ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

A Hybrid Approach to Preserving Privacy in Data Mining using L-K-C and personalized Privacy

Rahul Pandey[1], Shilpa Jain[1]

B.E Student, Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg, India[1]

ABSTRACT: Owing to the massive amounts of data being collected in every fraction of second, there lie threats to the confidentiality, privacy and authenticity of the information being transmitted. Due to ease of access to the data by anyone, a serious concern for privacy has arisen. Thus, in order to safeguard the information leakage to hackers, there are several techniques being practiced. The central issue with these techniques is that while maintaining the privacy, they tend to also increase the information loss.In this paper we present a new and hybrid approach to preserve privacy in data mining. Our model discusses about different privacy models and make a hybrid version out of it. The evaluation of the model's results show that we have obtained a significant decrease in information loss that occurs due to privacy compared to many other existing models. Thus, we propose a methodology to preserve data, without any hindrance in privacy, and at the same time maintaining the information and the utility balance.We will take data from hospital's database which will contain the data of the patients who got their treatment there. We assume that the intruder is having access to AADHAR list. As we will show in this paper it is not very difficult to link Aadhar list with patient's data and get access to sensitive information of the patient.

KEYWORDS: Preserving Privacy, Data Mining, Quasi Identifier, Personalized privacy, Sensitive information, Aadhar, Hybrid, Utility

I. INTRODUCTION

Data in 21st century is escalating day by day. In today's technological thriving environment, data acquisition is not as big of an issue as it was before. Advancement in data mining has made it possible for various analysis or decision making. But today's generation is much more conscious about their privacy being preserved while use of their data in any possible way. By privacy we mean not giving away any information about a person which is considered sensitive by them. Preserving privacy is that branch of data science and information security, which helps us in hiding the privacy and still maintaining the data utility for research.

For most of the applications like hospital, insurance, Schools and Universities, the data is stored in columnar way.

The attributes can be divided into following categories [1]:

i. Identifying attributes: These attributes like name, email id or aadhar number can explicitly identify or point out the person.

ii. Quasi identify attributes: The attributes like age, sex, pin code when linked with some other database.

or attributes can easily reveal a person's identity.

iii. Sensitive attributes: This includes the data which should not be shown or published against a person's identity. For e.g. disease, the patient's identity should not be revealed against any product.

iv. Non-Sensitive attributes: These are the columns which if published publicly do not lead to any problem.

To achieve the Privacy Preserving Data Mining the techniques being used are [2]:

i. Anonymization: It is also called generalization, it is done to remove the personally identifiable information from the data sets by generalizing the attributes to the parent node.

DOI: 10.15680/IJIRCCE.2018.0602098

ii. Perturbation: Perturbation refers to disturbances in the data. It is done with the help of a noise. This noise can be divided into two- additive and multiplicative noise.
iii. Randomization: Here the records are shuffled vertically. It doesn’t change any record just changes it position in the database. It helps to preserve the semantic meaning of the database and hides the correct identity.
iv. Cryptography: The sensitive information and identity is not changed but certain protocols are used to protect the message itself. Here data is preserved with the help of encryption of the records.
v. Condensation: In condensation process, the privacy is achieved by forming the clusters in such a way that the size of each cluster is different from that of any other cluster.

These methods are good at achieving the Privacy but still suffer from many limitations. In today’s world privacy concerns are real and it comes at top of the priority list of data miners and researchers . Privacy Preserving Data Mining techniques can prove to be useful and efficient in achieving the goal of gaining trusts by preserving identity and privacy of the individuals. The application developed here for research purpose is taken from a hospital located in Hyderabad. The attributes are categorized into:

Attribute	Category
Name	Identifying
Age	Quasi identifier
Gender	Quasi identifier
Pin code	Quasi identifier
Education	Quasi identifier
Disease	Sensitive Information

Table 1. Attributes

The Quasi identifiers are the weakest point for any linkage attack. It is very easy for the attacker to get patient’s identity even when his or her name or other personal information are not publicly published. This can be done by linking publicly available database and the values of these attributes in the two databases. For example, there are two applications stored on the Hospitals database and publicly available e- aadhaar data or voting list, both the applications contain these three fields- age, gender and zip code. If the attacker is able to locate and link the fields of Hospital’s database and e-addhar list he can easily breach the privacy of our patients.
Although these information in the fields may seem very harmless and individually may not present any harm but by linking them from each other attackers and fraud people can steal, change, misuse or sabotage the information. In order to hide these original data, we need to hide and secure these data which may in turn present us with another challenge, information loss. It will be discussed later in this paper. To preserve the identity of customers it is necessary to hide original values in the quasi identifier attributes and be able to minimize the information loss.

