A SURVEY ON PRIVACY PRESERVING DATA MINING

[1]GEETIKA GOPI, [2]SANTHI K

[1,2]Vellore Institute of Technology. Vellore, Tamil Nadu, India - 632014
E-mail: gopi.geetika@gmail.com

Abstract - With the rapid innovation of technologies, large amounts of data is generated day by day. Organizations and businesses usually seek new techniques and methodologies to extract useful knowledge from this massive amount of data. However, the data generated is of various types. The problem arises when data mining techniques are applied to sensitive and critical data. The major objective of privacy preserving data mining is to ensure that sensitive data is not compromised and at the same time make sure that extracted knowledge does not get misused by unauthorized users.

Index Terms - Privacy, data, knowledge, sensitive, hiding, mining

I. INTRODUCTION

Data mining is a new field that has emergeddue to the advances in technology and the rapidly increasing sizes of databases. The primary motivation behind data mining is the very notion that useful information is hidden within the vast masses of data. The extractionof this useful information can help in various fields and application for database ownersand organizations. The concept of data mining can sometimes easily be confused with data analysis. Data analysis is merely the study of a stream of data which data mining deals with the issues arising from the growing databases and externalfactors.

In today's world, the most valuable commodity a person can own is knowledge. Useable knowledge that can be extracted using data mining tools and techniques can generate much money in the commercial world. Data mining can be applied in various fields such asastronomy, banking sector, climate prediction models, molecular biology and much more.

II. KDD: A PROCESS MODEL

Knowledge discovery in databases (KDD), is the broad process of finding knowledge in data. This is done using data mining algorithms to identify knowledge that can be of any help and extracting it by applying suitable steps.

Step 1: The domain of the application is analyzed extensively, and the necessary steps are taken to understand the goals of the end user.

Step 2: Target data set is created by only selecting datasets that could be of any help inidentifying and extracting useful knowledge.

Step 3: Data is cleaned and prepossessed by removing any unnecessary noise, outliers and accounting for any missing field values.

Step 4: Transformation techniques are used to reduce the size of the data set, while only keeping data that can help with the goal of the task.

Step 5: Selecting appropriate data mingalgorithms or techniques depending on the goal and type of datasetinvolved

Step 6: Data mining algorithm(s) are applied to find any interesting data pattern and is put into some form of visual representation like clusters, trees, and graphs.

Step 7: Extracted patterns are observed and interpreted by data scientists.

Step 8: all observations and knowledge discovered through the KDD model are consolidated.

III. IMPORTANCE OF PRIVACY IN DATA MINING

The era of technology, companies, and organizations are highly dependant on data mining to extract useful knowledge that could help them in making more money, improving their services to customers and make better decisions that before. Such companies and organizations collect a massive amount of data from their customers and other resources. When data mining techniques and methods are applied to this data to extract useful knowledge, they not only extract private and sensitive data, but they also reveal them. This being inevitable, the real problem arises when this sensitive data gets in the hands of the wrong people. The possibilities are endless when sensitive data gets misused. Because of this threat, people are who own data are quite hesitant to provide their data for any data mining related processes. This however creates a significant obstacle in the field of data mining itself. Privacy preserving data mining is relatively a new and promising area of research that deals with inventing techniques and algorithms to collect, preserve, analyze and extract useful knowledge, while at the same time making sure that nokind sensitive or privateinformation is revealed during the process of it. PPDM primarily takes care of two objectives. First being, the very aim to protect the privacy of the knowledge extracted during data mining, while the other, proposing solutions that makes sure that the entireprocess of

KDD itself does not get banned due to the issue ofprivacy.

