International Journal of Scientific Research in Computer Science and Engineering **Survey Paper**
Vol.5, Issue.5, pp.21-25, October (2017) **E-ISSN:** 2320-7639

A Survey on Heuristic Based Approach for Privacy Preserving in Data Mining

Aniket Patel[1], Patel Shreya[2*], Kiran Amin[3]

[1] Dept. of IT, Silver Oak College of Engineering and Technology, GTU, Ahmedabad, India
[2*] Dept. of CE, Silver Oak College of Engineering and Technology, GTU, Ahmedabad, India
[3] Dept. of CE, UV Patel College of Engineering, GTU, Kherva-Mehsana, India

**Corresponding Author: patelshreya1706@gmail.com, Tel.: +91 9712535482*

Available online at: www.isroset.org

Received 10 Sep 2017, Revised 19th Sep 2017, Accepted 15th Oct 2017, Online 30th Oct 2017

Abstract— Data Mining has been the most researched area for researchers because of the possibilities of extension at each application of it. When the data becomes massive in volume, many problems strike for security and privacy breach. Some applications like sharing of such data to a particular user have threats of preserving the original data so that the injection of such data can be prohibited. So it is a timely need to secure the data while handling them to the known or unknown users. The requirement of not losing the essence of data and still publishing it with the actual information is a challenge. Such troubles prompted the advancement of Privacy Preserving Data Mining (PPDM) Techniques. Privacy Preserving has become an important issue in the development progress of Data Mining techniques. Methods like k-Anonymity, l-Diversity have been explored well by researchers but still, there are holes that force us to develop a more effective method and using such approach one can get better accuracy with minimum loss of data.

Keywords— Data Mining, Heuristic Based Approach, Privacy Preserving Data Mining.

I. Introduction

Data Mining is the computational process of discovering patterns in large data set involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The overall goal of the Data Mining process is to extract information from a data set and change it into an understandable structure for further use. Data Mining is the analysis step of the KDD process. From last few years, we have seen IT technology stores and record personal data about product, service or people with the help of recorded data they study, analyze and improve business tactics. Large amounts of such data are used for different Data Mining purposes such as knowledge discovery, decision-making and analysis etc. There are different opinions/points of view of privacy to different people. Some may think about/believe whole personal information as private while some may think certain attribute value should not be available directly or indirectly to the personal domain.

II. Need For Privacy Preserving

In today’s world information is the most important resource. Privacy Preserving in Data Mining has become increasingly popular because it allows sharing of privacy sensitive data for analysis purpose. Large numbers of detailed personal data are regularly collected and analyzed by an application like credit records, shopping data, medical data and criminal records among others. Analyzing such data opens new threats to privacy.As some sensitive data can also be revealed to people which the person doesn’t want to reveal. So there comes the need for Privacy Preserving Data Mining (PPDM).

III. Privacy Preserving In Data Mining(PPDM)

Generally, when privacy is the discussion it is a major concern of users, firstly privacy is a right or freedom of intrusion. Which includes Information *privacy which* is the right to have some control over how your personal information is collected and used. It is the ability of an individual or group to stop information about themselves from becoming known to people other than those they choose to give the information to.The privacy of the personal data and information must be maintained while sharing of data among another untrusted party. Privacy is sometimes related to anonymity although it is often most highly valued by people who are publicly known. Privacy can also be seen as an aspect of security—one in which there are trade-offs between the interests of one group and another can become particularly clear and privacy is a key issue when comes in discussion of social networking sites it involves the unwarranted access of private information, don’t necessarily have to involve security breaches. Someone can gain access to confidential information by simply watching you type your password. But the potential harm to an individual user who really boils down to how much a user engages in a social networking site, as well as the amount of information they're willing to share. In other words, the Facebook user with 900 friends and 60 group memberships is a lot more likely to be harmed by a breach than someone who barely uses the site.

A. PPDM Framework

Framework for PPDM is shown in fig.1 In Data Mining or (KDD) process the data (mostly transactional) is collected by single/various organization/s and stored at respective databases. Then, it is transformed to a suitable format for analytical purposes then stored in large data warehouse/s and then Data Mining algorithms are applied on it for the generation of information/knowledge [1].

