Efficient Method for Privacy Preservation Using δ-Presence Based on Heuristic Approach

[1]Yamini N. Pandya,[2]Name of 2nd Author,[2] Prof. Niti Shah
[1]Student of Master of Engineering,[2]Assistant Professor
Computer Engineering Department,
Silver Oak College of Engineering, Ahmedabad, India

Abstract : Now days Data and Knowledge extracted by data mining techniques represents a key asset driving research, innovation and policy making activities. the data publication and data security are still very difficult. Data offense contains personally identifiable information and therefore releasing such data may result privacy breaches. we presented a new privacy metric, δ-Presence, that clearly links the quality of anonymization to therisk posed by inadequate anonymization.In these paper work on medical data using proposed model for improving Privacy of Data.

***Keywords* – k-Anonymity, K- Medoid, delta presence, medical databases,Privacy Preserving Data Mining (PPDM)**

I. INTRODUCTION

Privacy Preserving Data Mining is an emerging technology which performs data mining operations on centralized and distributed data in a secured manner to preserve sensitive data. Enormous amount of precise personal data is regularly possessed and considered by application like shopping patterns, criminal reports, medical document, credit history, among others. Carefully studying such data opens new risks to privacy. As some sensitive data can also be reveal to people which the person doesn't want to reveal. So there comes the need for PPDM. Everyone wants to keep their personal information to themselves only. As most of the information are personal. If any other person gets that information, they can misuse them so there comes need for PPDM.

II. PRIVACY PRESERVING DATA MINING (PPDM)

The term Privacy means it is the ability of an individual or group to seclude themselves, or information about themselves, and thereby express themselves selectively. PPDM is a model used for sensitive data. The main goal is to keep the data private is to block the corruption of private data. Once critical data is revealed then it is impossible to block the corruption of data. If data owner published their data, they be afraid of corruption. So, this blocks them to divide their data. Various people have various context of privacy, for some people private data is privacy while for some people only some of the sensitive attribute is privacy. Different approaches based in PPDM basically the methods are branched into three major groups such as Heuristic based approach, Reconstruction based approach and Cryptographic based approach [9] which are as shown in the Fig-1