IV. PROCESS OF PRIVACY PRESERVING DATA MINING

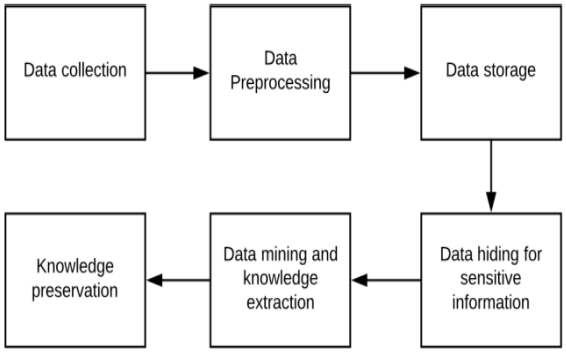

Figure 1. PPDM Process Model

From the process model for PPDM it is quite clear that PPDM techniques are applied majorly at two phases :

1. **Data Hiding:** As data is being extracted from various resources, it is very much possible that the dataset could contain some sensitive information which could be misused if it falls into wrong hands. Hence this stage is essential to protect the privacy for confidentialdata
2. **Knowledge hiding:** heextracted knowledge from target data set is highly valuable, and knowledge hiding techniques must be applied to ensure the integrity of the knowledge thusobtained.

V. DATA HIDING TECHNIQUES

a. Secure Multi Party Computation Cryptography
The ultimate goal of SMC is the carry out computations using private data from multiple sources. For example, if Participants p1, p2, p3 have private data say d1, d2, d3. Ifthe computation needed is to calculate the maximum of the data set, the function would be givenas

F = MAX (d1, d2, d3)

The model PPDM is quite similar to that of the KDD process, the only addition to it being, the integration of methods and modules that deals with the issue ofprivacy. In the first stage, data is collected from various sources relating to the specific domain. The essential preprocessing techniques are applied and later stored in a data warehouse.

To ensure that sensitive data from the preprocessed data is not misused, data hiding techniques are applied. Data mining algorithms are then applied, and useful knowledge isextracted. The extracted knowledge is then preserved further to prevent any unauthorizedaccess.

VI. STAGES WHERE PPDM IS APPLIED

A solution could be to have trustable third say X, to calculate the function; however, this is not a 100% securemethod.

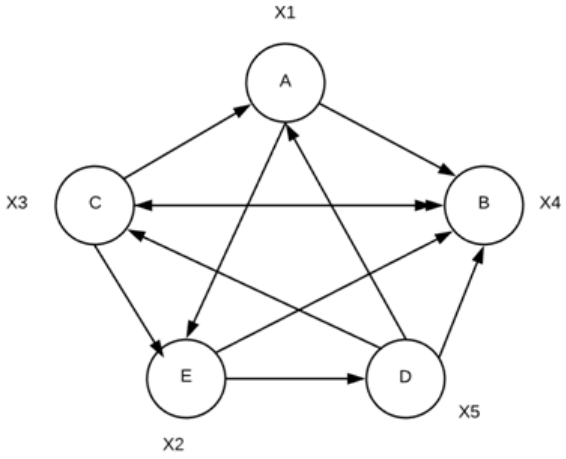

Figure 2. SMC Representation

As presented in the above diagram, at the end of the computation, the parties can onlylearn:

F(X1,X2,X3,X4,X5)

Correctness: The trusted parties taking partin the communication get to know only thefinal output.
Privacy: Only the output is leaked, and input data is keptsecure.

The goal of SMC is to design an algorithm where the participants did have to trust a third party to solve computations. Multiple algorithms have been proposed to carry out SMC cryptography.

Secure set sum and secure set unions are some of the solutions to SMC that could be used in data mining. The first is used to find the summation of local values but in a private manner. The latter is used to find the union of all private data without revealing the owner ofa particular dataset.

In the context of distributed data mining, SMCcan be used to protect the privacy of partial transactions.

b. DataCondensation

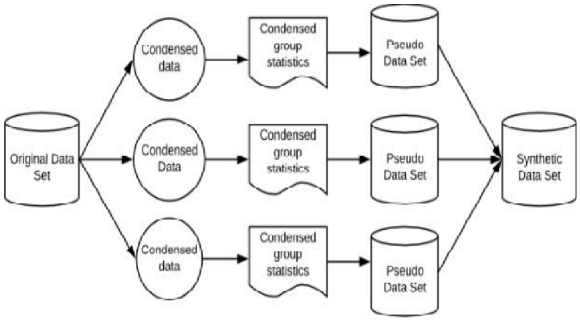

Figure 3. SMC Representation

In data condensation approach, data is splitand condensed into a set of equal-sized groups. For each group, some amount of statistical information is also maintained. From these statistics, a pseudo data set is obtained. The pseudo data set is used to generate a synthetic dataset which looks exactly like the original data set. This technique provides high data privacy since data mining algorithms will be applied to pseudo data and not the original data. However, a significant disadvantage is that since the data is condensed and only the pseudo data is used, it affects the accuracy of data mining results.

c. Anonymization

Anonymization aims at making a particular data indistinguishable amongst a group of data. Generalizationand suppression techniques are used to achievethis.

