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 9, September 2015

Analysis of Privacy Preserving in Data Mining

DharmikVasiyani, Hiral Desai, Jay Gandhi

PG Student, Dept. of Computer Engineering, B.H.Gardi College of Engineering & Technology, Rajkot, India

PG Student, Dept. of Computer Engineering, B.H.Gardi College of Engineering & Technology, Rajkot, India

Assistant Prof., Dept. of IT, B.H.Gardi College of Engineering &Technology, Rajkot, India

ABSTRACT*:* Here provides an overview of the new and rapidly emerging research area of privacy preserving data mining. It becomes more popular because it allows you to share your private data for analysis purpose. Sometimes Enterprise needs to share their data to gain mutual benefit with collaboration to the other enterprise data. This collaboration may cause attack on shared private data. So to privacy preserving on those data are necessary. Here given the overview of the privacy preserving in data mining with its techniques and different algorithms and its applications.Also here include latest challenges in this field and also describe that in future work we can apply L-diversity anonymity along with clustering to reduce information loss. There are few efforts are given regarding this approach and they all used different clustering approach to get good utility in anonymity.

KEYWORDS *:*Data mining, Privacy Preserving,Cryptography, Association Rules,Classification, Clustering.

I. INTRODUCTION

Basically, Data mining is to extracting or mining knowledge from tremendous amount of data set[1]. There are so many techniques to discover the meaningful pattern and rules like as association rule, classification, clustering and evaluation pattern etc.

We use all of these techniques to extract knowledge from various types of data, but there are some issues related to mine that data, those are information network analysis, discovery usage, and understanding of patterns and knowledge, stream data mining, mining moving data like RFID data and data from sensor networks, spatiotemporal and multimedia data mining, mining text and Web and other unstructured data, data cube oriented multidimensional online analytical mining, visual data mining, and data mining by integration of sophisticated scientific and engineering domain knowledge.

Now a days the tremendous growth of data in every field[1]. This increment of the data created lots of challenges in privacy. Privacy preserving in data mining becomes too important due to share this data for our benefit purpose[2]. This shared data may contain sensitive attributes. So to prevent those private data privacy preserving in data mining becomes more emerging field.

The main consideration in privacy preserving data mining is twofold. Initially, sensitive raw data like identifiers, names, locations ought to be adjusted or trimmed out from the first database, all together for the data's beneficiary not to have the capacity to bargain someone else's privacy.

Second, sensitive knowledge which can be mined from a database by utilizing data mining algorithms, ought to additionally be prohibited, in light of the fact that such a knowledge can similarly well trade off data privacy, as we will show. The fundamental target in protection saving data mining is to create algorithms for altering the first data somehow so that the private data and private knowledge stay private even after the mining procedure.

Privacy preserving becomes more popular because it allows you to share your private data for analysis purpose. Sometimes Enterprise needs to share their data to gain mutual benefit with collaboration to the other enterprise data. This collaboration may cause attack on shared private data. This report, provides an overview of the privacy preserving in data mining with its techniques and different algorithms.

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 9, September 2015

There have been several approaches found for privacy preserving in data mining are data distribution, data modification, data mining algorithm, rule hiding, privacy preservation[3].
Methodology of privacy preserving in data mining: original to which we wants to preserve firstly transformed in appropriate format. Then this transformed data encrypted or modified with the use of specific algorithms. And finally we have original data that prevents to privacy attacks.

In this paper section 2.contains the privacy preserving techniques, section 3.contains survey table of privacy preserving technique. Section 4.contains privacy preserving algorithms, section 5.contains applications of privacy preserving in data mining, section 6.contains future work and conclusion and also given proposed approach for anonymity l-diversity method to makes it efficient.

II. TECHNIQUES OF PRIVACY PRESERVING

1.Perturbation Method

Perturbation is technique of data distortion. There are mainly three methods of it random noise addition, random rotation and random projection.
Randomization technique is an cheap and efficient approach for privacy preserving data mining (PPDM). Random noise addition is value based method that is most common perturbation method[4]. Suppose x is original data and r is the random value. So r value is added to the x data. In this way whole data set are disturbed with this types of addition of the random value to the original data.

