Privacy Preserving Data Mining – Optimization In K-Anonymity using Machine Learning Approach

Patel Ankit K[1] Prof. Keyur N. Brahmbhatt[2] Prof. N. B. Prajapati[3]
[1]Student [2]Assistant Professor [3]Associate Professor
[1,2,3]Department of Computer Engineering
[1,2,3]Birla Vishvakarma Mahavidyalaya Vallabh Vidhyanagar

Abstract— In now days the information sharing is very important. One organization shares the information of user to another organization for the better survey purpose. But the sensitive data of user will not be disclosed. So for that purpose we have to hide some sensitive data of user for that the data must be encrypted. K-anonymity algorithm is one of the ways to encrypt data so that data cannot be stealing and the information in the data will not modify. But there is some way to attack on the k-anonymity encrypted data. One of the way is background knowledge attack, in this if the attacker knows some basic information about the use then he can get the detail from database. If we can add some more data in the original database and the apply k-anonymity algorithm so that the attacker is get more rows of data and he will confuse so the data should be protected from the attacker.

Key words: Anonymization, K- Anonymity, Machine Learning Algorithms

I. INTRODUCTION

We investigate privacy preservation from the anonymity aspect. The sensitive information, even after the removal of identifying attributes, is still susceptible to linking attacks by the authorized insiders [2]. This problem has been studied extensively in the area of micro data publishing and privacy definitions, e.g., k-anonymity [2] The anonymity techniques can be used with an access control mechanism to ensure both security and privacy of the sensitive information. However, privacy is achieved at the cost of accuracy and imprecision is introduced in the information provided by permissions under an access control policy. We use the concept of imprecision bound, for each permission to set a threshold on the amount of imprecision that can be tolerated. To exemplify our approach, role-based access control is assumed. However, the approach is generic and can be applied to any security policy, e.g., discretionary access control or mandatory access control. The heuristics proposed in this paper for privacy-preserving access control are also relevant in the context of workload-aware anonymization. There are many types of attacks are make on k- anonymity algorithm one most common is background knowledge attack in this type of attack the attacker is know some basic data of user so it easily identify the user data.

The issue of protecting the privacy of individuals before releasing micro data containing personal information has been studied extensively during the last decade [5]. Anonymization algorithms use suppression and generalization of records to satisfy privacy requirements, e.g., k-anonymity and with minimal distortion of micro data. Workload-aware anonymization techniques that minimize information loss for data mining tasks or for a given set of queries have been developed [6], [7]. However, the problem of satisfying accuracy constraints set by the multiple users of micro data has not been studied.

II. K-ANONYMITY

For the purpose of preventing record linkage through QID, k-anonymity is proposed that at least k records share the same qid. Thus, for one qid, we get k satisfying records and these records are indistinguishable from each other. Furthermore, the possibility for an attacker to link to the correct data owner is 1 / k.

III. MACHINE LEARNING ALGORITHMS

Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data. For example, a machine learning system could be trained on email messages to learn to distinguish between spam and non-spam messages. After learning, it can then be used to classify new email messages into spam and non-spam folders.

The core of machine learning deals with representation and generalization. Representation of data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the property that the system will perform well on unseen data instances; the conditions under which this can be guaranteed are a key object of study in the subfield of computational learning theory [5].

There is a wide variety of machine learning tasks and successful applications. Optical character recognition, in which printed characters are recognized automatically based on previous examples, is a classic example of machine learning.

IV. PROPOSED ALGORITHM

- Step 1: Get old data and train machine learning algorithm for get a good data se to be insert into data
- Step 2: Make a ranking model to add into main dataset and give a more complex and efficient data.
- Step 3: Get the data from main dataset and pass it to the short manager
- Step 4: Get data from the ranking model and Data manager and add the data like that the frequency of the data is not getting more impact and the data is become more complex and apply k-anonymity algorithm.
- Step 5: Pass data for the survey

By implementing this algorithm we can get the more confusing data. In first data make training dataset using machine learning so we can get a data that can be inserting into the dataset. In the data set we get the frequency of data occurrence is almost same as the data so the data for surviving is not getting effected by this algorithm.

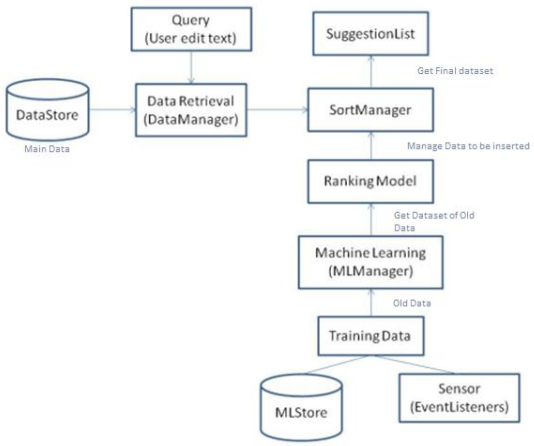

Fig. 1: Flowchart

A. *Results*

Name	AGE	ZIPCODE	SEX	NATIONLITY	DISEASE
Suresh	29	382028	Male	Indian	Cancer
Ramesh	31	382029	Male	Indian	Heart Attack
Jhon	45	525962	Male	American	Swine flue
Charli	25	896547	Female	Chinese	Cancer
Pankaj	22	484028	Male	Indian	Normal Fever
Janak	35	382028	Male	Indian	Swine flue
Dhaval	35	382028	Male	Indian	Swine flue
Ankit	28	382028	Male	Indian	Heart Attack