In generalization, Every single data record has a set of attributes or features that distinguish it from the other records. Some of them are:

1. **Explicit identifier:** It is an attribute that directly connects to the data record. It is usually removed before data miningtechniques areapplied.
2. **Pseudo Identifier:** It is an attribute thatdoes not uniquely identify the record, but when combined with some additional information or attributes it can directly identify the datarecord.
3. **Sensitive attribute:** This attribute usually talks about the sensitive information of a person.

Anonymization in PPDM is used to hide any identification and sensitive information. This implies that explicit identifiers need to be removed. However, this is not enough since attackers can combine pseudo identifiers to identify a data record and carry out attacks uniquely. Hence the ultimate goal ofanonymization is to make sure that any identifying information is removed before data mining techniques areapplied.

K anonymization is a conventional technique used in data mining. In this method, if every row in a data table can be uniquely identified butlooking at the attributes of at least k-1 rows, then the data table isk-anonymized

d. DataPerturbation

This method is typically used to modify data before any data mining techniques are applied. Various methods are available to do so. However, Noise adding (randomization)anddataswapping techniques are used in datamining. In the noise addition techniques, a random number(noise) is added to data. This random a number is selected from a group that has 0 mean and minimal standard deviation. This is done so that the nature of the original data set is preserved. This new data set is afterapplying noise is then sent for datamining.

In the method data swapping, the original data set is kept the same, except similar attribute values are swapped. This way data re-identification becomes difficult. This technique is found to be quite advantageous since original data is preserved while only thepositions of attribute values are interchanged.

VII. KNOWLEDGE HIDING TECHNIQUES

a. Association rule hiding

In association rule hiding, all the sensitive association rules are removed. This does not affect the non-sensitive rules or the results of data mining. The very notion behind thismethod is that it is the sensitive rules that create a security breach and not the data itself.

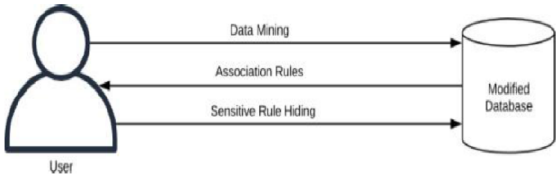

Figure 4. Association Rule Hiding Model

According to tooth's method, association rule hiding algorithms parse the transactions in the database and calculate the support and confidence value for each rule. The user provides a minimum threshold value for both support and confidence. All transaction rulesthat have a value less than the minimum threshold value are considered as sensitive rules and are removed. Many algorithms and methods are available In order to reduce the support and confidence values of sensitive rules. The two main methods used in data miningare data distortion and datablocking.

Data distortion: In this method, the attribute values of data records are replaced with new values. For example, in a binary database, an item can be removed, during which the valueis changed from 1 to 0. When an item is added,thevalueischangedfrom0to1.Thisway the support and confidence valuescan be altered. A significant disadvantage of this technique is that at times it leads to artificial or missing associationrules.

Data Blocking: There are situations whenadding or deleting attribute values could be quite dangerous. For example, medical records, bank records, etc. In this case instead of data distortion, data blocking is carried out. In datablocking, instead of deleted and inserting, attribute values and replaced with an unknown symbol such as '?'. This is a better method to hide sensitive association rules.

a. Query Auditing

Query auditing is the process of examining the past queries that have been answered to determine if the answers could have helped any user to determine confidential information that is against thepolicies.

Offline Auditing: Done by offline auditors. They examine the queries and the answers to determine if any sensitive information has been disclosed.
Online Auditing: it is used to detect any disclosure of sensitive information and also prevent such disclosures in the future based on the previously answeredqueries.

