International Journal of Recent Trends in Engineering & Research

ISSN (ONLINE) : 2455 – 1457

IMPACT FACTOR : 4.101

National Conference on Convergence of Emerging Technologies in Computer Science & Engineering (CETCSE-2K17)

An Enhanced Approach for Privacy Preserving Data Mining (PPDM)

[1]CH Venkata Lakshmi [2]Swapna Siddamsetti

[1]Assistant Professor, Department of Computer Science and Engineering, NNRESGI, Hyderabad.

[2]Senior Assistant Professor, Department of Computer Science and Engineering, Auroras Technological and Research Institute, Hyderabad.

***Abstract*—With the development of network, data collection and storage technology, the use and sharing of large amounts of data has become possible. Once the data and information accumulated, it will become the wealth of information. However, traditional data mining techniques and algorithms directly operated on the original data set, which will cause the leakage of privacy data. At the same time, large amounts of data implicate the sensitive knowledge that their disclosure cannot be ignored to the competitiveness of enterprise. In order to overcome these problems, Privacy Preserving Data Mining (PPDM) techniques are developed. Traditional PPDM techniques suffer from different types of attacks and loss of information. In this paper an alternative method was proposed which provides less information loss and more privacy.**

Keywords—storage technology, leakage of privacy data, sensitive knowledge, PPDM techniques.

I. Introduction

Due to World Wide Web, there is vast amount of information available over the network. This information could be accessed through data mining. Data mining is concerned with the extraction of non-trivial, novel and potentially useful knowledge from large databases. In order to extract the knowledge various data mining techniques are used as per the application domains like health care, Cyber security, banking, e-commerce etc.. [1]. These domains contain confidential data which should not be disclosed to all users which lead to the development of privacy preserving techniques in data mining [2].

PPDM is a new research in data mining where data mining functionalities are analyzed with respect to privacy in data. Privacy Preserving Data Mining [2][3]. (PPDM) is defined as getting valid data mining results without learning the underlying data values or extraction of knowledge from large datasets by preventing the access of sensitive information. The main consideration of PPDM is twofold. First, sensitive raw data like identifiers, names, addresses etc., should be hidden or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, the data which has to be mined from the database by using data mining algorithms should also be precluded, because such data will compromise data privacy. i.e., the main purpose of PPDM is to develop techniques for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process

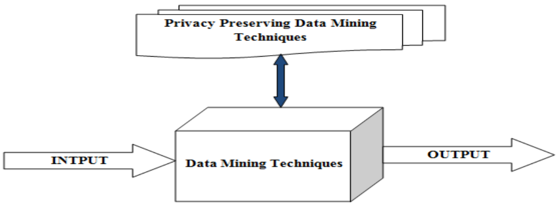


Fig. 1: PPDM model

The key aspects in PPDM model are:

- Privacy-preserving data publishing.
- Modifying the outputs of data mining algorithms to preserve privacy.
- Query auditing.
- Cryptographic methods for distributed privacy.
- Theoretical challenges in high dimensionality.

Privacy-preserving data publishing

These techniques study different transformation methods associated with privacy, e.g., randomization [3], k-anonymity [4], l-diversity [5] and also handles like how perturbed. Data can be used in conjunction with association rule mining approaches.

Modifying the outputs of data mining algorithms to preserve privacy

The outputs of data mining algorithms such as association rule mining are modified in order to preserve privacy of data. Example: association rule hiding.

Query auditing

The results of queries are either modified or restricted. Example: output perturbation and query restriction.

Cryptographic methods for distributed privacy

In many cases, the data may be distributed across multiple sites, and the owners of the data may wish to compute a common function. A variety of cryptographic protocols may be used in order to communicate between the sites, so that secure function computation is possible.

Theoretical challenges in high dimensionality

Real data sets are high dimensional making the process of privacy preservation difficult both from computational and effectiveness point of view. Example: optimal k- anonymization is NP-hard.

