International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

Encryption based Privacy Preservation on Big Data using Dynamic Data Encryption Strategy

Johny Antony P[1], Dr Antony Selvadoss Thanamani[2]
[1]*Research Scholar, Department of Computer Science, NGM College, Pollachi 642001*
[2]*Associate Professor & Head, Department of Computer Science, NGM College, Pollachi*

Abstract: ***Big data is termed as huge volume of both structured and unstructured dataset. When dealing with these huge dataset several challenges are encountered by the users such as analysis, capture, storage, search, transfer, sharing, and visualization. To handle these complex data sets numerous technologies have been developed. With the developing technologies and all its connected devices, it is predicted that vast amount of data is produced in the last couple of years. There are many algorithms available for clustering the big data. This section focus on the density based clustering algorithms and its related works. Technology is enhancing each and every day especially in the field of Information Technology and data is very momentous elements. The large volume of data generated through devices is a major obstacle to handle in real time. The accomplishment of data ciphering is a crucial problem during the data progression and dissemination. In order to achieve pursuance, many application disregards data encryption. This article represents a concern about data privacy and suggests a novel data encryption approach known as Dynamic Data Encryption Strategy (DDES).***

Keywords: ***Big data, encryption, security, cryptography, Information hiding, Dynamic Data Encryption Strategy.***

I. INTRODUCTION

Privacy is one of the most important properties that an information system must satisfy. For this reason, several efforts have been devoted to incorporating privacy preserving techniques with data mining algorithms in order to prevent the disclosure of sensitive information during the knowledge discovery. Consider separate medical institutions that plan to conduct a joint research while preserving patient‘s privacy. Here, protecting privileged information is necessary, but it must be enabled for research or other purposes. While heuristic based techniques are mainly conceived for centralized datasets, cryptography-based algorithms are designed for protecting privacy in a distributed scenario by using encryption techniques. Heuristic-based algorithms recently proposed aim at hiding sensitive raw data by applying perturbation techniques based on probability distributions [2]. Moreover, several heuristic-based approaches for hiding both raw and aggregated data through a hiding technique (k-anonymization, adding noises, data swapping, generalization and sampling) have been developed. These algorithms succeed in confirmable privacy protection and improve DM performance. Privacy methods like k-anonymity use generalization/suppression techniques to hide an individual‘s identifiable information [10]. K-anonymity is accomplished by generalization/suppression of attributes in original dataset. To reduce data loss during transformation using generalization/suppression, trade-off between privacy and information loss and flexible data generalizations which provide better solutions by optimizing an aggregated value over all features/records are needed. PPDM approaches can be classified in to five dimensions; such as data distribution, data modification, data mining algorithm, data or rule hiding and privacy preservation. Many privacy calculations approaches utilise some transformation on data to perform privacy preservation. These approaches reduce representation granularity to reduce privacy. Recent years have seen unprecedented growth in applicability of Computer Science in day-to-day activities. Organizations, community and individuals show an augmented trend of storing their data electronically. The huge amount of data collected can be used for analyzing trends of markets and individual or society. Data mining activities involve extracting knowledge from this massive pool of data. The sensitive information about the individuals may be disclosed creating ethical or privacy issues. Many individual therefore don’t share their data publicly, creating data unavailability. Privacy of individual should not be compromised under any case. PPDM has gained popularity so as to address the privacy concerns while data mining is being carried out. Different PPDM techniques include data perturbation, blocking based, cryptographic techniques etc are using for privacy preservation. The remainder of this article is organized as follows. Section 2 reviews different PPDM methods focusing on the evaluation criteria. Section 3 describes the proposed work we have implemented for privacy preservation. Section 4 shows the experimented result. We compared Encryption based Privacy Preservation algorithm with optimal result and proved encryption based privacy preservation is best method. Finally, Section 5 presents conclusion, future extensions and promising directions in the context of privacy preserving data mining.

