A SURVEY ON PRE AND POST MINING AND PRESERVING PRIVACY HEALTH CARE DATASETS

B. LAXMIKANTHA
Research Scholar
OPJSU, Churu, Rajasthan, India.

Dr. ARVIND KUMAR SHARMA
Associate professor
Department of CSE
OPJSU, Churu, Rajasthan, India.

ABSTRACT:

Data mining mechanisms have widely utilized for numerous business activities and various manufacturing companies across many industry sectors. The raw data is shared or sharing the extracted information in a form of rules it becomes a trend among business partnerships, as it is supposed to be a mutually benefit way of increasing productivity for all parties involved. The problem of protecting sensitive knowledge mined from databases. The sensitive knowledge is represented by a special group of association rules called sensitive association rules. These rules are paramount for strategic decision and must remain private (i.e., the rules are private to the company or organization owning the data). Data owners have to know in advance some knowledge (rules) that they want to protect. Such rules are fundamental in decision making, so they must not be discovered.

Keywords: *Data mining, privacy preserving, sequential pattern*

INTRODUCTION

Data mining is a method that helps to extract useful information from large databases. It is the technique of extracting relevant data from giant data bases through the utilization of data mining algorithms. As the quantity of information doubles each year, data mining is becoming an increasingly important tool to transform this data into information. Data mining deals with massive databases which can contain sensitive information. It needs data preparation which can discover information or patterns which may compromise confidentiality and privacy obligations. Advancement of efficient data mining technique has enlarged the risks of revealing sensitive data. Providing security to sensitive information against unauthorized access has been a long term objective for the database security research community and for the government statistical agencies. Hence, the protection issue has become an important area of research in data mining. The releasing of personal data in its most specific state poses a threat to the privacy of an individual. The threat to an individual's privacy happens when anyone who has access to the newly-compiled data set is able to identify specific individuals. The disclosure of extracted patterns open up the danger of privacy breaches that may reveal sensitive information to malicious users thus causing privacy violation in data mining. Hence there is a need for privacy preservation in data mining. Privacy preservation in data mining (PPDM) is the process of providing security to sensitive data in a database against unauthorized access. PPDM is implemented in order to protect the sensitive information and to prevent violation of privacy.

LITERATURE SURVEY

2.1 PRIVACY PRESERVING DATA MINING

Jian Wang et al (2009) intends to reiterate several privacy preserving data mining technologies clearly and then proceeds to analyze the merits and short comings of these technologies. In the recent years the advances in the technology has lead to assimilation of huge amount of data. These data can be stored and used for a various purposes. They may be processed in such a way it may lead to misuse of data .this impose a keen interest in the area of privacy preserving in data mining.

Privacy Preserving Data mining Analysis is (Lindel et al 2000) an amalgamation of the data of heterogeneous users without disclosing the private and susceptible details of the users. They proposed randomized response techniques to solve the DTPD (Building Decision Tree on Private Data) problem. The basic idea of the randomized response is to scramble the data in such a way that the central place cannot tell with probabilities better than a predefined threshold whether the data from a customer contain truthful information or false information. It introduced the condensation approach which constructs contained clusters in the data set and then generates pseudo data from the statistics of these clusters. The constraints to the cluster are defined in the terms of sizes of the clusters which are chosen in a way so as to preserve k-anonymity (Berberidis et al 2005). Utilization of various methods like data transformation, applications to preserve privacy, cryptographic methods, and higher dimensionality challenges.

Maryam Khan et al (2010) has presented medical tourism: outsourcing of healthcare. Their study examined whether the growth in medical tourism will eventually have a result in the outsourcing of U.S. healthcare services. They showed that as long as people in developed countries lack affordable healthcare, medical tourism continues to grow. There were already an outsourcing of manufacturing, technology and service related jobs. The U.S. healthcare maintained its status quotient so that the healthcare services may also be outsourced.

Ken Barker et al (2009) gave the idea of data privacy taxonomy. They offered an explicit definition of data privacy which was suitable for ongoing work in data deposits such as, a DBMS or data mining. Their work contributed by briefly providing the larger context for the way privacy was defined legally and legislatively but primarily providing taxonomy that is capable of thinking of data privacy technologically. They demonstrated the taxonomy's utility by illustrating how this perspective makes it possible to understand the important contribution made by researchers to the issue of privacy. They declared privacy was indeed multifaceted so no single current research effort adequately addressed the true breadth of the issues necessary for fully understanding the scope of this important issue.

