© 2018 IJSRSET | Volume 4 | Issue 6 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

A Review on Privacy Preservation in Data Mining

Mahesh Dumbere, Roshani Talmale
Department of Computer Science and Engineering, TGPCET, Nagpur, Maharashtra, India

ABSTRACT

The main focus of privacy preserving data publishing was to enhance traditional data mining techniques for masking sensitive information through data modification. The major issues were how to modify the data and how to recover the data mining result from the altered data. The reports were often tightly coupled with the data mining algorithms under consideration. Privacy preserving data publishing focuses on techniques for publishing data, not techniques for data mining. In case, it is expected that standard data mining techniques are applied on the published data. Anonymization of the data is done by hiding the identity of record owners, whereas privacy preserving data mining seeks to directly belie the sensitive data. This survey carries out the various privacy preservation techniques and algorithms.

Keywords : Data Mining, Privacy Preserving, Anonymization

I. INTRODUCTION

The huge amount of data available in information databases becomes worthless until the useful information is extracted. Mining knowledge from the data is said to be data mining. The two steps are analyses and extract useful information from database is mandatory for further use in different work environments like market analysis, fraud detection, science exploration, etc. Information extraction carried out the following duties such as cleaning data, integration of data, transformation of data, pattern evaluation, and data presentation. The boom of data mining relies on the availability of high in data quality and effective sharing. The figure 1 explains the process of privacy preservation technique.

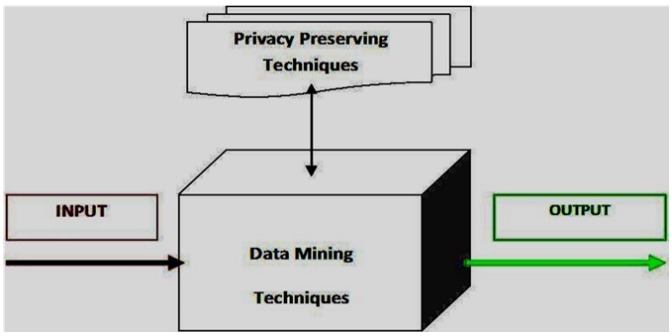

Figure 1. Privacy Preservation Technique

The data mining works with generation of association rules, the modification in support and confidence of the association rule for masking sensitive rules is done. A concept named „not altering the support" is deployed to hide an association rule. There are two approaches in privacy preserving data mining. The data Perturbing values for preservation of customer privacy is the first approach. The other approach is Cryptographic tools to build data mining models. Privacy preserving [12] is said to be worked out when the attacker is not able to learn anything extra from the given data even though with the presence of his background knowledge obtained from other sources.

II. OVERVIEW

Privacy Preservation

The main focus of privacy preserving data publishing was to enhance traditional data mining techniques which mask the sensitive information by modifying the data. The major issues were how to modify the data and how to rediscover the data mining result from the modified data. The data Perturbing values for preservation of customer privacy is the first approach. The other approach is Cryptographic tools to build data mining models. Privacy preserving [12] is preferred to be go out when the attacker is unable to know anything extra from the given data even though with the presence of his background knowledge obtained from other sources.

Anonymization

Anonymization of the data is done by hiding the identity of record owners, whereas privacy preserving data mining seeks to directly belie the sensitive data. The problem of privacy preservation in social networks is a major problem. The goal is to arrive at an anonymized view of the network which is unified without flat out to any of the data holder‟s information apropos links amid nodes that are controlled by other data holders. The anonymization algorithm and SaNGreeA algorithm [1] used for sequential clustering. Anonymity parameters are used for sequential clustering algorithms for anonymizing social networks.

III. LITERATURE SURVEY

Sequential Clustering for Anonymization of Centralized and Distributed Social Networks

The complication of privacy-preservation in social networks is a major problem. The goal is to arrive at an anonymized view of the unified network without eloquent to any of the data holder‟s information about links between nodes that are controlled by other data holders. The anonymization algorithm and SaNGreeA algorithm [1] used for sequential clustering. Anonymity parameters are used for anonymizing social networks by using sequential clustering algorithms. Several algorithms produce anonymizations by means of clustering which have an efficient utility than those achieved by existing algorithms.

On the Design and Analysis of the Privacy-Preserving SVM Classifier

SVM classifier without exposing the private content of training data is preferably said as Privacy Preserving SVM Classifier [2]. Data mining algorithm, Classification classifier for public use or deliver the SVM classifier to clients will bare the private content of support vectors. This violates the privacy-preserving needs for some legal or commercial account. Privacy violation problem, and propose an approach as a base technique for the SVM classifier to revamp it to a privacy preserving classifier which does not announce the private content of support vectors.

