International Journal of Engineering & Technology, 7 (3.34) (2018) 331-334

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Privacy-Preserving Technique for Incremental Dataset on Cloud by Synthetic Data Perturbation

[1] Vigneswari D [2]Komal Kumar N [3]Dr.R.Lakshmi Tulasi

[1]Assistant Professor, Department of Computer Science and Engineering,QIS Institute of Technology, Ongole, Andhra
[2]Assistant Professor, Department of Computer Science and Engineering,QIS College of Engg and Tech, Ongole, Andhra
[3]Professor, Department of Computer Science and Engineering, QIS College of Engg and Tech, Ongole, Andhra Pradesh,
**Corresponding author E-mail: vigneswari121192@gmail.com, Pradesh, komalkumarnapa@gmail.com, ganta.tulasi@gmail.com*

Abstract

Cloud is an impetus technology revolution, allowing data providers to store their privatized Electronic Health Record (EHR) for further analysis and outlook with a compromised privacy, where the shared data being exposed to various adversary attacks and malware threats. There are several masking and randomization techniques that provide hints of privacy. In this paper, the Electronic Health Record (EHR) values are perturbed using logarithmic data perturbation and outsourced to the cloud.

Aims.To develop a privacy-preserving technique for effective sharing ofElectronic Health Record (EHR) on a cloud by logarithmic data perturbation.

Methods.Synthetic logarithmic transformation is applied to the sensitive values in a record before they are outsourced and for better privacy.

Results. Synthetic Logarithmic Transformation along with the incremental anonymization produces effective result compared to classical anonymization..

***Keywords*:** EHR, SLT, Privacy, distortion, performance.

1. Introduction

Electronic Health Records or EHR holds the patient's demographics, medical history, encounter history, medications, and immunizations, lab test results subjected to diseases and indices and also the diagnostic method for specific encounters. EHR also holds sensitive information like diagnose codes where the patients hesitate to associate with it. Example of such attributes includes HIV, cancer or mental disorders. Revealing of such sensitive information publically to the outside world involves privacy right of an individual [1] [2]. Outsourcing of such sensitive information leads to the misuse or misinterpretation of the individual. Over recent years different encryption approaches has been developed for preserving the privacy of these sensitive datasets [3-5]. Encryptions like homomorphism tend to encrypt a dataset without decryption, failed to protect the data against the adversary attacks. Another technique for preserving privacy is through anonymization techniques like K-anonymity [6], l-diversity [7], t-closeness [8], and generalization. Privacy-preserving techniques face rapid growth of data from time to time. For instances consider healthcare application where the database holds several thousands of patients attribute values and sensitive information, this sensitive information is updated continuously [9] [10]. Sensitive datasets are anonymized with better privacy to the data providers but compromise privacy on the newly added data.

In recent years, for protecting electronic data from unauthorized access data perturbation is considered a relatively easy and productive technique. Data perturbation has been acknowledged as a more effective application for data protection in health care than re-identification due to the high probability that attacks could take place which links public datasets to original identifiers or subjects. For this reason, data perturbation is hailed as a more solid privacy-preserving approach when it comes to EHR security. Before conducting data mining operation, the miner should reconstruct the perturbed version to obtain the original data distribution [11].

The perturbation has two approaches

1. Probability distribution method [12] where the data is replaced with the sample distribution of from the same distribution. Perturbationis a mathematical model thatrestricts the access for the data server to learn or recover the records. This method never reconstructs the original data but only reconstructs its distributions.

2. Value distortion method [13], where the data is perturbed either by adding, multiplying noise or by other randomized processed. It is more effective than the former type. Many approaches include building a decision tree classifier where each element is assigned a random noise from the mathematical distributions, by determining the distribution applied; the original data distribution is rebuilt from the perturbed version.

This paper proposes Synthetic Logarithmic Transformation (SLT) based privacy preservation model where the sensitive data of EHR are perturbed before outsourcing to cloud for future analysis and outlooks to achieve better privacy.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Related Work

This section briefly discusses the recent works on incremental dataset privacy preservation on cloud computing and perturbation techniques employed on the cloud platform.

