**Privacy in Data Mining: A Review**

|  |  |
| --- | --- |
| **Sharmistha Dutta** | **Ankit Kumar Gupta** |
| Deptt of Computer Science and | Deptt of Computer Science and |
| Engineering, ASET | Engineering, ASET |
| Amity University, Noida | Amity University, Noida |
| Uttar Pradesh, India  Email ID: sharmisthadutta333@gmail.com | Uttar Pradesh, India  Email ID: ankit.gpt112@gmail.com |
|  |  |  |

***Abstract: In todays’ technology cloud computing and big data is being explored to quite a great extent which results in extraction of data i.e. data mining. During the process of data mining maintaining security of data is a must and this is where privacy in data mining comes into play. This paper mainly focuses on the framework and overview of privacy in data mining and also discusses its various uses and applications as a foundation for various fields of research.***

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

1. INTRODUCTION

Data mining is a process which uses different data analysis tools to investigate the relationships and patterns which can be used for creating assumptions. Recently, there has been an upsurge in preserving the privacy of data mining [1, 2]. Privacy of data mining [18] mainly refers to protection of useful data in cloud and providing data security against disclosure. Here in this paper, we will deal with the privacy concepts and outline its features.

While hearing the word, ‘privacy’, one thing that comes to everybody’s mind is to *‘keep my information unavailable to others’.* It is essential as it prevents the data from getting misused by an intruder. Also the gap between confidentiality [19] of data and data mining can be filled by privacy preserving data mining [20]

(PPDM). PPDM also takes into consideration the *corporate privacy* which is preventing the release of information from a collection of data rather than individualistic data. It is useful in places where big data is stored like in business centres, corporates and institutions.

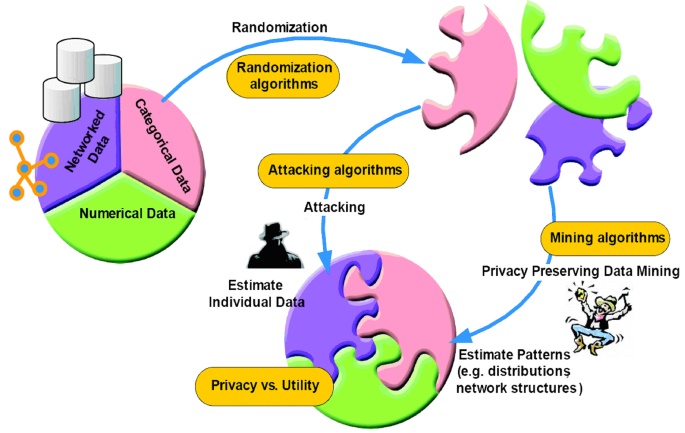
In the next sections, we go into the depth of privacy in data mining techniques, its pros and cons and other characteristics.

1. APPROACHES TO PRIVACY IN DATA MINING

The algorithms based on privacy preserving data mining has no proven guarantees of privacy and henceforth, technically weak. By its nature, PPDM claims to protect the privacy of all types of datasets from malicious attacks or from getting disclosed. Let us now discuss the latest approaches to data mining privacy.

1. **Randomization:** For instance, let us consider that there is a central server and customers each having a small piece of information. The server collects the information and performs data mining to construct an aggregate data model [16]. As the name suggests, this approach then protects the customers’ data by shuffling the information of the respective customer or swapping the attributes of the records and introducing some noise before sending it to the server. Noise can be inserted by adding or multiplying random values to the numeric attributes [3] or by removing some real items and instead adding falsified items to the records. [4] After receiving the information, the statistical data is employed to recover the aggregates required for data mining.

Given the amount of randomization and right choice of perturbing the information, it is possible to protect individual and corporate values and estimating the aggregate model [16] with relatively high accuracy.



The advantage of this technique is that one does not need to know the entire concepts of data mining but only knowing the randomization operations and desired algorithms. The results generated are an approximate and also high amount of accuracy requires a large amount of randomized data which proves it to be a demerit.