Varun Yadav (2012) they suggested about the security and information needed because of the wide availability of huge amount of data and the imminent need for such data into extracting the information and knowledge. The information and knowledge gained can be used for applications ranging

from market analysis, fraud detection and customer retention, to production control and science exploration. With and more information accessible in electronic forms and available on the web, and increasingly powerful data mining tools being developed and put into use, data mining may pose a threat to our privacy and data security . The real privacy threats (Tass et al 2014) are with unconditional access of individual records, like credit card, banking applications, customer ID, which must access privacy sensitive information. In this work they investigated the issue of data mining, as data shared before mining the means to shield it with Unified Modeling Language diagrams. Describing the privacy preserving definition, problem statement privacy preserving data mining technique, Architecture of the proposed work. They propose an amalgamated scaffold for Privacy Preserving Data Mining that ensures that the mining process will not trespass Privacy up to a certain degree of security.

Richard Huebner et al (2012) discussed about the various issues related to the privacy. They have discussed about the implementing privacy preserving data mining (Jian Wang 2009) solutions must first address internal privacy-related policies by investigating certain criteria's about privacy such as existing privacy policy, responsible person for making the policy, whom does the policy refer? Whether the policy address legal and ethical concerns and so on. The scope Data mining resources and processes must be defined. Algorithms for data mining must become more generalize. As of this writing, algorithms are designed to solve one specific task. Since there are so many different data mining tasks, it would be beneficial if Privacy Preserving Data Mining techniques were able to be applied to a variety of data mining tasks.

Elisa Bertino et al (2005) the order to use the medical data is to predict and accurate analysis for the quality health care. Two important issues pertaining to this sharing of data have to be addressed: protecting the data of an individual, the other is copyright protection over the data. In this work, they present a unified framework that flawlessly combines techniques of binning for privacy and digital watermarking for copy right to attain goal. Proposed binning method is built upon an earlier approach of generalization and suppression by allowing a broader concept of generalization. To ensure data usefulness, they proposed constraining binning by usage metrics that define maximal allowable information loss, and the metrics can be enforced off-line. The watermarking algorithm watermarks the binned data in a hierarchical manner by leveraging on the very nature of the data. The method is resilient to the generalization attack that is specific to the binned data, as well as other attacks intended to destroy the inserted mark. They proved that watermarking could not adversely interfere with binning, and implemented the framework.

Aaron et al (2010) The ultimate aim of their study was to shed light on the motivations and experiences of whistle-blowers in cases of major health care fraud. They conducted various interviews with whistle-blowers who were key informants in recent prosecutions brought against pharmaceutical

manufacturers. Enforcement actions against pharmaceutical manufacturers have become the most lucrative type of health care fraud litigation on the basis of recovery amounts.

2.2 PRIVACY PRESERVING ON DATABASES

Jaideep Vaidya (2003) put forth that privacy and security concerns can prevent sharing of data, derailing data mining projects. Distributed knowledge discovery, if done correctly can alleviate this problem. They presented a method for k-means clustering when different sites contain attributes for a common set of entities .each site learns the cluster of each entity, but learns nothing about the attributes at other sites. The secure multi party computation framework, computing a function privately is equivalent to computing it securely. Closest cluster computation returns the index of the closest cluster. To find important data points or patterns locally and utilize these to compute the global patterns. Assembling these into efficient privacy preserving data mining algorithms and proving them secure. They demonstrated how these can be combined to implement a standard data mining algorithms with provable privacy and information disclosure properties. The key is to obtain valid results, while providing guarantees on the non disclosure of data .The protocols holds in a situation where the parties do not collude to discover information.

