Efficient Method for Privacy PreservationUsing δ-PresenceBased on Heuristic Approach

[1]Yamini N. Pandya,[2]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. In these paper work on medical data privacy using proposed model for improve efficiency and time complexity using δ-Presence with K-Medoid for allow specific data and also secure it.

Keywords–*Privacy Preserving Data Mining (PPDM), LSB Embedding, K-Medoid,δ-Presence*

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.

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 have fear 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

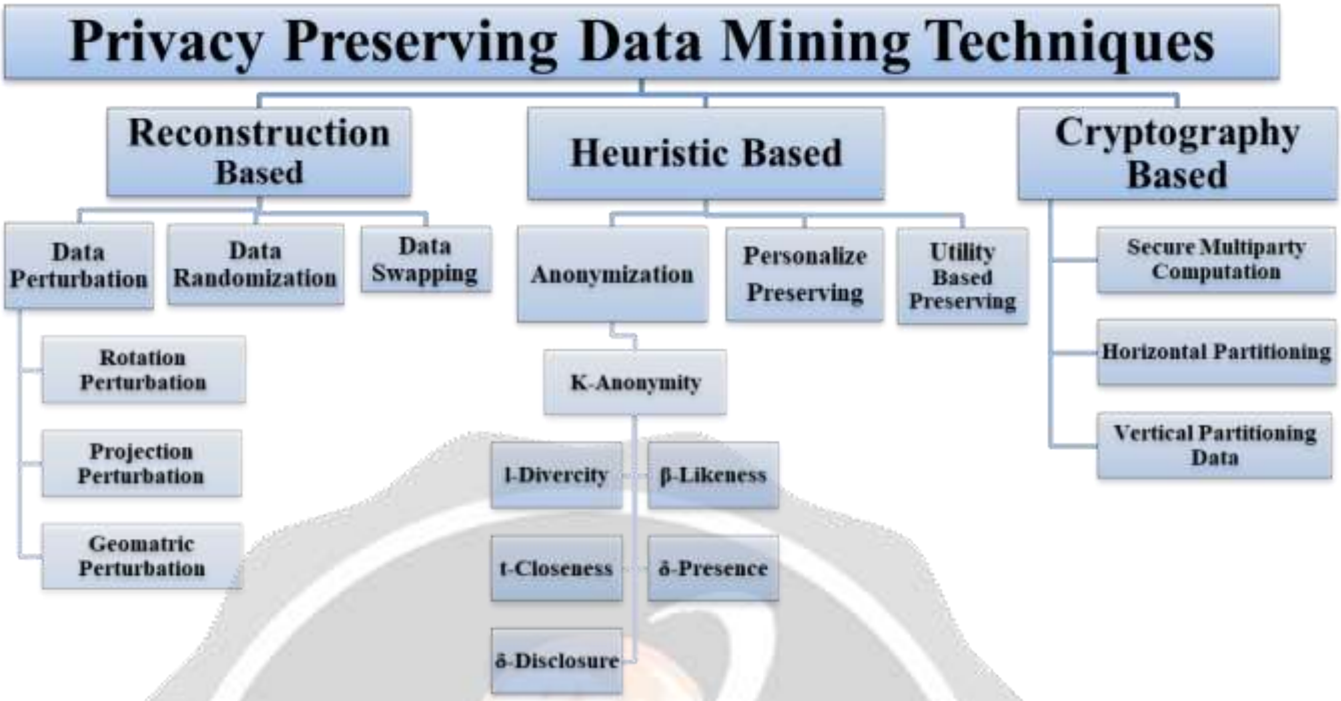


Fig 1: PPDM Techniques

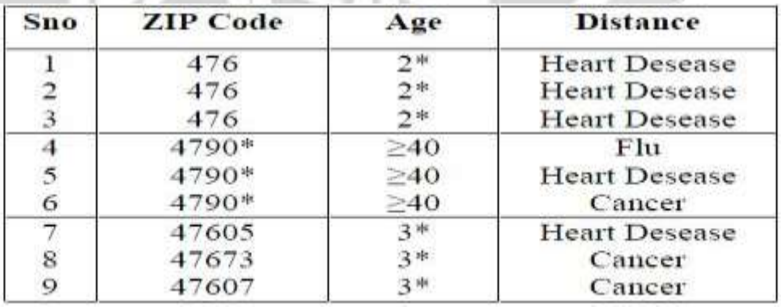
II. HEURISTIC BASED METHODS

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].