Random rotation transformation is dimension based perturbation method. It decrease the loss of the privacy without affecting the quality of mining. Rotation matrix can be described as:

$$g(X)=RX,$$

where R is the rotation matrix and X be the original data[4]. Rotation is handled such a way to preserve the multi-dimensional geometric properties, such as Euclidian distance of the original data set. There are few changes near the rotation centre.
Random projection is also dimension-base method.to reduce the dimensionality of original data set with the usu of projecting the set of data points from a high-dimensional space to a randomly chosen lower-dimensional subspace there are several properties provided by the Johnson and Linden Strauss.

2. Anonymization

To protect individuals' identity when releasing sensitive information, data holders often encrypt or remove explicit identifiers, such as names and unique security numbers. However, unencrypted data provides no guarantee for anonymity. In order to preserve privacy, k-anonymity model has been proposed by Sweeney which achieves k-anonymity using generalization and suppression[1].The basic idea of k-anonymization is to specify a number k to state required protection, so that each record is indistinguishable from at least k-1 other records.

k-anonymization provides a certain level of privacy preservation, but it is vulnerable to several attacks such as homogeneity attack and background knowledge attack, background knowledge attack.

To overcome these attacks there is extended anonymity is L-diversity. L-Diversity: An equivalence class is said to have l-diversity if there are at least l-"well-represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity[7]. There are other improved anonymity methods are t-closeness, k^m-anonymity, (α,k) anonymity, p-sensitive k-anonymity, (k,e) anonymity which are described in.
Generalization and suppression based anonymity suffers from information loss. Here proposed approach uses clustering in anonymity to reduce information loss.

3. Secure multi-party computation

Secure Multi-party Computation (SMC) is a key system in information mining where mining is directed by together multi parties. It is cryptography based method. A substitute methodology in light of the multiparty calculation is that all

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 9, September 2015

aspects of private information is validly known to one or more parties. Uncovering private information to parties, for example, by whom the information is claimed or the person to whom the information alludes to is not a state of disregarding security. The issue emerges when the private information is uncovered to some other outsiders. To deal with this problem, we use a specialized form of privacy preserving distributed data mining. Parties that each knows some of the private data contribute in a protocol that generates the data mining results, that guarantees no data items is revealed to other parties. Thus the process of data mining doesn't cause, or even increase the chance for breach of privacy[1].

4. Sequential pattern hiding

Sequential pattern hiding method is important to hide sensitive patterns that can generally be extricated from distributed information, without basically influencing the information and the non-sensitive intriguing patterns. Sequential pattern hiding is a challenging problem, because sequences have more composite semantics than item sets, and calls for efficient solutions that offer high utility[8].

III. SURVEY TABLE OF PRIVACY PRESERVING ALGORITHM

Sr. No.	Paper Name	Technique	Method	Remark
1	"k-Anonymity: A Model for Protecting Privacy-2002[6]	K-Anonymity	Generalization, Suppression, Permutation	A record from a dataset can't be recognized from in any event k-1 records whose data is additionally in the dataset.
2	L-Diversity: Privacy Beyond k-Anonymity - 2006[7]	I-Diversity	Generalization, Suppression, Permutation	l-diversity is extended version of k-anonymity, if there is a dataset that is l-diverse then there should l different values for sensitive attribute for specific class. It is better than k-anonymity.
3	On the Privacy Preserving Properties of Random Data Perturbation Techniques-2003[4]	Perturbation	Adding -Noise, Swapping	Here noise is randomly added to the dataset to preserve privacy, then after it gives meaningful knowledge.
4	A condensation approach to privacy preserving data mining-2004[4]	Condensation	Aggregation	This approach meets expectations with pseudo-data instead of with modifications of unique data, this helpes in preferred preservation of privacy over systems which basically utilize modifications of the first data.
5	Privacy Preserving Data Mining: An Extensive Survey-2013[1]	Secure Multiparty Computation	Cryptography	Generally SMC technique use where data are coming from different sites and it uses trusted third party to preserve privacy.
6	Privacy Preserving Data Mining: An Extensive Survey-2013[1]	Pseudonymization	Cryptography	It is an approach that breaks the link between personal and medical information. It provides a form of traceable anonymity of health records.

