Journal of Statistics & Management Systems

Taylor & Francis
Taylor & Francis Group

Journal of Statistics and Management Systems

ISSN: 0972-0510 (Print) 2169-0014 (Online) Journal homepage: http://www.tandfonline.com/loi/tsms20

An illustration to secured way of data mining using privacy preserving data mining

Richa Purohit & Deepshikha Bhargava

To cite this article: Richa Purohit & Deepshikha Bhargava (2017) An illustration to secured way of data mining using privacy preserving data mining, Journal of Statistics and Management Systems, 20:4, 637-645, DOI: 10.1080/09720510.2017.1395183

To link to this article: https://doi.org/10.1080/09720510.2017.1395183

Published online: 16 Nov 2017.

Journal of Statistics & Management Systems
Vol. 20 (2017), No. 4, pp. 637–645
DOI : 10.1080/09720510.2017.1395183

Taylor & Francis
Taylor & Francis Group

An illustration to secured way of data mining using privacy preserving data mining

Richa Purohit *
MIT Arts, Commerce & Science College
Pune 412105
India

Deepshikha Bhargava [†]
Amity University
Jaipur 303002
Rajasthan
India

Abstract

The aim of data mining is to extract useful information from huge source of multiple data. But during the process of data mining, intentionally or unintentionally the data becomes visible and thus vulnerable while handling. Privacy Preserving is a new concept in the area of data mining taking the security issues of users' data being mined as prime concern. It ensures that privacy of sensitive data will be preserved even after mining by multiple parties. There are various existing methods for privacy preserving data mining which are based on data distortion, clustering, intersection, data distribution etc. The paper discusses few of such privacy preserving data mining techniques. At the end of the paper, a simple mathematical approach for this, is also discussed which is applicable for a number of sites, sharing data on distributed database environment.

***Keywords**: Association Rules, Clustering, Data Distortion, Data Mining, Decision Tree, Privacy Preservation.*

***Mathematics Subject Classification 2010:** 68M99*

1. Introduction

Data mining is the process of finding relevant information and discovering patterns from a huge data store and process it to make it

**E-mail:* `richapurohit81@gmail.com`
[†]*E-mail:* `deepshikhabhargava@gmail.com`

useful for the intended users. It may be considered as a computation method of obtaining required limited information from a large data set and then to transform it into an understandable format so that interested users can use it as per their requirement. The data store or warehouse, from where data mining is being performed, can contain data of multiple users. The data can be very crucial as it may contain personal information like social security number, PAN, bank account number, bank customer ID, license number, Personal Identification Number, confidential data of perspective clients of an organization, critical patient data etc. and user may not wish to disclose them to other users while process of mining is being carried out. Also maliciously the users of data store can handle this crucial and private data of other users for their own benefit. In today's scenario where organization has its offices at multiple cities all over the world, this distributed data travels among all these various sites and thus becomes quite insecure when it is transmitted over internet because many intruders and attackers are interested in obtaining private information that is being transmitted over media. Thus, it is always desirable to maintain the privacy of data while transmitting as well as while mining a huge database of other owners.

2. Privacy Preserving Data Mining

Whenever data is disclosed to a user other than owner, there arises a risk of losing privacy and confidentiality of those data parts that were not even being fetched by users. And in such case, this data can be misused. Various online activities like online banking, information transfer, and research for contents, E-Commerce and like are exposing humans and their private data to many intruders, resulting in huge financial loss, intellectual loss and emotional suffering. This has increased the demand of secured data mining and then to make available only the part that owner wishes to see by others. Thus Privacy preservation has become a requisite with an increase of volume of available online data. Privacy Preserving Data Mining is a method and concept of data mining which ensures data will not be disclosed to unconstitutional users while fetching vital data from data warehouse. Thus it lets enjoy the features of data mining that too without compromising privacy aspect. There are various PPDM techniques that modify data during the process of mining such that private information remains private even after mining.

The overall aim of all Privacy Preserving Data Mining Techniques is to design efficient Frameworks and suitable algorithms that can preserve the privacy of confidential data and keep them safe from deliberately or inadvertently exploitation. It includes three main aspects- Association rules, Classification and Clustering [1]. At present There are many techniques for data mining process in existence such as clustering, associative classification, association rule, distortion, taxonomy tree, hide association rule, distortion, distributes, k-autonomy, out-sourced data mining, randomization, condensation, L-diverse etc. Each technique has its own pros and cons.