II. Types of PPDM Techniques

In this paper, privacy preserving techniques [1],[6],[7], has been classified based on the data lifecycle phases such as data collection, data publishing, and at the output of data mining.

1. Data Collection

To ensure the privacy at data collection period, the sensory device transforms the raw data by randomizing the values, before sending it to the collector. In order to perform this method used is randomization.

Randomization

In this method the original data values are modified by adding noise using known statistical distribution, such that when data mining algorithms are implemented the original data distribution may be reconstructed but not the original data values. This randomization method is known as randomization with additive noise. Another way to add noise is by multiplying noise with a known statistical distribution known as randomization using multiplicative noise.

Assume X be the original data distribution i.e., X={x1,x2,x3,…xn}, Y be the publicly known noise distribution independent of X i.e., Y={y1,y2,..,yn} and Z be the result of randomization. The simplest additive noise randomization approach can be described as: Z=X+Y

The collector estimates the Z distribution from the received n samples z1,z2,..,zn, then X may be reconstructed using noise distribution Y i.e.,

X=Z-Y

The below fig.(1) depicts the output after adding noise to non random data.

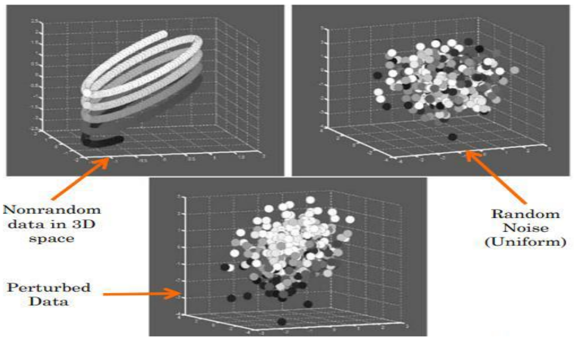


Fig. 2: After adding additive noise

Since original data values are modified into perturbed data as shown in the fig. randomization method requires specific data mining algorithms that can leverage knowledge discovery from distribution of data but not from original values.

Additive noise is not the effective method for preserving privacy as shown by

[9]. The alternate randomization method namely multiplicative noise [8] is used.

Data modification may be applied at other phases apart from data collection and they use different methods rather than additive noise and multiplicative noise. But in this method, the collector is not trusted. So the original data must not be stored, nor buffered after transformation.

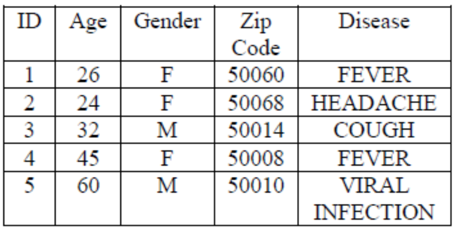
2. Data Publishing Privacy

Entities may release the data either publicly or to third parties for data analysis without hiding the ownership of the sensitive data. In this scenario, privacy can be preserved by anonymizing the records before publishing. Hence PPDM at data publishing is known as Privacy Preserving Data Publishing (PPDP).

Some of Privacy Preserving methods at data publishing:

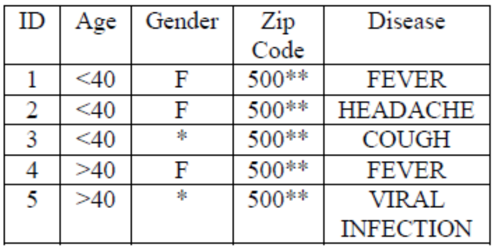
a. k-anonymity if the identifiable attributes of any database record are undistinguishable from at least other k-1 records, then the dataset is known as k-anonymous [4]. In other terms, with a k-anonymized dataset, an attacker could not identify the identity of a single record since there exist k-1 similar records.