II. RELATED WORK

A method of random perturbation for Privacy Preservation Data Mining was proposed by XiaolinZhang, Hongjing Bi[5] which maintained data privacy by replacing the attribute values with the code values(1,2,3…,n) and arranging these values in a square matrix which were further randomly perturbed. The data is extracted from these matrices using if then rules after pruning it with post pruning algorithms. The stated technique helped in improving accuracy and achieving better data mining results.Khaled M. Khan [6] portrayed the trust issues in the cloud environment and the reason behind them. It is difficult for anyone to follow and trust the system where they have very less power and where there is no transparency. By providing access control, maintaining privacy and reimbursing for data leakages, any organization can certainly win their customer’s trust.To overcome the Curse of Dimensionality, Mohammed et al. stated that most of the intruders would not be aware of all QID attribute values. So, this can be best modelled using the technique where no user knows more than L of QID attributes, where L is an integer parameter. This privacy measure is known as LKC privacy.

DOI: 10.15680/IJIRCCE.2018.0602098

In case of numerical sensitive attributes, Zhang et al. proposed a technique known as (k,e)-anonymity. Here, it is stated that an equivalence group for any QID contains at least k different sensitive values with a range of at least e.Xiao and Tao [12],2006, proposed the notion of personalized privacy to allow each record owner to specify her own privacy level. This model assumes that each sensitive attribute has a taxonomy tree and that each record owner specifies a guarding node in this tree.

Arshveer Kaur [13]proposed a hybrid privacy preserving data mining technique with the aid of suppression and perturbation over the centralized server environment where simulation and implementation has proven the method to be successful. Majid Bashir Malik [14] and team has preserved privacy and minimized information loss using soft computing techniques. Using neural networks, they have been able to reach a certain level of privacy even with the fuzzified data. By merging entropy-based partition and combined data distortion, Putri, Awalia W. proposed a model to preserve privacy. The model basically focusses on data utilities and maintain data privacy better.

III.ANONYMISATION

Anonymization technique means that we use taxonomy tree to generalize data.It helps to hide information without abruptly changing the records. The records are defined in k-1 different *forms* hence the data is k anonymized as there are k data set with same value in each quasi field. The generalization of the data is usually done to anonymize the original data. The process of anonymization is always done with respect to the quasi identifiers. [3]All the quasi-identifiers are suppressed or changed with the same character or represented in the form of same intervals for the purpose of uniformity in the data set. The problem with generalization process is that it suffers from a high information loss when trying to obtain the original data and the original values of the records is lost forever as it can never be retrieved again. The figures below show the anonymized data achieved by generalizing the records.

GENDER	AGE	PINCODE	EDUCATION	DISEASE
Male	21	500001	5th standard	HIV
Female	24	500002	B.E	Viral Infection
Male	37	500004	12TH	HIV
Male	31	500006	M.Sc	Viral Infection
Female	23	500001	8th	HIV

Table 2. Original data

GENDER	AGE	PINCODE	EDUCATION	DISEASE
ANY	[20-25)	500***	Primary	HIV
ANY	[20-25)	500***	U.G	Viral Infection
ANY	[35-40)	500***	Primary	HIV
ANY	[30-35)	500***	P.G	Viral Infection
ANY	[20-25)	500***	Primary	HIV

Table 3. Anonymised data

So the generalization rule followed here are such :

