exactness [3]. However, a large portion of these efforts focused on identifying interruptions at the system or operating system level. They are not equipped for distinguishing malicious data corruptions, i.e., what specific information in the database are controlled by which particular malicious database transaction(s). Without this info, quick harm evaluation & recovery can't be accomplished.

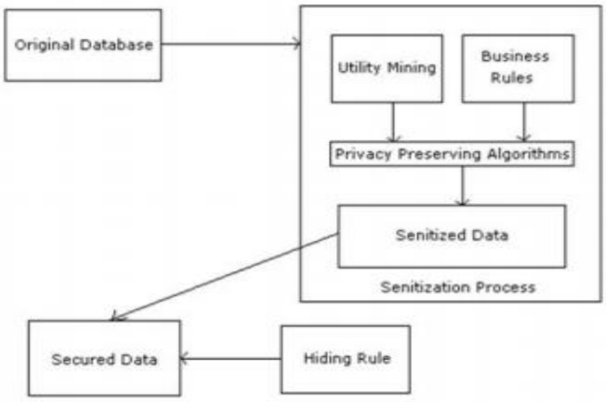

Fig.1 Block diagram of privacy preserving data mining technique

II. ANONYMIZATION BASED PPDM

The basic form of the data in a table consists of following four types of attributes: (i) Explicit Identifiers is an arrangement of attributes containing data that recognizes a record owner clearly for example, name and SS number etc. (ii) Quasi Identifiers is an arrangement of attributes that could possibly distinguish a record proprietor when joined with publicly accessible information. (iii) Sensitive Attributes is a set of attributes that contains sensitive person specific information such as disease, salary etc. (iv) Non-Sensitive Attributes is an arrangement of properties that makes no issue if uncovered even to conniving gatherings . Anonymization alludes to an approach where character or/and delicate information about record proprietors are to be covered up. It even assumes that sensitive data should be retained for analysis. It's clear that unambiguous identifiers ought to be expelled yet at the same time there is a risk of security interruption when semi identifiers are connected to openly accessible data. Such attacks are called as linking attacks. For example attributes such as DOB, Sex, Race, and Zip are available in public records such as voter list.

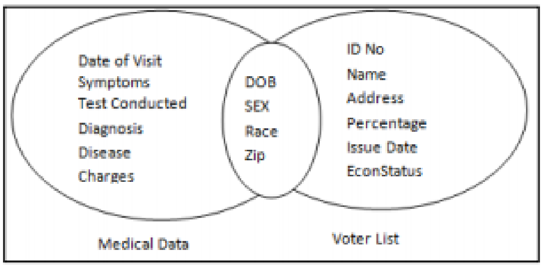

Fig.2 Linking Attack

These records are accessible in medicinal records likewise, when connected, can be utilized to induce the personality of the relating individual with high possibility as shown in fig.2 [4].

III. CLASSIFICATION OF CLUSTERING

Clustering is the main task of DM & it is done by the number of algo. The most ordinarily utilized algo in clustering are hierarchical, partitioning, density & grid based algo.

a) Hierarchical algorithms

Hierarchical clustering is a technique of cluster analysis which looks to construct a chain of clusters. It is the connectivity based clustering algo. The hierarchical algo make clusters slowly. Hierarchical clustering commonly divided into two sorts i.e. in hierarchical clustering the data are not partitioned into a specific cluster. It takes a progression of segments, which may run from a solitary cluster containing every objects to n clusters each containing a solitary object. Hierarchical clustering is partitioned into agglomerative techniques, which continue by arrangement of combinations of the n objects into gatherings & troublesome strategies, which isolates n objects progressively into better groupings.

a) Advantages of hierarchical clustering

- Embedded adaptability in regards to the level of granularity.
- Ease of dealing with any types of comparability or separation.
- Applicability to any attributes sort.

b) Disadvantages of hierarchical clustering

- Vagueness of execution criteria.
- Most hierarchal algo do not return to once built clusters with the reason of development [5].

