A Novel approach for Preserving Data Privacy in Data Mining

Divya Rathod[1] Gurucharan Sahani[2]
[1,2]Department of Computer Engineering
[1,2]Sardar Vallabhbhai Patel Institute of Technology, Vasad, Gujarat, India

Abstract— In this Paper we uses a clustering algorithm as a pre-process for privacy preserving methods to improve the diversity of anonymized data. T-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). We review Paillier`s Encryption and application to privacy preserving computation outsourcing and secure system (e.g. Online voting). Our construction begins with a somewhat homomorphic encryption scheme that works when the function is the scheme's own decryption function. We will show how, anonymization and encryption works together for better privacy preserving in data mining.

Key words: Anonymization, K-Means, L-Diversity, T-Closeness, Encryption (Partial Homomorphic, Fully Homomorphic

I. INTRODUCTION

Privacy preserving is one of the most important research topics in the data security field and it has become a serious concern in the secure transformation of personal data in recent years. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. Current studies of PPDM mainly focus on how to reduce the privacy risk brought by data mining operations, while in fact, unwanted disclosure of sensitive information may also happen in the process of data collecting, data publishing, and information (i.e., the data mining results)delivering.

"Privacy" refers to extraction of sensitive information using data mining. On the other hand, the excessive processing power of intelligent algorithms puts the sensitive and confidential information that resides in large and distributed data stores at risk. Recent developments in information technology have enabled the collection and processing of vast amounts of personal data, such as criminal records, shopping habits, credit and medical history, and driving records. Undoubtedly, this information is very useful in many areas, including medical research, law enforcement and national security. Privacy is commonly seen as the right of individuals to control information about themselves.

In contrast to groups of privacy models trying to control access and encrypt data in the process of information transformation, privacy preservation issues target this problem by protecting identity and preventing sensitive information disclosure, while publishing original data and for general use. These techniques try to publish data keeping as much detail in the data as possible while still making sure the information is sufficiently de-identified. This is the main challenge, since the more details in the published data, the more possible breach in data, which leads to reveal of protected identity and sensitive information. In other words, to keep the utility of data to be as much as possible, it is more difficult to guarantee that individuals cannot be identified or their sensitive information will not be disclosed.

II. RELATED WORK

Pelin Canbay and Hayri Server, they have use anonymization with the attributes of anonymity and clustering. Anonymity means that no one except from authorized user can identify records belonging to a specific individual. Anonymity is the most widespread approach in privacy protection systems. It aims at protecting datasets from identity disclosure, which means that an adversary can learn sensitive information about individual by linking to a specific form of data item. In this system, the diversity difference between original and clustered data was compared in terms of anonymization.[1]

Mohamed Nassar, Abdelkarim Erradi and Qutaibah M. Malluhi, they have use Encryption with the attributes of security and privacy. In this they review homomorphic encryption and present an efficient implementation of the Paillier's additive homomorphic encryption. In this the system review on encryption (Partial homomorphic encryption and Fully homomorphic encryption). Partially homomorphic encryption schemes are simpler and support only one kind of computation. They are more practical and have a wide range of applications ranging from secure voting and collision resistant. Fully Homomorphic Encryption (FHE) which supports both addition and multiplication.). FHE effectively allows the construction of programs which may be run on encryptions of their inputs to produce an encryption of their output.[2]

Mohamdmad-Reza Zare-Mirakabad, fatemes Kaveh-Yazdy and Mohammad Tahmasebi they have use K-anonymization with the attributes of time series, privacy preservation and Ngrams. *K*-anonymization process can be done on wide range of data types, such as census data, social network linkage data, gene expression data and Medical data records. In this they proposed a novel privacy preservation framework for publishing Ngram models of time series.[3]

Tsubasa Takahashi, Koji Sobataka, Takao Takenouchi, Yuki Toyoda, Takuya Mori and Takahide Kohro they have use Complex data, which has single-valued attributes and set-valued attributes, enables us to associate these attribute values and analyze these relationships. They proposed a top-down itemset recoding which transforms itemsets into the generalized ones to ensure *k*-anonymity along with the top-down recoding manner for complex data. Complex data having both single-valued attributes and set-valued attributes are uniformly anonymized.[4]

Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian they have use t-closeness with the attributes of k-anonymity and l-diversity. K-anonymity protects against identity disclosure. L-diversity attempts to solve this problem by requiring that each equivalence class has at least _ well-represented values for each sensitive attribute. In this they propose a novel privacy notion called t-closeness, which requires that the distribution of a sensitive

attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). They use the Earth Mover Distance measure for our t-closeness requirement. [5].