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 9, September 2015

IV. REVIEW OF PRIVACY PRESERVING TECHNIQUES

1. Association Rule Mining

The most vital information mining procedure utilized as a part of information mining is Association Rule in numerous genuine sales. It is utilized to uncover unpredicted dealings in the information. An affiliation standard is a ramifications of the structure A⇒B where A⊆I, B⊆I and A∩B =∅. Where I = {i1, i2, . . . , in} be an arrangement of things. The rule A⇒B has support S in the transaction database DB if S% of transactions in DB contains A∪B. The association rule holds in the transaction database DB with confidence C if C% of transactions in DB that contain A also contains B. An itemset X with k items is called a k-itemset[12].

2. Clustering

For privacy preserving in clustering most commonly utilized algorithm is k - means clustering methodologies separated from other privacy preserving information mining ones is imperative because of the utilization of this algorithm in critical different territories, similar to image and signal processing where the issue of security is unequivocally postured. A large portion of workings in privacy preserving clustering are made on the k-means algorithm by applying the model of secure multi-party calculation on distinctive disseminations (vertically, on a level plane and discretionary apportioned information).

3. Classification Data Mining

Classification is a standout amongst the most widely recognized applications found in this present reality. The objective of classification is to construct a model which can predict the estimation of one variable, taking into account the values of alternate variables. Decision tree classification is one of the best known arrangement approaches. The decision tree in ID3 is constructed top-down in a recursive manner. In the first cycle it discovers the characteristic which best arranges the information considering the objective class attribute[8].

4. Bayesian Data Mining

Bayesian networks are an effective data mining tool. A Bayesian network comprises of two sections: the network structure and the network parameters. Bayesian networks can be utilized for some assignments, for example, hypothesis testing and automated scientific discovery. A Bayesian network (BN) is a graphical model that encodes probabilistic connections among variables of hobby[8].

V. APPLICATIONS OF PRIVACY PRESERVING ALGORITHMS

1. Medical Database

Suppose in scrub system Clinical information would be in the type of content, which has information of patients like his relatives, location, blood groups and telephone number. Old-style procedures have been used just for worldwide hunt and swap forms in order to maintain privacy. According to creator Sweeny L findings, Scrub system used numerous detection algorithms in order to support the privacy[12].

2. Bioterrorism Application

It is crucial to investigate the medical data for privacy preservation in the bioterrorism application. For instance, Biological operators are broadly found in the common habitat, for example, Bacillus anthracic. It is imperative to discover the Bacillus anthracic attack from the typical attack. It is important to track rates of the normal diseases. The comparing data would be accounted for to the general health organizations. The respiratory diseases were not reportable-diseases. This gives an answer for more identifiable information as per general health law[12].

4. Multiparty data sharing

Suppose more than one party wants to collaborate their data to mine it for their mutual benefit then those parties need to share their data. This sharing of the data cause privacy issue in picture. Secure Multi Party Computation used to secure it.

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 9, September 2015

5. Social network

May be government agencies wants social data for their private matter and the owner of the social data not wants to reveal privacy of their customers. So to maintain the privacy to the social data we have to use privacy preserving techniques to achieve it.

VI. CONCLUSION AND FUTURE WORK

Here presented a classification and an extended description of various privacy preserving data mining algorithms. The work presented here indicates the ever increasing interest of researchers in the area of securing sensitive data and knowledge from malicious users. This paper gives the overview to the privacy preserving in data mining with its techniques, algorithms of privacy preserving in data mining, applications of privacy preserving in data mining. With this paper we can get an overview to the emerging topic of privacy preserving data mining.

Numerous PPDM strategies have been proposed to ensure touchy information in each PPDM layer. In any case, a few issues still stay to be tended to later on. First, personalized PPDM should be studied. Since privacy is a subjective concept regarded as a personal issue, mostly privacy preservation needs to be personalized. Second, the trust of data miner should be evaluated in order to optimize PPDM. Apart from personalized privacy preservation for data providers, multi-level trust of data miners is also an open issue that affects the degree of modification of raw data for privacy preserving purpose. Third information loss especially in anonymization techniques and balancing information loss and privacy. With using clustering techniques in l-diverse anonymity we can get good utility in result. We proposed an approach efficient anonymity using l-diversity along with clustering to get good utility. There are few efforts are given regarding this approach and they all used different clustering approach to get good utility in anonymity[13][14][15][16].