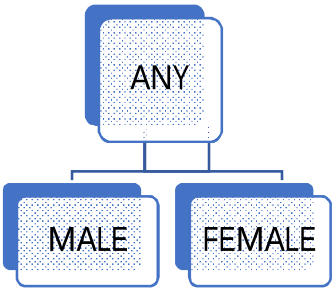


Fig 1.Gender anonymisation

DOI: 10.15680/IJIRCCE.2018.0602098

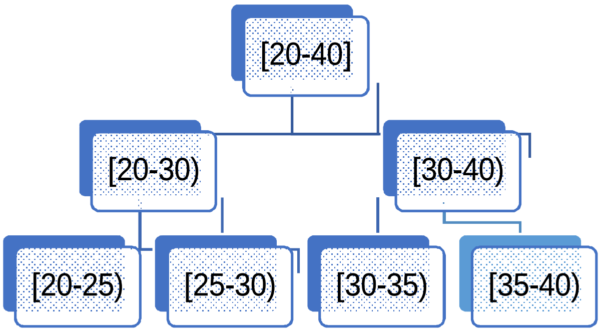


Fig 2. Age anonymization

IV. PROPOSED METHOD

In our hybrid approach we tend to solve the privacy issue and be able to secure information simultaneously. Privacy and Utility are not mutually exclusive. We can have a balance. That takes oversight, regulation, and Privacy Enhancing Technologies. The method in this paper is the hybrid approach for Privacy Preserving Data Mining based on L-K-Personalized privacy, Suppression and Perturbation implemented over the centralized server. The data is taken from hospitals data base and the name field which specifically identifies the patients is removed then further the data sets Qids are changes by adding noise and using k-e anonymity.

The biggest problem was to deal with gender field because of its Boolean nature. It adds more difficulty to the challenge of preserving privacy. Using generalization rules for privacy preservation anonymizes the values in the fields to the level where there is complete loss of information and the original data cannot be retrieved. Achieving privacy by generalizing the numeric fields again becomes hindrance in the reverse process. The proposed method is an initiative to preserve privacy on the online centralized data repository with minimum information loss. The method in this paper is the mixed approach for Privacy Preserving in Data Mining based on personalized privacy, Perturbation, (k-e) and L-K-C implemented over the centralized server. The approach focuses on preventing their identification being revealed through quasi identifier attributes.

Let L be selected QID, this L takes in account the adversary’s background knowledge. So not all the QID in the table are perturbed or suppressed or generalized. This actually helps us to reduce the information loss. Here as we know that the attacker has access to e-adhaar. The selected qids are-{gender, age, pin code}.Note that education field is not mentioned in Aadhar card hence it is not included in the selected QIDs.

Another feature proposed in this paper is personalized privacy. In Personalized privacy we trade off fields of QID with sensitive information. A survey is conducted for each patient whose record is in the database. They are asked whether they want to hide their sensitive information i.e disease information. If they say yes, then they have the guarding node which will not tell their exact disease but will give a generalized result. It assumes that each sensitive attribute has a taxonomy tree and that each record owner specifies a guarding node in this tree. The record owner’s privacy is violated if an adversary is able to infer any domain sensitive value within the subtree of her guarding node with a probability, called breach probability, greater than a certain threshold. Suppose HIV and SARS are nodes of Infectious Disease in the taxonomy tree. An HIV patient Ram can set the guarding node to Infectious Disease, meaning that she allows people to infer that she has some infectious diseases, but not any specific type of infectious disease. Another HIV patient, Shyam, does not mind disclosing his medical information, so he does not set any guarding node for this sensitive attribute.

The stepwise execution of the approach is explained below:
i. Check personalized privacy
ii. Select Guarding nodes
iii. apply guarding nodes to respective fields

iv. Remove the identifying field
v. Choose two numeric values.
vi. Suppress the gender field :
Male → First value
Female → Second value
vii. Calculate appropriate range for age through (k-e) anonymity
viii. change the age attribute:
x. Apply perturbation on zip code:
Zip code = Zip code * noise

The formula used to calculate the information loss of the perturbed data in the reverse process is **Minimal Distortion (MD)**[4]. The MD metric is a simple counter that increments every time a value is generalized to the parent value. The higher the MD value, the more generalized is the data, and consequently, more information was lost. The MD counter approach can be implemented in any privacy model as it is very easy to implement and to work on.