Anonymization can be done by methods such as generalization, generalization, suppression, data removal, per-mutation, swapping [13]. Sweeny [14] and Samarati [15] shows that the removal of the personally identifying information from a data is insufficient for data protection. Both studies identified that the de-identification of dynamic generalization is not sufficient to make that identification anonymous and the minimal generalization which captures the property without disturbing the data.The work [14] states that their approaches produce k-anonymization with less generalization compared toprevious approaches. They conclude that a bottom-up approach for k anonymizationis preferable for a small number of quasi-identifying attributes. A task independent approach is presented in [16] [17] involved in masking data before applying mining algorithms. [18] proposed a modified entropy l-diversity model to protect medical datasets. K-anonymity based methods are illustrated in [19] used to search the optimal feature partition. [20] proposed a data reconstruction method where potentially identifying attributes are mapped to numerical data and swapped for nominal data. An approach based on geometric perturbation is presented in [21].

3. Proposed Methodology

We worked on an EHR dataset contained in the University of California at Irvine repository, the dataset contained patient demographics, lab results, diagnostics, images, encounters, medications, disease indices, alerts, of 556 patients. The logarithmic transformation is based on the use of multiple transformations where the sensitive data value is taken the log and added with some random noise with zero mean and some variance, and then antilog of the resultant value is taken to the previous output as a final perturbed.

Algorithm: SLT
Initiate parameters(x_i, e_i, y_i, Σ)
BEGIN
Let x_i={x_1,x_2,x_3,....x_n) be sensitive values
for all x_i
GENERATE e by normal distribution with 0 mean and some variance
COMPUTE y_i by taking log on x_i and adding w
COMPUTE antilog on y_i
RETURN z_i
END

Where $x_i$are the sensitive values in a database D containing multiple records, R={atrval,S} where s={s_1,s_2,s_3,….s_n} y_i is the intermediate output and z_i or D^p is the perturbed value.

In our proposed system data are partitioned into a various number of small blocks that are stored in the cloud. The diving of the small blocks is based upon the perturbation parameter P. let D be the database containing datasets, D= {D_1, D_2, D_3...D_n). Each dataset contains a combination of identifiers and sensitive values. Let the sensitive values s={s_1,s_2,s_3,….s_n} which are going to be perturbed using the SLT algorithm. Let D and D^p are the datasets being uploaded to the database. In order to apply the perturbation each and every time, the perturbation has to be computed which becomes an overhead, to do whole perturbation like (D_1+ b_1+b_2+…. $b_{n)}$,. Rather it makes use of the K in order to update the values without the overhead.

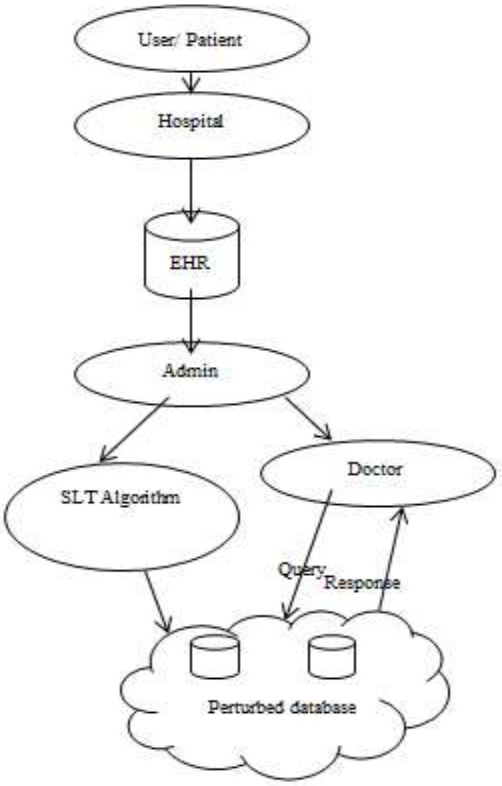

Figure 1:. Proposed framework

The proposed framework for privacy preserving by synthetic logarithmic transformations is shown in Fig 1. The work flow starts from the user or the patient, where the demographics, medical encounters and other details are collected and stored in a digital record called as Electronic Health Record (EHR). An electronic health record (EHR) is a digital version of a patient's paper chart. EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users. The framework contains an administrator, who takes care of the synthetic logarithmic transformation, where the sensitive information of the patient are perturbed and then stored in a cloud environment, since the cloud environment is dynamic the values of the patients or the user changes from time to time, so the administrator plays a prominent role in the privacy preservation. The administrator applies the SLT to the sensitive data and the perturbed values are exported to a database. The doctor retrieves the data values stored in the database for future considerations.