II. LITERATURE REVIEW

Privacy and security are two terms used interchangeably under different contexts. But both are related to each other and at the same time entirely separate issues. The three fundamentals of security are Confidentiality, Integrity and Availability. In context of Census data, security can be termed as the facility for controlling person-specific access information, protect it from unauthorized disclosure, modification, loss or destruction of his information. Security can be accomplished through controls based on operational and technical knowhow. In contrast privacy is very specific. It can be termed as a right of an individual to keep his/her personal information from being disclosed. Privacy can be accomplished through policies and procedures. Person's personal information which may lead to his identification may not be disclosed under ethical grounds. PPDM is extensively studied by researchers to address these issues for privacy. The security aspects can be taken care by enforcing vigorous methods for protection of sensitive data. Elisa Bertino and Dan Lin et al [2]., surveyed different approaches used in evaluating the effectiveness of privacy preserving data mining algorithms. A set of criteria is identified, which are privacy level, hiding failure, data quality and complexity. As none of the existing PPDMalgorithms can outperform all the others with respect to all the criteria, we discussed the importance of certain metrics for each specific type of PPDM algorithms, and also pointed out the goal of a good metric. There are several future research directions along the way of quantifying a PPDM algorithm and its underneath application or data mining task. One is to develop a comprehensive framework according to which various PPDM algorithms can be evaluated and compared. It is also important to design good metrics that can better reflect the properties of a PPDMalgorithm, and to develop benchmark databases for testing all types of PPDM algorithms. Alpa Shah, Ravi Gulati [1], in the article have tried to classify the PPDM techniques available in the literature and showed its implications best suited under various scenarios. Currently no such technique that provides the best solutions under different scenarios exists. A study to find a new technique altogether or combination of these techniques best suited is an open research area still. Prajwala T R, Sangeeta V [7], made Comparative Analysis of EM Clustering Algorithm and Density Based Clustering Algorithm Using WEKA tool. Machine learning is type of artificial intelligence wherein computers make predictions based on data. Clustering is organizing data into clusters or groups such that they have high intra-cluster similarity and low inter cluster similarity. The two clustering algorithms considered are EM and Density based algorithm. EM algorithm is general method of finding the maximum likelihood estimate of data distribution when data is partially missing or hidden. In Density based clustering, clusters are dense regions in the data space, separated by regions of lower object density. The comparison between the above two algorithms is carried out using open source tool called WEKA, with the Weather dataset as it's input.

III. PROPOSED WORK

Data mining methods are used to extract the meaningful information from the large amount of data. Clustering is one of the most significant methods of extraction of knowledge. It also plays a very important role in investigating big data. The clustering is a process where large volume data should be grouped. Clustering is a method to organize the things into a set in a way that they have high intra-cluster similarity and low inter cluster similarity. Since clustering is the key of Big Data analytics, it is measured as an vital pre-processing step for the detection of information from enormous data [7]. Clustering algorithms can be classified into various types, such as;

1) *Partitioning Based Clustering* - Partitioning method segregates the given data objects into number of divisions known as clusters. In this approach each cluster requires to contain a minimum of one data object and each data object should belong to exactly one group. There are several algorithms for partitioning the data objects. K-means, K-medoids, k-mods, PAM, CLARA, CLARANS and FCM are examples of partitioning based clustering approach.
2) *Hierarchical Based Clustering* - In hierarchical based clustering data objects are clustered in a hierarchical manner. There are two approaches in hierarchical partitioning. They are (i) Top – Down approach (Agglomerative) (ii) Bottom – Up approach (Divisive). Top down approach starts with single data object and then merges into the complete cluster. Bottom up approach starts with the entire cluster and splits them into single data object. AGNES (AGglomeratice NESting) is an example of top down approach where DIANA (DIvisive ANAlysis) is an example of bottom up approach. BIRCH and Chameleon are some other types of hierarchical approach.
3) *Density Based Clustering* - The prior clustering methods are not capable of finding clusters of arbitrary shapes. To satisfy this condition density based clustering algorithms are evolved which performs grouping that depends on density. Density based clustering method has the capability of finding the clusters of arbitrary shapes. It also prevents from outliers. DBSCAN, DENCLUE and OPTICS are example of density based clustering algorithm [6].
4) *Grid Based Clustering* - It divides the space of data objects into predetermined number of cubicles that forms an organization of grids. The major benefit of this method is its high-speed processing time because it only depends on numbers of cells and

independent of number of things. It is an efficient method to many spatial data mining problems. STING and CLIQUE algorithms are example of this method.

5) *Model Based Clustering* - Model based clustering approach is based on the improvement of relationship between the predefined mathematical model and the specified model. MCLUST, EM and COBWEB are example of such type of algorithms.