3. Study of Various PPDM Methods

Basically PPDM techniques protect data by masking the sensitive information at the time of mining and transmitting. These approaches hide the sensitive data from unauthorized users, check the authenticity of user before receiving data, and prevent the data loss while mining. One approach of this could be deleting crucial information before it is sent over the cloud based architectures. But deletion alone cannot solve this issue. At one end we may reach to a point with no transmission at all because of deletion of every information considered as crucial or at another end, we may negotiate with security aspect of information treating as non-crucial from owner’s aspect but being taken as quite informative from intruder’s perspective.

In the following section, few Privacy Preserving Data Mining techniques are discussed:

3.1 *PPDM Based on Data Distortion*

These techniques involve mechanism of making some changes into data before sharing it with others to keep it secure. In this, some or all individual data values are distorted before data mining application to imply privacy protection. Various privacy parameters [12] are used for depicting the extent up to which protection has been achieved. Many researchers proposed their mechanisms under this. In 1985, a method for data distortion using probability distribution was proposed by Liew et. al. [13]. He devised three steps in his algorithm: (a) Underlying density function identification (b) use this density function to generate a distorted series and (c) map this output of previous step on to input series. In 2010, Peng et al. [14] combined four different methods to form a more secured PPDM strategy using data distortion. He used (a) partitioning of attribute

(b) decompose the single value (c) factorize the non-negative matrix and (d) Discrete wavelet transformation. In 2012, Kamakshi [2] proposed that on the basis of threshold limit of sensitivity of attributes, they can be classified dynamically. Further without changing basic properties of data, it is modified. This mechanism takes huge time for execution. Zang et. al [3] proposed a new scheme using data distortion named Association Probability based Noise Generation Algorithm. This is comparatively an expensive mechanism because it involves associative probability on noise implications on data.

3.2 *PPDM Based on Association Rules*

This is an improved version of distortion technique, proposed by Shrivastava et. al [4] is based on association rules. Two significant factors for this are support and confidence (also called strength). An association implication rule takes the form $X \Rightarrow Y$ ($X \cap Y = 0$) assuming s% of the transactions in T contains XUY and c% of transactions that contain X also contain Y. Here c is known as confidence or support s. If we consider association rule of the form $X => Y$, the percentage of total transactions in $X \cup Y$ is the support. The ratio of the total transaction numbers and X gives us the confidence or strength of this rule. The algorithm works on two parameters- fp and nfp and it gives the best results when number of frequent items is very less in total available population of items. Association rule hiding [15] is another process in which original database is modified such that certain sensitive association rules disappear without effecting whole database or remaining non-sensitive association rules.

Heuristic Approach, Border-Based Approach and Exact approach are three main approaches for association rule hiding algorithms [16]. In heuristic approach, the transactions set from a database are carefully cleansed to hide sensitive information. In border based approach, the original border in the lattice of various intermittent patterns of dataset, are modified in order to hide sensitive rules. The exact approach works as non-heuristic method and hides sensitive association rules in the manner similar to constraint satisfaction problem in linear programming. In 2011 Jain et al. [5] proposed a new algorithm to reduce the support of LHS or RHS to secure or hide association rule. This algorithm came out to be faster and more secure than the previous one.

3.3 *PPDM Based on Clustering*

Clustering is a mechanism of placing data on the basis of values of attributes, into various groups, known as clusters. Yi and Zhang presented k-means clustering in 2013 for privacy preserving data mining [6]. This technique can be used where data has been partitioned or fragmented vertically. In this, each site involved in distribution of vertical fragments encrypt k values (each value is taken from the point of site and distance from coordinator site) using a common public key for every cluster. The final output after all encryptions is sent instead of any intermediate value. Another mechanism for clustering is- EM algorithm [7]. It can work for both attributes- continuous and discrete. The privacy preserving model of EM algorithm works well for data that is horizontally partitioned or fragmented. In this mechanism, we assume that there are nl data items at each site. At each step cluster membership for the ith cluster is multiplied with a secret value yj and is added with previous steps sum. After the last round only final value is shared with all, without revealing intermediate results of sum or yj of any site. It prevents unwanted data sharing across sites. In 2011, Kumar et.al. [17] [21] used fuzzy based algorithms for PPDM in particular to data classification, as fuzzy based data transformation provides accurate results in classification problems. It keeps privacy preserving also because transformed data sets cannot reconstruct the original data set.

