IJournals: International Journal of Software & Hardware Research in Engineering
ISSN-2347-4890
Volume 5 Issue 1 January, 2017

A Survey on Heuristic Based Approach for Privacy Preserving on Data Stream

Palak Godhani: Akshay Kansara[1]; Aniket Patel[2]; Dr. Kiran Amin[3]

Department of Computer Engineering[1 2 3]

S. P. B. Patel Engineering College[1]: Silver Oak College of Engineering & Technology[2]; U V Patel College of Engineering[3]

palakjsca@gmail.com: akshay.kansara@saffrony.ac.in[1]; aniketpatel.it@gmail.com[2]; kiran.amin@ganpatuniversity.ac.in[3]

IJournals

ABSTRACT

Privacy preserving becomes an important issue in the development progress of data mining techniques. Privacy preserving data mining has become increasingly popular because it allows sharing of privacy-sensitive data for analysis purposes. So, people have become increasingly unwilling to share their data, frequently resulting in individuals either refusing to share their data or providing incorrect data. In turn, such problems in data collection can affect the success of data mining, which relies on sufficient amounts of accurate data in order to produce meaningful results. In recent years, the wide availability of personal data has made the problem of privacy preserving data mining an important one. A number of methods have recently been proposed for privacy preserving data mining of multidimensional data records. This paper intends to reiterate several privacy preserving data mining technologies clearly and then proceeds to analyze the merits and shortcomings of these technologies. Using such approaches the data accuracy and preservation can be achieve. An effective approach is heuristic based approach. It has aim to achieve the privacy of static data with minimum information loss.

Keywords: Data Mining, Privacy preserving data mining, heuristic based approach.

1. INTRODUCTION

Data mining is nothing but extracting meaningful knowledge from the large amount of data. We can classify data mining techniques as follows: classification, association rule mining, clustering, sequential pattern analysis, data visualization, prediction. In recent years, simple transactions like using credit card, browsing the web, phone database, sensor network lead to wide and automated data storage. All these have large flows of data continuously and dynamically. This type of large volume data leads to many mining and computational challenges.

Huge volumes of detailed personal data are regularly collected and analyzed by applications using data mining. Such data include shopping habits, criminal records, medical history, credit records, among others. On the one hand, such data is an important asset to business organizations and governments both to decision making processes and to provide social benefits, such as medical research, crime reduction, national security, etc. [2] The threat to privacy becomes real since data mining techniques are able to derive highly sensitive knowledge from un classified data that is not even known to database holders. Worse is the privacy invasion occasioned by secondary usage of data when individuals are unaware of “behind the scenes” use of data mining techniques [3]

A. Data Stream Mining

Data stream is new type of data that is different than traditional static database. Data stream is continuous and dynamic flow of data. It is sequence of real time data with high data rate and application can read once. The characteristics of data streams are different than traditional static database which are as follows [4]: (1) Data has timing preference (2) Data Distribution changes constantly with time (3) The amount of data is enormous (4) Data flows in and out with fast speed (5) Immediate response is required. Because of these, data stream mining is challenging.

We have many data mining algorithm for traditional database where data is static and continuous flow. Use of traditional data mining algorithm is not appropriate in data stream mining because of no control over dataflow. If data will change, then we have to rescan the database. This will take more computational time. In data stream mining data is not persistent but rapid and time varying.

In mining of data stream, solution is categorized in data based and task based. In data based solution, data transform horizontally or vertically. In task based solution, different techniques have been adopted to achieve time and space efficient solution. Figure 1.1 shows simple data stream mining process. Once element of data stream is processed, it is discarded. So, it is not easy to retrieve it unless if we explicitly store them in memory.

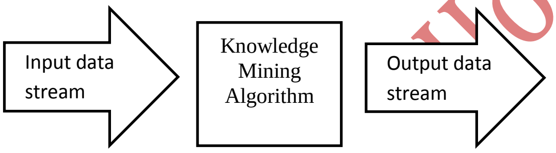

Fig. 1. Data Stream Mining Process

2. Need for Privacy in Data Mining

Information is today probably the most important and demanded resource. We live in an internetworked society that relies on the dissemination and sharing of information in the private as well as in the public and governmental sectors. Governmental, public, and private institutions are increasingly required to make their data electronically available [5][6]. To protect the privacy of the respondents (individuals, organizations, associations, business establishments, and so on). Although apparently anonymous, the deidentified data may contain other data, such as race, birth date, sex, and ZIP code, which uniquely or almost uniquely pertain to specific respondents (i.e., entities to which data refer) and make them stand out from others[7].By linking these identifying characteristics to publicly available databases associating these characteristics to the respondent's identity, the data recipients can determine to which respondent each piece of released data belongs, or restrict their uncertainty to a specific subset of individuals.