Consider the below example of k-anonymization:



ID	Age	Gender	Zip Code	Disease
1	26	F	50060	FEVER
2	24	F	50068	HEADACHE
3	32	M	50014	COUGH
4	45	F	50008	FEVER
5	60	M	50010	VIRAL INFECTION

Table 1: original medical diagnosis dataset



ID	Age	Gender	Zip Code	Disease
1	<40	F	500**	FEVER
2	<40	F	500**	HEADACHE
3	<40	*	500**	COUGH
4	>40	F	500**	FEVER
5	>40	*	500**	VIRAL INFECTION

Table 2: After applying k-anonymization on medical diagnosis dataset

In the above example, entries in the age attribute are replaced by using generalization technique and the attributes Gender and Zip code are replaced by using suppression technique.

b. l-diversity l-diversity [5] is based on the observation that if the sensitive values in one equivalence class lack diversity, then no matter how large the equivalence class is, attacker may still guess the sensitive value of an individual with high probability.

By intuition, a table is l-diverse if each equivalence class contains at least l “well represented” sensitive values, that is, at least l most frequent values have very similar frequencies. Consider a table T = (A1,A2,…….An, S) and constant c and i,

where (A1,A2,….An) is a quasi-identifier and

S is a sensitive attribute. Suppose an equivalence class EC contains values s1,s2….sm with frequency f(s1),f(s2)….f(sm)

(appearing in the frequency non-ascending order) on sensitive attribute S [10], EC satisfies (c, i)-diversity with respect to S if

$f(s1) < c\sum m \ =1 \ f(si)$

3. Data Mining Output privacy

The outputs of the data mining algorithms may reveal the sensitive information without explicitly accessing the original dataset. The most common techniques to preserve the privacy to the output of data mining are:

a. Association rule hiding In data mining the association rules may explicitly reveal private information about an individual or group of individuals. So Association rule hiding technique will mine the only non-sensitive data. That is, in this technique data is perturbed to prevent mining of sensitive rules.

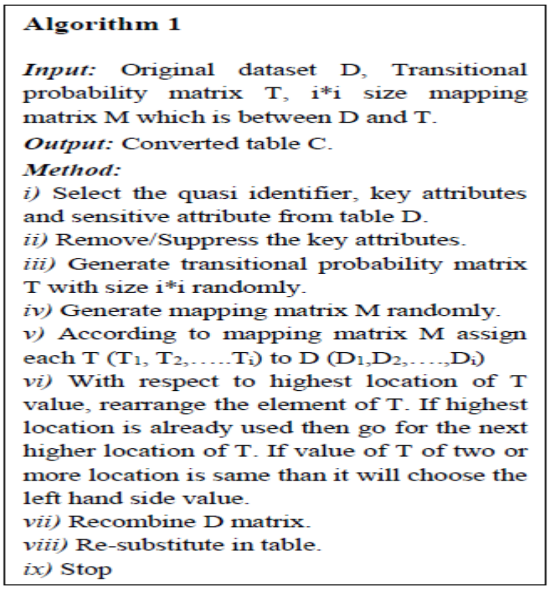
b. Query auditing In some cases, some entities may provide access to the original dataset, allowing exclusively statistical queries to the data. Users can only query the aggregate data from the dataset but not the individual records. Still some queries may reveal original or private data [2]. This can be overcome by query auditing.

Query auditing technique provides privacy by using two approaches: Query inference control and Query auditing. In Query inference control, either the original data or the output of the query is perturbed. In Query auditing, one or more queries are denied from a sequence of queries.

III. Proposed Technique

The above listed Privacy preserving algorithms have drawbacks like Information loss and data utility. This paper mainly focuses on integration of randomization [3] and k-anonymity [4] techniques to preserve privacy and reduce the information loss and increase the privacy gain.

