International Journal of Trend in Innovative Research (IJTIIR)
www.ijtiir.com | Volume- 2 | Issue- 3

Recent Trends and Techniques in Privacy Preserving Data Mining

Ritika Jani
Asst. Prof., Computer Engineering Department, DJMIT, Mogar, Anand, Gujarat, India.

Abstract-- In this current age, computing is getting larger amount of information, analysis of that information is require. This information is having some private data. Every organization is having different style of preserving their data. Transformation of the data is reducing the utility and it's having inaccurate results and extraction of information through privacy preserving data mining. In general goal of data mining is to find the patents from the large amount of data. Those patterns may contain the important and confidential information. It is central for any data manager to protect that data from getting escaped. Data mining processes should complete such a way that it can achieve the stuff of getting useful evidence as well as the possessions of privacy protection of the data. There are several applications or we can say the uses of data mining it can be used for perceiving the rare facts or the patents like fraud detection, abnormal behavior or in data analysis as for the prediction.

Keywords-- Data mining, privacy in data, security

I. INTRODUCTION

Generally data mining is the process to evaluating the data from different kind of perceptions and summarizing that data with the useful information which can be used for different applications and for different analysis purposes. Previously this all processes have been done on the paper but as the invention of the computer and technology, we move towards the digitization of the data and storing that data is became so much easy for any of the user. By using the tools for mining the data have been became more popular in those past few years. Tools are containing mathematical equations, algorithms, machine learning methods, statistical methods etc. Also mining of data includes the association rules, classification of data, clustering of the data. Most of the organization uses the mining software's or we can say the tools for getting information. There are many changes in the existed software are being made for enhancing the quality of the extracted data [1].

From recent years people are going towards the big data and cloud computing technology, in that case there is a huge amount of data can be processed and maintain via distributed systems for various applications. Lack of the security in that database case the huge drawback for the data holding organization. There is a dramatic change in storing customers data, that is results to increasing the complexity for privacy in classification, clustering, association rule mining. For some of the organizations the data mining or the data is centralized so therefore for any of the centralized or the distributed database system we need the privacy so that it can prevent anyone for directly sharing the data [11]. Companies are releasing their data for data mining and permit unlimited access to it. Data alteration is used to get the privacy. And organizations do not release their data sets but still allow data mining tasks. Cryptographic techniques are basically used for privacy preserving [6].

Remember, data security and privacy are two different things, data security deals with the confidentiality, integrity and availability, while the data privacy deals with the appropriate use of data. The information cannot be sold to any of the other companies without getting approval from the data holders. There are many data mining tools are available now a days so any one can extract the information from anywhere. Many of the data holders put prices for the data and then use it with the available tools so there is a need to control these type of activities and protect the data. [8]. Some off the times this privacy preserving problem is seen as the interface problem, using interface controller program we can protect the data being stolen or leaked [8]. Data mining tools allows user to analyze the privacy of data. For that many of the researchers are focusing on the privacy preserving data mining systems. There are many challenges are affecting on maintaining the privacy preserving systems for data mining.

II. ISSUES IN DATA MINING

One of the common issue arise in data mining is the privacy preservation. Analyzing the buying habits of a customer, analyzing the frequent transactions made from the organization may cause the problem of the fraud [1]. Data integrity can also be the problem for the privacy preservation in data mining. A key challenge for data integrity is the conflicting nature of the data collected from the different sources. Quality of the data is one of the thing users are concern about. The quality in data mining is defining the accuracy and correctness of the data. To deal with privacy issue in data mining the sub field of data mining is introduced which is known as privacy preserving data mining. It preserves the efficacy of data, as data contains sensitive information, the main goal of the PPDM is to hide those confidential data. As the KDD process is having different kind of phases but preserving privacy is in collecting the database or in data preprocessing step. This sensitive information can only see by the authorized person, no other one will be allow getting or seeing the sensitive information. If two of the organizations are having some sensitive data and they want to exchange the data than they need some of the distributed secure environment via which they can interchange the data.

III. CHALLENGES IN PRIVACY PRESERVING DATA MINING

Privacy in data mining is protected through different methods such as data modification and secure multiparty computation (SMC) [6].There are many huddles and challenges are arriving in privacy preserving data mining. Some of them are listed below.