GENDER	AGE	PINCODE	EDUCATION	DISEASE
5	[20-25)	1000002	5th standard	HIV
16	[20-25)	1000004	B.E	Viral Infection
5	31	1000008	12TH	*Infectious disease*
5	[30-35)	10000012	M.Sc	Viral Infection
16	[20-25)	1000002	8th	HIV

Table 4Anonymized perturbed Data by Hybrid approach
In the table above table:
I. QIDs-Gender, Age, Pincode
II. Non-Sensitive – Education
III. Sensitive – Disease
Health records are considered to be extremely private,as much of this data is considered sensitive. However, theincrease in the amount of data, combined with the favorableproperties of the cloud has led health services to store and exchange medical records through this infrastructure. Thus, to protect from unwanted disclosures privacy-preserving the proposed approaches can be considered.In, a survey on the state-of-the-art privacy-preservingapproaches employed in the e-Health clouds is given, wherethe authors divide PPDM techniques in either cryptographicand non-cryptographic. The cryptographic techniques are usually based on encryption, whereas non-cryptographic approaches are based on policies and/or some sort of restrictedaccess. The proposed method is a non-cryptographic approach.

DOI: 10.15680/IJIRCCE.2018.0602098

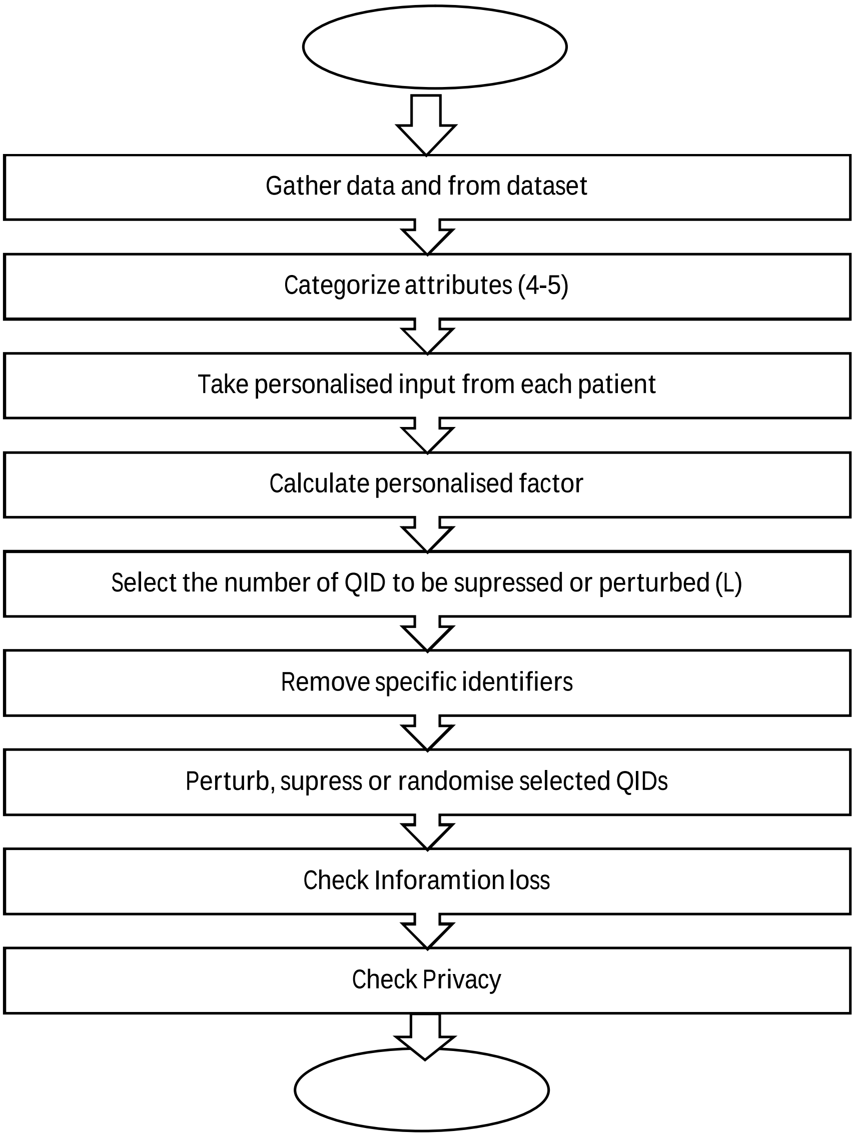

Gather data and from dataset
Categorize attributes (4-5)
Take personalised input from each patient
Calculate personalised factor
Select the number of QID to be supressed or perturbed (L)
Remove specific identifiers
Perturb, supress or randomise selected QIDs
Check Inforamtion loss
Check Privacy

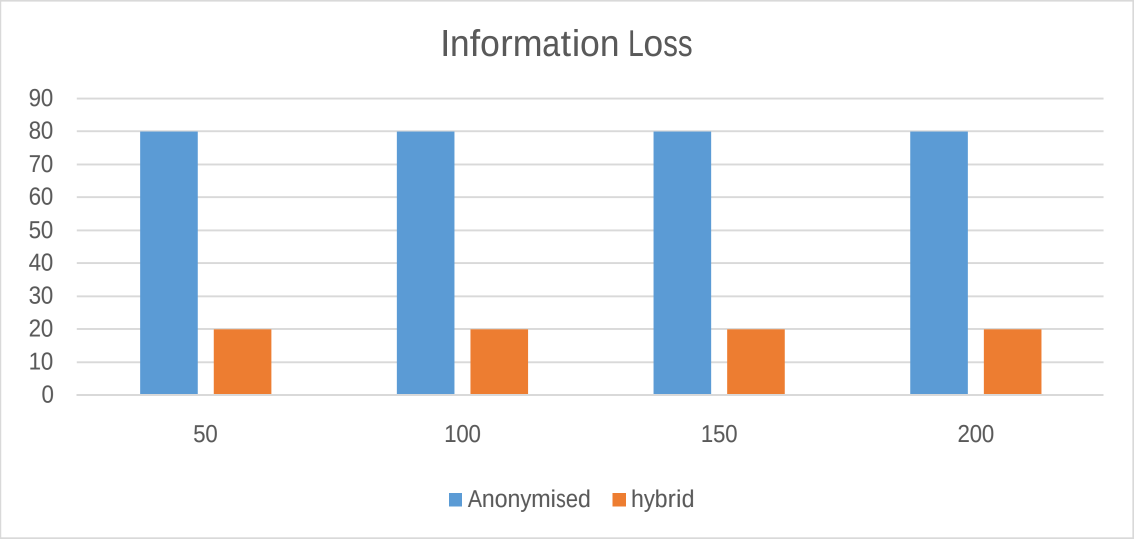
Fig 3. Algorithm's flowchart

V. TECHNOLOGY USED

The experiment is carried by running a local server on the windows system. A virtual internet like environment is setup and the data is displayed on the web browser with the help of local server apache tomcat, the way it is displayed on the server side. Apache tomcat is the open source software used for the implementation of the JAVA and JavaScript based projects, the Java Server Pages and Servlet. JAVA and JavaScript language is used for the development and the implementation part of the proposed approach. MS-Access is used to store the original as well as perturbed data on the machine.

VI. RESULT ANALYSIS

Hospitals record are created in database as shown in the table 1.This table contains original data set taken from hospitals located at Hyderabad area. In Table 2, the data is anonymized using generalization approach as sown in Figure1 and Figure 2. These generalization causes a lot of information loss and reduces data utility, though It successfully implants the privacy in the data. In table 3 the data is perturbed by applying the hybrid approach, whose flowchart is shown in Fig 3.Initially, the tables contain all QID attributes are generalized to the most general values, and the sensitive attributes remain ungeneralized. At each iteration, the algorithm performs a top-down specialization on a QID attribute and, for each qid group, performs cell generalization on the sensitive attribute to satisfy the personalized privacy requirement; the breach probability of inferring any domain sensitive values within the subtree of guarding nodes is below certain threshold.