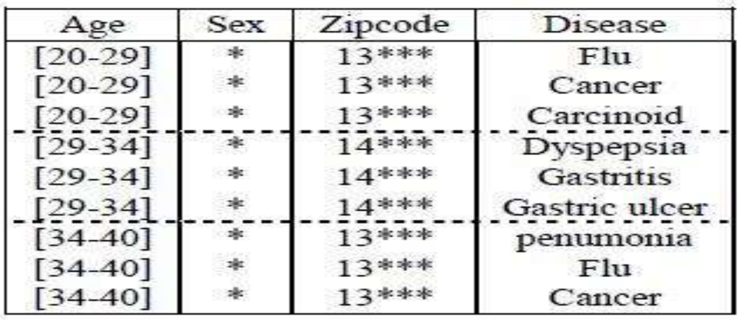
Homogeneity Attack: Here all the value of sensitive attributes in an EC are same. So, it is easy for the adversary to predict that the person is in which equivalence class.

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. In order to be able to calculate these probabilities, users need to specify a population table. [6]

III. Literature Review

4.1Applying Dynamic Verification Tagging to the k-Anonymity Model [12]

Year: 2017 IEEE (International Conference for Internet Technology and Secured Transactions)
Author: Ahmad Bennakhi, Mohamed A. Jeragh
In this paper hashing method is used for tagging the k- anonymity model. This tagging system can also use trusted third party to verify the data in case the data miner had any doubts on it. According to paper using this method there is no privacy preservation for Transactional data so this is the main issue in this paper.

4.2 Attribute Based Diversity Model for Privacy Preservation [11]

Year: 2017 IEEE (International Conference on Information Technology)
Author: Salah Bindahman1, Muhammad Rafie Hj. Mohd. Arshad, Nasriah Zakaria
Here the author describes privacy disclosure risk on the Anatomy model that is based on gender and age attribute was analyzed. It shows some vulnerability regarding that matter. This paper presents an approach that can eliminate all the possibilities for such kind of disclosure risk by clustering the original data based on gender or age criteria

4.3 Secure Techniques of Data Anonymization for Privacy Preservation [22]

Year: 2017 IJARCS (International Journal of Advanced Research in Computer Science)
Author: Disha Dubli and D.K Yadav
This paper describes the various secure methods of Data anonymization are suppression, generalization, Bucketization and perturbation. There are various constraints with these techniques like suppression reduces the quality of data drastically, generalization is inadequate in handling high dimensional data, Bucketization needs to have a clear difference between QIs and SAs, perturbation reduces utility of data. The slicing technique which involves partitioning of data both horizontally and vertically is one of the best methods of anonymization.

4.5 Privacy and data mining: evaluating the impact of data anonymization on classification algorithms [23]

Year: 2017 IEEE (European Dependable Computing Conference)
Author: Hebert O. Silva, Tania, Regina Moraies
Here the anonymization techniques such as generalization are used, the accuracy of the classifiers tend to be increased.in this paper author suggest the k- anonymity model, it is recommended to suppress the data that do not meet the criterion established by the model.

4.5 t- Closeness through Micro aggregation: Strict Privacy with Enhanced Utility Preservation [35]

Year: 2015 IEEE (IEEE Transactions on Knowledge and Data Engineering)
Author: Jordi Soria-Comas, Josep Domingo-Ferrer
Here author Describe the new method is Micro aggregation instead of generalization and suppression for generating k-anonymized t-close datasets.it improves the data utility, increasing the data granularity, reducing the impact of outliers.t- closeness gives best result.