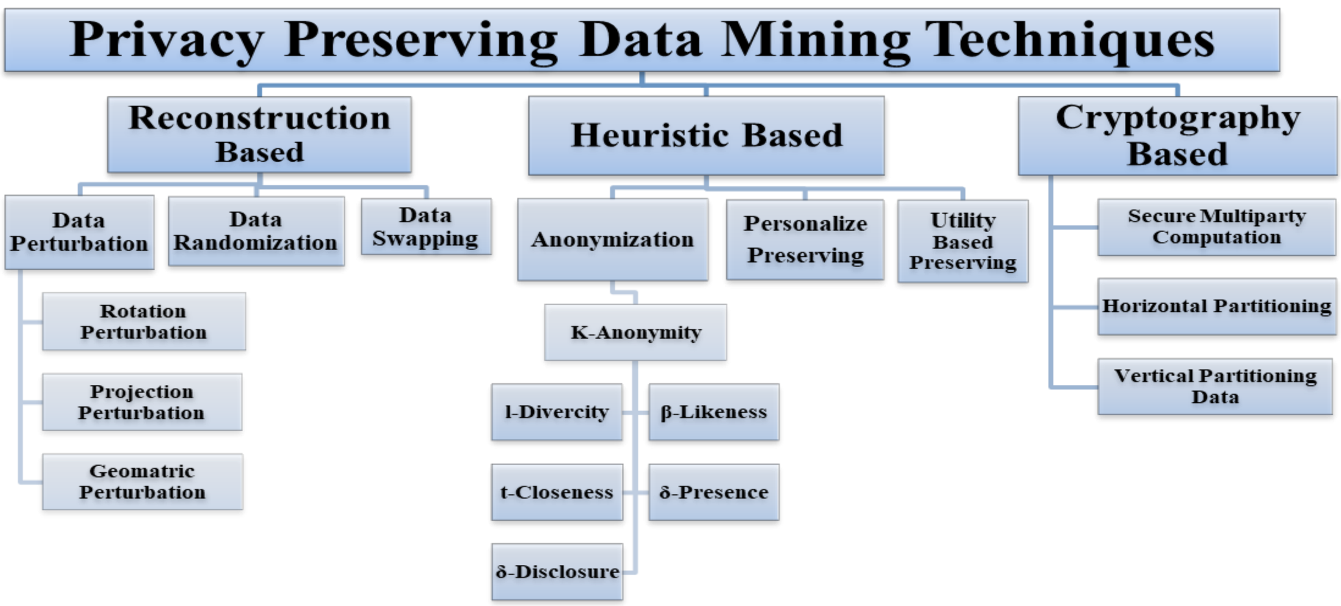

Fig 1: PPDM Techniques

III. Heuristic Based Methods

Heuristic based approach processes the records in "group based" manner. It protects the database by anonymize the data so that the adversaries cannot understand which data belongs to whom. This whole process is called as privacy-preserving data publishing.

A. k-Anonymity

To overcome with these disclosure Samarati and Sweeney [25] introduced k-anonymity in which each record is different to k-1[26][39] other records with respect to the QI i.e. every EC should contain k records in k-anonymity [18]. And is achieved through Generalization and suppression [27]

Table 3.1: 3- Anonymous Version [18]

Sno	ZIP Code	Age	Distance
1	476	2*	Heart Desease
2	476	2*	Heart Desease
3	476	2*	Heart Desease
4	4790*	≥40	Flu
5	4790*	≥40	Heart Desease
6	4790*	≥40	Cancer
7	47605	3*	Heart Desease
8	47673	3*	Cancer
9	47607	3*	Cancer

There are basically two types of attack in k-anonymity [18].

Homogeneity Attack: Here all the value of sensitive attributes in an EC are same. So, it is easy for the adversary to predict that the person is in which equivalence class.

Background Knowledge Attack: Here attacker link the quasi-attribute which they know to the Sensitive attribute to get the information [18].

B. l-Diversity

As identity disclosure is secured by k-anonymity, but it will not secure attribute disclosure. [27] To conquer this drawback of k-anonymity, Machanavajjhala et al. [28] introduce L-diversity, in which each EC contain well represented distinguish values of sensitive attributes [29].

Table 3.2: 3-Diverse table [27]

Age	Sex	Zipcode	Disease
[20-29]	*	13***	Flu
[20-29]	*	13***	Cancer
[20-29]	*	13***	Carcinoid
[29-34]	*	14***	Dyspepsia
[29-34]	*	14***	Gastritis
[29-34]	*	14***	Gastric ulcer
[34-40]	*	13***	penumonia
[34-40]	*	13***	Flu
[34-40]	*	13***	Cancer

Skewness Attack: If a record has 1000 number of patients with and without cancer then that sensitive attribute is 2-diverse and there will be 50% of chances for the adversary to understand that whether that person have cancer or not.

Similarity Attack: In a record if the value of sensitive attributes is l-diverse but semantically similar so there are chances of similarity attack.

C. t-closeness

The distance between the sensitive attribute of an EC should not be more than threshold t [30] [31]. It prevents attribute disclosure. There are many methods to find the t-closeness of sensitive attribute like earth mover's distance and variational distance formula etc. While EMD formula satisfies the two properties of t-closeness they are the generalization and subset property [32].

D. δ-Disclosure

It enforces a restriction on the distances between the distributions of sensitive values but uses a multiplicative definition which is stricter than the definition used by t-closeness. [41]

Hellinger's Distance formula is used to quantify the similarity between two probability distributions. For two discrete probability distributions P and Q.

$$1 - H^2(P,Q) = \sum_{i=1}^{k}(\sqrt{p_i q_i})$$

Now here for the same Age example one gets the minimum range compared to the EMD here if the value is 45 then one gets the value 40-45-50 which is stricter range value compared to EMD.Here one can't get better information gain in order to do so one can use Beta likeness

E. β-Likeness

Here beta likeness aims to overcome limitations of prior models by restricting the relative maximal distance between distributions of sensitive attribute values, also considering positive and negative information gain.

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i).$$

The expected information needed to classify a tuple in D is given by where pi is the probability that an arbitrary tuple in D belongs to class Ci and is estimated by jCi, Dj/jDj. A log function to the base 2 is used, because the information is encoded in bits.