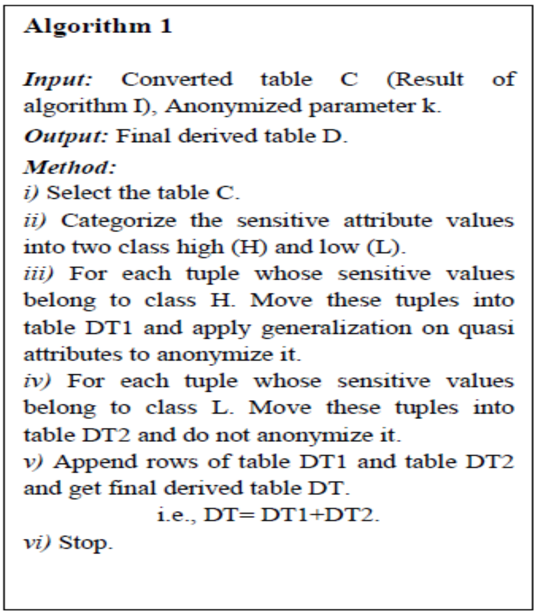
This method is divided into two algorithms. Algorithm 1 is used to perform randomization on dataset using attribute transitional probability matrix and algorithm 2 is used to perform k-anonymity on randomized output dataset. In other words, algorithm 2 uses the result of algorithm 1 as input.



```
Algorithm 1

Input: Original dataset D, Transitional probability matrix T, i*i size mapping matrix M which is between D and T.
Output: Converted table C.
Method:
i) Select the quasi identifier, key attributes and sensitive attribute from table D.
ii) Remove/Suppress the key attributes.
iii) Generate transitional probability matrix T with size i*i randomly.
iv) Generate mapping matrix M randomly.
v) According to mapping matrix M assign each T ($T_1$, $T_2$,.....$T_i$) to D ($D_1$,$D_2$,....,$D_i$)
vi) With respect to highest location of T value, rearrange the element of T. If highest location is already used then go for the next higher location of T. If value of T of two or more location is same than it will choose the left hand side value.
vii) Recombine D matrix.
viii) Re-substitute in table.
ix) Stop
```

In Algorithm 1, first the quasi identifier, sensitive attribute and key attribute are selected from table D. since key attribute uniquely identifies the individuals, it can be suppressed or removed. Once T and M are generated randomly, elements of D are rearranged w.r.t the highest value of location of T. If the highest location is already used then it takes the next highest location. If two are more location of T are same then it will chose the left hand side value. All values are re-substituted in table D. The result of Algorithm 1 will be table C on which k-anonymity is applied.



```
Algorithm 1

Input: Converted table C (Result of algorithm I), Anonymized parameter k.
Output: Final derived table D.
Method:
i) Select the table C.
ii) Categorize the sensitive attribute values into two class high (H) and low (L).
iii) For each tuple whose sensitive values belong to class H. Move these tuples into table DT1 and apply generalization on quasi attributes to anonymize it.
iv) For each tuple whose sensitive values belong to class L. Move these tuples into table DT2 and do not anonymize it.
v) Append rows of table DT1 and table DT2 and get final derived table DT.
        i.e., DT= DT1+DT2.
vi) Stop.
```

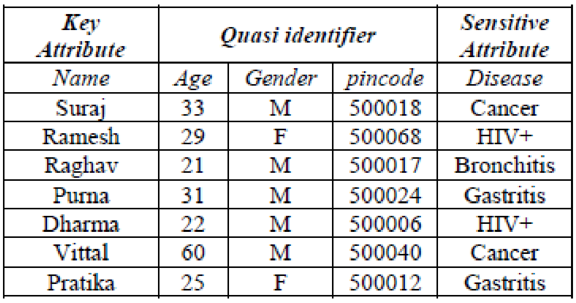
In k-anonymity technique, consider all tuples as equally sensitive, so all the tupes get anonymized resulting in information loss. To overcome this problem, k-anonymity method is modified in which we categorize the sensitive attribute values into high sensitive class and low sensitive class.

IV. Result

Consider an example medical dataset as shown in table 3.