4.6 Hiding of User Presence for Privacy Preserving Data Mining [37]

Year: 2012 IEEE
Author: Takao Takenouchi, Takahiro Kawamura and Akihiko Ohsuga
In this paper, privacy preserving data mining Technique like K- anonymity Here δ-site-presence, which indicates the probability of the presence of a user being revealed in distributed environment, and introduce Dummy user protocol. evaluation results show that Dummy user protocol can anonymize users in accordance with the policy of hiding users’ presence and user anonymity with small impact on data mining results. [37]

Table 4.1 Papers summary Table

TITLE	YEAR	APPROACH	OPEN ISSUE	MY OBSERVATION
Applying Dynamic Verification Tagging to the k-Anonymity Model [12]	2017 IEEE	PPDM Technique Hashing Method L-Diversity	privacy preservation For Transactional Data	Hybrid approach of Hashing Method and L- Diversity.
Attribute Based Diversity Model for Privacy Preservation [11]	2017 IEEE	K- Anonymity, L-Diversity, S-clustering Anatomy	Background Knowledge Attack, Identity Disclosure	One can used the L-diversity and S cluster method for preventing from the Background Attack
Secure Techniques of Data Anonymization for Privacy Preservation [22]	2017 IJARCS	Anonymization, Generalization, Bucketization	Privacy of Sensitive Data	Here used the Slicing Algorithm for Attribute Disclosure.
Privacy and data mining: evaluating the impact of data anonymization on classification algorithms [23]	2017 IEEE	Anonymization, Classification, Data Utility	Private Data Accessible to the third party.	Classification Algorithms, ZeroR, KNN
t- Closeness through Micro aggregation: Strict Privacy with Enhanced Utility Preservation [35]	2015 IEEE	Several micro aggregations algorithms for k-anonymous t-closeness are presented	To satisfy t-closeness in micro aggregation with desired k seems to be time consuming	Overcomes Data utility loss with the help of hybrid approach of Entropy l-diversity and t-closeness.
Hiding of User Presence for Privacy Preserving Data Mining [37]	2012 IEEE	Many types of technique such as k-Anonymity, Distributed anonymization and privacy preserving data publishing used	One needs to improve the heuristic function to make it more efficient and extend the protocol for supporting the multiple sites.	Here using distributed anonymization method, which combines the personal information and anonymize it to prevent identifying specific user records and calculating relative error of anonymized data.

IV. PROBLEM STATEMENT AND DEFINITION OF WORK

Motivated by the privacy concerns on Data Mining, a research area called PPDM. The main issues are how to change the data and how to regain the data mining outcome from the changed data. The objective is to change a given data set D into changed version D' that amuse a given privacy demand and protect as much information as possible for the meant data analysis task.

Basically, this dissertation aims that the Heuristic Based Technique to provide. Security totheData. HereHeuristicBasedtechniqueisusedforthePPDM which contains δ-Presence and k-medoid with Embedding technique.