Minimality Attack [3]: For trying to minimize information loss and such an attempt provide a loophole for attacks is a Minimality attack. The Minimality attack occurs when conditioning on A increases the posterior belief in a particular QI value being associated with a particular SA value,

i.e. Pr[t[SA] = s|A,D] > Pr[t[SA] = s|D][4]

DeFinetti Attack: Aims to learn the correlation between SA values and QI values by building a Bayesian network. it starts by assuming a random permutation to assign each SA value to a QI value in each EC, and builds a Naive Bayes classifier out of all such assignments. [5]

F.δ-Presence

This model can be used to protect data from membership disclosure. A dataset is (δmin, δmax)-present if the probability that an individual from the population is contained in the dataset lies between δmin and δmax.[6]

IV.PROPOSED METHOD

Step 1: Initialise Load dataset

Step 2: Select attributes and Identifiers.

Table: 4.2.1 Original Dataset

Sr. no	Zip code	Age	Disease
1	234721	53	Carcinoid
2	338409	28	lung cancer
3	284582	37	stomach cancer
4	16107	49	fever
5	209642	52	brain tumour
6	45781	31	ulcer
7	159449	42	blood cancer
8	280464	37	Flu
9	141297	30	pneumonia
10	122272	55	Flu
11	16107	49	fever

Step 3: Apply Normalization on Attributes.

Table:4.2.2 Normalization on Attribute

Sr. no	Zip code	Age	Disease
1	234721	53	Carcinoid
2	338409	28	lung cancer
3	284582	37	stomach cancer
4	16107	49	fever
5	209642	52	brain tumour
6	45781	31	ulcer
7	159449	42	blood cancer
8	280464	37	Flu
9	141297	30	pneumonia
10	122272	55	Flu

Step 4: Apply Clustering methods

Add attribute partitioning (**K-Medoid**) on dataset.

Partitioning of Selected Attribute value in Two Clusters from step 3.

Table:4.2.3 K-medoid on Dataset

Sr. no	Zip code	Age	Disease	
1	234721	52	Carcinoid	C1
2	338409	53	lung cancer	
3	284582	55	stomach cancer	
4	16107	49	Fever	
5	209642	37	brain tumour	C2
6	45781	30	ulcer	
7	159449	37	blood cancer	
8	280464	31	Flu	
9	141297	42	pneumonia	
10	122272	28	Flu	

Step 5: Apply privacy base method

1. Randomization

This method is applying on the age attribute for Privacy gain.

Table:4.2.4 Randomization on Dataset

Sr. no	Zip code	Age	Disease	
1	234721	53	Carcinoid	C1
2	338409	52	lung cancer	
3	284582	30	stomach cancer	
4	16107	49	fever	
5	209642	55	brain tumour	C2
6	45781	55	ulcer	
7	159449	37	blood cancer	
8	280464	37	Flu	
9	141297	42	pneumonia	
10	122272	28	Flu	

2. Suppression

Here this method is applying on Zip code for less information loss.

Table:4.2.5 Suppression on Dataset

Sr. no	Zip code	Age	Disease	
1	2347**	53	Carcinoid	C1
2	3384**	52	lung cancer	
3	2845**	30	stomach cancer	
4	161**	49	fever	
5	2096**	55	brain tumour	C2
6	457**	55	ulcer	
7	1594**	37	blood cancer	
8	2804**	37	Flu	
9	1412**	42	pneumonia	
10	1222**	28	Flu	

Step 6: Dataset Calculate of **δ-Presence**

Select Appropriate class for δ-Presence and apply Maximum Possibilities Between 0 and 1 for different Minimum and Maximum Possibilities of Desire Matrix and balance it with equal Distribution

$$\boldsymbol{\delta_{min} \leq P(t \in T \setminus T^*) < \delta_{max}}$$

Table:4.2.6 Delta Presence on Dataset

Sr. no	Zip code	Age	Disease	Probability Distribution	
1	2347**	53	Carcinoid	5.0	C1
2	3384**	52	lung cancer	5.0	
3	2845**	30	stomach cancer	3.0	
4	161**	49	fever	4.0	
5	2096**	55	brain tumour	5.0	C2
6	457**	55	ulcer	3.0	
7	1594**	37	blood cancer	3.0	
8	2804**	37	Flu	4.0	
9	1412**	42	pneumonia	3.0	
10	1222**	28	Flu	5.0	

Step 7: Get Anonymized Data.

