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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

A Survey on Novel Approach for Non-cryptographic Privacy Preservation

Vasudha Deokate

M.E Student, Dept. of Computer, Dattakala college of Engineering, Swami- Chincholi(Bhigwan), Daund, India.

ABSTRACT: Through data mining system can extract knowledge from large amount of data in many large companies and organizations. So from such a large collection of data there generates some problem with related to the privacy. The privacy is main factor used in many large organizations like medical database, business interests and personal interests. The collected data may contain private and sensitive data which should be protected from world .The privacy protection is an important issue of any organization. If any organization release data of outside world for sharing purpose. Privacy preserving data mining approaches allow publishing data for the mining purpose while at that time preserve the private data of the personalize. The proposed system is a very simple and efficient approach for Privacy preserving data mining. Cryptographic techniques protect sensitive data with less information loss. Not only the data usability but also complexity increase and also protect the sensitive data for various types of unauthorized user.

KEYWORDS: Data mining, Sensitive data, Privacy preserving.

I. INTRODUCTION

There are several large organizations like credit card companies, real estate companies, search engines, hospitals collects, Institutes and hold large amount of information. The data are further used by the data mining for the analysis purpose which helps the organizations for gaining useful knowledge. These data may include sensitive or valuable information of any individuals, For example, organizations such as hospitals contain medical records of the patients, and they provide these dataset or information to the researchers or data miner for the purpose of research. Data analyzer analyzes the various medical records to gain useful global health statistics. However, in this process the data miner may able to obtain sensitive information and in combination with an external dataset may try to obtain personal attribute of an individual’s privacy is become an important issue when data that includes sensitive information. To solve this, an interesting new topic in the field of data mining has been known as privacy preserving data mining (PPDM). The goal of this technique is extraction of useful knowledge from very large of data, while protecting the sensitive information simultaneously. Privacy preserving data mining techniques are separated into two parts:

1) Data hiding technique and
2) Knowledge hiding technique

Data hiding is changes or edit of confidential information from the data before disclosing to others. Knowledge hiding is based on hiding the sensitive knowledge which can be retrieving from the database using any data mining algorithm. In this Proposed System mainly focus on non-cryptographic techniques. In proposed method, first apply randomization on original data and then after randomization categorize the sensitive attribute values into high sensitive and low sensitive class. Secondly apply k-anonymization on those tuples who belongs to high sensitive class and those tuples who belongs to low sensitive remain as it is. So it reduces the information loss and improves the data usability. The combination of anonymization technique and randomization technique is made not easy for the attacker to attack on database.

II. RELATED WORK

There are various methods for privacy preserving data mining:

DOI: 10.15680/IJIRCCE.2015. 0312012

1) First Randomization Method: The randomize method is very simple & effective in privacy preserving data mining. Randomize method is easy and very popular method in current data mining process. It‘s data mining process the noise data is include in original dataset to mask the attribute value of datasets. The actual data cannot be recovered at very short time interval because the very large amount of unwanted data including in original dataset [8]. To collecting a data in a randomize method in a two steps:

a) **Step first:** Original Dataset provides randomize their data and transmit randomized data to data receiver.

b) **Step Second:** Data receiver estimates original distribution reconstruction algorithm. Suppose there is central servers perform the role of data collector for example, of a college, and many students, each having its own of information. So server collects the information from the student & performs data mining process to create an aggregate data model [9].