Ahmed Ali Mubark, Emad Elabd and Hatem Abdulkader, they have use semantic anonymization with the attribute of l-diversity. In this they have propose an approach to categorical data preservation based on Domain-based of semantic rules to overcome the similarity attacks. [6]

III. METHODOLOGIES

A number of algorithms have been proposed for extracting knowledge from privacy preserving. In this section, we review K-Anonymity, L-Diversity, T-Closeness, and Encryption Method are discussed.

A. Anonymity

Anonymity is the most widespread approach in privacy protection systems. It aims at protecting datasets from identity disclosure, which means that an adversary can learn sensitive information about individual by linking to a specific form of data item. In order to reduce the risk of identification, the k-anonymity technique was proposed by those concerned [1]. k-anonymity on preserving sensitive information leads to adding extra anonymization criteria on original data New Anonymity algorithms are discussed as under:

- K-Anonymity [1] is anonymization algorithm which consists of two phases generalization and suppression approach. In first phase that is generalization phase which the attribute values are generalized to a range in order to reduce specification. For example, the date of birth could be generalized to a range such as year of birth, so as to reduce the risk of identification. In second phase suppression the values of attribute are completely removed. It is clear that such methods reduce the risk of identification with the use of public records, while reducing the accuracy of applications on the transformed data [13].
- Explicit identifiers (I) [1] is containing information that explicitly identifies record owner and are typically removed from the released data such as name, social security number and cell-phone number.
- Quasi-identifiers (QID) [1] containing information that could potentially identifies record owner and typically transformed in the released data such as date-of-birth, gender and ZIP code.
- Sensitive attributes (S) [1] containing sensitive information about data owner such as salary or disease which should be protected.
- K-Means [1] is a simple and straightforward algorithm that is frequently used in clustering. It clusters a group of data into a predefined K value. Clustering process starts with randomly selected initial cluster centers and keeps reassigning the data object in dataset to cluster center based on the distance between cluster centers and data object. This process continues until a condition is satisfied. In our experiments, to observe the alteration on anonymized dataset based on the alteration of cluster numbers they use K-means algorithm for clustering [1].

B. L-diversity

The definition of *k*-anonymity, as well as the variety of algorithms available to perform the anonymization, can make this model a very appealing choice for potential data publishers. Nevertheless this technique is proven vulnerable to different attacks, especially when the attacker has access to background knowledge. New extended model for protecting privacy: *l*-diversity. The authors of the paper exposed the vulnerabilities of *k*-anonymity in two different attack models: Homogeneity attack and Background-Knowledge attack.

- Homogeneity Attack: In this attack, all the values for a sensitive attribute within an equivalence class are the same. Therefore, even though the data is *k*-anonymous, the value of the sensitive attribute for any record in that group of size *k* can be predicted with 100% accuracy.
- Background Knowledge Attack: In this attack, the adversary can use an association between one or more quasi-identifier attributes with the sensitive attribute or public knowledge of the target in order to eliminate possible values of the sensitive attribute.

For example, if a young individual's *QI* can be linked to an equivalence class, where all values of the sensitive attribute "disease" are either Arthritis, Alzheimer syndrome or Flu, it can be inferred that the target's sensitive info is probably "Flu" since the first values are highly unlikely to occur to a young person.

Defination: A group of records that belong in the same Equivalence Class q* is l-diverse, if it contains at least l "well-represented" values for the Sensitive Attribute S. A table T is considered l-diverse if every Equivalence Class q* in T is l-diverse.