Algorithm: SLT based incremental dataset
Initiate parameters(D,D^p,kold)
Resultant output: $(D+b_i)^p$
b_i^p= Generalization (b_i,kold)
D=(D^p+b_i^p)
Knew=testk(D^p+b_i^p)
While(knew>kold)
(D^p+b_i^p)'=Specialization (D^p+b_i^p)
Knew=testk((D^p+b_i^p)
End
If(knew=kold)
Export(D) to the database
Else return
End
$(D+b_i)^p$=k perturb($D+b_i$)
Export D to the database
Return
end

The main steps in perturbating value by SLT and incremental

dataset are as follows

Step 1: Initialization. In this step, the parameters are initialized to find out the D^p

Step 2: To perturb the sensitive value in a database

Let $s=\{s_1,s_2,s_3,\ldots s_n\}$ be the sensitive values in a record. The SLT algorithm is applied to the sensitive values in a record containing in a dataset, the perturbation is done in two steps, first, the logarithmic transformation of a sensitive value is calculated and the error is added to the transformed value, later the antilog is taken in order to get the perturbed value D^p

Step 3: Applying D^p function to the incremental dataset. In this step, D^p function is applied to the incremental dataset by identifying the perturb parameter K

Step 4: Updating database values

In this step, the SLT is applied to the incremental data and exported to the database.

Step 5: Termination

Terminate the updating until all dataset values are exported to the database.

4. Experimental Analysis and Findings

This section contains the experimental results and analysis of SLT algorithm in preserving privacy and performance overhead. Association rule mining algorithms are evaluated on the proposed model to find the privacy of the rules mined. The experimental evaluation takes place in two layers. The first layer of privacy preservation is by SLT algorithm and the second layer is by reducing the execution time of SLT in incremental dataset compared to the classical perturbation technique.

Table 1. No of rules generated with varying confidence

Algorithm	No of rules generated	
	Original	Perturbed
Apriori	6	5
FP growth	6	4
Predictive apriori	6	3
Tertius	6	4
Association rules	6	4

Note: Varying confidence as (20, 40, 60)

Table 2. No of hidden rules

Algorithm	No of hidden rules	
	Original	Perturbed
Apriori	NA	1
FP growth	NA	2
Predictive apriori	NA	3
Teritus	NA	2
Association rules	NA	2

Table 3:. No of ghost rules

Algorithm	No of ghost rules	
	Original	Perturbed
Apriori	NA	0
FP growth	NA	1
Predictive apriori	NA	1
Teritus	NA	2
Association rules	NA	1

Fig 2. Shows no of rules generated with Apriori, FP growth, Predictive apriori, Tertius, and association rules, where the rules generated with D is not same as D^p

Fig 3. Shows no of hidden rules with Apriori, FP growth, Predictive apriori, Tertius, and association rules.

Fig 4. Shows no of ghost rules with Apriori, FP growth, Predictive apriori, Tertius, and association rules.