The x-axis in this chart is number of patient and the y-axis represents the amount of information loss.Inthe above hybrid method it successfully reduces the Information loss to 20 % from 80% and hence makes the data much more useful . Due to the reverse process most of the data in the hybrid model is saved and only age field is not completely recovered as privacy weight attached to it is quite high hence it is encrypted in a stronger way.

VII. CONCLUSION

The paper proposes a hybrid Privacy Preserving Data Mining technique using suppression and perturbation over the centralized server environment. The real values of the data can also be retrieved back while performing the reverse process so there is no information loss. The important issue of securing the Boolean gender field without information loss is resolved by the described algorithm. The major challenge of information loss in the process of privacy preservation has been successfully achieved. The proposed method resolved the critical conflict between the privacy preservation and information loss.Diseaseand treatment data is highly identifiable and can be extremelysensitive and personal, revealing health conditions and individual traits. Furthermore, this type of data also revealsinformation about blood relatives, thus involving not only asingle individual. It is, therefore, critical to preventunwanted disclosure of this type of data, while preservingmaximum utility.

REFERENCES

[1] Zhu, J. (2009, August). A new scheme to privacy-preserving collaborative data mining. In Information Assurance and Security, 2009. IAS'09. Fifth International Conference on (Vol. 1, pp. 468-471). IEEE.

[2]M.Suriyapriya, A. Joicy, Attribute Based Encryption with Privacy Preserving In Clouds. International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321- 8169 Volume: Issue: 2

[3] Zhan, J., &Matwin, S. (2006, December). A crypto-based approach to privacy-preserving collaborative data mining. In Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on (pp. 546-550). IEEE.

[4]P.Samarati, “Protecting respondents identities in microdata release,” IEEE transactions on Knowledge and Data Engineering, vol. 13, no. 6, pp. 1010–1027, 2001.

[5]Zhang, X., & Bi, H. (2010, October). Research on privacy preserving classification data mining based on random perturbation. In Information Networking and Automation (ICINA), 2010 International Conference on (Vol. 1, pp. V1-173). IEEE.

[6] Khaled M. Khan and QutaibahMalluhi (2010) “How can cloud providers earn thir customers’ trust when a third party is processing data and technologies used to address these challenges” by IEEE Computer Society in IEEE Setember/October 2010.

[7]Malik, M. B., Ghazi, M. A., & Ali, R. (2012, November). Privacy preserving data mining techniques: current scenario and future prospects. In Computer and Communication Technology (ICCCT),2012 Third International Conference on (pp. 26-32). IEEE

[8]N. Mohammed, B. C. M. Fung, P. C. K. Hung and C.-K. Lee. Anonymizing Healthcare Data: A Case Study on the Blood Transfusion Service, KDD’09, June 28–July 1, 2009, Paris, France, 1285-1293.

[9] N. Mohammed, B. C. M. Fung, P. C. K. Hung and C.-K. Lee. Anonymizing Healthcare Data: A Case Study on the Blood Transfusion Service, KDD’09, June 28–July 1, 2009, Paris, France, 1285-1293.

[10]B. C. M. Fung, K. Wang, A. W.-C. Fu and P. S. Yu, Introduction to rivacy-Preserving Data Publishing - Concepts and Techniques, CRC Press, Taylor & Francis Group, 2011.

[11] [6] X. Xiao and Y. Tao. Personalized privacy preservation. In Proc. Of ACM International Conference on Management of Data (SIGMOD), Chicago, IL,2006.

[12]Arshveer Kaur A Hybrid Approach of Privacy Preserving Data Mining using Suppression and Perturbation Techniques International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2017)

[13] Majid Bashir Malik M. Asger A model for Privacy Preserving in Data Mining using Soft Computing Techniques 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)

[14]A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo,“Protection of big data privacy,” IEEE Access, vol. 4, pp. 1821–1834, 2016.

[15] A. Abbas and S. U. Khan, “A review on the state-of-the-art privacy-preserving approaches in the e-health clouds,” IEEE Journal of Biomed-ical and Health Informatics, vol. 18, no. 4, pp. 1431–1441, 2014.