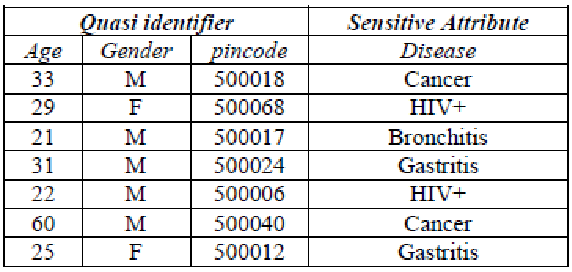
In table 3, name is considered as key attribute, so it has to be removed from the dataset since it uniquely identifies the individuals. After removing the name attribute, we get table 4. In table 4. Age, Gender and pincode attributes are considered as quasi identifiers and disease attribute is considered as sensitive attribute. Assume D1 is Age, D2 is Gender and D3 is pincode and randomly generate 7*7 size matrices (T1,T2,T3) respectively because the number of tuples are 7.

Table 3. Medical Dataset



Key Attribute	*Quasi identifier*			*Sensitive Attribute*
Name	*Age*	*Gender*	*pincode*	*Disease*
Suraj	33	M	500018	Cancer
Ramesh	29	F	500068	HIV+
Raghav	21	M	500017	Bronchitis
Purna	31	M	500024	Gastritis
Dharma	22	M	500006	HIV+
Vittal	60	M	500040	Cancer
Pratika	25	F	500012	Gastritis

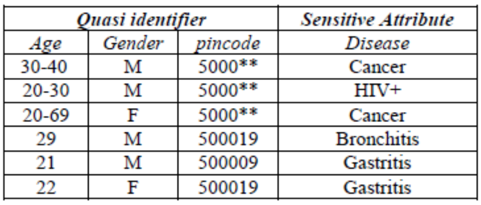
Table 4. After removing key attribute (e.g., Name)



Quasi identifier			*Sensitive Attribute*
Age	*Gender*	*pincode*	*Disease*
33	M	500018	Cancer
29	F	500068	HIV+
21	M	500017	Bronchitis
31	M	500024	Gastritis
22	M	500006	HIV+
60	M	500040	Cancer
25	F	500012	Gastritis

Table 4 dataset is given a input to algorithm 1, it processes the dataset by considering randomly generated values of Probability matrix (T) and Mapping matrix (M) and produces a converted table C as shown in table 5.

Table 5. Final Derived Table DT



Quasi identifier			*Sensitive Attribute*
Age	*Gender*	*pincode*	*Disease*
30-40	M	5000**	Cancer
20-30	M	5000**	HIV+
20-69	F	5000**	Cancer
29	M	500019	Bronchitis
21	M	500009	Gastritis
22	F	500019	Gastritis

In Table 5, the sensitive attribute Disease contains four values namely cancer, HIV+, Bronchitis, and Gastritis. Among which HIV+ and Cancer are categorized into High Sensitive class (H) and Bronchitis and Gastritis are classified into low sensitive class (L). Generalization is applied only for high sensitive class, so in the table 5 cancer and HIV+ values are generalized using suppression techniques. That is the pincode 5000** indicates the values from 500006 to 500019. The whole tuple is not generalized indicating less information loss and more privacy gain

The performance of the proposed methodology is evaluated in terms of two data metrics namely information loss [11] and privacy gain. The following formulae are used to measure information loss ILoss and privacy gain.

ILoss(g)= (|VG|-1)/|DA|

Table 6. Results comparison with different methods

Methods	*ILoss*	*PG*
Randomization	1.6	9
k-anonymity	1.18	16
Proposed method	0.73	14

Where ILoss indicates Information Loss

PG indicates Privacy Gain

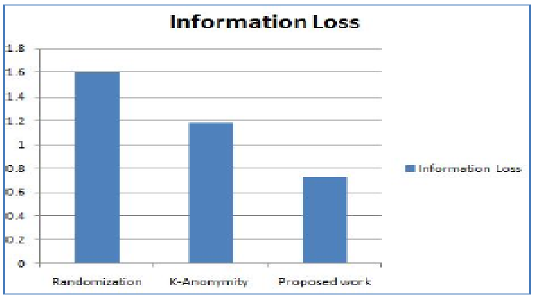


Fig.3(a)