Fig 5. Shows comparison of execution time between classical and perturbation

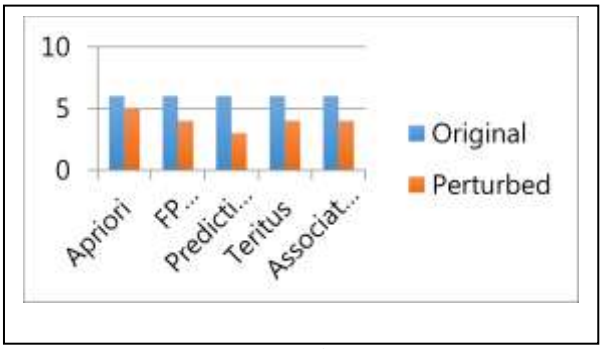

Figure 2:. No of rules generated

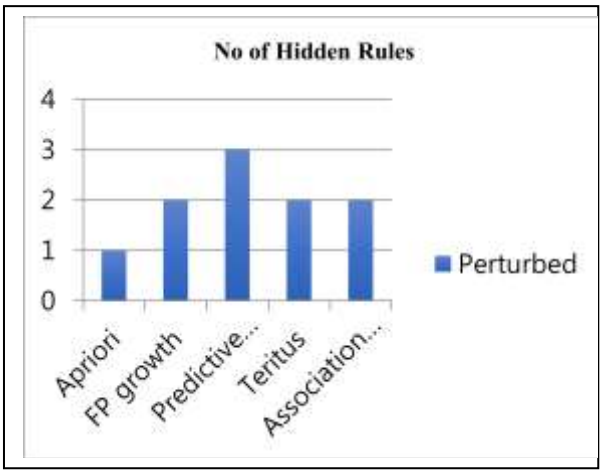

Figure 3:. No of hidden rules

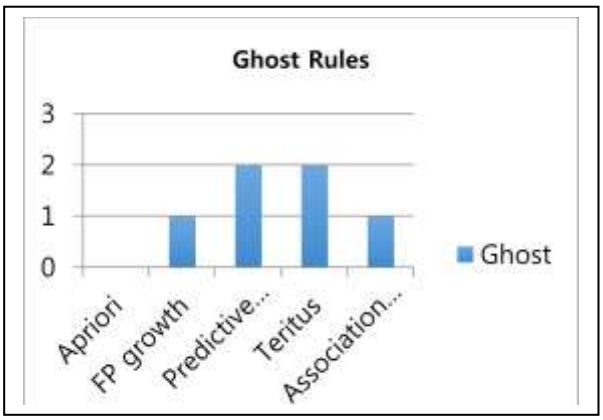

Figure 4:. No of ghost rules

The scenario depicts that the rules in D are not same as D^p, because the values are perturbed resulting in rule hiding, the original rules generated on five algorithms are same, but the rules generated on D^p are totally different and fig 2. depicts that the rules are hidden in case of Apriori, FP growth, Predictive Apriori, Tertius, and association rules association rule mining algorithms. This provides a single layer of privacy to the datasets.Predictive apriori algorithm sustains to produce more hidden rules when compared to all the association rule mining algorithms under study and apriori algorithm produces less number of ghost rules because of no candidate generation in the algorithm. Table 4. Shows the execution time in incrementing data for both classical and perturbation algorithms. The perturbation algorithm outperforms in the execution time as it take lesser time compared to the classical algorithm. Fig 6. Shows the graphical representation of

execution time when adding new records.

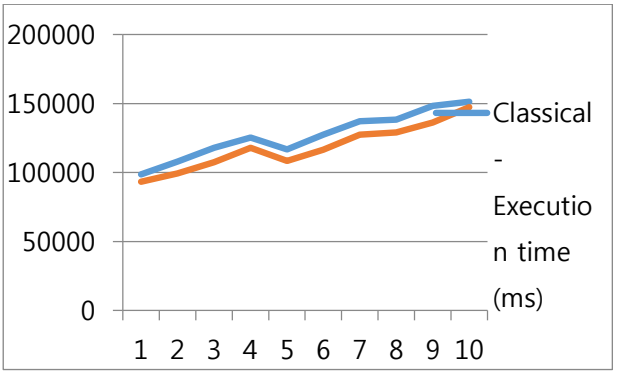

Figure 5.: Comparison of execution time between the classical and perturbation

Table 4:. Execution time in addition to new records

Added records	Classical (ms)	Perturbation (ms)
5000	32764	53241
10000	12736	4635
15000	36272	15342
20000	23646	21635
25000	15364	12636`
30000	42536	56836
35000	56322	8742
40000	57827	43626
45000	45626	34261
50000	42538	21663