Improved MASK Algorithm for Privacy Preserving Association Rules on Data Mining

A data perturbation strategy is implemented through the MASK algorithm, which leads to a debased privacy-preserving degree. In a while, it is challenging to handle the MASK algorithm into real time due to long execution time. A hybrid algorithm encapsulated with data perturbation and query restriction (DPQR) [3] to maximize the privacy-preserving degree by multi-parameters perturbation. Data Perturbation and Query Restriction (DPQR) algorithm are used to improve privacy-preserving degree and time-efficiency is achieved. The proposed DPQR is more suitable for Boolean data, and it cannot deal with numerical data or other types of data.

Privacy-Preserving Gradient-Descent Methods

Gradient descent [4] aims to minimize a target function in order to reach a local trace. In data mining, this function accords to a decision model that is to be discovered. The author present two technical approaches stochastic approach and least square approach. Languages modeling smoothing parameters, weight parameter are used to measure the performance of the system. The proposed secure building blocks are scalable and the proposed protocols permit us to determine an efficient secure protocol for the applications for each scenario. The author will extend PPGD to vertically partitioned data implementing the least square approach for N-number of parities.

Crowd sourcing Database for K-Anonymity

Author suggested integrating the crowdsourcing techniques [5] into the database engine. It addresses the privacy concern, as each crowdsourcing job requires revealing of some sensitive data to the anonymous human trader. In this paper, the study focused how to guarantee the data privacy in the crowdsourcing scenario. A probability-based matrix model is inaugurated to estimate the lower bound and upper bound of the crowdsourcing certainty for the anonymized data. The model exhibits that K-Anonymity approach needs to solve the trade-off between the privacy and the accuracy. Propose a novel K-Anonymity approach. Experiments show that the solution can cultivate high accuracy results for the crowdsourcing jobs.

Privacy Preserving Decision Tree Learning Using Unrealized Data Sets

Author suggested a privacy preserving approach that can be applied to decision tree learning [6], without loss of accuracy. It deploys the strategies to the preservation of the privacy of collected data samples. It converts the original sample data sets into a group of unreal data sets, from which the original samples cannot be, reestablish without the entire group of unreal data sets. In a while an unreal data sets which directly built an accurate decision tree. It can be applied directly to the stored data as soon as the first sample is collected. The approach is better than the other privacy preserving approaches, such as cryptography, for extra protection.

Traffic Information Systems Based On Secure and Privacy-Preserving Smartphone

Author leverage state-of-the-art cryptographic schemes [7] and readily available telecommunication infrastructure and presented a comprehensive outperform for traffic estimation on smartphone that is tried and true to be secure and privacy preserving. A localization algorithm, suitable for GPS location samples, and evaluated it through realistic simulations. Results confirm it is attainable to build accurate and trustworthy smartphone-based TIS.

A Data Mining Perspective in Privacy Preserving Data Mining Systems

The PPDM systems deployed the key exchange process by cryptographic manner and the key computation process accomplished by a third party. The Key Distribution-Less Privacy Preserving Data Mining (KDLPPDM) [9] system is designed. The system novelty is that no data is published in a same while the association rules are reported to achieve effective data mining results. Commutative RSA cryptographic algorithms are suggested for key exchanging process. It overcomes the sustentation arising due to key exchange and key computation by applying the cryptographic algorithm.

Privacy-Preserving Data Analysis

The existing PPDA techniques [12] cannot prevent participating parties from modifying their private inputs. It is difficult to check whether the parties participating are reliable about their private input data. Proposed model first develop key theorems, then based on these theorems, they analyze certain important privacy-preserving data analysis tasks that

telling the truth is the optimized opinion for any participating party. Deterministically non-cooperatively computable (DNCC) parameter used for measure the system performance. Claim 5.1, as long as the last step in a PPDA task is in DNCC, it is always possible to make the entire PPDA task satisfying the DNCC model.

Random Nonlinear Data Distortion for Privacy-Preserving Outlier Detection

The data owner has some private or sensitive data and needs a data miner to access them for speculating important patterns by which the sensitive information [20] is not revealed. Privacy preserving data mining desired to solve this problem by transforming randomly the data prior to be allowed to the data miners. Previous works only focused towards the case of linear data perturbations. Author defines nonlinear data distortion through nonlinear random data transformation.

