IJASRM
International Journal of Advanced Scientific Research and Management, Vol. 1 Issue 5, May 2016.
www.ijasrm.com
ISSN 2455-6378

A Survey on Security Techniques in Data Mining

Nidhi Saxena [1], Dr. Priya Gupta [2] and Onkar Singh [3]

[1] Computer Science Department, Maharaja Agrasen College, University of Delhi, New Delhi, India
[2] Asst. Professor , Maharaja Agrasen College, University of Delhi, New Delhi, India
[3] Asst. Professor, Shaheed Sukhdev College of Business Studies, University of Delhi, New Delhi,India

Abstract
Data mining is a powerful means of extracting useful information from data. With the increase of ease of availability of digital data, the potential for misuse of raw as well as mined data grows.. A fundamental challenge is to develop privacy and security methods appropriate for data mining. This is the reason due to which privacy preservation in data mining has gained momentum in recent times. But still no PPDM algorithm outperforms the other. This paper discusses about privacy preservance in data mining in both centralised and distributed systems.

Keywords: *Privacy preserving data mining, cryptography, distributed data mining, DAG Privacy preserving data mining, cryptography, distributed data mining, DAG*

1. Introduction

Privacy preservation in data mining has emerged as new discipline in the world of increasingly massive datasets. Automated data collection devices (credit cards, mobile phones, etc) and services (emails, online financial transactions, etc) used in business and sciences are capable of generating terabytes of data per hour and thus render existing inference methods obsolete. Data mining technology is born out of these requirements of handling avalanche of data automatically to obtain insightful pattern that might be useful to the organisation doing this process or person whose experiments generate such data. A pattern might be a simple data summary, a data segmentation, or a model of dependencies within the data. Data mining as a knowledge discovery process is intended to lead all way from raw data to 'documented knowledge' [10].

Data mining is now widely used in many areas including science, business, politics, nuclear and astrophysics. So it becomes very crucial to maintain the safety of data. The raw data as well as extracted information can be highly susceptible to various attacks. The problem does not lie in data mining but in the way data mining is done. To overcome this problem, the topic of privacy preserving data mining or PPDM is emerged. The aim of PPDM algorithms is to reduce the risk of malicious use of data maintaining the efficiency of results. Thus privacy preservation must be an important aspect of data mining.

2. Concept of Privacy-Preserving Data Mining

Privacy preserving data mining or PPDM is technique of data mining done, keeping in mind the security of data. PPDM has become an important topic for research in data mining. Privacy techniques must not be applied to just one step in data mining, rather it should be kept in mind while performing every step in data mining [6]. Different stages of data mining are shown in figure 1.

In the process of data mining the information from one or more sources is collected, selectively extracted and organised on data warehouses (very large databases), and finally analysed by different algorithms for patterns or useful information. Data input to pre-processing is the stage 1 in which raw data (raw data here refers to the data directly taken from the user which may include confidential information) is collected and some processing is done on it to make it

IJASRM International Journal of Advanced Scientific Research and Management, Vol. 1 Issue 5, May 2016.
www.ijasrm.com
ISSN 2455-6378

useable. Security in the process of data mining must begin from this stage itself. Security at stage 1 means having secure databases to prevent any unauthorised access to raw data.

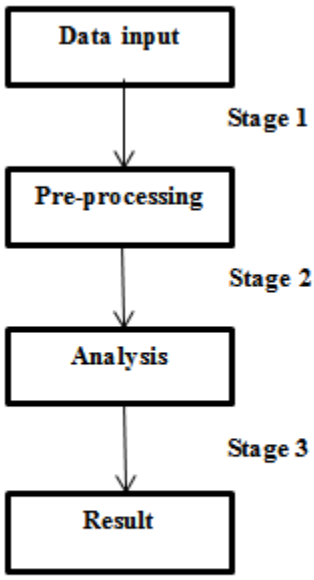

Figure 1. Process of data mining

At stage 2, the data is processed in warehouse to make it suitable for analysis. Data cleaning is done which includes removing redundancy and unusable data. Security at stage 2 is done by sanitizing data after which it can be revealed even to untrustworthy miners. Many methods can be applied at this stage like modelling, modification, sampling, blocking, etc. Then data mining algorithms are applied for analysing or predicting the results.

At the 3rd stage, the results obtained by applying data mining algorithms are checked for their sensitiveness towards disclosure risks. Thus privacy preserving techniques must be applied at any stage according to requirement of application. The majority of existing algorithms can be classified into following broad categories: (i) techniques that prevent unprocessed data in data mining process, and (ii) techniques that protect the data mining results [6].

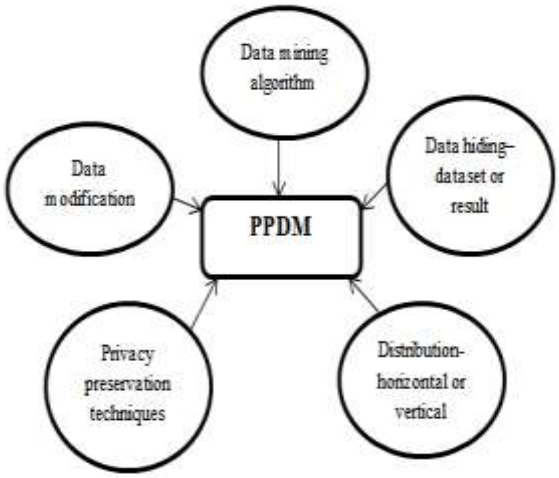

Figure 2. Dimensions of Privacy-preserving Data Mining

3. Preserving Datasets

When data resides in database it is still prone to attacks [3], as it consists of users or organisation's information (may be confidential) which they may not like to share. Therefore this information is protected by applying some techniques such as perturbation, sampling, transformations, etc. on the data collected to change it in a way that it can be safely disclosed to untrusted parties. In this case main purpose is to make sure that any particular record cannot be identified uniquely.