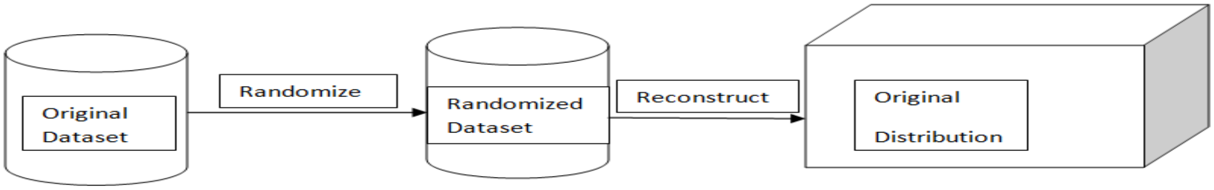


Fig: Randomization Method

Now to protect the student’s data by selecting they randomly reorder their records before sending them to the server, taking some true information and introducing some noise or unwanted data. At the server‘s side, statistical estimation over noisy data is students to recover the aggregates needed for data mining. Noise can be introduced, for example, by adding some data or multiplying random values to numerical attributes or by deleting some real items and adding wrong information to set-valued records[2][4].

Advantages:
i. The randomization method is very simple method and easily applied when collect the data.
ii. It is protecting individual’s privacy.
iii. It is more efficient and very simple working anyone can use.

Disadvantages:
i. Multiple attribute databases are used but they are not suitable.
ii. When data collector collect the data from data provider the data provider adds some noise in data and to reorder that data it takes more time that why it is very slow technique.

I. Methods of Anonymization

Selecting information is removed from the original dataset to protect personal or private information is also called as Anonymization. There are many ways to perform data anonymization basically this method uses k-anonymization approach. If each row in the table cannot be different from at least other k-1 rows by only showing a set of attributes, then this table is K-anonymized on these attributes [7].

The simple example of data anonymization is college student. In the students record information of a single patient is stored in a single line is also called as tuple. i.e. tuple, and database is store confidentially at the server side. The users may be a medical researchers they have the access to Database. Since Database is anonymous because the one important part is to protect the privacy of students. Such this part is guaranteed through the use of anonymization. If the database is anonymous, it is not possible to identify the student’s record.

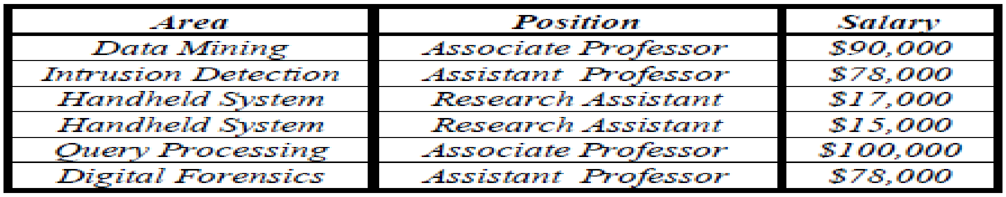
There are two anonymization methods:

A.**First Suppression-based k-anonymization:**

Assume that the content in table T={t_1,…..,t_n} all over the attribute set A. The idea is to form subsets of same tuples by masking the values of some well-chosen attributes. In special, when using a suppression-based anonymization method and mask with the special value‘*’. In this method uses following notations: Quasi-Identifier (QI): using external information to identify a specific individual and it contain a set of attributes. T [QI]: T [QI] is the projection of T to the set of attributes including in QI [9].

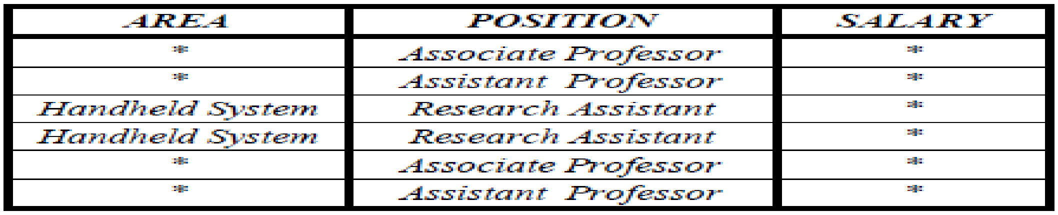
B.Second Generalization-based k-anonymization:

In generalization-based anonymization method, original values are change by more general ones in the database, according to a priori established value generalization hierarchies (VGHs)[1][6].