As we know that data mining is a widely accepted technique for vast range of organizations. Data mining is included in day to day operational activities of every organization. During the whole process of data mining we get the data. These data may contain sensitive information of individual one. This information may expose to several other entities including data collector, users and miners. Disclosure of such information is results breaking the individual privacy. We take a simple ex: exposed credit card details of a user will affect its social and economic life. Private information can also be disclosed by linking multiple databases belong to giant data warehouse and accessing web data [6].Using fake identity to create the deceptive information to access the database and get the confidential information is unauthorized [5]. Sometimes the fake user's activity might be getting different than the activity of the original users identity. Using security masks will reduce the risk of leaking the data. Also using the sock-puppets in which member of a group is creating the fake identity and represent his/her self as another person's identity and get the access to all the rights and information [5]. The scope of delicate data is not restricted to medical or financial data it may be phone calls made by an individual, buying patterns and many more [6]. No one wants that any one's personal data is traded to any other party lacking their prior consent. Some entities become uncertain to share their data which results other efforts to obtaining correct information. Public responsiveness is so much significant if personal information is shared between altered entities. Public awareness about privacy and lack of public confidence in organization may present extra complexity to data collection. Strong public anxiety may force government and law forcing organizations to present new privacy protecting regulations [6].

Currently there are several privacy preserving techniques are available. These techniques are including k-anomaly detection, distributed privacy preservation, L-dimensions etc. it is important to protect data before its get distributed in cloud or in overall system [12] Data must be protect at the storage side and it must be translated over a secure protocols using some of the security encryption algorithms. Privacy preserving in classification is define by two process step, first is the learning process step in which the classifier is built and then it's learn from the historical data, train the classifier and then the classification applied. The second one is the classifier is used for the classification for predict the data. Usually classification includes Bayesian classification, decision tree and support vector machine. Generally the privacy-preserving data mining methods are applied via transformation which decreases the practicality of the basic data when it is applied to data mining methods or algorithms [4].

IV. PRIVACY PRESERVING TECHNIQUES

Objective of privacy preserving data mining algorithms or techniques is to decrease the risk of corruption of the data. Privacy is measure in terms of confidence intervals. The randomize scheme needs to be implemented for preserving the data. The best approach to protect the identity is to applying the k-anomaly algorithm. In k-anomaly it is difficult to archive the identity of individual in the group of the data. Patten hiding method is also useful when the organization is more concern about the privacy of the data [7]. By using association rules, a one can construct the rules for sensitive data by which they can hide the data from another party. Cryptographic based techniques are securing the data atvery high level. Their solution is based on the theory that each party first encrypts its own item sets using commutative encryption, then the previously encrypted item sets of every further party. Later on, an introducing party conveys its frequency count, plus a random value, to its neighbor, which adds its frequency count and permits it on to other parties. As a final point, a secure link takes place between the final and beginning parties to control if the final result is greater than the threshold plus the random value [2].The value of the data mining results estimates the alteration in the information that is take out from the database after the privacy preservation process, on the beginning of the intended data use [2].

Data swapping is one of the techniques in which the values of individual are changed with another one's value. A new data swapping techniques is introduced for privacy preserving data mining. This technique focus on the pattern preservation instead of obtaining unbiased statistical parameters. Basically it preserves the most classification rule even if they are obtained by different classification algorithm [6]. A Secure Multi-party Computation (SMC) technique encrypts the data sets, while still allowing data mining operations. SMC techniques are not supposed to disclose any new information other than the final result of the computation to a participating party. These techniques are typically based on cryptographic protocols and are applied on distributed data sets. Parties involved in a distributed data mining encrypt their data and send to others parties. These encrypted data are used to compute the aggregate data, belonging to the joint data set, which is used for data mining purpose [6]. For publishing the privacy protection sensitive data does not publish or release with the lower accuracy. Popular anomaly detection technique includes the k-anomaly algorithm, in which any other records cannot be distinguished for other k-1 records. Data publishing methods are using k-anonymity, i-diversity, t-closeness, generalization techniques which reduces accuracy and data utility [10].

V. RANDOMIZATION IN PRIVACY PRESERVING DATA MINING

The randomization method is one of the popular methods in privacy preserving data mining. It generally adds random noise in the data so that the original data is

covered. Randomization method must be chosen before collecting the database. The value of the added noise will be so large enough so that the original sensitive data could not be recovered or seen by any unauthorized user. Statistical database is usually chosen for randomization methods. There are basically two randomization methods ne is numerical randomization and another one is item set randomization. In a particular organization presume there are many clients, each one is having some personal information, and one server, which is involved only in aggregate, statistically significant, properties of this information. The clients can protect privacy of their data by perturbing it with a randomization algorithm and then submitting the randomized version. The randomization algorithm is chosen so that aggregate belongings of the data can be recovered with sufficient precision, while individual entries are significantly distorted [3].