The l-diversity principle ensures the existence of l "well-represented" values in every block of records (equivalence class), without further clarifications on what exactly "well- represented" means [21].

C. T-Closeness

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to satisfy t-closeness if all its equivalence classes have t-closeness. [5]

D. Encryption Methods

- Partial Homomorphic
- Fully Homomorphic

1) Partial Homomorphic

Partially homomorphic encryption schemes are simpler and support only one kind of computation. However they are more practical and have a wide range of applications ranging from secure voting and collision resistant hash functions to private information retrieval and secure computation on the cloud.

a) Encryption Function

$$c = g^m \cdot r^n \bmod n^2$$

Where g = nonzero integer number
r = random number
n = product of prime number
m = data (Which we have encrypt)

b) Decryption Function

$$m \equiv L(c^{\lambda(n)} \bmod n^2) \cdot \mu \bmod n$$

Where m = plain text

Carmichael's function,

$\lambda(n) = lcm[(p-1)(q-1)]$

n = prim number,

$\mu \equiv k^{-1} \bmod n$,

$L(c^{\lambda(n)} \bmod n^2) = k$

2) Fully Homomorphic

A cryptosystem which supports both addition and multiplication (thereby preserving the ring structure of the plaintexts) is known as Fully Homomorphic Encryption (FHE). FHE effectively allows the construction of programs which may be run on encryptions of their inputs to produce an encryption of their output. Since such a program never decrypts its input, it can be run by an untrusted party without revealing its inputs and internal state. This would have great practical implications in the outsourcing of private computations, for instance, in the context of cloud computing.

a) Encryption Function

$$c = m+(2*r)+(2*p)$$

Where p = Random Prime number r,

xi = Random number

p = 2*sum(n) (product of prime number)

pk=sum(xi) (keygen)

m = data(Which we have encrypt)

b) Decryption Function

$$m \equiv c \bmod 2\ [c/p] \bmod 2$$

Where m = plain text

p = Random prim number

IV. CONCLUSIONS

Privacy is the major concern to protect the sensitive data. People are very much concerned about their sensitive information which they don't want to share. The Anonymization algorithm are used reduced information loss and increase the privacy protection. T-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). Hereby it is concluded that fully homomorphic encryption algorithm is used to provide more secure to the system. It will provide better privacy, security and encryption to the system.

From the above analysis of the different types of algorithms I conclude that the different algorithms for different techniques have their own criteria and own prior knowledge of the data. The different algorithms analysis shown in the tables as below:

Sr. No.	Methods	Merits	Demerits
1.	K-Anonymity	– K-anonymity model is simple, intuitive, and well-understood. It appeals to non-expert, the model's end client. – This protects respondents' identities while releasing truthful information	– K-anonymity is difficult to achieve before all data are collected in one trusted place. – k-Anonymity does not provide privacy if Sensitive values in an equivalence class lack diversity
2.	L-Diversity	– l-diversity works one step ahead of k-anonymity in preventing attribute disclosure	– l-diversity is more difficult to achieve and also it is not able to provide sufficient protection for privacy
3.	T-Closeness	– It is a further enhancement of l-diversity group based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data representation.	
4.	Partial Homomorphic	– It private information retrieval	– There is not much overhead involved in performing the computations, at least for those presented.
5.	Fully Homomorphic	It Provide Privacy in following: – Banking transactions – Voting systems – Cloud computing applications – Private information retrieval	– It`s Complexity – Only fully homomorphic cryptosystem is lattice based

Table 1: Comparative Analysis of Methods based on various Parameters

Parameters	Equations
Information Gain	Information Gain(v) = $I(Rv) - \sum_c \left\| \frac{Rv}{Rc} \right\| I(Rc)$ Where $I(Rx)$ is the entropy of Rx: $I(Rx) = -\sum_{cls} \frac{freq(Rx,cls)}{\|Rx\|} \times \log 2 \frac{freq(Rx,cls)}{\|Rx\|}$
Anony Loss	$Aonoy\ Loss(v) = avg\{A(VIDj) - Av(VIDj)\}$
Score	$Score(v) = \begin{cases} \frac{Info\ Gain(v)}{Anony\ loss(v)} & if\ Anony\ Loss(v) \neq 0 \\ InfoGain(v) & Otherwise \end{cases}$

Table 2: Comparative Analysis based on various Parameters & Equations

A. Example

The specialization on ANY Edu refines the 34 records into 16 records for Secondary and 18 records for University. The calculation of Score (ANY Edu) is shown below.