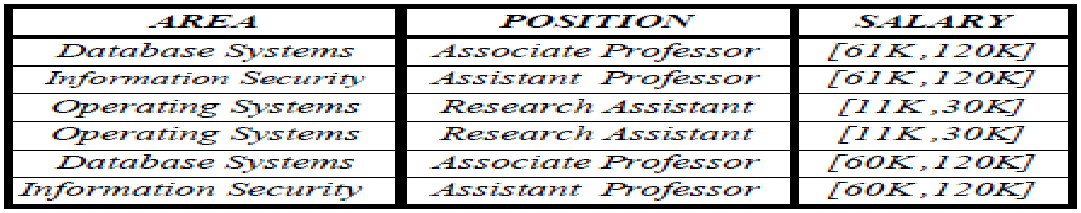
Area	Position	Salary
Data Mining	Associate Professor	$90,000
Intrusion Detection	Assistant Professor	$78,000
Handheld System	Research Assistant	$17,000
Handheld System	Research Assistant	$15,000
Query Processing	Associate Professor	$100,000
Digital Forensics	Assistant Professor	$78,000

Table 1: Original dataset



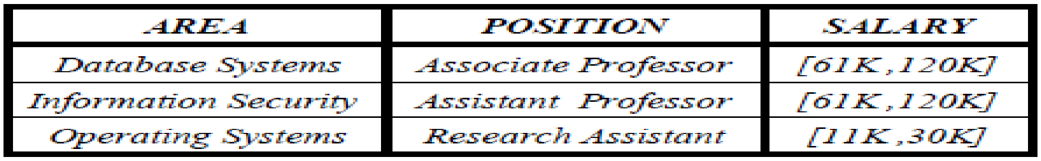
AREA	POSITION	SALARY
*	Associate Professor	*
*	Assistant Professor	*
Handheld System	Research Assistant	*
Handheld System	Research Assistant	*
*	Associate Professor	*
*	Assistant Professor	*

Table 2: Suppressed Data with k=2



AREA	POSITION	SALARY
Database Systems	Associate Professor	[61K,120K]
Information Security	Assistant Professor	[61K,120K]
Operating Systems	Research Assistant	[11K,30K]
Operating Systems	Research Assistant	[11K,30K]
Database Systems	Associate Professor	[60K,120K]
Information Security	Assistant Professor	[60K,120K]

Table 3: Generalized Data with k=2



AREA	POSITION	SALARY
Database Systems	Associate Professor	[61K,120K]
Information Security	Assistant Professor	[61K,120K]
Operating Systems	Research Assistant	[11K,30K]

Table 4: The Witness Set

Table 1 contains the original dataset that include all the actual information in the form of tuple. Then after applying suppression based technique on original dataset the original dataset is anonymized and display the anonymized records it make a changes in two QI and the value of k=2 in table 2 .Now in table 3 shows the generalized method result with replacing the value after the data mining process is applied. The Data Mining is important important because the generalized to more specific value with Database Systems. So like this the remaining values is replacing in table and more general value the original dataset is anonymized by applying generalized method. Finally When T is k-anonymous, and then replacing duplicate tuples, and call the resulting set the witness set of T. Table 4 presents a witness set of Table 3[9].

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

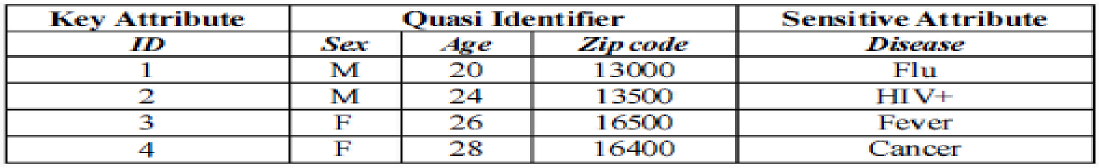
Advantages:

i. More general value is place in actual value and it becomes very difficult to find out or guess actual data.
ii. K-anonymous techniques is very fact and efficient as compared to previous techniques.
iii. By replacing actual value with * symbol. Because unauthorized user get confused and it creates many possible combination related to original dataset.