Fig. 2: Hospital Data

Name	AGE	ZIPCODE	SEX	NATIONLITY	DISEASE
Suresh	29	382028	Male	Indian	Cancer
Ramesh	31	382029	Male	Indian	Heart Attack
Jhon	45	525962	Male	American	Swine flue
Charli	25	896547	Female	Chinese	Cancer
Pankaj	22	484028	Male	Indian	Normal Fever
Janak	35	382028	Male	Indian	Swine flue
Dhaval	35	382028	Male	Indian	Swine flue
Ankit	38	382028	Male	Indian	Heart Attack
Ankit	29	382028	Male	Indian	Normal Fever
Hetal	31	382029	Female	Indian	Swine flue
Parbat	45	382030	Male	American	Swine flue
Dhaval	25	382023	Male	Chinese	Cancer
Maulik	22	382028	Male	Indian	Normal Fever

Fig. 3: data after adding machine learning data

Name	AGE	ZIPCODE	SEX	NATIONLITY	DISEASE
*	<30	3820**	Male	*	Cancer
*	>30	3820**	Male	*	Heart Attack
*	>30	5259**	Male	*	Swine flue
*	<30	8965**	Female	*	Cancer
*	<30	4840**	Male	*	Normal Fever
*	>30	3820**	Male	*	Swine flue
*	>30	3820**	Male	*	Swine flue
*	>30	3820**	Male	*	Heart Attack
*	<30	3820**	Male	*	Normal Fever
*	>30	3820**	Female	*	Swine flue
*	<30	3820**	Male	*	Swine flue
*	<30	3820**	Male	*	Cancer
*	<30	3820**	Male	*	Normal Fever

Fig. 4: After applying proposed algorithm

So, the data is become more so the attacker is confused between more occurrence of data. Here attacker is get three result so it is easy to find the user information and now for The proposed algorithm attacker get the 4 data and the main disease cancer and heart attack frequency is same. Attacker in now confuse in Cancer, Normal Fever, Heart Attack.

Name	AGE	ZIPCODE	SEX	NATIONLITY	DISEASE
*	<30	3820**	Male	*	Cancer
*	<30	3820**	Male	*	Normal Fever
*	<30	3820**	Male	*	Cancer
*	<30	3820**	Male	*	Heart Attack

Fig. 5: proposed algorithm attacker data

V. ADVANTAGE

- Make a more complex data
- Get more security on background knowledge attack then normal k-anonymity algorithm
- Data is added but the frequency of occurrence of data is not more change

VI. DISADVANTAGE

- Take a more time to make a survey data
- Database size is become large
- Database is not accurate like k-anonymity algorithm

REFERENCES

[1] Sridhar Mandapati, Dr. Raveendra Babu Bhogapathi, Ratna Babu Chekka" A Hybrid Algorithm for Privacy Preserving in Data Mining" I.J. Intelligent Systems and Applications, 2013, 08, 47-53 Published Online July 2013 in MECS (http://www.mecs-press.org/)

[2] Zahid Pervaiz Center for Education and Research Information Assurance and Security Purdue University, West Lafayette, IN 47907-2086" Privacy-preserving Access Control" IEEE, Arif Ghafoor Fellow, IEEE, Nagabhushana Prabhu[2013].

[3] LATANYA SWEENEY School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA E-mail: latanya@cs.cmu.edu " Achieving K-Anonymity Privacy Protection using Generalization And Suppression"2002

[4] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati Universitµa degli Studi di Milano, 26013 Crema, Italia fciriani, decapita, foresti, samaratig@dti.unimi.it" k-Anonymity"[2007]

[5] Shijun Wang, Ronald M. Summers Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Building 10 Room 1C224D MSC 1182, Bethesda, MD 20892-1182, United States" Machine learning and radiology"[2012]

[6] Peter C. Austina,b,c,*, Jack V. Tua,b,d, Jennifer E. Hoe,f,g, Daniel Levye,f,h, Douglas S. Leea,b,i" . Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes" 2013.

[7] Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati,"k-Anonymity Data Mining: A Survey", Springer US, Advances in Information Security (2007)

[8] www.wikipedia.com/neuralnetwork/

[9] www.wikipedia.com/machinelearning/

[10] Ashwin Machanavajjhala, Johannes Gehrke Daniel Kifer, Muthuramakrishnan Venkitasubramaniam, "ℓ-Diversity: Privacy Beyond k-Anonymity".
[11] https://en.wikipedia.org/wiki/L-diversity
[12] M.Saranya, "A Survey on Privacy Preservation for Anonymzing Data", International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015