IV. DATA ENCRYPTION STANDARD

DES was the consequence of an research project setup via international business machines (IBM) corporation in the late 1960‟ s which brought about a cipher known as LUCIFER. The changed edition of LUCIFER was advanced as a proposition for the new national encryption standard ask for via the national bureau of standards (NBS). It was lastly received in 1977 as the data encryption standard (DES). DES depend on a cipher called as the feistel block cipher. This was a block cipher created via the IBM cryptography researcher horst feistel in the mid 70‟s. It consists of a number of rounds where each round contains bit-shuffling, nonlinear substitutions (S-boxes) and exclusive OR operations. Once a plain-text message is customary to be encoded, it is organized into 64 bit squares required for input. In the event, that the quantity of bits in the message isn't equitably distinct by 64, after that the last square will be padded [6] [7]. DES plays an original permutation on the whole 64-bit square of information. It is then divided into 2, 32 bit sub-squares, Li & Ri which are then passed into 16 adjusts (the subscript i in Li and Ri demonstrate the current round). Every rounds is indistinguishable & the impacts of expending their number are double - the algo, security is expended & its transient effectiveness diminished. Clearly, these are two conflicting outcomes and a compromise must be made. For DES the number picked was 16, most likely to ensure the removal of any connection among the cipher text & either the plaintext or key. Toward the finish of the sixteenth round, the 32 bit Li & Ri yield amounts are swapped to make what is known as the pre-outcome. This [R16, L16] connection is permuted utilizing a capacity which is the correct reverse of the primary permutation. The output of this final permutation is the 64 bits cipher text [8] [9].

V. LITERATURE SURVEY

Mina Sheikhalishahi et al. [2017] This paper displayed a structure for building a hierarchical categorical clustering algo on horizontal & vertical dividation databanks. It is accepted that data is dispersed among two groups, such that for general advantages both will are identify the clusters on entire databanks, however for privacy concerns, they decline to share the primary databanks. To this end, we propose algorithms based on secure weighted average protocol and secure number comparison protocol, to securely compute the desired criteria in constructing clusters' scheme [10].

Qingchen Zhang et al. [2016] This paper proposed a high-order PCM algo (HOPCM) for huge data clustering via optimizing the target function in the tensor space. Furthermore, we outline a disseminated HOPCM strategy in view of map-reduce for a lot of heterogeneous information. Lastly, we devise a privacy-preserving HOPCM algo (PPHOPCM) to ensure the private information on cloud by applying the BGV encryption plan to HOPCM, In PPHOPCM, the functions for refreshing the enrollment framework and grouping focuses are approximated as polynomial capacities to help the safe figuring of the BGV conspire. Experimental results demonstrate that the PPHOPCM can efficiently cluster an expensive number of heterogeneous information utilizing cloud computing without exposure of private information [11].

Kai Xing et al. [2016] This paper proposed the issue of common security assurance in social participatory detecting in which people contribute their private data to construct a (virtual) network. Especially, we propose a mutual privacy preserving k-means clustering plan that neither reveals person's private data for release the community's characteristic data (clusters). Our plan contains two privacy-preserving algo called at every cycle of the k-means clustering. The first one is working by every member to discover the closest cluster while the cluster centers are held mystery to the members & the second one registers the cluster centers without releasing any cluster center info to the member while keeping every member from making sense of different individuals in a similar cluster. A broad execution investigation is completed to demonstrate that our approach is successful for k-means clustering, can oppose complicity attacks & can give mutual privacy protection even when the data analyst colludes with all except one participant [12].

Zakaria Gheid et al. [2016] In this paper, proposed a novel privacy-preserving k-means algo in view of a straightforward yet secure & productive multiparty additive method that is cryptography-free. We composed our explanation for horizontally partitioned data. Additionally, we show that our plan opposes against adversaries passive model [13].

Jiawei Yuan et al. [2016] This paper proposed a practical privacy-preserving K-means clustering plan that can be proficiently outsourced to cloud servers. Our plan permits cloud servers to perform clustering straightforwardly over encoded databanks, while accomplishing practically identical complexity & precision compared with clusterings over decoded ones. We also examine secure integration of map-reduce into our plan, which makes our plan amazingly appropriate for cloud computing environment. Thorough security analysis & numerical analysis complete the execution of our plan as far as security & efficiency. Experimental evaluation over a 5 million objects dataset further validates the practical performance of our scheme [14].

Vadlana Baby et al. [2016] in this paper they proposed an efficient distributed threshold privacy-preserving k-means clustering algorithm that use the code based threshold secret sharing as a privacy-preserving method. development include code based approach which enables the information to be partitioned into numerous offers and handled independently at various servers. Our protocol takes less number of iterations compare with existing protocols and it do not require any trust among the servers or users. We additionally furnish explore results with examination & security investigation of the proposed plan [15].