Yingpeng Sang et al (2009) deals with the horizontal partitioned databases (Mingquan Ye et al 2010) among N parties, each every party have a private share of the database and tuple have the same set of attributes. By calculating the total exps (modular exponentiations) and muls (modular multiplications), and total communication bits, their PPDTM protocol (Yingpeng Sang et al 2009) for the semi honest model has a lower computation cost than the related solution in and by trading off communication cost. Since the computation cost dominates the whole cost, their protocol is faster than the solution the existing solution. The PPTAM protocol (Kantarcioglu et al 2004) for the semihonest model has lower computation and communication costs than the related solution derived by the techniques. By constructing the required zero-knowledge proofs, to extend the PPDTM and PPTAM protocols on malicious model. Some of these zeroknowledge proofs were also mentioned, but detailed constructions were not given. In addition, the Proof of Correct Polynomial Evaluation (POCPE) is not considered, without which an adversary can ask for decryptions of any useful information for itself. In comparison with the solutions derived from the techniques in the PPDTM protocol in malicious model has the same magnitude of costs, and the PPTAM protocol in malicious model has lower costs both in computation and communication.

Yonghong YU et al (2010), have suggested the integrated privacy protection and access control over outsourced database services. Moreover, a solution to enforce data confidentiality has also been explained the data privacy, user privacy and access control over outsourced database services. They started from a flexible definition of privacy constraints on a relational outline, applied encryption on information in a parsimonious

way and mostly rely on attribute partition to protect sensitive information. Their approximation algorithm for the minimal encryption attribute partitioned with quasi-identifier detection (Gudes et al 2006), they allowed storing the outsourced data on a single database server and minimized the amount of data represented in encrypted format. Here, they applied cryptographic technology on the auxiliary random server protocol that can solve the problem of private information retrieval to protect data privacy, user privacy and access control for outsourced database services. Their analysis showed the new model can provide efficient data privacy protection and query processing, which is efficient in computational complexity without increasing the cost of communication complexity of user privacy protection and access control.

Zhenmin lin (2009) Data perturbation is one of the commonly used models for privacy preserving data mining. The data owners change the data values in some way to hide the sensitive information while trying to maintain the utility of the data. They publish the distorted data instead of the original one. Random rotation is one of the popular approaches for data perturbation. It can preserve the data privacy without affecting the accuracy for rotationinvariant classifiers and clustering. They generalize this idea for vertically partitioned data sets. The proposed work rotate each sub-matrix randomly and independently and prove that it will preserve the geometric properties of the data matrix and thus the rotation invariant classifier and clustering techniques will achieve similar accuracy on the transformed data as on the original data. This method enables us to develop efficient centralized data mining algorithms while preserving privacy. Experiments on two real data sets show that this generalization is effective for vertically partitioned data sets. They note that such generalization only works for vertically partitioned data sets, but not for horizontally partitioned data sets.

Xun Yin & Yanchun Zhang (2012) this work proposes k-means clustering algorithm is a method of cluster analysis which aims to partitions n objects into k clusters. The clusters are formed to optimize an objective partitioning criterion based on a similarity function, such as distance, so that the objects within a cluster are similar, whereas the objects of different clusters are dissimilar in terms of the database attributes. The protocol is built on Lloyd‟s algorithm for k-means clustering, where the cluster centers are iteratively refined until no change in each cluster.

Security analysis has shown that the protocol protects privacy (Lindell et al 2000) in each round of iteration of k-means clustering with overwhelming probability as long as the underlying cryptographic building blocks are secure. To determine the index of minimal distance, each data site re-encrypts average k/2 cipher- texts, decrypts average k/2 cipher texts and posts k/2 intermediate cipher texts in PETs. The computation cost for one data site is average k/2 ElGamal‟s encryptions and average k/2 ElGamal‟s decryptions while the communication cost for one data site is the transmission of mk modular exponentiations. After updating cluster centers, to check if the termination criterion

is satisfied, each data site needs to perform two encryptions plus one decryption.

Latanya Sweeney (2002) provided a formal protection model named k-anonymity protection if the and a set of accompanying policies for deployment release provides k-anonymity protection if the information for each person contained in the release cannot be distinguished from at least k-1 individual whose information also appears. They also examines re identification attacks that can be realized on release that adhere to kanonymity unless accompanying policies are respected .The k-anonymity protection model is important because it forms the basis on which the real world systems known as data fly µ-Argus and k-similar provide guarantees of privacy protection.

Jingquan Li et al (2012) provided a technique for safe guarding the privacy of electronic medical records. They developed a formal privacy policy for safeguarding the privacy of EMRs. They described the impact of EMRs and HIPAA on patient privacy. They proposed access control and audit logs policies to protect patient privacy. To illustrate the best practice in the healthcare industry, they presented the case of the University of Texas M. D. Anderson Cancer Center. Where it has been demonstrated that it was critical for a healthcare organization to a formal privacy policy in place.