V. RESULTS AND DISCUSSION

The performance of the proposed algorithm is evaluated in terms of two data metrics namely information loss and privacy gain. The proposed method and three existing methods namely k-anonymity (k=3), ℓ diversity(l=3) and t-closeness are experimented with the same data set and their performance were compared in terms of information loss and privacy gain. The following formulae are used to measure information loss ILoss and privacy gain PG [12].

1. Information Loss

$$ILOSS(vg) = \frac{|vg| - 1}{|DA|}$$

where; |vg| is the number of domain values that are descendants of vg. DA is the number of domain values in the attribute A of vg. ILOSS(vg)=0 if vg is an original data value in the table. In words, ILOSS(vg) measures the fraction of domain values generalized by vg. The loss of a generalized record r is given by

$$ILoss(r) = \sum_{Vg \in r} (w_i \times ILoss(v_g))$$

Where wi is a positive constant specifying the penalty weight of attribute Ai. The overall loss of a generalized table T is given by

2. Privacy Gain

$$ILoss(T) = \sum_{r \in T} Iloss(r) .$$

$$PG = avg\{A(QIDj) - As(QIDj)\} .$$

Where, A(QIDj) and as(QIDj) denote the anonymity of QIDj before and after specialization. The Principle of information/privacy trade-off can also be used to select a generalization g, in the which case it will minimize.

$$ILPG = \frac{IL(g)}{PG(g)}$$

Table: 5 Comparison with Other Techniques

Methods	Information Loss	Privacy	ILPG
K anonymity	1.488	12	0.124
l-diversity	1.488	10.5	0.1417
t-closeness	0.990	10.5	0.094
Proposed method	0.8	10.2	0.0674

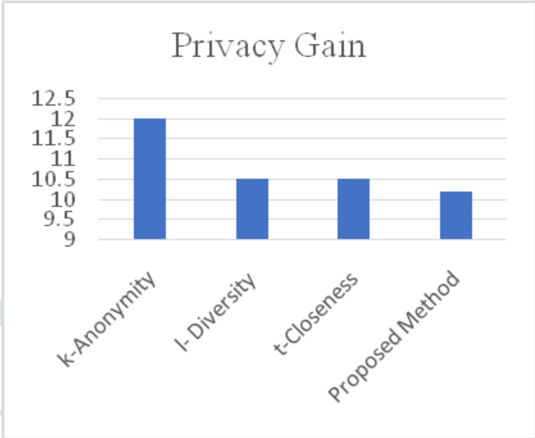

Fig:5.1 Masure the information loss Fig:5.2 Masure the Privacy with with Sample DatasetSample Dataset