IJournals

3. PRIVACY PRESERVING DATA MINING TECHNIQUES (PPDM)

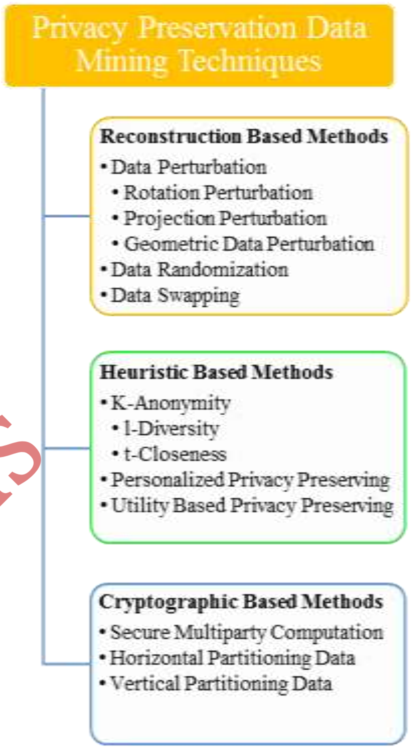
Privacy Preservation Data Mining Techniques
Reconstruction Based Methods
•Data Perturbation
•Rotation Perturbation
•Projection Perturbation
•Geometric Data Perturbation
•Data Randomization
•Data Swapping
Heuristic Based Methods
•K-Anonymity
•l-Diversity
•t-Closeness
•Personalized Privacy Preserving
•Utility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

Figure 1: Classification of Privacy Preserving Data Mining [22]

The term privacy means keep information about me from being available to others. PPDM is producing valid model and pattern without disclosing sensitive information. The main aim 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. On other side, sharing of these data will useful in industries and business organization. They collect and analyze the data to know market policy. Sharing of data will help in improving the business strategy and to know customer behavior. If owner share the data, privacy must not breach.

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 various techniques for privacy preserving for entire dataset modification or modification of some sensitive attributes. PPDM techniques are classified into four types: data partitioning, data modification, data restriction, data ownership. Because of personal data, privacy preserving becomes important in recent years. Now a day advanced technology provides the capacity to store large number of personal data. Suppose cancer research institutes in different areas need to collaboratively find the environmental factor related to certain type of cancer [8]. These distributed databases contain sensitive information.

4. RELATED WORK

4.1 Reconstruction based technique

Reconstruction-based techniques perturb the original data to achieve privacy preserving. The perturbed data would meet the two conditions. First, an attacker cannot discover the real original data from the issuance of the distortion data. Second, the distorted data is still to maintain some statistical properties of the original data, namely some of the information derived from the distorted data are equivalent to data obtained from the original information [13]. So, for each data mining techniques separate algorithms need to be developed. Data perturbation is that the value of each property is transformed into other value of the property domain by given probability [12].

Data perturbation: Basically, change the original data into another.

- **Rotation perturbation:** Rotate the data according to a given angle θ with the origin as the center [9] [10].
- **Projection perturbation:** Projecting a set of data points from the origin multidimensional space to another randomly chosen space [9].
- **Geometric perturbation:** Transform the data into sequence of random geometric transformations [9].

Data randomization: Randomly chose the data from the record and place it at the place of another data. [11]

Data swapping: Swaps the data into two half [10].

4.2 Heuristic based approach

Various techniques have been developed to sanitize or modify selective data for data mining techniques like association rule mining, classification and clustering. Selective data sanitization or modification based mining problem is NP-hard and for this reason, heuristic can be used to address the complexity issues. The concept of protecting respondent identity through micro data release using k-anonymity was first proposed by P. Samarati in [16], and subsequently many techniques have been proposed based on it, such as l-diversity [15], t-c1oseness [15], Incognito [14], and so on. K-anonymity protects against identity disclosure; it does not provide sufficient protection against attribute disclosure.