VIII. CHALLENGES

Because of the broadness of the term, characterizing privacy is very testing. Indeed, even in the limited scope of information privacy, a few definitions have been exhibited. There is dependably a decent measure of abstraction because of people's privacy idea, convictions and hazard affirmations. It is, in this manner, essential to create systems that execute the idea of personalized privacy. This thought enables users to have a dimension of control over the explicitness of their information.
Notwithstanding, personalized privacy is trying to execute. One hand, it has been demonstrated that Users' worries about privacy do not reflect clients' activities, that is, users' will in general exchange their privacy for utility. Accordingly, a personalized privacy arrangement could give users authority over the data, yet that control can wind up unsafe, exceptionally when clients are ignorant of the Privacy dangers of information divulgence. Then again, the way that users have no access to the general dispersion of sensitive values can prompt progressively defensive choices over their data, hence adversely influencing the utility of data. Looking for novel answers for this outstanding exchange off among privacy and utility is subsequently required with regards to personalized privacy arrangements.
Information publishing privacy is accomplished with privacy models that clear information. In any case, because of the entrance to other openly accessible sources, foes can attempt to de-anonymize or to gather sensitive information. As the measure of Published information keeps on developing in both amount and multifaceted nature, demonstrating foundation learning of foes shows a few challenges, for example, the recognizable proof of what Data can be utilized to de-anonymize and the measure of Public Data sources that can be connected.

CONCLUSION

Data mining is a vital and critical fieldthat helps millions of companies, and individuals to increase their profits, improve their service to customers, as well as help in making better decisions. However, privacy and security are a significant threatto data mining especially in an era where knowledge is power and money. PPDM has tackled this problem of data mining withvarious techniques and methodologies, eachof them advantageous in theirway.

REFERENCES

[1] Hand, D. J. (2006). DataMining.
[2] Encyclopedia of Environmetrics,
[3] Raj, R., & Kulkarni, V. (2015). A Study on Privacy Preserving Data Mining: Techniques, ChallengesandFutureProspects. International Journal of Innovative Research in Computer and Communication Engineering, 3(11).
[4] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., &Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. ACM Sigmod Record, 33(1), 50-57.
[5] Qi, X., &Zong, M. (2012). An overview of privacy preserving data mining. Procedia Environmental Sciences, 12,1341-1347.
[6] Parmar, K., & Shah, V. (2016). A review on data anonymization in privacy preserving data mining. Int J Adv Res ComputdCommunEng, 5(2).
[7] Agrawal, R., &Srikant, R. (2000, May). Privacy-preserving data mining. In ACM Sigmod Record (Vol. 29, No. 2, pp. 439-450). ACM
[8] Lindell, Y., &Pinkas, B. (2000, August). Privacy preserving data mining. In Annual International Cryptology Conference (pp. 36-54). Springer, Berlin,Heidelberg.
[9] Agrawal, D., & Aggarwal, C. C. (2001, May). On the design and quantification of privacy preserving data mining algorithms. In Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 247-255).ACM.
[10] Lindell, Y. (2005).Secure multiparty computation for privacypreserving data mining. In Encyclopedia of Data Warehousing and Mining (pp. 1005-1009). IGIGlobal.
[11] Pinkas, B. (2002) . Cryptographic techniques for privacy-preserving data mining. ACM Sigkdd Explorations Newsletter, 4(2),12-19.
[12] Aggarwal, C. C., & Philip, S. Y. (2004, March). A condensation approach to privacy preserving data mining. In International Conference on Extending Database Technology (pp. 183-199). Springer, Berlin, Heidelberg.
[13] Bertino, E., Fovino, I. N., &Provenza, L. P. (2005). A framework for evaluating privacy preservingdata mining algorithms.Data
[14] Mining and Knowledge Discovery, 11(2), 121-154
[15] Kenthapadi, K., Mironov, I., &Thakurta,A.
[16] G. (2019, January). Privacy-preserving Data Mining in Industry. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (pp. 840-841). ACM.
[17] Li, L., Lu, R., Choo, K. K. R., Datta, A., &Shao, J.(2016). Privacy-preserving- outsourced association rule miningon vertically partitioned databases. IEEE Transactions on Information Forensics and Se

★★★