As these techniques are manipulating datasets some inaccuracies in the results can be expected. The aim of such techniques is to improve privacy as well as maintaining the validity and accuracy of results obtained by data mining algorithms.

Secure Multiparty Computation (SMC), a fundamental field in cryptography, allow multiple parties to collectively mine their data without revealing their inputs (datasets) to each other. SMC is used successfully in applications. Though SMC can be an ad-hoc, a much simpler and robust model, i.e. the DAG model [11] is proposed by Sin G. Teo, Jianneng Cao, Vincent C.S. Lee, which uses simple mathematical operators +, - *, /, max and their compositions.

4. Distributed Data Mining

Distributed data mining refers to the process where data is collected from multiple sources. This creates partitioning in data. There are two types of partitions [2]:
A. Horizontal Partitioning: Different sources have different records containing same attributes.

B. Vertical Partitioning: Different sources may have different attributes of the same set of records.

In distributed data mining cryptography is generally used to achieve privacy [1, 8]. Both types of data partitions have algorithms for security. Murat Kantarcioglu, in his paper [5], discusses and analyses some privacy preservation techniques on horizontally partitioned data. Jaideep Vadiya, in his paper performed a detailed survey on privacy preservation techniques on vertically partitioned data.

Some algorithms for cryptography can be applied to data irrespective of its partition.] N. Abitha, G. Sarada, G. Manikandan, Sairam .N, in their paper [1] presented modified approaches of Rail-fence algorithm and Vigenere cipher algorithm. According to them, modified Vigenere Cipher algorithm had more complex procedures than modified Rail fence algorithm. On the contrary the encryption of records with the modified Vigenere cipher produces highly efficient results than the modified Rail fence algorithm. Hence depending on the purpose, execution resource in hand, deadline and quantity of data sets the user can select the preferred algorithm.

In case the data is stored on different clouds different techniques are used for privacy preservation. One such approach is presented by Maria Luisa Merani, Cettina Barcellonay, Ilenia Tinnirelloy, in their paper [7] which derives and analyses two methods to secure multi-cloud data.

On clouds generally data is in encrypted form so data mining algorithms cannot be applied. Researches have been done to make data mining algorithms work on encrypted data on cloud. One such work is done by Bharath K. Samanthula, Yousef Elmehdwi, and Wei Jiang [9]. They described k-NN (a data mining algorithm) to extract the data from relational database when it is encrypted.

5. Conclusion

Privacy preservation in data mining is a relatively new field. The continuous growth of digital data results in the need for privacy of that data. Not enough work has been done in this field. This are needs to be grown at a extremely high level as security is very crucial.

6. Future Scope

This is an emerging field and requires a lot of research work to follow. More algorithms are needed to analyse encrypted data and store user's information more securely. Also more work is required to reduce the inaccuracy when transformed data is mined or analysed. In a nutshell, this area is growing and is open to more researches.

References

[1] Abitha N, Sarada G, Manikandan G., Sairam, A Cryptographic Approach for Achieving Privacy in Data Mining. International Conference on Circuit, Power and Computing Technologies, IEEE, (2015)

[2] Aggarwal C, Yu P, An Introduction to Privacy-Preserving Data Mining, Chapter 2 in Privacy Presrving Data Mining: Models and Algorithms, Springer, NY, USA, Pg-11 to Pg-27, (2012)

[3] IMPERVA - top ten security threats, (2015).

[4] Janbandhu, Chaware, Survey on Privacy Preservation in Data Mining, Sachin Janbandhu et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5, (Issue – 4) , (2014).

[5] Kantarcioglu, A Survey of Privacy-Preserving Methods Across Horizontally Partitioned Data, , Chapter 13 in Privacy Presrving Data Mining: Models and Algorithms, Springer, NY, USA, Pg-313 to Pg-332, (2012).

[6] Malik M.B., Ghazi M.A, Ali R, Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects, 3rd International Conference on Computer and Communication Technology, (2012)

[7] Merani M., Barcellonay C, Tinnirelloy I, Multi-Cloud Privacy Preserving Schemes for Linear Data Mining, IEEE, Communication and Information Systems Security, (2015).

[8] Pinkas B., Cryptographic techniques for privacy-preserving data mining, SIGKDD Exploration, Vol.4 (Issue - 2), Pg-12 to Pg-19.

[9] Samanthula, Elmehdwi, Jiang, k-Nearest Neighbour Classification over Semantically Secure Encrypted Relational Data, IEEE Transactions on Knowledge and Data Engineering, (2013).

[10] Soman K.P., Diwakar S, and Ajay V. Insight into data Mining Theory and Practice, 1 Edn, PHI Learning Pvt Ltd, (2014).

[11] Teo S, Cao J., and Lee, DAG: A Model for Privacy Preserving Computation, IEEE, International Conference on Web Services, (2015).

[12] Vadiya, A Survey of Privacy-Preserving Methods Across Vertically Partitioned Data, , Chapter 14 in Privacy Presrving Data Mining: Models and Algorithms, Springer, NY, USA, Pg-337 to Pg-356, (2012).