REFERENCES

1. Jisha Jose Panackal and DrAnitha S Pillai ‘Privacy Preserving Data Mining: An Extensive Survey’ Proc. of Int. Conf. on Multimedia Processing, Association of Computer Electronics and Electrical Engineers, 2013.
2. XinjunQi ,MingkuiZong ‘An Overview of Privacy Preserving Data Mining’ International Conference on Environmental Science and Engineering, 2011.
3. Xuyun Li, Zheng Yan, Peng Zhang ‘A Review on Privacy Preserving Data Mining’ IEEE International Conference on Computer and Information Technology, 2014.
4. HillolKarguptaand SouptikDatta, Qi Wang andKrishnamoorthySivakumar ‘On the Privacy Preserving Properties of Random Data Perturbation Techniques’ Proceedings of the Third IEEE International Conference on Data Mining (ICDM’03).
5. Charu C. Aggarwal and Philip S. Yu ‘A Condensation Approach to Privacy Preserving Data Mining’ IBM T. J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532.
6. L. Sweeney, ‘k-Anonymity: A Model for Protecting Privacy’ in proceedings of Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002.
    a. Machanavajjhala, J.Gehrke, D. Kifer and M. Venkitasubramaniam, ‘I-Diversity: Privacy Beyond k-Anonymity’, Proc. Int'l Con! Data Eng. (ICDE), p. 24, 2006.
7. TamannaKachwala, SwetaParmar‘ An Approach for Preserving Privacy in Data Mining’International Journal of Advanced Research in Computer Science and Software Engineering , Volume 4, Issue 9, 2014 .
8. Santosh Kumar Bhandare ‘Data Transformation and Encryption Based Privacy Preserving Data Mining System’International Journal of Advanced Research in Computer Science and Software Engineering ,Volume 4, Issue 7, 2014 .
9. Supriya S. Borhade, Bipin B. Shinde‘Privacy Preserving Data Mining Using Association Rule With Condensation Approach’ International Journal of Emerging Technology and Advanced Engineering , Volume 4, Issue 3, 2014.
10. S.Gokila, Dr.P.Venkateswari ‘A Survey On Privacy Preserving Data Publishing’ International Journal on Cybernetics & Informatics (IJCI) Vol. 3, No. 1, 2014
11. Mahmoud Hussein, Ashraf El-Sisi, and Nabil Ismail ‘Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Data Base’,Springer-Verlag Berlin Heidelberg, pp. 607–616, 2008.
12. MdEnamulKabir, Hua Wang, Elisa Bertino&Yunxiang Chi ‘Systematic Clustering Method for l-diversity Model’ Twenty-First Australasian Database Conference (ADC2010), Brisbane, Australia,Conferences in Research and Practice in Information Technology(CRPIT), Vol. 103, 2010.
13. Xianmang He, HuaHui Chen, YefangChen,Yihong Dong, PengWan, and ZhenhuaHuan ‘Clustering-Based k-Anonymity’,Springer-Verlag Berlin Heidelberg, pp. 405–417, 2012.
14. Gaoming Yang, Jingzhao Li, Shunxiang Zhang, Li Yu ‘An Enhanced l-Diversity Privacy Preservation’ 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD),2013.
15. Pingshui Wang, JiandongWang ‘L-diverse Anonymity Algorithm Based onClustering Techniques’ Journal of Information & Computational Science 9: 9 2012.

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 9, September 2015

BIOGRAPHY

DharmikVasiyani is a ME student in the Computer Engineering Department in B.H. Gardi college of engineering & Technology, Rajkot, Gujarat, India. He received Bachelor of Information Technology degree in 2014 from L.E. College, Morbi, Gujarat, India. Her research interests are Data Mining and Security etc.

Hiral Desai is a ME student in the Computer Engineering Department in B.H. Gardi college of engineering & Technology, Rajkot, Gujarat, India. She received Bachelor of Information Technology degree in 2014 from L.E. College, Morbi, Gujarat, India. Her research interests are Data Mining, Big data and Artificial intelligence etc.

Jay Gandhi is an Assistant Professor of Information Technology Department in B.H. Gardi college of engineering & Technology, Rajkot, Gujarat, India. He received Master of Technology in information Technology from Charusat University, Changa.