1. **Cryptography:** In this method, the clients exchange their information using cryptographic protocol [16], which means that the parties exchange encrypted messages which makes the data in it secured and computationally intractable only to the concerned parties. The cryptographic technique was first adapted by Lindell and Pinkas [5] for the decision tree construction problem over horizontally partitioned data followed by certain assumptions which include permitted exposure and constraints on input data. One of the essential results of this method is that all functions are efficiently computable in a secured, privacy preserving manner. It is well recognized in application fields like electronic auctions.

The main disadvantage of cryptographic approach is even though all functions can be calculated securely; the resulting protocols are not that efficient. This method is also difficult to scale when few parties are involved. Thus, scalability turns out to be an interruption in cryptographic approach. The other issue which arises is that it does not report the question of whether the exposure of the final information or data-mining result may break the privacy of individual records.

1. **Suppression:** As the name suggests, in this technique all sensitive data are suppressed before any exposure. Here, we can suppress the data attributes in particular records as verbalized by the privacy policy, in a given database. Apart from suppressing the sensitive data attributes, we can also suppress the sensitive values of individual records (like the identity of a person).

The process of changing the data set to restrict identity relation is called as “de-identification”. One popular technique of this process is *k-anonymity* formulated by Samarati and Sweeney [6]. The aim of k-anonymity [6] is to roughly release data where for all likely queries, at least k results will be returned. In suppression techniques, the data points which results in too much generalization may be removed or a column which contains detailed information is eliminated. Suppression can also be used to protect certain statistical features like sensitive association rules while decreasing the misrepresentation of other data mining results.

This approach is applicable for individual data and cannot be used where data mining involves full admission to sensitive data.

1. **Summarization:** It is one of the approaches to data mining privacy where the data is released in the form of a summary of that permits evaluation of particular classes of cumulative queries while preserving the individual records. In a way, summarization is an extension of randomization (as discussed in (i)) but is relatively shorter in length as compared to the actual data set.

Various types of summarization techniques include sampling, tabular representation [7,8] of data and clustering. In the former one, the data set is replaced with a set of smaller samples of its records combined with perturbation and suppression of values in order to avoid re-identification of the data set thus, preserving the privacy in data mining. Tabular representation of data encapsulates data into aggregate quantities like averages, counts or sums over some range of attributes keeping other attributes fixed. Clustering [15] is used in decision tree construction to analyse data net-flow for the detection of cyber-attacks. Presently, further research work is in progress to combine summarization technique with cryptographic approach.

1. **Secure Multiparty Computation:** It is also a type of PPDM where the concerned parties study only the results and not how it is acquired. SMC allows delivery of data without using any third party or an intermediate. A considerable amount of communication may occur between the sender and the receiver but none of them get to know anything about the communication but only receive the final output. For better understanding, let us consider an example. Suppose a decision tree is built using the SMC technique from two sites-firstly the people having diseases with high risk and secondly with lower risks. Assuming that the non-sensitive data is open and the sensitive data cannot be exposed, this technique will not reveal the sensitive data but the decision tree classifier will enable to estimate the sensitive data value. Here the SMC is not broken but the data itself cannot be exposed.

Perfect privacy in SMC ensures that there is no publicizing of meaningful information to the intermediary. Therefore, it gives a strong sense of privacy and has been used in PPDM related works [5, 9].

1. DEMERITS ON GENERATED RESULTS

Various researches are being worked upon to reduce the cons of PPDM like concealing specific association rules [10] or limiting self-reliance in the process of data mining [11]. The research work is mainly conducted to improve the algorithm efficiency of the above stated methods by providing up-front restraints on the interested result. Henceforth, we suggest some recommendations which have the possibility of reducing the limitations on the results, when implemented.