3.4 *PPDM Based on Distribution*

Existing PPDM techniques that are based on distribution, can be grouped into three categories [8], as- (1) Secure multi-party communication (2) Restricted Query and (3) Perturbation. This can be applied by data owner before they publish their data. For this they can perform additive Perturbation or matrix multiplication perturbation [18]. In Additive Perturbation, a sufficiently large noise is added to the data values so that they cannot be recovered. In Matrix multiplicative Perturbation, The original data A is replaced with C = AB, where B is a $n' \times n$ matrix, such that for any column a1, a2 in A, their corresponding columns $c1$, $c2$ in C satisfy $a1 - a2 = c1 - c2$.

Considering the security issue, while mining data on cloud, Chow et al. [9] presented one more approach that involves data classification, disintegration and distribution. In this approach, the data is split into various groups and distributed over various suitable clouds available. Although it provides security from mining based attacks yet it degrades the performance when whole data is required by a user at a very high frequency, as it is obtained from multiple clouds.

3.5 *PPDM Based on Intersection*

This method is used when we need to obtain size of common items from the set of items of each party [10]. All parties generate a public key for encryption. Every party encrypts its own data set with its public key and then forwards it to other parties. Next party, after receiving the encrypted part, first encrypts its own part of data along with previous input and then permutes the order of output before forwarding it to next party in the link. At the end, any party can find out total number of values present in the encrypted set without finding the exact part of any of the party due to encryption involved.

3.6 *PPDM Based on Decision Tree*

This mechanism works for multiparty computation for security purpose and also reduces total bits transferred between communicating parties [11]. A Decision tree algorithm is applied on a join database and does not reveal unnecessary information during join. ID3 is a well established Decision tree based algorithm. To make a decision tree the best attribute is selected at each level and data is partitioned at that level. This process is repeated till there is no further attribute to be selected for partitioning. Attribute is selected by information gain theory which maximizes the information gain and minimizes partition entropy. Information gain is calculated on whole database set rather than data at individual site. ID3 works on very less communication cost. Quinlan proposed a C4.5 [19] decision tree classifier which is an extension of ID3. C4.5 is among the best algorithms for handling various numeric continuous attributes. It finds the best numeric attribute and best splitting point of numeric continuous attributes. Later in 2009, Li Liu presented modified C4.5 [20] which builds the decision tree from perturbed data. He also considered noise element while considering splitting point of the attribute.

4. Proposed Method for Privacy Preserving Data Mining

The paper proposes a mathematical mechanism for privacy preservation in data mining process from distributed databases. We assume that n parties are involved in execution of a common transaction.

The parties may compute a data value in a secured manner in following way:

Assume that there are n sites involved in a communication in a distributed environment. Also assume, they share a common public key *K*. Each site has something to share with others.

Let x = sum of all xi,

where xi is the data that is to be shared by a site with remaining n-1sites.

Assume a site A is designated as initialization site and it possesses a secret key denoted as Ka.

Site A will compute the value S using modular arithmetic as::

$$S = ((Ka + xA) \bmod N)/K$$

Here, N is the range of xi

Site A will send this calculated value Y to next site B. Similarly, the site B will compute next value in the same manner. Thus, every other site (except site A) receives:

$$S = ((Ka + \text{sum of xi of other site}) \bmod N)/K$$

Or,
$$S = \left(\left(Ka + \sum_{j=1}^{n} x_j\right) \bmod N\right) / \mathrm{K}$$

As site A knows its secret key Ka, it can easily computer sum S. This method can be used as a simple tool for securing data accessing and transferring during mining across distributed sites.

5. Conclusion

Privacy preserving data mining is considered a need of today's scenario where multiple parties are interested in mining data from various data warehouses. If privacy is not preserved, the data can be misused. The paper discussed various mechanisms for privacy preserving data mining such as data distortion, association rules, hide association rules, clustering, distribution etc. Further a simple mathematical approach is also described that can specifically be used for distributed databases.