5.1Proposed System

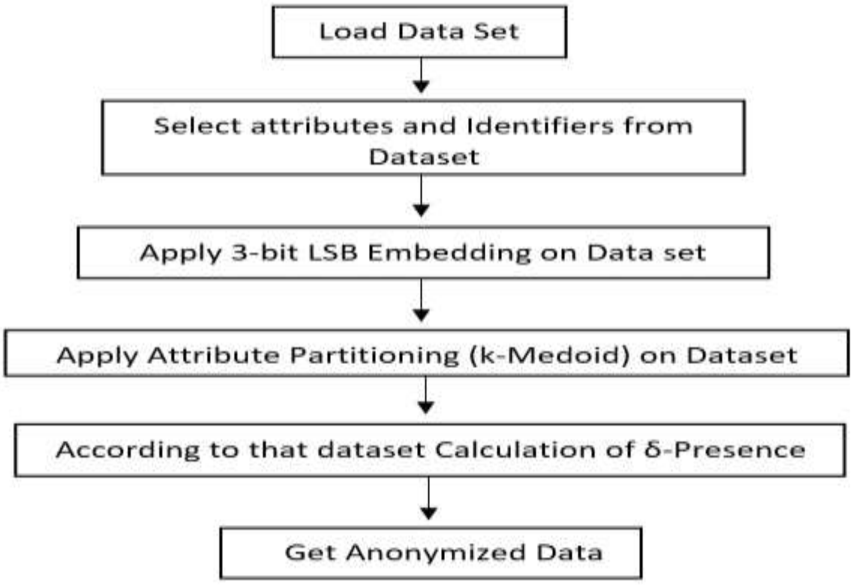

Load Data Set
Select attributes and Identifiers from Dataset
Apply 3-bit LSB Embedding on Data set
Apply Attribute Partitioning (k-Medoid) on Dataset
According to that dataset Calculation of δ-Presence
Get Anonymized Data

Fig 5.1: Flow Diagram

5.2 Proposed Methodology

Step 1: Take dataset D.

Step 2: Select the attribute and identifiers from dataset.

Step 3: Select Quasi identifiers from data set and apply 3- bit **LSB embedding**.

Word Embedding

A word embedding is a learned representation for text where words that have the same meaning have a similar representation.

three techniques of word embedding:

1. Embedding Layer
2. Word2Vec
3. GloVe

Step 4: Apply **attribute partitioning** (k-medoid) on dataset.

Attribute Clustering (K- Medoid)

The partitioning method is then performed based on the principle of minimizing thesum of the dissimilarities between each object and its corresponding reference point.

absolute-error criterion

$$E = \sum_{j=1}^{k} \sum_{p \in C_j} |p - o_j|$$

PAM (Partitioning Around Medoid)

Step 5: δ-Presence Calculation.

Given an external public table P, and a private table T, we say that δ -presence holds for a generalization T* of T, with δ= (δmin, δmax) if:

δmin ≤P (t ϵ T\ T*) < δ max

Step 6: Get Anonymized Data.

V. 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 MINING RESEARCH*", 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: Analyzing 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
7) Chu F., "*Mining Techniques for Data Streams and Sequences*", Doctor of Philosophy Thesis, University of California, 2005.
8) Ann C., Data Mining: Staking a Claim on Your Privacy, Information and Privacy Commissioner/Ontario. Iman Elyasi, and Sadegh Zarmehi, "*Elimination Noise by Adaptive Wavelet Threshold*", World Academy of Science, Engineering and Technology 56 2009.
9) Tapasya Dinkar, Aniket Patel and Dr. Kiran R. Amin, "*Preserving the Sensitive Information Using Heuristic Based Approach*", IEEE 2016.
10) Krupali N. Vachhani, Dinesh B. Vaghela "*Geometric Data Transformation for Privacy Preserving on Data Stream Using Classification*" International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 6, June2015
11) Salah Bindahman1, Muhammad Rafie Hj. Mohd. Arshad2, Nasriah Zakaria3, "*Attribute Based Diversity Model for Privacy Preservation*",8th ICIT 2017
12) Ahmad Bennakhi, Mohamed A. Jeragh, "*Applying Dynamic Verification Tagging to the k-Anonymity Model*" 12th ICITST-2017
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 Preserving Datamining*" (IJARCE) Vol.1, No.2, March 2014.
14) Aniket Patel, Hirva Divecha, "*A Study of Data Perturbation Techniques for Privacy Preserving Data Mining*", IJSHRE Feb 2014.
15) Christy Thomas, Diya Thomas, "*An enhanced method for privacy preservation in data publishing*", 4th ICCCNT, Tiruchengode, India, July 4 - 6, 2013.
16) Nagendra Kumar's, Aparna.R, "*Sensitive Attributes based Privacy Preserving inData Mining using k-Anonymity*", IJCA International Journal of Computer Applications (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) [1R. Indhumathi, S. Mohana, "*Data Preserving Techniques for Collaborative DataPublishing*", IJERT International JournalEngineering Research & Technology, Vol.2, Issue 11, pp.3449-3454, November 2013.
19) [19]Pierangela Samarati, Latanya Sweeney, "*Protecting Privacy when disclosing information: k-Anonymity and its enforcement through generalization andsuppression*", The work of Pierangela Samarati was supported in part by National Science Foundation and by DARPA, pp.1-19.
20) W.T. Chembian, Dr. J. Janet, "*A Survey on Privacy Preserving Data MiningApproaches and Techniques*", Proceedings of the Int. Conf. on Information Science and Applications ICISA, Chennai, India, 6 February 2010, pp.60-63.
21) Haisheng Li East China Jiaotong University, Nachang330013, China," Study ofPrivacy Preserving Data Mining" 978-0-7695-4020-7/10 © 2010 IEEE