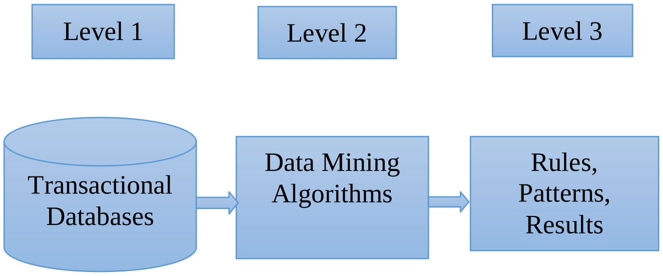

Fig.1 PPDM Framework [2]

B. Purpose of PPDM

PPDM is used to extract knowledge from huge amount of data with the help of data hiding and rule hiding. Data hiding consists to develop new treatment methods of the original data in order that sensitive data stays protected during and after the mining process [3].While rule hiding(knowledge hiding) try to hide some sensitive knowledge in order that they cannot be discovered by Data Mining techniques [3].

C. PPDM Algorithms

PPDM uses a modified version of the standard algorithm.The main aim of the PPDM algorithm is to extract suitable information from large dataset and protecting the data at same time. The PPDM algorithm is specifically on the tasks of classification, association rule mining and clustering classification is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of the model is used for prediction the class of objects whose label is unknown clustering analysis concerns the problem of separating a data set in one group which are similar top each other and are different as possible in another group [4].

D. PPDM Techniques

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 information private is to prevent the misuse of private information. Once important data is disclosed then it is impossible to prevent the misuse of data. If data owner published their data, they have fear of misuse. So, this prevents them to share their data. Different people have different perspective of privacy, for some people personal information is privacy while for some people only some of the sensitive attribute is privacy. There are many different approaches based in Privacy Preserving in Data Mining basically the techniques are divided into three major groups such as Heuristic based approach, Reconstruction based approach and Cryptographic based approach which are as shown in the Fig-2

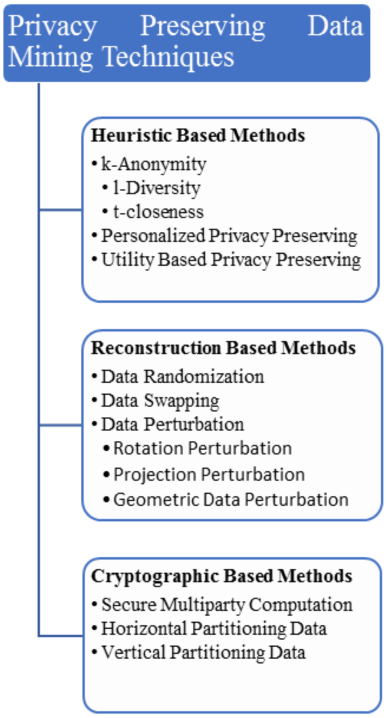
Privacy Preserving Data Mining Techniques
Heuristic Based Methods
• k-Anonymity
• l-Diversity
• t-closeness
• Personalized Privacy Preserving
• Utility Based Privacy Preserving
Reconstruction Based Methods
• Data Randomization
• Data Swapping
• Data Perturbation
• Rotation Perturbation
• Projection Perturbation
• Geometric Data Perturbation
Cryptographic Based Methods
• Secure Multiparty Computation
• Horizontal Partitioning Data
• Vertical Partitioning Data

Fig.2 Techniques of Privacy Preserving Data Mining.

IV. RELATED WORK

A. Heuristic Based Approach

A Heuristic Based is a technique designed for solving a problem more quickly when classic methods are too slow, or for finding a close solution when classic methods do not find any exact solution. The goal of a heuristic based approach is to produce a solution in a reasonable time frame that is good enough for solving the problem that is important now. This solution may not be the best of all the actual solutions to this problem, or it may simply come close to the exact solution. But it is still valuable because finding it does not require a way too much long time. This approach processes the records in group based manner [5]. It protects the database by anonymizing the data so that the attacker cannot understand which data belongs to whom. By different anonymization, the data is changed and it holds good enough utility and that can be released to other parties safely. This whole process is called as privacy-preserving data publishing. Data is stored mainly in the tabular form [6]. And mostly data is published in two ways microdata and macrodata. In past data are published mostly in precomputed statistical and tabular form Such type of data is called macrodata [6]. Various organizations (e.g., Medical authorities and Government agencies) are in need of releasing a person specific data which is often called micro data for public health researchers and demographic analysis [1].