References

[1] Y. A.S. Aldeen, M. Salleh and M. A. Razzaque. "A Comprehensive Review on Privacy Preserving Data Mining". Springer Plus 2015 4:694. 2015. DOI: 10.1186/s40064-015-1481-x

[2] P. Kamakshi, A. Vinaya Babu. “Automatic detection of sensitive attribute in PPDM”. Published in: 2012 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) , pp. 37-41, 2012. DOI: 10.1109/ICCIC.2012.6510183

[3] G. Zhang, Y. Yang, J. Chen. “A Historical Probability based Noise Generation Strategy for Privacy Protection in Cloud Computing”. *Journal of Computer and System Sciences*. Vol 78(5), pp- 1374-1381, September 2012.

[4] V. K. Shrivastava, P. Kumar and K. R. Pardasani. “Extraction of Interesting Association Rules using GA Optimization”. *Global Journal of Computer Science and Technology*. Vol 10(5), pp. 81-85, 2010.

[5] Y.K. Jain, V.K.Yadav and G.S.Panday. “An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining”. *International Journal on Computer Science and Engineering.* Vol. 3(7), pp- 2792-2798, 2011.

[6] Huh, Myung-Hoe, and Yong B. Lim. “Weighting variables in K-means clustering.” *Journal of Applied Statistics* 36.1 (2009): 67-78.

[7] Stanley R.M. Oliveira and Osmar R Zaiane. “Towards Standardization in Privacy Preserving Data Mining”. In Proceedings of 3rd Woorkshop on Data Mining Standards (DM-SSP 2004), in conjuction with KDD 2004.

[8] D. Thakur and H. Gupta. “An Exemplary Study of Privacy Preserving Association Rule Mining Techniques”. I*nternational Journal of Advanced Research in Computer Science and Software Engineering*. Vol 3(11), 2013.

[9] R. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, and J. Molina. “Controlling data in the cloud: Outsourcing computation without outsourcing control.” In Proceedings of the 2009 ACM workshop on Cloud computing security. pp.85–90, 2009.

[10] Xu, Shuting, and Mable Qiu. “A privacy preserved data mining framework for customer relationship management.” *Journal of Relationship Marketing* 7.3 (2008): 309-322.

[11] Y. Lindell and B. Pinkasy. “Privacy Preserving Data Mining”. In Proceedings CRYPTO ‘00 Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology. pp. 36-54.

[12] S. Xu, J. Zhang, D. Han, J. Wang. “Data Distortion for Privacy Protection in a Terrorist Analysis System”. Proceeding of IEEE

International Conference on Intelligence and Security Informatics, pp. 459-464, 2005.

[13] C. K. Liew, U.J. Choi and C. J. Liew. "A Data Distortion by Probability Distribution". ACM Transaction on Database Systems (TODS), Vol. 10(3), pp. 395-411, 1985.

[14] B. Peng, X. Geng, J. Zhang. "Combined Data Distortion Strategies for Privacy-Preserving Data Mining". Proceeding of IEEE International Conference on Advanced Computer Theory and Engineering (ICACTE), pp. VI-572- VI-576, 2010.

[15] U. Sahu and A. Singh. "Approaches for Privacy-Preserving Data Mining by various Associations Rule Hiding Algorithms- A Survey". *International Journal of Computer Applications*, Vol 134(11), pp- 21-26, 2016.

[16] S. Narmadha, S. Vijayarani. "Protecting Sensitive Association rules in Privacy-Preserving Data Mining using Genetic Algorithms". *International Journal of Computer Applications*, Vol 33(7), pp- 37-43, 2011.

[17] P. Kumar, K.I. Verma, A. Sureka. "Fuzzy Based Clustering Algorithm for Privacy-Preserving Data Mining". *International Journal of Business Information Systems,* Vol. 7(1), pp. 27-40, 2011.

[18] X. Ge and J. Zhu. "Privacy-Preserving Data Mining, New Fundamental Technologies in Data Mining." Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-547-1, InTech, DOI: 10.5772/13364. Available at hhtp://www.intechopen.com/books/new-fundamental-technologies-in-data-mining/privacy-preserving-data-mining

[19] J.R.Quinlan. "C4.5: Programs for Machine Learning". Morgan Kaufmann, 1993.

[20] Li Liu. "Privacy Preserving Decision Tree Mining from Perturbed Data". Proceedings of the 42nd Hawaii Conference on System Sciences-2009.

[21] Kumar, B. Suresh, Deepshikha Bhargava, and Ramesh C. Poonia, "Fuzzy keyword search and ranking frame work of DRS based file information management system using TF-RDF ranking strategy", Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies, ACM; 2014.