Education	Sex	Work Hrs	Class	# of Recs.
9th	M	30	0Y3N	3
10th	M	32	0Y4N	4
11th	M	35	2Y3N	5
12th	F	37	3Y1N	4

Bachelors	F	42	4Y2N	6
Bachelors	F	44	4Y0N	4
Masters	M	44	4Y0N	4
Masters	F	44	3Y0N	3
Doctorate	F	44	1Y0N	1
			Total:	34

Table 3: (Compressed) table

1) Calculation

$$I(RANY\_Edu) = -\frac{21}{34} \times \log_2 \frac{21}{34} - \frac{13}{34} \times \log_2 \frac{13}{34} = 0.9597$$

$$I(RSecondory) = -\frac{5}{16} \times \log_2 \frac{5}{16} - \frac{11}{16} \times \log_2 \frac{11}{16} = 0.8960$$

$$I(RUniversity) = -\frac{16}{18} \times \log_2 \frac{16}{18} - \frac{2}{18} \times \log_2 \frac{2}{18} = 0.5033$$

$$InfoGain(ANY\_Edu) = I(RANY\_EDU) - (\frac{16}{34} \times I(RSecondory) + \frac{18}{34} \times I(RUniversity)) = 0.2716$$

AnonyLoss(ANY Edu) = avg {A(V ID1) ¡ AANY_Edu(V ID1)} =(34 – 16)/1 = 18

$$Score\ (ANY\ Edu) = \frac{0.2716}{18} = 0.0151$$

Parameters	Information Gain	Anony Loss	Score
K-Means	0.60991	2084	0.00029266
Generalization	0.022141	2494	8.8777e-06
T-Closeness	0.78526	1045	0.00036161

Table 4: Analysis of based on various Parameters

This are the results of using the following parameters: Information Gain, Anony Loss and Score and using these parameters we find results of K- Anonymization, Generalization, T-Closeness etc.