Disadvantage:

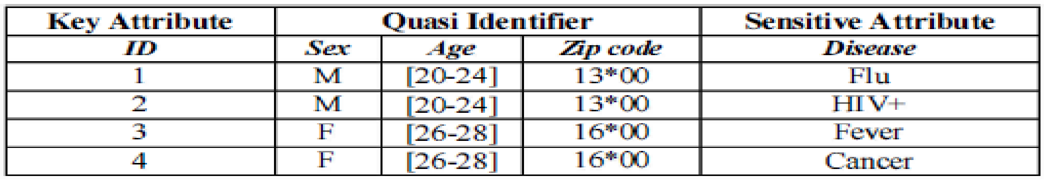
i. generalization main problems are it fails on high-dimensional data due to the curse of dimensionality it causes too much information loss due to the uniform distribution consideration.
ii. The database with the tuple data is very difficult and does not be maintained confidentially [2].

When small amount of data is released for the research purpose, one needs to limit disclosure risk while large amount the utility of data. Sweeny introduced the k-anonymity technique to some limitations the disclosure risk [4]. K - anonymity requirements that a data set is k anonymous (k 2: I) if each record in the data set is in different from at least (k-l) other records within the same data set. This k-anonymity requirement is generally achieve by using generalization technique and suppression technique [5]. In generalization the attribute values are generalized in a particular two values interval [6][7]. In suppression the attribute values are changes or edit with some other values. Suppression includes minimum loss of information so it is generally avoided. K-anonymous table include three types of attributes. First one is key attributes like name, SSN No, ID etc. which can be used to the individuals uniquely identification . Second attribute is quasi identifier (QI) attribute which are generally access with publically available database to re-identify the individuals. This is called linking attack [8]. Third attributes are sensitive attribute which needs to be protected. In table 5 see the diagnosis data set. Table 6 shows the 2-anonymous view of table 5.



Key Attribute	Quasi Identifier			Sensitive Attribute
ID	*Sex*	*Age*	*Zip code*	*Disease*
1	M	20	13000	Flu
2	M	24	13500	HIV+
3	F	26	16500	Fever
4	F	28	16400	Cancer

Table 5: DIAGNOSIS DATA SET



Key Attribute	Quasi Identifier			Sensitive Attribute
ID	*Sex*	*Age*	*Zip code*	*Disease*
1	M	[20-24]	13*00	Flu
2	M	[20-24]	13*00	HIV+
3	F	[26-28]	16*00	Fever
4	F	[26-28]	16*00	Cancer

Table 6: ANONYMOUS VIEW-2 OF TABLE 5

In table 6 sex, age and zip code is considering that as a quasi-identifier attributes group. Age is applied generalized in particular intervals and zip code is apply suppressed. While k-anonymity protects identity disclosure but it suffers from attack which leads to attribute disclosure. Several techniques are representing for privacy preserving in data mining process but they have some shortcomings like information loss and data utility. This research work is mainly focus on combined randomization technique and k-anonymity technique to preserve the privacy and increase utility of data and minimum loss of information.

III. **PROPOSE ALGORITHM**

In this Proposed System mainly focus on non-cryptographic techniques. The proposed approach uses the combined techniques of randomization and k-anonymization. It contains three main advantages:

1. It protects private data with minimum loss of information.
2. Increased the data utility.
3. Data can also be reconstructed using those techniques.

In proposed system is dived into two algorithms.

I. Randomization Algorithm
II. k-anonymity algorithm

In randomization algorithm is performed on dataset using attribute transitional probability matrix and in k-anonymity algorithms performed on result of randomized algorithm. In proposed method, first apply randomization on original data and then after randomization classified the sensitive attribute values into high sensitive and low sensitive class. Secondly apply k-anonymization on those tuples who belongs to high sensitive class and those tuples who belongs to low sensitive remain as it is. So it reduces the minimum information loss and improves the data usability. The combination of randomization technique and k- anonymization is made difficult for the attacker to attack on database.