1. RECOMMENDATIONS

We recommend the subsequent steps to increase the potential of data mining privacy:

1. Develop methods of privacy that gives a profitable usability
2. Develop a proper definition of privacy that is not only related to individual data but also includes preservation of corporate data as well.
3. Use big data to authenticate new algorithms on data mining.
4. Develop mitigation schemes that can give privacy enhancement and also preserve online data transactions simultaneously which are in trend nowadays like Paytm, Freecharge.
5. Maintain a balance between data analysis and PPDM by considering the flow of information in data analysis.
6. Support interdisciplinary researches that will lead to an advancement of computer science and other disciplines.
7. Develop new methodologies for data cleaning and data access.
8. Promote graph mining algorithms with multiple nodes and edges for complex and dynamic networks [14].
9. Define privacy breaches in graphs
10. TYPES OF PRIVACY

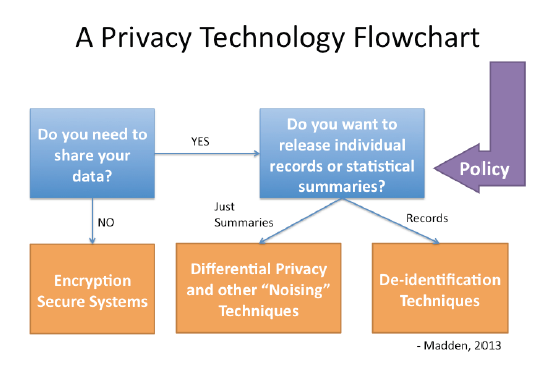
*Privacy,* in a general sense can be broadly categorized into two major types, i.e., Individual Privacy and Corporate Privacy.

1. **Individual Privacy:** In common terms, we usually consider privacy as protecting personal or individual data but what does it actually mean? According to the European Community, personal data [12] has been defined as:

**“** personal data shall mean any information relating to an identified or identifiable natural person (‘data subject’); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity. ”

One method to ensure individual privacy is to obscure data where we make the private data public but add enough noise so that the actual data cannot be determined. We can then reconstruct the original distribution of actual values by constructing decision trees [1,2] enabling the data collected at the web survey to be obscured from the source. Thus, the real values will never leave the respondent’s system.

1. Corporate privacy- In the corporate world, the protection of data is not confined to individual data alone but protecting data from learning against a collection. Privacy of individual data eventually leads to the preservation of corporate data. Concerned participants may want to hide their information from unknown sources but their maybe a probability that the trust can be broken. Thus, exercising secrecy of data will help in expanding the applicability and growth of data mining.



1. VARIOUS APPLICATIONS
2. **Medical Domain:** Medical data are considered as one of the sensitive type of private data. PPDM is essential for research works in pharmaceutical, environmental and medical lines. For instance, to study the effects of certain gene X on a drug Y, the medical histories and DNA sequences are stored at different repositories of data for privacy concerns. So, PPDM over vertically partitioned data can be utilized to calculate aggregate counts keeping the records preserved.
3. **Conduction of Surveys and Data Collections:** Companies do collect personal information of their clients to conduct surveys for business planning or for recommendation of products. In such cases data is preserved using randomization approach.
4. **Social Networking:** Nowadays, everything seems to be available on the Web through social media whether it is searching for a job, finding new connections, knowing our whereabouts or be it matchmaking. In all such scenarios, personal data are given and thus have the tendency to get publicized. Thus, data mining privacy techniques are utilized for allowance of limited disclosure preferences.
5. **Monitoring for Emergencies:** Terrorist attacks, outbreak of a disease, natural disasters can be detected before harming the entire population through data mining as it is quite hard to look into individual cases.
6. **Tracing a Product:** Before reaching to the customer, a product goes through a series of steps that includes manufacturing, packaging, transportation, storage and sale. Radio Frequency Identification Tag (RFID) [13] will be used in upcoming days so that the product gets registered automatically at every step thus creating a vast network which can be mined or extracted for detecting market trends and business patterns and wrongdoings like counterfeiting and theft.
7. FUTURE PROSPECTS

The main challenge of privacy in data mining is to increase the efficiency of its algorithms and to achieve higher security and accuracy rates ensuring privacy guarantees. The existing techniques are not that flexible enough to fully utilize the approaches of PPDM.

One of the newer techniques is RFID which has been discussed in earlier sections that can reduce criminal activities to a greater extent and also to plan business strategies. The other is the development of “secure coprocessors” which consists of a computer in a sealed box and is able to withstand hardware and software attacks and gets destroyed when the data in it is getting tampered.

PPDM in health fields also needs to be further developed like working on the mining of semi-structured or ontology based data like annotated images.