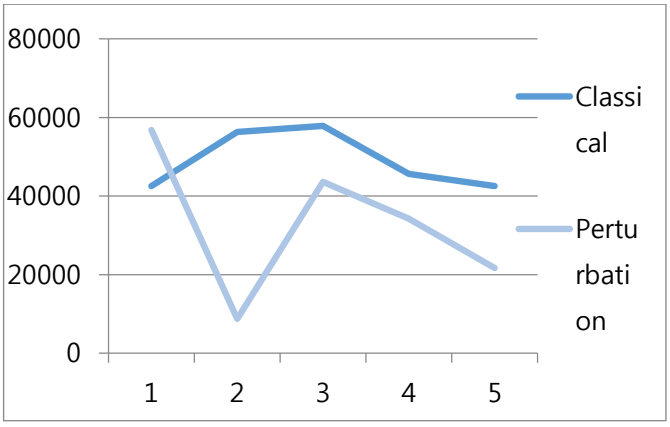

Figure 6:. Execution time in adding new records

5. Conclusion and Future Directions

Cloud being a versatile technology has more influence on the ICT industries and communities. Recent privacy preserving technique proposes a significant challenge when more records added to the stored record. The integration of SLT algorithm with the incremental data anonymization can overcome the privacy issues and performance overhead. The privacy preservation by SLT algorithm produces effective results than a classical anonymity. Improved data privacy algorithms providing more confidentiality can be proposed in the future.

References

[1] "The Privacy Torts" (December 19, 2000). Privacilla.org.
[2] http://faculty.uml.edu/sgallagher/Brandeisprivacy.htm
[3] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, “Privacy-Preserving Multi-keyword Ranked Search over Encrypted Cloud Data,” no. 1, 2014.
[4] M. Li, S. Yu, N. Cao, and W. Lou, “Authorized Private Keyword Search over Encrypted Personal Health Records in Cloud Computing,” System, pp. 1–12.
[5] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, “Enabling efficient fuzzy keyword search over encrypted data in cloud computing,” {IACR} Cryptol. {ePrint} Arch., no. September 2015, pp. 1–16, 2009.
[6] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570
[7] Machanavajjhala, Ashwin; Kifer, Daniel; Gehrke, Johannes; Venkitasubramaniam, Muthuramakrishnan (March 2007). "L-diversity: Privacy Beyond K-anonymity". ACM Trans. Knowl. Discov. Data. 1 (1). doi:10.1145/1217299.1217302. ISSN 1556-4681
[8] Li, Ninghui; Li, Tiancheng; Venkatasubramanian, S. (April 2007). "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity". IEEE 23rd International Conference on Data Engineering, 2007. ICDE 2007: 106–115. doi:10.1109/ICDE.2007.367856
[9] X. Zhang, C. Liu, S. Nepal, and J. Chen, “An efficient quasi identifier index based approach for privacy preservation over incremental data sets on cloud,” J. Comput. Syst.Sci., vol. 79, no. 5, pp. 542–555, Aug. 2013.
[10] P. Bhatotia, A. Wieder, R. Rodrigues, U. a Acar, and R. Pasquin, “Incoop: MapReduce for incremental computations,” Proc. 2nd ACM Symp. Cloud Comput. - SOCC ’11, pp. 1–14, 2011.
[11] N. Adam and J. Wortman. Security control methods for statistical databases. ACM Computing Surveys, 21(4), 1989.
[12] Kargupta, H., Datta, S., Wang, Q., and Sivakumar, K. (2003). Random Data Perturbation Techniques and Privacy Preserving Data Mining, Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), Melbourn, Florida, USA, December 2003
[13] K. Chen and L. Liu. Privacy-preserving data classification with rotation perturbation. In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM’05), pages 589–592, Houston, TX, November 2005.
[14] Latanya Sweeney. Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):571–588, 2002.
[15] P. Samarati. Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027, November/December 2001.
[16] J. A. Nelder and R. Mead. A simplex method for function minimization.Computer Journal, 7:308–313, 1965.
[17] S. R. M. Oliveira and O. R. Za¨ıane. Privacy preserving clustering bydata transformation. In Proceedings of the 18th Brazilian Symposium onDatabases, pages 304–318, Manaus, Amazonas, Brazil, October 2003
[18] S. Guo, X. Wu, and Y. Li. Deriving private information from perturbeddata using IQR based approach. In Proceedings of the Second InternationalWorkshop on Privacy Data Management (PDM’06), Atlanta, GA, April2006.
[19] L. Sweeney. "Database Security: k-anonymity". Retrieved 19 January 2014
[20] M Stephens, NJ Smith, P Donnelly, “A new statistical method for haplotype reconstruction from population data”- The American Journal of Human, 2001 – Elsevier
[21] K Chen, G Sun, L Liu, “Towards attack-resilient geometric data perturbation” - proceedings of the 2007 SIAM international conference …, 2007 – SIAM