VI. PROPOSED METHOD

Advanced Encryption Standard (AES) is an algo in which key block cipher utilize a solitary key to encode & decode the data for both the sender and receiver. In spite of the fact that, the block length of Rijndael can be 192, 256, or 128 bits, the AES algo just received the block length of 128 bits. For the time being, the key length can be 192, 256, or 128 bits. The AES algo internal task are performed on a two dimensional cluster of bytes known as state & every byte comprises of 8-bits. The state comprises of 4 lines of bytes & every line has Nb bytes. Every byte is signified by Si, j ($0 \leq i < 4$, $0 \leq j < Nb$). Since the block length is 128 bits, each line of the state contains Nb = 128 / (4 x 8) = 4 bytes. The 4-bytes in every segment of the state array form a 32-bit word, with the line number as the index for the four bytes in every word. Toward the start of encryption or decryption, the array of input bytes is mapped to the state array, accepting a 128-bit block can be communicated as 16 bytes: in0, in1, in2, … in15. The encryption/ decryption are executed on the state, toward the finish of which the last esteem is mapped to the output bytes array out0, out1, out2, … out15. The input of the AES algo can be mapped to 4 rows of bytes comparatively, with the exception of the quantity of bytes in each line meant by Nk can be 4, 6, or 8 when the length of the key, K, is 128, 192, or 256 bits, individually. The AES algo is an iterative algo. Every cycle can be known as a round. The aggregate number of rounds, Nr, is 10 when Nk = 4, Nr = 12 when Nk = 6, and Nr = 14 when Nk = 8.

In the initial step, we get the data from the database in which the operations can be performed. The overall dataset has been fetched from the database and divides into group which is known as clusters. For the clustering procedure, hierarchical clustering have been utilized. The items are grouped by calculating the distance among the items and if the data has minimum distance then the clusters merge. Then new and large cluster has been formed and update the distance of the items. Now the data has been encrypted using AES Algorithm where it contains various operations for the key generation and then the optimal results obtained.

Proposed Algorithm:

```
Step:1    Start
Step:2    Input dataset from the database
Step:3    Apply Hierarchical clustering
a.        Compute distance between dataset
b.        Put items into cluster
c.        If Distance between two clusters is min
d.        Then merge both cluster
e.        Update distance
Step:4    Apply AES Algorithm over the output
a.        Key selection
b.        Generation of multiple key
c.        Encryption
d.        Decryption
Step:5    Get optimal result
Step:6    Exit
```

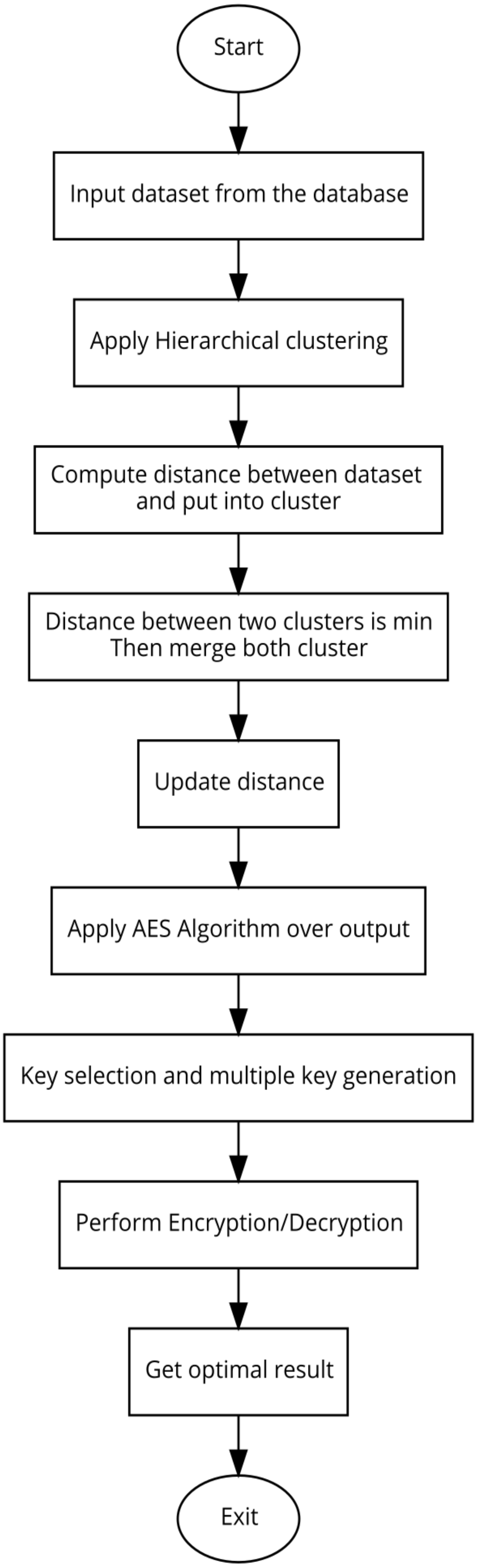
Start
Input dataset from the database
Apply Hierarchical clustering
Compute distance between dataset and put into cluster
Distance between two clusters is min Then merge both cluster
Update distance
Apply AES Algorithm over output
Key selection and multiple key generation
Perform Encryption/Decryption
Get optimal result
Exit