The database contains various types of attributes A set of non-sensitive attributes {Q1, …, Qp} of a table is called a quasi-identifier if these attributes can be linked with external data to uniquely identify (can be called as candidate key) at least one individual in the general population [7]. Age, Gender, State is a set of QI attributes.[8].Sensitive attributes example Medical records, salaries, etc and these attributes are what the researchers need, so they are always released directly [9]. An attribute K consists of values which is the most unique value for to identify the individual from set S. Denote by K. Key attributes that can be used to identify a record, such as Name and Social Security Number [5][7]. Equivalence Class (EC):- Each group that shares the same values on every QI for example Birthdate and Age [5]. While releasing the sensitive information it must require to preserve them from disclosure. There are mainly two types of Information Disclosure. Identity disclosure: An individual is linked to a particular record in the published data. Attribute disclosure: Sensitive attribute information of an individual is disclosed [7][10][11].

k-Anonymity: It is an anonymizing approach proposed by Samarati and Sweeney [12].This technique is used to limit the disclosure risk. K - anonymity requirements says that a data set is k anonymous (k 2: I) if each record in the data set is indistinguishable from at least (k-l) other records within the same data set [13]. This k-Anonymity requirement is generally achieved by using generalization and suppression [14]. While k-Anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure [15]. To address this limitation of k-Anonymity, Machanavajjhala et al. [16] recently introduced a new notion of privacy, called L-Diversity.

l-Diversity: This technique enhances K-Anonymity. This technique removes explicit identifiers and generalizes the QID values to ensure that the data users cannot specify each individual’s sensitive values with a confidence greater than 1/*l* [17].This technique is used for maintaining the minimum size of k and for preventing the homogeneous attack. Machanavajjhala et al. [18] gave a number of interpretations of the term “well represented” in this principle. Here two types of attacks are addressed they are skewness attack and similarity attack.

t-closeness: An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t [18].k-Anonymity does not protect against attribute disclosure while t-closeness seeks to prevent attribute disclosure from happening [19]. The t parameter in t-closeness enables one to the tradeoff between utility and privacy. There are mostly two ways to find the probability distribution, first is variational distance formula and second is earth mover distance formula. While EMD formula satisfies the two properties of t-closeness they are the generalization and subset property.

B. Personalized privacy preservation

This technique performs the minimum generalization for satisfying everybody's requirements, and thus, retains the largest amount of information from the microdata [20].

C. Utility based privacy preservation

A utility based metric should capture two aspects: the information loss caused by the anonymization and the importance of attributes. Such utility-aware anonymization may help to improve the quality of analysis afterward [21].

V. Comparision

The basic idea of k-Anonymity is easy and simple to understand. k-Anonymity was able to protect against identity disclosure but it cannot protect against attribute disclosure so l -diversity works one step ahead of k-Anonymity in preventing attribute disclosure. But l-Diversity is more difficult to achieve and also it is not able to provide sufficient protection for privacy. l -diversity also does not provide protection against attribute disclosure while t-closeness protect attribute disclosure and in order to calculate the distance between attributes in each equivalence class which is solved by Earth Mover's Distance.

VI. Conclusion

Heuristic Based Approach as an alternative method to Privacy Preserving Data Mining has been reviewed. The heuristic based technique includes the k-anonymization method, l-Diversity, t-closeness methods. Out of with it considers sensitive attribute as a dependent attribute and remaining attributes of dataset except as independent attributes. The process we have passed through is to examine the research papers and gather the valuable information for solving different approaches in many different ways. We are trying to adopt the best of them to develop the research flow. We have observed that by eliminating the weakness of t-closeness and developing a new heuristic approach with k-Anonymity and l-Diversity, we can achieve better results in future, we will develop an algorithm and implement it to find expected results.