We know that the database contains different types of attributes. Explicit Identifiers: Are attributes which can identify the person uniquely example identity number, PAN card number etc. [17] Quasi-Identifier (QI): The attributes which cannot alone identify the person uniquely but by collecting quasi-identifiers any one can easily recognize any person, example Zip-code, Birth date etc. Sensitive Attributes (SA): The information which everyone tries to hide from adversaries example Salary, Disease etc[18][19]. Non-Sensitive Attributes: The attributes which does not relate to any other categories and have no importance when disclose to anyone. Each group that shares the same values on every QI is called Equivalence Class (EC). While releasing the sensitive information, it is required to preserve them from disclosure. There are mainly two types of Information Disclosure: Identity Disclosure and Attribute Disclosure. [18] [19] [20] [21].

k-Anonymity: The database is said to be K-anonymous where attributes are suppressed or generalized until each row is identical with at least k-1 other rows. K-Anonymity thus prevents definite database linkages and guarantees that the data released is accurate. One of the emerging concept in microdata protection is k-anonymity, which has been recently proposed as a property that captures the protection of a microdata table with respect to possible re-identification of the respondents to which the data refer. K-anonymity demands that every tuple in the microdata table released be indistinguishably related to no fewer than k respondents.

l-Diversity: The next concept is “l-diversity”. Say you have a group of k different records that all share a particular quasi-identifier. That’s good, in that an attacker cannot identify the individual based

IJournals

on the quasi-identifier. But what if the value they're interested in, (e.g. the individual's medical diagnosis) is the same for every value in the group. The distribution of target values within a group is referred to as "l-diversity". [23]

t-Closeness: t-closeness that formalizes the idea of global background knowledge by requiring that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). This effectively limits the amount of individual-specific information an observer can learn. Intuitively, privacy is measured by the information gain of an observer.

Personalized privacy preservation: minimize the generalization for satisfying everyone's requirements so discard the maximum amount of information from the microdata (raw data is called micro data). [25]

Utility based privacy preservation: To improve the query answering accuracy on anonymized tabled. [26]

4.3 Cryptography based technique

In a distributed environment, the primary issue to achieve privacy preserving is the security of communications, and encryption technology just to meet this demand. Therefore, privacy preserving based on data encryption technology commonly applies to distributed applications. Lindell & Pinkas [24] first proposed Secure Multi-Party Computation protocol for data mining classification techniques. Cryptography based techniques offer a well-defined model for privacy, which includes methodologies for proving and quantifying it. Cryptography-based techniques have more time complexity compare to other method for data updating. Cryptography techniques are used to preserve privacy.

Distributed data mining provides different algorithms to perform computation in distributed manner without pooling the whole data into one place.

The secure multiparty computation is one of the distributed computing examples which is using worldwide for data distributed across the network. Firstly Yao [16] introduced the secure multiparty computation technique.

IJournals

Table 1: Comparison table of techniques

Technique	Advantages	Disadvantages
Reconstruction Based Technique	Data is transformed to achieve greater security. Different attributes are preserved independently.	Reduce the granularity loss of effectiveness of data.
Cryptography Based Technique	Encryption provide security to data.	More time complexity Security & attacks. Difficult to scale multiple parties are involved.
Heuristic Based Technique	Handle data in group based manner	Handling sensitive Data. Linking Attacks.

5. CONCLUSION

Heuristic based approach is applied for privacy preserving static data mining. Proposed approach is tried to keep the relationship between the sensitive data and anonymized data. So, that no one can forge the sensitive information from the data set. Here all attributes are independent attribute except the sensitive attributes. Here the risk analysis of anonymized data is decreased. Here we only use l-diversity method for future purpose we do work on data also including t-closeness i.e. we extend our algorithm using t-closeness.

6. REFERENCES

[1]. Aniket Patel, HirvaDivecha, Samir Patel, "A Study of Data Perturbation Techniques For Privacy Preserving Data Mining",2014.

[2]. L.Golab and M.T.Ozsu ,Data Stream Management issues-"A Survey Technical Report",2003.

[3]. Majid,M.Asger,Rashid Ali, "Privacy preserving Data Mining Techniques:Current Scenario and Future Prospects",IEEE 2012.