1. CONCLUSION

Privacy preserving has the potential to flourish and reach the merits of data mining technology. In this paper, we have reached to a conclusion that privacy in data mining has two essential uses: ensuring the privacy rights of individuals and in personal and corporate world and analysis of data in order to provide better amenities. As each day a new set of big data is stored, be it a trillion searches on Google or the comments posted on social sites like Twitter, Facebook or some other applications, data privacy in cloud is a must which henceforth increases the usage of privacy in data mining. Also by knowing the disadvantages of PPDM techniques, one can lead to creation of newer techniques, hybrid approach of the older techniques or improving the previous ones.

REFERENCES

[1] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 247–255, Santa Barbara, California, USA, May 21-23 2001.ACM.

[2] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD Conference on Management of Data*, pages 439–450, Dallas, TX, May 14-19 2000. ACM.

[3] Agrawal, R., & Srikant, R. (2000).*Privacy preserving data mining.* Proceedings of ACM SIGMOD Confer­ence on Management of Data (SIGMOD’00), Dallas, TX.

[4] Evfimievski, A., Srikant, R., Agrawal, R., & Gehrke, J. (2002).Privacy preserving mining of association rules. *Proceedings of 8*th *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’02)* (pp. 217-228).

[5] Lindell, Y., & Pinkas, B. (2000).Privacy preserving data mining. In *Lecture notes in computer science: Vol. 1880. Proceedings of Advances in Cryptology: Crypto’00* (pp. 20-24)*.* Springer-Verlag.

[6] Samarati, P., & Sweeney, L. (1998).*Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.*Proceedings of the IEEE Symposium on Research in Security and Privacy, Oakland, CA.

[7] Adam, N. R., & Wortmann, J. C. (1989). Security-con­trol methods for statistical databases: A comparative study.*ACM Computing Surveys, 21*(4), 515-556.

[8] Willenborg, L., & de Waal, T. (2001). Elements of sta­tistical disclosure control.In *Lecture notes in statistics* (Vol. 155)*.*Springer-Verlag.

[9] W. Du and M. J. Atallah. Privacy-preserving cooperative scientific computations. In *14th IEEE Computer Security Foundations* *Workshop*, pages 273–282, Nova Scotia, Canada, June 11-13 2001.

[10] L. Sweeney. Computational Disclosure Control: A Primer on Data Privacy Protection. PhD thesis, Massachusetts Institute of Technology, 2001.

[11] A. C. Yao. How to generate and exchange secrets. In Proceedings of the 27th IEEE Symposium on Foundations of Computer Science, pages 162–167. IEEE, 1986.

[12] Directive 95/46/ec of the european parliament and of the council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European* *Communities*, No I.(281):31–50, Oct. 24 1995.

[13] Garfinkel, S., & Rosenberg, B. (2005). *RFID: Ap­plications, security, and privacy.* Addison-Wesley Professional.

[14] Kokkinos, Y., Margaritis, K., 2013, Distributed privacy-preserving P2P data mining via probabilistic neural network committee machines, Fourth International Conference on Information, Intelligence, Systems and Applications (IISA), 2013, pp: 1-4

[15] Tsiafoulis, S.G. Zorkadis, V.C., 2010, A Neural Network Clustering Based Algorithm for Privacy Preserving Data Mining, International Conference on Computational Intelligence and Security (CIS), 2010, pp: 401-405

[16] Shu Qin Ren , Khin Mi Mi Aung ; Jong Sou Park, 2010, A Privacy Enhanced Data Aggregation Model, Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on, pp: 985 – 990

[17] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita, 2014, A Review on Privacy Preserving Data Mining: Techniques and Research Challenges, International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, pp: 2310- 2315.

[18] Xueyun Li, Zheng Yan and Peng Zhang, 2014, A Review on Privacy-Preserving Data Mining, IEEE International Conference on Computer and Information Technology (CIT), 769 – 774.

[19] A. Machanavajjhala and J.P. Reiter, “Big Privacy: Protecting Confidentiality in Big Data,” ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

[20] M. B. Malik, M. A. Ghazi and R. Ali, “Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects”, in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.