Sometimes it needs to be change the data from getting leaked, it is usually done by the data administrator. The goal of modify the data is known as privacy preserving data publishing. Privacy preserving generally having two categories. One is able to identify the record of another sources like linked records. Another one is having the capability of carrying the background activities for any privacy attack. It generally consists about if the record is existed in the table and the sensitive information in the table [5]. Protecting data for social aspects are difficult because sometimes it needs to be published from third party side, and the information which needs to be published is the confidential information. As the social media keeps evolving so sometimes it is require to change or publish the social network periodically.

Now a day there is a location sharing features are available in many of the devices. It can track the data from the device and sometimes store it in the form of cookies in a central system also. Providing the location based services in commercial as well as at the industrial level will collect the massive amount of the trajectory data [5]. In corporate systems, protecting only individual data is not enough, there is a need to protect entire data or we can say the information.

VI. SECURE MULTIPARTY COMPUTATION TECHNIQUE

The SMC technique is nothing but if two of the parties are wish to communicate and want to interchange their data then one of the way is to relay on the trusted third party who can gives the surety of interchanging data in a secure way [9]. If the organizations are the medical organizations than they cannot directly give the information of the raw data to other parties. If there is a no third party than how the communication will go to be done. For that one way is using the secure multiparty computation. Which suggest that there is a SMC solution for each and every polynomial function. If at the end of the computation no parties are knowing about other data rather than having information about their data than the computation is the secure computation [9]. This technique is also developed for the association rules, clustering and classification and for making decision trees [9]. Building the protocols for secure transfer of the sensitive data can be done with the use of SMC. Generally numbers of cryptographic blocks protocols used for transferring data. Homophobic encryption is one of the techniques in SMC which uses the encryption algorithms like RSA, RSA256 etc, and same as the threshold decryption which uses multiparty functionality for getting the original information from the other party [9]. The secure multiparty computation technique makes the outcome of data mining accurate without data loss. The defect of the technique is the calculation and communication overhead of protocol is very high, specifically for the large database, which hinder its application in practice [9].

CONCLUSION

Privacy preserving data mining is having a wide range of applications in many fields, in this we focus on how to protect or preserve the data in an organization. Privacy preserving methods involves many other technologies as well as the algorithms. Although there is a need to be do work in the mobile devices and data stream field for preserving the data streams for preserving the data security. As we know data mining is a very important issue in distributed privacy preserving data mining. We should try to develop more efficient algorithms and achieve a balance between computation, communication and disclosure cost.

References

1) Dileep Kumar Singh, V. S. (2013). Data Security and Privacy in Data Mining: Research Issues & Preparation. IJCTT.
2) Elisa Bertino, D. L. (n.d.). A Survey of Quantification of Privacy Preserving Data Mining Algorithms. Retrieved from http://web.mst.edu/~lindan/publication/privacy_metric.pdf.
3) Evfimievski, A. (2004). Randomization in Privacy Preserving Data Mining. ACM.
4) Hina Vaghashia, A. G. (2015). A Survey: Privacy Preservation Techniques in Data Mining . IJCA.
5) LEI XU, C. J. (2014). Information Security in Big Data:Privacy and Data Mining. IEEE.
6) Manish Sharma, A. C. (2013). A Review Study on the Privacy Preserving Data Mining Techniques and Approaches. IJCST.
7) Nivetha.P.R, T. s. (2013). A Survey on Privacy Preserving Data Mining Techniques. IJCSMC.
8) Thuraisingham, B. (n.d.). Privacy-Preserving Data Mining:. IGP.
9) Xinjing Ge, J. Z. (n.d.). Privacy Preserving Data Mining. Retrieved from http://cdn.intechopen.com/pdfs/13287/InTech-Privacy_preserving_data_mining.pdf
10) Xinjun Qi, M. Z. (2011). An Overview of Privacy Preserving Data Mining. ICESE.
11) Yong Yin, I. k. (2011). Privacy Preserving Data Mining.

12) Yousra Abdul Alsahib S. Aldeen, M. S. (2015). A comprehensive review on privacy preserving data mining. Springer.