IV. COMPARISONS ON DIFFERENT PRIVACY PRESERVATION TECHNIQUES

TITLE	ALGORITHM	PARAMETER	CONCLUSION
Anonymization of Centralized and Distributed Social Networks by Sequential Clustering	Anonymization algorithm and SaNGreeA algorithm used for sequential clustering	Clustering coefficient, Diameter, Average distance, Effective diameter, Epidemic threshold.	The presented sequential clustering algorithms for anonymizing social networks. Those algorithms produce anonymizations by means of clustering with better utility.
Data Mining for Privacy Preserving Association Rules Basedon Improved MASK Algorithm	Data Perturbation and QueryRrestriction (DPQR)	Multi-parameters perturbation	The privacy-preserving degree and timeefficiency is achieved. The DPQR is more suitable for Boolean data.
K-Anonymity for Crowdsourcing Database	K-Anonymity algorithm	No. Of Tuples And Data spaces are used for measure the performance of the system.	The Outperforms standard K-Anonymity approaches on retaining the effectiveness of crowdsourcing.
Privacy Preserving Decision Tree Learning Using Unrealized	Tree learning Algorithm, decision tree generation are used.	Temperature Humidity, Wind Play.	The decision tree algorithm is compatible with other privacy preserving approaches, such as cryptography, for extra protection.

Data Sets			
Secure and Privacy-Preserving Smartphone-Based Traffic Information Systems	KeyGen(*n*) algorithm	GSC(group signature center)Accuracy,Simulation, time stamp	A localization algorithm, suitable for GPS location samples, and evaluated it through realistic simulations.
On the Design and Analysis of the Privacy-Preserving SVM Classifier	Data mining algorithm, Classification algorithm, kernal adatron algorithm and datafly algorithm.	Cost parameter, Kernalparameter are used to measure the performance of the system.	PPSVC can achieve similar classification accuracy to the original SVM classifier. By protecting the sensitive content of support vectors.
Privacy-Preserving Gradient-Descent Methods	Genetic Algorithms	Languages modeling smoothing parameters, weight parameters are used to measure the performance of the system.	The secure building blocks are scalable and the proposed protocols allow us to determine a better secure protocol for the applications for each scenario.
A Data Mining Perspective in Privacy Preserving Data Mining Systems	1.C5.0 data mining algorithm, Commutative RSA cryptographic algorithm.	Area covered by roc, curve data set id, sensitivity, specificity-1	Overcomes the overheads arising due to key exchange and key computation by adopting the cryptographic algorithm.
Incentive Compatible Privacy-Preserving Data Analysis	Data analysis algorithms	Deterministically noncooperatively computable (DNCC).	Claim 5.1, as long as the last step in a PPDA task is in DNCC, it is always possible to make the entire PPDA task satisfying the DNCC model.

Privacy and Quality Preserving Multimedia Data Aggregation for Participatory Sensing Systems	Outlier detection anomaly detection algorithm, secure hash algorithm.	Detection rate, data range. Indices, anomaly score	A general method for computing the bounds on a nonlinear privacypreserving data-mining technique with applications to anomaly detection.

V. CONCLUSION

Review on data mining privacy preserving in social network. Main objective of this review on privacy preservative technique is to protect different users and their identities in the social network along with obtaining originality. To achieve this goal, there is a need to develop perfect privacy models to specify the expected loss of privacy under different attacks, and deployed anonymization techniques to the data. So, the various techniques are surveyed.

VI. REFERENCES

[1] Tamir Tassa and Dror J. Cohen"Anonymization Of Centralized And Distributed Social Networks By Sequential Clustering", *IEEE Transactions On Knowledge And Data Engineering*, Vol. 25, No. 2, February 2013.

[2] Keng-Pei Lin and Ming-Syan Chen, "On The Design And Analysis Of The Privacy-Preserving SVM Classifier", *IEEE Transactions On Knowledge And Data Engineering*, Vol. 23, No. 11, November 2011.

[3] Haoliang Lou, Yunlong Ma, Feng Zhang, Min Liu, Weiming Shen "Data Mining For Privacy Preserving Association Rules Based On Improved Mask Algorithm" ,*Proceedings Of The 2014 IEEE 18th International Conference On Computer Supported Cooperative Work In Design*.