[4]. Golab, L. And Ozsu, M., "Issues in Data Stream Management," ACM SIGMOD Record, Vol. 32, pp. 5-14(2003).

[5]. Chen K, and Liu, “Privacy Preserving Data Classification with Rotation Perturbation”, proc.ICDM,2005,pp.589-592.

[6]. K.Liu, H Kargupta, and J.Ryan,” Random projection – based multiplicative data perturbation for privacy preserving distributed data mining .” IEEE Transaction on knowledge and Data Engg,Jan 2006,pp 92-106.

[7]. Keke Chen,Gordon Sun , and Ling Liu. Towards attack-resilient geometric data perturbation.” In proceedings of the 2007 SIAM international conference on Data mining,April 2007.

[8]. Keke chen, Ling Liu, Privacy Preserving Multiparty Collaborative Mining With Geometric Data Perturbation, IEEE TRANSACTION ON PARALLEL AND DISTRIBUTED COMPUTING, VOL.XX, NO. XX. JANUARY 2009.

[9]. Dasseni E., Verykios V., Elmagarmid A. and Bertino E., Hiding association rules by using confidence and support, Proceedings of the 4th international workshop on information hiding, pp. 369-383,2001.

[10]. Domingo-Ferrer J. and Mateo-Sanz J., Practical data-oriented micro aggregation for statistical disclosure control, IEEE transaction on knowledge and data engineering, pp. 189-201, 2002.

[11]. Dutta H., Kargupta H., Datta S. and Sivakumar K., Analysis of privacy preserving random perturbation techniques: further explorations, Proceedings of the workshop on privacy in the electronic society (in association with the 1 o•h ACM conference on computer and communications security), 2003.

[12]. Xiaolin Z. and Hongjing B., Research on privacy preserving classification data mining based on random perturbation,International Conference on lnfonnation, Networking and Automation (IC!NA), Vol. I, No. I, pp. 173-178, 2010.

[13]. Haisheng L., Study of privacy preserving data mining, 3rd Internat ional Symposium on Intelligent Information Technology and Security Informatics, pp. 700-703, 2010.

[14]. LeFevre K., DeWitt D. and Ramakrishnan R., Incognito: Efficient full domain k-anonymity., Proceedings of the ACM SIGMOD international conference on management of data, pp. 49-60, 2005.

[15]. Li N., Li T. and Venkatasubramanian S., !-closeness: Privacy beyond k-anonymity and !-diversity, Proceedings oLthe IEEE 23rd International conference on data engineering, 2007,

[16]. P.Samarati ,“Protecting respondents’ identities in microdata release,” In IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 13, issue 6, pp 1010-1027, 2001.

[17]. S. Fienberg and J. McIntyre, “Data Swapping: Variations on a Theme by Dalenius and Reiss,” Technical Report, National Institute of Statistical Sciences, 2003.

[18]. P. Samarati and L. Sweeney, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,” Technical Report SRI-CSL-98-04, 1998. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[19]. Christy Thomas, Diya Thomas,Dept of Computer Science & Engineering Rajagiri School of Engineering and Technology, Kochi,” An Enhanced Method for Privacy Preservation in Data Publishing”IEEE 2013.

[20]. Jian Wang,Yongcheng Luo, Yan Zhao,Jiajin Le College of Information Science and Technology, Donghua University Shanghai, China “A Survey on Privacy Preserving Data Mining” 2009 IEEE.

[21]. P.Samarati ,“Protecting respondents’ identities in microdata release,” In IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 13, issue 6, pp 1010-1027, 2001.

[22]. Machanavaijhala A., Gehrke J., Kifer D. and Venkitasubramaniam M., !-diversity: Privacy beyond k-anonymity, Proceedings of the 22nd international conference on data engineering, 2006.

[23]. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. “l-diversity: Privacy beyond kanonymity”. In ICDE, 2006.

[24]. Y. Lindell and B. Pinkas , “Privacy preserving data mining, Journal of

IJournals

Cryptology," vol. 15, issue 3, pp 177-206, 2002.

[25]. X. Xiao and Y. Tao, "Personalized Privacy Preservation," ACM SIGMOD Conference, 2006.

[26]. J. Xu, W. Wang, J. Pei, X. Wang, B. Shi and A. W. C. Fu, "Utility Based Anonymization using Local Recoding," ACM KDD Conference, 2006

IJournals