REFERENCES

[1] Hina Vaghashia, Amit Ganatra, "*A Survey: Privacy Preservation Techniques in Data Mining* ", IJCA International Journal of Computer Applications (0975 – 8887) Vol.119, No.4, pp.20-26, June 2015.

[2] A.S.Shanthi, Dr. M. Karthikeyan, "*A review on privacy preserving Data Mining* ", IEEE,2012.

[3] Sarra Gacem, Djamila Mokeddem and Hafida Belbachir, "*Privacy Preserving Data Mining: Case of association rules*", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 1, pp.91-96, May 2013.

[4] N.Punitha, R.Amsaveni, "*Methods and Techniques to Protect the Privacy Information in Privacy Preservation Data Mining* ", IJCTA Int. J. Comp. Tech. Appl., Vol 2, No.6, pp.2091-2097, Nov-Dec 2011.

[5] Tapasya Dinkar, Aniket Patel and Dr. Kiran R. Amin ," *Preserving The Sensitive Inofrmation Using Heuristic Based Approach* ", IEEE 2016.

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

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

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

[9] 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.

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

[11] Mohana Chelvan P., Perumal K., "*On Privacy Preserving Data Mining Techniques: Merits and Demerits*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.9, pp.210-214, 2017.

[12] Pierangela Samarati, Latanya Sweeney, "*Generalizing Data to Provide Anonymity when Disclosing Information*", The work of Pierangela Samarati was supported in part by National Science Foundation and by DARPA, pp.1-13.

[13] Nivedita Bairagi, Punit K. Johari, "*A novel approach for privacy preserving using Animal Migration Optimization and RSA algorithm*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.6, pp.193-197, 2017.

[14] S.Vijayarani, 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.

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

[16] 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.

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

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

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

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

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

[22] Marie Fernandes, "*Data Mining: A Comparative Study of its Various Techniques and its Process*", IJSRCSE International Journal of Scientific Research in Computer Science and Engineering Volume-5, Issue-1, pp.19-23, February 2017.

Authors Profile

Aniket R Patel has received BE Degree in Information Technology from L.D. College of Engineering, Gujarat University, Ahmedabad in 2005-2009, M.Tech in Information Technology from U V Patel College of Engineering, Ahmedabad in 2011-2013.One of the achievement is he has received gold award for M.Tech.Presently, he is pursuing ph.D from Ganpat University. His area of interest are Core Java, Advanced Java, J2ME, Privacy Preserving in Data Mining, Object Oriented Concepts, Data Structure, Software Engineering, Android. He has worked as Assistant Prof. U. V. Patel College of Engineering, Kherva, Mehsana in 2009, Saffrony Institute of Technology SPBPEC,Mehsana in 2014. HE is presently a Assistant Prof. at Silver Oak College of Engineering & Technology, Ahmedabad since 2016 till date. Email Id-aniketpatel.it@gmail.com .

Patel Shreya B has received BE Degree in Computer Engineering from Silver Oak College of Engineering & Technology, Ahmedabad in 2012-2016. Presently, she is pursuing ME Degree from Silver Oak College of Engineering & Technology, Ahmedabad. Her area of interest is Data Mining. She is presently a PG student at Silver Oak College of Engineering & Technology, Ahmedabad since 2016 till date. Email Id-patelshreya1706@gmail.com .

Kiran Amin has received BE Degree in Computer Engineering from L. D. College of Engineering, Gujarat University, Ahmedabad in 1993, ME in Computer Engineering from S.P. University, Vallabh Vidyanagar, Gujarat in 2005, PhD in Computer Engineering completed under Faculty of Engineering &Technology, Ganpat University in 2013. His area of interest are Data Mining, Data Warehousing, Artificial Intelligence, Computer Networks. His merits are Managing as Associate Dean, Vice Principal, Head Senior most faculty at U. V. Patel College of Engineering (15 Years), IEEE-Gujarat Section, Chair, Technical Activities, Group Instructor-Computer at Department of Employment & Training, Govt. of Gujarat (1994-1999), More than 20 years of Experience in teaching. More than 8 years of teaching experience in PG teaching. Presently, he is Principal of Institute of Technology Ganpat University. Email Id-kiran.amin@ganpatuniversity.ac.in.