22) Disha Dubli and D.K Yadav, "*Secure Techniques of Data Anonymization forPrivacy 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. Vijaya ani, A.Tamilarasi, M.Sampoorna, "*Analysis of Privacy Preserving K-Anonymity Methods and Techniques*", Proceedings of the International Conference 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 Narasimha Rao, J.S. Venu Gopalkrisna, R.N.V. Vishnu Murthy, Ch. RajaRamesh, "Closeness *Privacy Measure for Data Publishing Using Multiple Sensitive Attributes*", IJESAT InternationalJournal of Science & Advanced Technology, Vol.2, Issue-2, pp.278 – 284, Mar-Apr 2012.

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

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

30) Ninghui Li, Tiancheng Li, Suresh Venkata Subramanian, "*t-closeness: Privacy Beyond k-Anonymity and l –Diversity*", IEEE 2007, pp.106-115.

31) Jordi Soria-Comas, Josep Domingo-Ferrer, David S´anchez and SergioMart´ınez, "t-*closeness through Micro aggregation: Strict Privacy with Enhanced Utility Preservation*", IEEE 2016, pp.1464-1465.

32) Rashmi B. Ghate, Rasika Ingle, Y.C.C.E Nagpur, India," *ClusteringmmBasedAnonymization for Privacy Preservation*", IEEE, 2015

33) Xiaokui Xiao Yufei Tao, "*Personalized Privacy Preservation*", SIGMOD 2006, Chicago, Illinois, USA, June 27–29, 2006.

34) Jian Xu, Wei Wang, Jian Pei, Xiao yuan Wang, Baile Shi, Ada Wai-Chee Fu, "*Utility-Based Anonymization Using Local Recoding*", KDD'06, Philadelphia, Pennsylvania, USA, August 20-23, 2006.

35) Jordi Soria-Comas, Josep Domingo-Ferrer, Fellow, IEEE, David Sanchez and Sergio Martínez, "*t-Closeness through Micro aggregation: Strict Privacy with Enhanced Utility Preservation*" IEEE Transactions on Knowledge and Data Engineering 2015

36) Rajesh N., Sujatha K., A. Arul Lawrence "*Survey on Privacy Preserving Data Mining Techniques using Recent Algorithms*" International Journal of Computer Applications (0975 – 8887) Volume 133 – No.7, January 2016

37) Takao Takenouchi, Takahiro Kawamura and Akihiko Ohsuga, "*Hiding of User Presence for Privacy Preserving Data Mining*" IIAI International Conference on Advanced Applied Informatics IEEE 2012.

38) Supriya Borhade Researcher, Department of Computer Engineering, Pune University, Pune, India "*A Survey on Privacy Preserving Data Mining Techniques*" IJETEA International Journal of Emerging Technology and Advanced Engineering Volume 5, Issue 2, February 2015.

39) Nivetha.P. R, Thamarai selvi. K "*A Survey on PPDM Techniques*" InternationalJournal of Computer Science and Mobile Computing IJCSMC, Vol. 2, Issue. 10, October 2013, PP.166 – 170

40) Mehmet Ercan Nergiz and Christopher Clifton, Senior Member, IEEE, "Delta- *Presence without Complete World Knowledge*" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 6, JUNE 2010

41) R. Natarajan1, Dr. Sugumar, M. Mahendran, K. Anbazhagan "*A survey on Privacy Preserving Data Mining*" International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 1, MARCH 2012, PP103-112.

42) Oliveira P. and Stanley M., "*Privacy - Preserving Data Mining. Encyclopaedia of Data Warehousing and Mining*", 2nd ed., (IGI Global, pp 1582-1588,2009).