Fig.3 Flowchart of Proposed Work

VII. Result Analysis

The simulation of the proposed work has done with MATLAB 2013. There are two graphs demonstrated below which show that the proposed technique has better accuracy and less error rate. Time graph also demonstrated below which is lesser than the existing techniques.

Table I: Comparison of Elapsed Time among Base and Propose Techniques

No. of records	100	200	300	400	500
Base	34.12	69.99	105.46	139.56	174.15
Propose	1.58	2.16	3.70	5.41	7.56

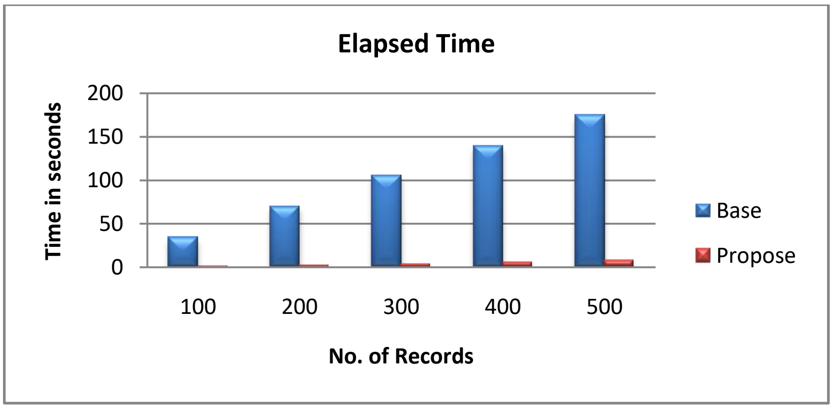

Fig. 4 Elapsed Time among the base and propose appraoch

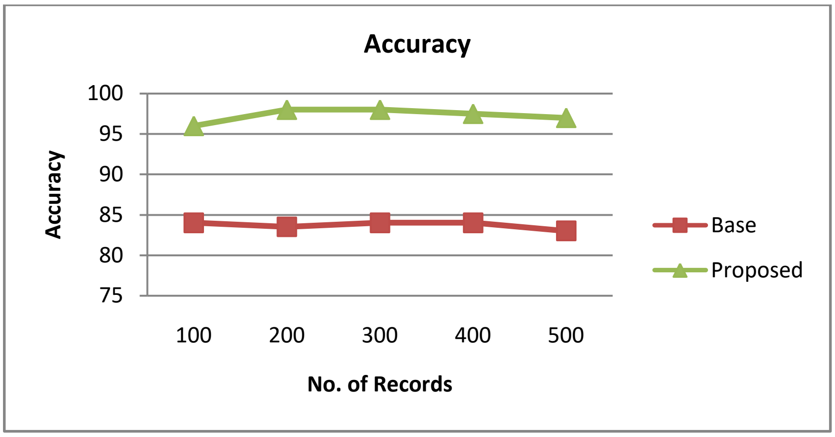

Fig. 5 Accuracy among the base and propose appraoch

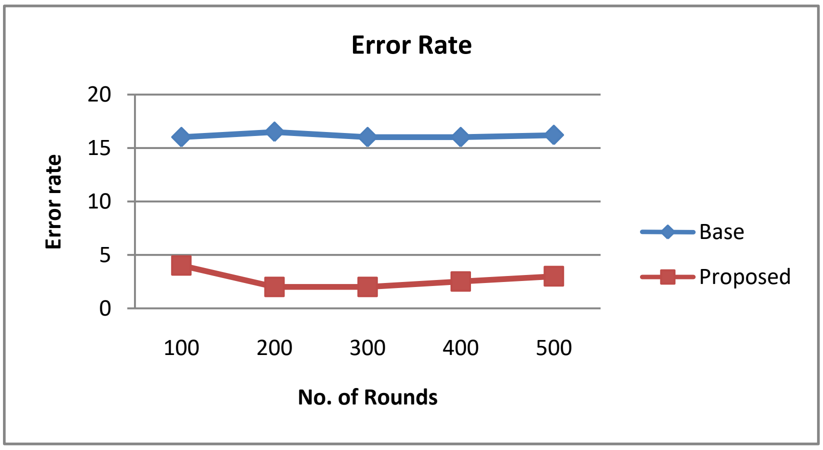

Fig. 6 Error rate among the base and propose appraoch

Conclusion

Data Mining deals with the production of formerly unidentified patterns automatically from huge quantity of data sets. These data sets usually include sensitive individual information or significant business information, which consequently get exposed to the other parties during Data Mining activities. This creates obstruction in Data Mining method. Solution to this problem is provided by Privacy preserving in data mining (PPDM). PPDM is a dedicated set of Data Mining activities where techniques are developed to protect privacy of the data, so that the knowledge detection process can be carried out without barrier. The principle of PPDM is to secure sensitive detail from leaking in the mining process along with precise Data Mining results. The goal of this paper is to discussed about the clustering with the introduction of hierarchical clustering and AES algorithm on privacy preserving technique which are helpful in mining large amount of data with reasonable efficiency and security.

References

[1] Krishna Pratap Rao, Adesh Chaudhary, Prashant johri “Survey on Privacy Preserving Data Mining” (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3342-3343

[2] Nivetha.P.R Nivetha.P.R et al, International Journal of Computer Science and Mobile Computing Vol.2 Issue. 10, October- 2013, pg. 166-170.

[3] Pei, J., Han, J., Pinto, H., Chen, Q, Dayal, U., and Hsu, M-C. PrefixSpan: Mining Sequential Patterns Efficiently by PrefixProjected Pattern Growth. In Proceeding of 2001