Tiancheng Li et al (2012) presented their technique (called slicing), which partitions the data both horizontally and vertically. They indicated that slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data. They demonstrated how slicing can be used for attribute disclosure protection and developed an efficient algorithm for computing the sliced data that obeys the „l"-diversity requirement. Their experiments confirm that slicing preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. Their experiments also demonstrated that slicing can be used to prevent membership disclosure.

Data mining technology (Murat Kantarcioglu et al 2009) has emerged as a means of identifying patterns and trends from large quantities of data. Recently, there has been growing concern over the privacy implications of data mining. Data mining is generally aimed at producing general models rather than learning about specific individuals, the process of data mining creates integrated data warehouses that pose real privacy issues. The collection of data is ubiquitous. With the rapid increase in computing, storage and networking resources, data is not only collected and stored but also analyzed. Indeed, data is often anonymized and released for public use.

Data that is of limited sensitivity by itself becomes highly sensitive when integrated, and gathering the data under a single roof greatly increases the opportunity for misuse. Even though some of the distributed data mining tasks protect individual data privacy, they still require that each site reveals some partial information about the local data One solution to this problem is to avoid disclosing data beyond its source, while still constructing data mining models equivalent to those that would have been learned on an

integrated data set. Since they prove that data is not disclosed beyond its original source, the opportunity for misuse is not increased by the process of data mining. Privacy-preserving distributed data mining algorithms require collaboration between parties to compute the results, while provably preventing the disclosure of any information except the data mining results. The definition of privacy followed in this line of research is conceptually simple: no site should learn anything new from the process of data mining. Specifically, anything learned during the data mining process must be derivable given one‟s own data and the final result. In other words, nothing is learned about any other site‟s data that isn‟t inherently obvious from the data mining result. The approach followed in this research has been to select a type of data mining model to be learned and develop a protocol to learn the model while meeting this definition of privacy.

2.3 PRIVACY PRESERVING ON SEQUENTIAL PATTERN MINING AND RULE HIDING

The manufacturers (Yu-Chiang Li et al 2006) have their own databases that record their patterns of stock and sale. For their mutual benefit, multiple companies decide to share their databases to cooperatively generate information and trends found in the shared large database. However, each company prefers to keep their own strategic patterns confidential from the others. Thus, a company can both uncover more interesting trends from the combined shared database than that available only from their own database, and can prevent sensitive information from being discovered by other companies In order to preserve strategic or sensitive intelligence and still share such knowledge among allied companies, a data sanitization process or privacy-preserving techniques must be applied. Sharing data or sharing mined rules has become a trend among business partnerships, as it is perceived to be a mutually benefit way of increasing productivity for all parties involved. Nevertheless, this has also increased the risk of unexpected information leaks when releasing data.

The sanitization process, which decreases the support values of restrictive itemsets by removing items from sensitive transactions essentially, includes four sub-problems. Identifying the set of sensitive transactions for each restrictive itemset. Selecting the partial sensitive transactions to sanitize. For each selected transaction, identifying an item from it to be deleted (called the victim item). Rewriting the modified database to disk. Employing distinct algorithms for the first or the fourth sub-problem does not affect the released database. Accordingly, users can apply an algorithm that deals with these two sub-problems as fast as possible. To solve the second sub-problem, most algorithms sort the sensitive transactions by transaction size or by the degree of conflict of the transactions.

Selecting a victim item in each sensitive transaction significantly affects the side effect. Consequently, this study also uses the simple sorting method on the second sub-problem. For the third sub-problem, this study proposes the maximum item conflict first (MICF) method to choose an

appropriate victim item for each designated sensitive transaction.

This study proposes the maximum item conflict first (MICF) algorithm to reduce the impact on the source database for preserving privacy in mining frequent itemsets. MICF employs the strategy of maximum degree of item conflict first to simultaneously decrease the support of the maximum number of restrictive itemsets. In the experimental results, MICF outperforms all other algorithms in several simulated and real datasets on misses costs for most cases. Furthermore, MICF always has the lowest sanitization rates than the other four methods.