After clustering process completes, the encryption process takes place on clustered result. Encryption method works in three phases, initially DDES encryption algorithm perform sorting process by using privacy weight values. In this method Pairs Matching Collision (PMC) mechanism is used to provide privacy protection. This mechanism is designed to avoid the scenario when two plain texts can release users' privacy even though leaking each plain text will not be harmful. The operating principle of PMC mechanism is to make sure that the two pre-defined pair data have at least one data encrypted. The paired data must contain privacy information when they are transmitted or operated in plain texts. Then in encryption operation second phase takes place of selecting data packages. Third phase produces as output an encryption strategy deriving from the correct outcomes of second phase. Those data with higher-level encryption priority will be selected for the encryptions under a certain constraints. The rest of data will not be encrypted such that plain texts operations are applied. But main disadvantage of this method is, it is computationally complex, it reduces the system speed as well as this method is not suitable for public databases.

The main algorithms used in our DDES model, which include Dynamic Encryption Determination (DED) algorithm, S Table Generation (STG) algorithm, and Weight Modelization (WM) algorithm. DED algorithm is designed to dynamically select data packages that can be encrypted under certain conditions when considering both timing constraints and resources capacities. STG and WM algorithms are designed for supporting DED algorithm. The following figure 2 shows our proposal work;

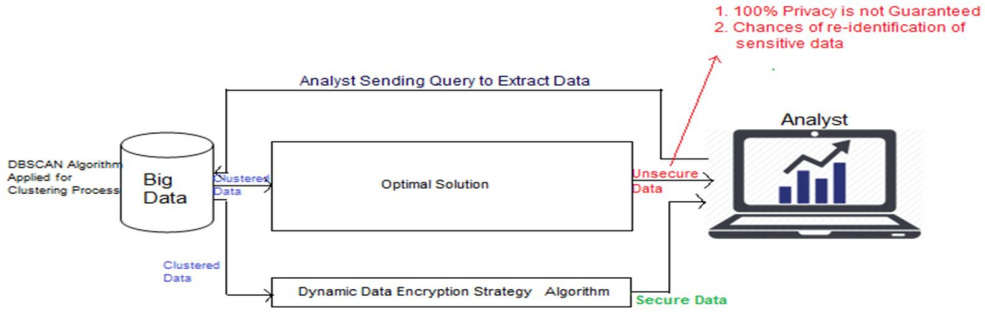

Figure 2: Dynamic Data Encryption Strategy

The following Dynamic Encryption Determination algorithm is used to dynamically select data packages that can be encrypted under certain conditions when considering both timing constraints and resources capacities.

Require: S Table, M Table, T_c, T_m
Ensure: P (Encryption Strategy Plan)
1: Input S Table, M Table, T_c, T_m
2: Initialize P ← ∅
3: Ts ← $[Tc - (Tm + \sum_{D_i \in S\,Table}(N_{D_i}^n \; X \; T_{D_i}^n) + \sum_{D_i \in \{W_{D_i}=0\}}(N_{D_i}^n \; X \; T_{D_i}^n))]$
4: while S Table is not empty do
5: Get D_i having the highest priority from S Table
6: for ∀ D_i, i=1 to N_{D_i} do
7: if $T_s > T_{D_i}^e - T_{D_i}^n$ then
8: Add one D_i to P
9: $T_s \leftarrow T_s - (T_{D_i}^e - T_{D_i}^n)$
10: else
11: Break
12: end if
13: end for
14: end while
15: Output P.

Figure 3: Dynamic Encryption Determination algorithm

The following S Table Generation (STG) algorithm is designed to support DED algorithm.

Require: M-Table'
Ensure: S Table, T_m
1: Input S Table
2: Initialize S Table ← ∅
3: Initialize $T_m = 0$
4: for ∀ D_i in M-Table do
5: if $W^e_{D_i} = \infty$ then
6: $T_m = T_m + N_{D_i} \text{ X } T^e_{D_i}$
7: else
8: if $W^e_{D_i} > 0$ then
9: Calculate $S_{D_i} = W_{D_i} / T^e_{D_i}$
10: Put S_{D_i} to S Table
11: end if
12: end if
13: end for
14: Sort S Table by S_{D_i} in a descending order
15: Return S Table, T_m

Figure 4: S Table Generation (STG) algorithm

IV. EXPERIMENTAL RESULT

In this section, the performance metrics used for evaluation of the optimal method compared with dynamic data encryption strategy algorithm and proved efficient privacy preserving algorithm performs better compared with optimal method. The dynamic data encryption strategy algorithm is implemented in MATLAB. The following figure 5 clearly depicts that Dynamic Data Encryption strategy provides better result for privacy preservation compared with normal optimal method.