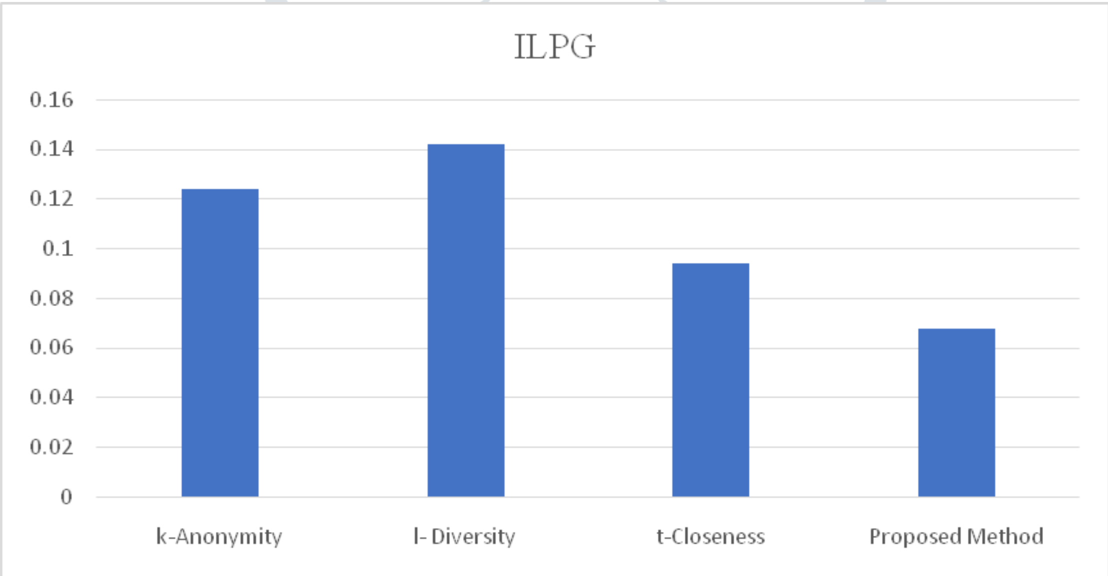

Fig:5.3 Masure the overall performance of ILPG with Sample Dataset

It is observed that the proposed method reduces theinformation loss compared to existing methods, as shown intable VI. It is also observed that the proposed methodperforms well in terms of privacy gain and ratio of informationloss to privacy gain(ILPG). The overall performance of themethods is shown in the last column as ILPG. The overallperformance of the proposedmethod is better than the existingtechniques as shown in figures 5.1 and 5.2

VI. CONCLUSION

As we all know Security become a prime concern for current generation because of high tech technology branches are there. Currently number of technologies works on medical database. in this paper dissertation works on medical database analysis and security using new scheme (proposed model) and try to achieve current issue which exact in current technology.

REFERENCES

[1] [Han J. and Kamber M., Data Mining: Concepts and Techniques,2nd ed., The Morgan Kaufmann Series in Data Management Systems. Elsevier, 2006, pg.1-36.

[2] Qiang Yang, Xindong Wu, "*10 CHALLENGING PROBLEMS IN DATA MININGRESEARCH*", International Journal of Information Technology &Decision-*Making* Vol. 5, No. 4 (2012) 597–604.

[3] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei. "*Minimality attack in privacy preserving data publishing*" In VLDB, pages 543–554, 2007.

[4] Graham Cormode, Divesh Srivastava "*Minimizing Minimality and Maximizing Utility: Analysing Method-based attacks on Anonymized Data*" VLDB Endowment, Vol. 3, No. 1, 2010

[5] Jianneng Cao, Panagiotis karras "*Publishing Microdata with a Robust Privacy Guarantee*" VLDB Endowment 2150-8097/12/07

[6] Takao Takenouchi, Takahiro Kawamura and Akihiko Ohsuga, "*Hiding of User Presence for Privacy Preserving Data Mining*", International Conference on Advanced Applied Informatics, IEEE- 2012

[07] Chu F., "*Mining Techniques for Data Streams and Sequences*", Doctor of Philosophy Thesis, University of California, 2005.

[08] Ann C., Data Mining: Staking a Claim on Your Privacy, Information and Privacy Commissioner/Ontario.ImanElyasi, and SadeghZarmehi, "*EliminationNoise by Adaptive Wavelet Threshold*", World Academy of Science, Engineeringand Technology 56 2009.

[9] Tapasya Dinkar, Aniket Patel and Dr. Kiran R. Amin ," *Preserving The SensitiveInofrmation Using Heuristic Based Approach* ", IEEE2016.

[10] Krupali N. Vachhani, Dinesh B. Vaghela "*Geometric Data Transformation for Privacy Preserving On Data Stream Using Classification*" International Journalof Innovative Research in Computer and Communication Engineering Vol. 3, Issue 6, June2015

[11] Salah Bindahman1, Muhammad RafieHj. Mohd. Arshad2, Nasriah Zakaria3, "*Attribute Based Diversity Model for Privacy Preservation*",8th ICIT 2017

[12] R. Mahesh, T. Meyyappan, "*Anonymization Technique through Record Elimination to Preserve Privacy of Published Data*" 2013 IEEE

[13] Samir Patel, Gargi Shah, Aniket Patel, Assistant Professor,Sigma Institute Of Engineering, U V Patel College of Engineering Baroda, Gujarat Kherva - Mehsana, India, "*Techniques of Data Perturbation for Privacy PreservingDatanMining*" , (IJARCE) Vol.1, No.2, March 2014.