V. RESULTS

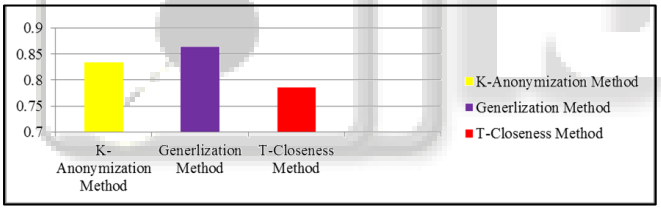

Fig. 1: Information Gain

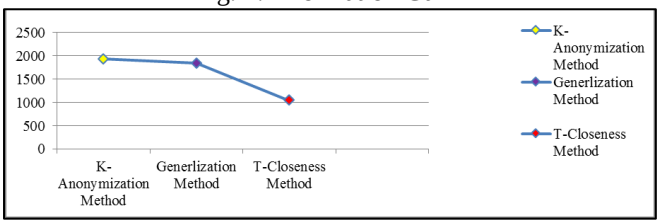

Fig. 2: Anony Loss

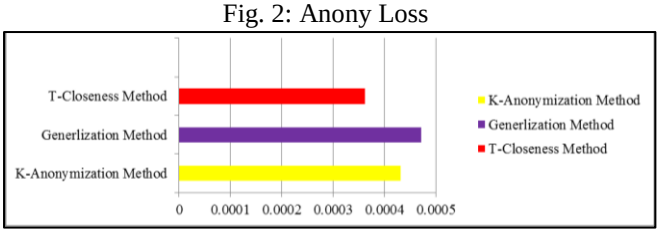

Fig. 3: Score

REFERENCES

[1] Pelin Canbay and Hayri Sever, "The Effect of Clustering on Data Privacy"2015 IEEE International Conference on. IEEE 2015.
[2] Mohamed Nassar, Abdelkarim Erradi, Qutaibah M. Malluhi, "Paillier's Encryption: Implementation and Cloud Applications" KINDI Center for Computing Research Qatar University Doha, Qatar.
[3] Mohammad-Reza Zare-Mirakabad, Fatemeh Kaveh-Yazdy, Mohammad Tahmasebi, "Privacy Peservation by k-anonymizing Ngrams of Time Series" Yazd University, Iran, Dalian University of Technology, Dalian.
[4] Tsubasa Takahashi, Koji Sobataka , Takao Takenouchi, Yuki Toyoda, Takuya Mori and Takahide Kohroy "Top-Down Itemset Recording for Releasing Private Complex Data" Cloud System Research Laboratories, NEC Corporation, Kawasaki, Kanagawa Japan, Jichi Medical University Hospital, Shimotsuke, Tochigi Japan. IEEE 2013.
[5] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, "T-Closeness: Privacy Beyond k-Anonymity and _-Diversity" Department of Computer Science, Purdue University, AT&T Labs – Research. IEEE 2007.
[6] Ahmed Ali Mubark, Emad Elabd, Hatem Abdulkader, "Semantic Anonymization in Publishing Categorical Sensitive Attributes" Ibb University Yemen, Menoufia University Egypt. IEEE 2016.
[7] Michael O`Keeffe, "The Paillier Cryptosystem" Department of Mathematics, College of New Jersey, April 2008.
[8] Nirav. U.Patel, Vaishali.R.Patel, "Anonymization of Social Networks for Reducing Communication Complexity and Information Loss by Sequential Cluatering", 2015.
[9] R. Mahesh and Dr.T.Meyyappan, "A New Method for Preserving Privacy in Data Publishing against Attribute and Identity Disclosure Risk" 2013.
[10] Benjamin C. M. Fung, Ke Wang, Philip S. Yu, "Top-Down Specialization for Information and Privacy Preservation.
[11] Grigorios Loukides and Aris Gkoulalas-Divanis, "Utility-preserving transaction data anonymization with low information loss".
[12] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Computing Surveys (CSUR), vol. 42, 2010.
[13] R. B. Ghate and R. Ingle, "Clustering based Anonymization for privacy preservation," in Pervasive Computing (ICPC), 2015 International Conference on, 2015.
[14] Sweeney. "k-anonymity: a model for protecting privacy". Int. J. Uncertain. Fuzziness Knowl.-Based Syst., Vol. 10, pp. 557-570, 2002.
[15] M. B. Malik, M. A. Ghazi, and R. Ali, ''Privacy preserving data mining techniques: Current scenario and future prospects,''inProc.3rdInt.Conf. Comput. Commun. Technol. (ICCCT), Nov. 2012, pp. 26–32.
[16] M.-J. Choi, H.-S. Kim and Y.-S. Moon. "Publishing time-series data under preservation of privacy and distance orders". International Journal of Innovative Computing, Information and Control (IJICIC), Vol. 8, pp. 3619-3638, 2012.
[17] Fung, B.C.M., Wang, K. and Yu, P.S.: Top-down specialization for information and privacy preservation. Proc. ICDE2005, pp. 205–216 (2005).

[18] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg (ICDE), 2006.
[19] X. Xiao and Y. Tao. Personalized privacy preservation. In Proceedings of ACM Conference on Management of Data (SIGMOD'06), pages 229–240, June 2006.
[20] C. C. Aggarwal and S. Y. Philip, A general survey of privacy-preserving data mining models and algorithms: Springer, 2008.
[21] Olga Gkountouna, A Survey on Privacy Preservation Methods, June -2011.
[22] Pierangela Samarati and Latanya Sweeney, Protecting Privacy When Disclosing Information: K-Anonymity and its Enforcement through Generlization and Suppression.
[23] Freny Presswala, Amit Thakkar and Nirav Bhatt, Survey on Anonymizati on in Privacy Preserving Data Mining, International Journal of Innovative and Emerging Research in Engineering (IJIERE), 2015.