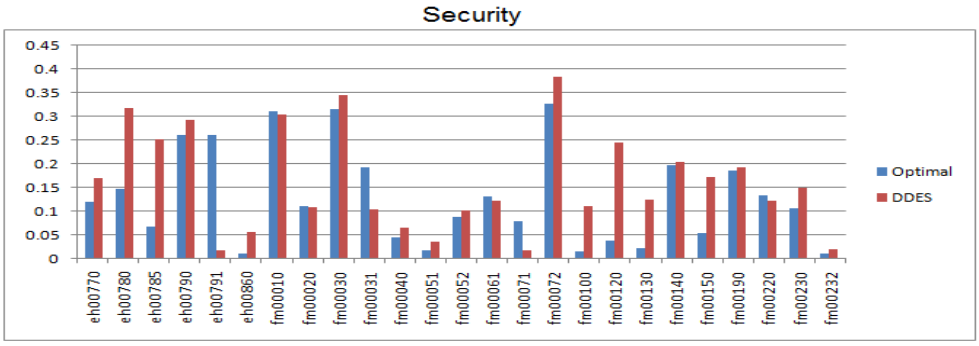

Figure 5: Security (Optimal vs DDES)

V. CONCLUSION

This article solely focused on the privacy issues of big data and considered the practical implementations in big data. The proposed approach, DDES, was used with DBSCAN algorithm to maximize the efficiency of privacy protections. Main algorithm supporting DDES model was DED algorithm that was developed to dynamically alternative data packages for encryptions under different timing constraints. The experimental evaluations showed the proposed approach had an adaptive and superior performance. In this article, the database privacy problems are addressed and a new technique for privacy preservation is proposed. A new heuristic method to hide the sensitive association rules is proposed. Data distortion technique is applied so that sensitive information cannot be discovered through data mining techniques. Confidence of the rules is represented as representative rules. Confidence of the rule is recomputed and compared with threshold level. The confidence of the sensitive rules might be reduced while maintaining the support. From the experimental results, it is observed that all the rules containing sensitive items are hidden. The algorithm is implemented and proved this approach is best approach compared with optimal solution. Further research is in progress to evolve a method which can avoid the computational overhead associated with confidence of the rules.

REFERENCES

[1] Alpa Shah, Ravi Gulati, "Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey", International Journal of Computer Applications (0975 – 8887) Volume 137 – No.12, March 2016.
[2] Dan Lin, and Wei Jiang et al., "A Survey of Quantification of Privacy Preserving Data Mining Algorithms", Purdue University, 305 N. University St., West Lafayette, IN, USA.
[3] Ester, Martin; Kriegel, et al, "A density-based algorithm for discovering clusters in large spatial databases with noise", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. CiteSeerX 10.1.1.121.9220, ISBN 1-57735-004-9.
[4] K.Sekar, M.Padmvathamma et al, "Privacy Preserving-Aware Over Big Data in Clouds Using GSA and Map Reduce Framework", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 5, Issue 8, August 2016, ISSN: 2278 – 1323.
[5] M. Suriyapriya, A. Joicy, "Attribute Based Encryption with Privacy Preserving In Clouds", International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 2, ISSN: 2321-8169, pp.231 – 236.
[6] N.S.Nithya, Dr.K.Duraiswamy, "A Survey on Clustering Techniques in Medical Diagnosis", International Journal of Computer Science Trends and Technology (IJCST) – Volume1 Issue2, Nov-Dec 2013.
[7] Prajwala T, Sangeeta V, "Comparative Analysis of EM Clustering Algorithm and Density Based Clustering Algorithm Using WEKA tool", International Journal of Engineering Research and Development, e-ISSN: 2278-067X, p-ISSN: 2278-800X, Volume 9, Issue 8 (January 2014), PP. 19-24.
[8] Sumit Vikram Tripathi, Ritukar et al., "Privacy-Preserving Data Encryption Strategy for Big Data in Mobile Cloud Computing Environment", International Journal of Innovative Research in Science, Engineering and Technology, 2nd National Conference on Recent Trends In Computer Science & Information Technology, ISSN : 2319 - 8753, Volume 7, Special Issue 6, May 2018.
[9] Sandhya Pradip Mohite, Dr. Sunita S. Barve, "Encryption based Cloud Data Search Technique for Privacy Preserving", International Journal of Computer Science and Information Technologies, Vol. 8 (3) , 2017, 330-334.
[10] Sweeney. L, “k-Anonymity: A Model for Protecting Privacy”, IEEE Security and Privacy, Vol.10, PP. 1-14, January 2002.