[14] Aniket Patel , HirvaDivecha, "*A Study of Data Perturbation Techniques For PrivacyPreserving Data Mining*", IJSHRE Feb 2014.

[15] Christy Thomas, Diya Thomas , "*An enhanced method for privacy preservationin data publishing*", 4th ICCCNT, Tiruchengode, India, July 4- 6, 2013.

[16] Nagendra kumar.S, Aparna.R, "*Sensitive Attributes based Privacy Preserving in Data Mining using k-Anonymity*", IJCA International Journal of ComputerApplications (0975 – 8887) Vol.84, No.13, pp.1-6, December 2013.

[17] Pu Shi, Li Xiong, Benjamin C. M. Fung, "*Anonymizing Data with Quasi-Sensitive Attribute Values*", CIKM'10, Toronto, Ontario, Canada,October 26–
30, 2010.

[18] R. Indhumathi, S. Mohana, "*Data Preserving Techniques for Collaborative Data*

Publishing", IJERT International Journal of Engineering Research &Technology, Vol.2, Issue 11, pp.3449-3454, November – 2013.

[19] Pierangela Samarati, Latanya Sweeney, "*Protecting Privacy when disclosinginformation: k-Anonymity and its enforcement through generalization and suppression*", The work of Pierangela Samarati was supported in part by NationalScience Foundation and by DARPA, pp.1-19.

[20] W.T. Chembian, Dr. J. Janet, "*A Survey on Privacy Preserving Data Mining Approaches and Techniques"*, Proceedings of the Int. Conf. on InformationScience and Applications ICISA, Chennai, India, 6 February 2010, pp.60-63.

[21] Haisheng Li East China Jiaotong University, Nachang330013,China ," *Study of Privacy Preserving Data Mining*" 978-0-7695-4020-7/10 © 2010 IEEE

[22] Disha Dubli and D.K Yadav, "*Secure Techniques of Data Anonymization for Privacy Preservation*" International Journal of Advanced Research in Computer Science Volume 8, No. 5, May-June 2017

[23] Hebert O. Silva, Tania, Regina Moraies "*Privacy and data mining: evaluating the impact of data anonymization on classification algorithms*" 13th European Dependable Computing Conference 2017

[24] Aniket Patel, Nisha Khurana, "*Preserving the Sensitive Information Using Heuristic Based Approach*" Mathematical Sciences International Research Journal: Volume 6 Issue 1 (2017)

[25] Manish Shanna, Atul Chaudhar, Manish Mathuria, Shalini Chaudhar, Santosh Kumar, "*An Efficient Approach for Privacy Preserving in Data Mining*", IEEE 2014, pp.244-249.

[26] S.Vijayarani, A.Tamilarasi, M.Sampoorna, "*Analysis of Privacy Preserving K-Anonymity Methods and Techniques*", Proceedings of the InternationalConference on Communication and Computational Intelligence – 2010, Kongu Engineering College, Perundurai, Erode, T.N., India.27 – 29 December 2010, pp.540-545.

[27] M V R NarasimhaRao, J.S.VenuGopalkrisna, R.N.V. Vishnu Murthy, Ch. Raja Ramesh, " *Closeness Privacy Measure For Data Publishing Using MultipleSensitive Attributes*", IJESAT International Journal Of Science & AdvancedTechnology, Vol.2, Issue-2, pp.278 – 284, Mar-Apr 2012.

[28] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, "*l-Diversity: PrivacyBeyond k –Anonymity*", Proceedings of the 22nd International Conference onData Engineering (ICDE'06), IEEE 2006.

[29] Yuichi Sei, Takao Takenouchi, Akihiko Ohsuga, *"(l1, ..., lq)-diversity forAnonymizing Sensitive Quasi-Identifiers*", IEEE 2015, pp.596-603.

[30] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, "*t-closeness: PrivacyBeyond k-Anonymity and l –Diversity*", IEEE 2007, pp.106-115.

[31] Jordi Soria-Comas, Josep Domingo-Ferrer, David S´anchez and Sergio Mart´ınez, "*t-closeness through Microaggregation: Strict Privacy with EnhancedUtility Preservation*", IEEE 2016, pp.1464-1465.