ALGORITHM I

Input Method: Original dataset T, Transitional probability matrix P, I * j size mapping matrix M this is between T & P.

Output Method: Convert table C.

Method:

a) The select quasi identifier (QI) and key attributes and sensitive attribute from table T.
b) Remove/Suppress the key attributes.
c) Generate transitional probability matrix P with size j*j randomly.
d) Generate mapping matrix M randomly.
e) According to mapping matrix M assign each P (Pl, P2,.....P j) to T (T1, T2, ...T j).
f) With respect to highest location of P value, rearrange the element of T. If highest location is used already then go to the next higher location of P. If value of P of two or more location is same than it will choose the left hand side value only.
g) Re-combine T matrix.
h) Re-substitute in table.
i) Stop.

ALGORITHM II

- **Input Method:** Converted table C (Result of algorithm I), Anonymized parameter k.
- **Output Method:** Finally derive table D.
- **Method:**

a) Select the table C.
b) Categorize the sensitive attribute values into two classes: high (H) class and low (L) class.
c) For each tuple whose sensitive values belong to class H - Move those tuples into table D1 and applying generalization on quasi attributes (QI) to anonymize it.
d) For each tuple whose sensitive values belong to class L - Move these tuples into table D2 and do not Sanonymized it.
e) Append rows of table DI and table D2 and get final derived table D. D = D1, D2.
f) Stop.

VI. Conclusion and Future Work

There are several privacy preserving techniques available but still they have some advantages and disadvantages. Anonymity technique gives privacy protection and data usability but it suffers from attack. Cryptography technique gives privacy protection but does not provide data usability and it requires more computational overhead. Randomized response technique preserve privacy but they are information loss. In the proposed method combined K-anonymity with randomization. It makes difficult for the attacker to identify background and homogeneity attack. In combination of two different techniques using that it protects private data with better accuracy and gives less loss of information which increases data utility. Data can also be reconstructed by using proposed approach but minimum loss of information.

REFERENCES

1. Manish Sharma , Atul Chaudhary, Manish Mathuria, Shalini Chaudhary, Santosh Kumar , “An Efficient Approach for Privacy Preserving in Data Mining”, IEEE 2014.
2. Tamanna Kachwala, Sweta Parmar , “An Approach for Preserving Privacy in Data Mining”, Research Paper, 2014.
3. L. Sweeny, "K-Anonymity: A Model for Protecting Privacy", International Journal on Uncertainty, Fuzziness and Knowledge based System, pp. 557 -570, 2002.
4. K. Chen and L. Liu,"Privacy Preserving Data Classification with Rotation Peturbation", Proceedings of the Fifth International Conference of Data Mining (lCDM' 05), pp. 582 - 589,2005.
5. K. Wang, P.S. Yuand S. Chakraborty,"Bottom Up Generalization: A Data Mining Solution to Privacy Protection", In International Conference on Data Mining, pp. 249 - 256, 2004.
6. Smita D Patel, Sanjay Tiwari, "Privacy Preserving Data Mining", International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 139 – 141.
7. H. Karagupta, S.Datta, Q. Wang and K. Sivakumar, "Random Data Peturbation Techniques and Privacy Preserving Data Mining", IEEE International Conference on Data Mining 2003.
8. "Top Down Specialization For information and privacy preservation”, International Conference on Data Engineering (lCDE' 05), pp. 205 - 216.
9. Mr. Mahesh T.Dhande1, Mrs. N.A.Nemade2, Mr. Yogesh V. Kolhe,“Privacy Preserving in K- Anonymization Databases Using AES Technique”, International Journal of Emerging Technology and Advanced Engineering,pp.1-4,March 2013.