Cheng Xiao-Hui & Gui Qiong (2009) described a privacy preserving framework privacy preserving distributed mining algorithm of association rules (PPTDM-ARBSM). Combining the advantages of both the RSA public encryption and homomorphic encryption. The PPDATA mining algorithm requires three indicators such as privacy, accuracy, and efficiency. they have simulated experiments in PPDM-ARBSM algorithm and ARSBM algorithm using approximately 50,000 sample data provided by Almaden research center of IBM in order to test the efficiency of the two algorithm.

Mohammad et al (2009) discussed the privacy breaches which occurred from certain type of association rules. The method developed in this work uses the binary transactional dataset as an input and modifies the 32 32 original dataset based on the genetic algorithms such that sensitive rules are hidden and minimum modification on the original dataset. They hide the rules which are not sensitive by lost rules and introduce the rules not supported by original database ghost rules. To minimize the undesired result, by suitable modification of original dataset.

CONCLUSION:

In this paper different pre and post mining algorithms were reviewed. A detailed survey of data mining algorithms were made it back basically lack of privacy issues and they are not suitable for the privacy preserving mining. in the privacy preservation on databases (pre mining) is reviewed the outcome of the database was trustworthy because the entire dataset is preserved for privacy issues. The privacy preserving sequential patterns and rule hiding (post mining) deals only on the privacy scope of the knowledge discovery is minimal. Finally a survey made on the balanced knowledge discovery for seeking an effective privacy preservation algorithm.

REFERENCES:

1. *Ahmed K. Elmagarmid; Panagiotis G. Ipeirotis ; Vassilios S. Verykios, 2007, "Duplicate Record Detection: A Survey", ISSN: 1041-4347, Volume 19, Issue 01, PP:1-16.*
2. *Anindya Ghose; Panagiotis G. Ipeirotis, 2011, "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics", ISSN: 1041-4347, Volume 23, Issue 10, PP: 1498-1512.*
3. *Anthony Bagnall ; Jason Lines ; Jon Hills ; Aaron Bostrom, 2015, "Time-Series Classification with COTE: The Collective of Transformation-Based Ensembles", ISSN: 1041-4347, Volume 27, Issue 9, PP: 2522-2535.*
4. *Boyu Li ; Yanheng Liu ; Xu Han ; Jindong Zhang, 2017, "Cross-Bucket Generalization*

for Information and Privacy Preservation", ISSN: 1041-4347, Volume PP, Issue 99, PP: 1-1.

5. *Ben D. Fulcher ; Nick S. Jones, 2014, "Highly Comparative Feature-Based Time-Series Classification", ISSN: 1041-4347, Volume 26, Issue 12, PP: 3026-3037.*
6. *Bin Cui ; Zhe Zhao ; Wee Hyong Tok, 2012, "A Framework for Similarity Search of Time Series Cliques with Natural Relations", ISSN: 1041-4347, Volume 24, Issue 3, PP: 385-398.*
7. *Chenping Hou ; Feiping Nie ; Hong Tao ; Dongyun Yi, 1998, "Multi-View Unsupervised Feature Selection with Adaptive Similarity and View Weight", ISSN: 1041-4347, Volume 29, Issue 9, PP: 1998-2011.*
8. *En Tzu Wang & Guanling Lee 2007, „An efficient sanitization algorithm for balancing information privacy and knowledge discovery in association patterns mining", Data and Knowledge Engineering, vol.65, pp. 463-484 doi:10.1016/j.datak.2007.12.005.*
9. *G. Adomavicius ; A. Tuzhilin, 2005, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions", ISSN: 1041-4347, Volume 17, Issue 06, PP: 734-749.*
10. *Goryczka S, Li Xiong & Fung, BCM 2014, „Mining Sequential patterns with regular expression constraints", IEEE Transactions on Knowledge and Data Engineering, vol.26, no.10, pp. 2520-2533.*
11. *Ganggao Zhu ; Carlos A. Iglesias, 2017, "Computing Semantic Similarity of Concepts in Knowledge Graphs", ISSN: 1041-4347, Volume 29, Issue 1, PP: 72-85.*
12. *Huan Liu ; Lei Yu, 2005, " Toward integrating feature selection algorithms for classification and clustering", ISSN: 1041-4347, Volume 17, Issue 4, PP: 491-502.*