[4] Hina Vaghashia, Amit Ganatra, “A Survey: Privacy Preservation Techniques in Data Mining” International Journal of Computer Applications (0975 – 8887) Volume 119 – No.4, June 2015

[5] Amandeep Kaur Mann, Navneet Kaur “Survey Paper on Clustering Techniques” ISSN: 2278 – 7798 International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[6] Gurjeevan Singh, Ashwani Kumar Singla,K.S. Sandha, ”Through Put Analysis Of Various Encryption Algorithms”, IJCST Vol. 2, Issue 3, September 2011.

[7] Ramesh, A. et.al.,, “Performance analysis of encryption algorithms for Information Security ” Circuits, Power and Computing Technologies (ICCPCT),March 2013 , pp. 840 – 844

[8] Shashi Mehrotra Seth, Rajan Mishra,” Comparative Analysis Of Encryption Algorithms For Data Communication”, IJCST Vol. 2, Issue 2, pp.192- 192 , June 2011.

[9] Agarwal, R. , Dafouti, D., Tyagi, S. “Peformance analysis of data encryption algorithms “, Electronics Computer Technology (ICECT), 2011 3rd International Conference , vol.5 , April 2011, pp. 399 - 403 .

[10] Mina Sheikhalishahi, Fabio Martinelli “Privacy Preserving Clustering over Horizontal and Vertical Partitioned Data” 2017 IEEE Symposium on Computers and Communications (ISCC), 978-1-5386-1629-1/17/$31.00 ©2017 IEEE.

[11] Qingchen Zhang, Laurence T. Yang, Zhikui Chen, and Peng Li “PPHOPCM: Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing” 2332-7790 (c) 2016 IEEE.

[12] Kai Xing, Chunqiang Hu, Jiguo Yu, Xiuzhen Cheng, Fellow, Fengjuan Zhang “Mutual Privacy Preserving k-Means Clustering in Social Participatory Sensing” 1551-3203 (c) 2016 IEEE.

[13] Zakaria Gheid, Yacine Challal “Efficient and Privacy-Preserving k-means clustering For Big Data Mining” EEE TrustCom/BigDataSE/ISPA 2324-9013/16 $31.00 © 2016 IEEE DOI 10.1109/TrustCom/BigDataSE/ISPA.2016.139 792 791 2016 IEEE TrustCom-BigDataSE-ISPA.

[14] Jiawei Yuan, Yifan Tian, Student “Practical Privacy-Preserving MapReduce Based K-means Clustering over Large-scale Dataset” 2168-7161 (c) 2016 IEEE

[15] Vadlana Baby, Dr. N. Subhash Chandra “Distributed threshold k-means clustering for privacy preserving data mining” 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India..