[4] Shuguo Han, Wee Keong Ng, Li Wan, and Vincent C.S. Lee "Privacy-Preserving Gradient-Descent Methods" *IEEE Transactions On Knowledge And Data Engineering*, Vol. 22, No. 6, June 2010.

[5] Sai Wu, Xiaoli Wang, Sheng Wang, Zhenjie Zhang, And Anthony K.H. Tung "K-Anonymity For Crowdsourcing Database", *IEEE Transactions On Knowledge And Data Engineering*, Vol. 26, No. 9, September 2014.

[6] Pui K. Fong And Jens H. Weber-Jahnke, "Privacy Preserving Decision Tree Learningusing Unrealized Data Sets", *IEEE Transactions On Knowledge And Data Engineering*, Vol. 24, No. 2, February 2012.

[7] Stylianos Gisdakis, Vasileios Manolopoulos, Sha Tao, Ana Rusu, And Panagiotis Papadimitratos, "Secure And Privacy-Preserving Smartphone-Based Traffic Information Systems",*IEEE Transactions On Intelligent Transportation Systems*, Vol. 16, No. 3, June 2015.

[8] Kinjal Parmar1, Vinita Shah2, "A Review On Data Anonymization In Privacy Preserving Data Mining", *International Journal Of Advanced Research In Computer And Communication Engineering*, Vol. 5, Issue 2, February 2016.

[9] Kumaraswamy Sȧ, Manjula S Hȧ, K R Venugopalȧ And L M Patnaikḃ "A Data Mining Perspective In Privacy Preserving Data Mining Systems",*International Journal Of Current Engineering And*

Technology India ,Accepted 20 March 2014, Available Online 01 April 2014, Vol.4, No.2 (April 2014)

[10] Jordi Soria-Comas, Josep Domingo-Ferrer, David Sanchez And Sergio Martınez "T-Closeness Through Microaggregation: Strict Privacy With Enhanced Utility Preservation ",*IEEE Transactions On Knowledge And Data Engineering*, Vol. 27, No. 11, November 2015 .

[11] Jung Yeon Hwang, Liqun Chen, Hyun Sook Cho, And Daehun Nyang"Short Dynamic Group Signature Scheme Supporting Controllable Linkability *",IEEE Transactions On Information Forensics And Security*, Vol. 10, No. 6, June 2015.

[12] Murat Kantarcioglu and Wei Jiang"Incentive Compatible Privacy-Preserving Data Analysis", *IEEE Transactions On Knowledge And Data Engineering,* Vol. 25, No. 6, June 2013.

[13] Lei Xu, Chunxiao Jiang, *Yan Chen,* Yong Ren, And K. J. Ray Liu, "Privacy Or Utility In Data Collection?A Contract Theoretic Approach" ,*IEEE Journal Of Selected Topics In Signal Processing,* Vol. 9, No. 7, October 2015.

[14] Huang Lin and Yuguang Fang "Privacy-Aware Profiling And Statistical Data Extraction For Smart Sustainable Energy Systems", *IEEE Transactions On Smart Grid,* Vol. 4, No. 1, March 2013.

[15] Depeng Li, Zeyar Aung, John Williams, and Abel Sanchez "P3: Privacy Preservation Protocol For Automatic appliance Control Application In Smart Grid", *IEEE Internet Of Things Journal*, Vol. 1, No. 5, October 2014.

[16] Fudong Qi, Ieee, Fan Wu, Guihai Chen, "Privacy And Quality Preserving Multimedia Data Aggregation For Participatory Sensing Systems",*IEEE Internet Of Things Journal,* Vol. 1, No. 5, October 2014.

[17] Günther Eibl, And Dominik Engel, "Influence Of Data Granularity On Smart Meter Privacy", *IEEE Transactions On Smart Grid,* Vol. 6, No. 2, March 2015.

[18] Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, And Hui (Wendy) Wang "Privacy-Preserving Mining Of Association Rules From Outsourced Transaction Databases", *IEEE Systems Journal,* Vol. 7, No. 3, September 2013.

[19] Siyuan Liu, Qiang Qu, Lei Chen, And Lionel M. Ni, "SMC: A Practical Schema For PrivacyPreserved Data Sharing Over Distributed Data Streams", *IEEE Transactions On Big Data,* Vol. 1, No. 2, April-June 2015.

[20] Kanishka Bhaduri, Mark D. Stefanski, and Ashok N. Srivastava, "Privacy-Preserving Outlier Detection Through Random Nonlinear Data Distortion", *IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics,* Vol. 41, No. 1, February 2011.