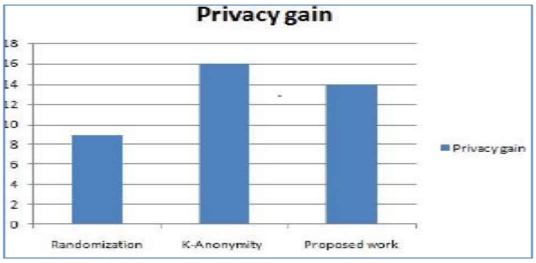


Fig. 3(b)

Fig. 3(a) & 3(b). Information Loss and Privacy Gain for sample dataset

In the above Fig. 3(a) and Fig. 3(b), it is observed that the information gain, privacy gain has improved in the proposed method compared to the traditional methods.

V. Conclusion

Public and Private Organizations and Educational Institutions gather data to provide their services. Those services may require collection, analysis and sharing of private sensitive data. Maintaining privacy to such sensitive data by using data mining techniques became challenge. Various PPDM techniques are proposed to extract knowledge by preserving privacy of individuals. In this paper, an overview of PPDM Techniques and an efficient method to overcome the problems faced by traditional PPDM techniques is provided. The proposed method is the combination of randomization and k-anonymity. It makes the attacker difficult to find the background and homogeneity attack. And also it provides less information loss and more Privacy for the sensitive data.

REFERENCES

[1] L. Cranor, T. Rabin, V. Shmatikov, S. Vadhan, and D.Weitzner, Towards a privacy research roadmap for the computing community, Comput. Commun. Consortium Committee, Comput. Res. Assoc., Washington, DC, USA, White Paper, 2015.

[2] C. C. Aggarwal and P. S. Yu, A general survey of privacy-preserving data mining models and algorithms, in Privacy-Preserving Data Mining, New York, NY, USA: Springer, 2008, pp. 11_52.

[3] Agarwal, R and Shrikant, R, Privacy Preserving Data Mining, Proceeding of Special Interest Group on Management of Data, 2000.

[4] Samarati P, Sweeney L. Protecting Privacy when Disclosing Information:k-Anonymity and its Enforcement through Generalization and Suppression. IEEE Symp. on Security and Privacy, 1998.

[5] A Machanavajjhala, D. Kifer, J Gehrke, and M. Venkata Subramaniam: l-diversity: Privacy beyond k-anonymity, ACM Trans. Knowl. Discovery Data, vol. 1, no. 1, p. 3, 2007

[6] S. Dua and X. Du, Data Mining and Machine Learning in Cybersecurity. Boca Raton, FL, USA: CRC Press, 2011.

[7] A. Shah and R. Gulati, ``Privacy preserving data mining: Techniques, classification and implications-A survey," Int. J. Comput. Appl., vol. 137, no. 12, pp. 40_46, 2016.

[8] K. Liu, H. Kargupta, and J. Ryan, Random projection based multiplicative data perturbation for privacy preserving distributed data mining, IEEE Trans. Knowl. Data Eng., vol. 18, no. 1, pp. 92_106, Jan. 2006.

[9] Kargupta H, Datta S, Wang Q, and K.Sivakumar, On the privacy preserving properties of random data perturbation techniques, in Proceedings, 3rd IEEE International Conference on Data Mining, Nov. 2003, pp. 99-106.

[10] Jaydip Sen: Privacy Preserving Data Mining – Applications, Challenges and Future Trends, in Proceedings 2nd ICCCT, MNNIT, Sep. 2011.

[11] Nissim Matatov, Lior Rokach and Oded Maimon "Privacy-preserving data mining: A feature set partitioning approach"2010.

[12] Langhreinrich L, Privacy in ubiquitous Computing, Boca Raton, FL, USA: CRC Press, 2009. Ch.3, pp.95-159.

[13] C.M. Bishop, Pattern Recognition and Machine Learning, vol.4, Newyork, USA, Springer, 2006.