International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 4 Issue 8, August 2017

ISSN: 2394-4404

A Survey On Privacy Preserving Data Mining Techniques

Yogyata Jain
Ph.D Research Scholar in CSE

Dr. Dinesh Kumar
Associate Professor in Computer Department

Abstract: ***Huge volume of detailed personal data is regularly collected and sharing of these data is proved to be beneficial for data mining application. Such data include shopping habits, criminal records, medical history, credit records etc. The development and penetration of data mining within different fields and disciplines, security and privacy concerns have emerged. Data mining technology which reveals patterns in large databases could compromise the information that an individual or an organization regards as private. The aim of privacy-preserving data mining is to find the right balance between maximizing analysis results that are useful for the common good and keeping the inferences that disclose private information about organizations or individuals at a minimum.***

Keywords: Data mining, privacy, perturbation, blocking, k-anatomization

IJIRAS

I. INTRODUCTION

Data mining deals with large database which can contain sensitive information. It requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. Advancement of efficient data mining technique has increased the disclosure risks of sensitive data. A common way for this to occur is through data aggregation. Data aggregation is when the data are accrued, possibly from various sources, and put together so that they can be analyzed. This is not data mining per se, but a result of the preparation of data before and for the purposes of the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when originally the data were anonymous.

What data mining causes is social and ethical problem by revealing the data which should require privacy? Providing security to sensitive data against unauthorized access has been a long term goal for the database security research community and for the government statistical agencies. Hence, the security issue has become, recently, a much more important area of research in data mining [1] and [2].

Data mining promises to discover unknown information. If the data is personal or corporate data, data mining offers the potential to reveal what others regard as private. This is more apparent as Internet technology gives the opportunity for data users to share or obtain data about individuals or corporations. In some cases, it may be of mutual benefit for two corporations (usually competitors) to share their data for an analysis task. However, they would like to ensure their own data remains private. In other words, there is a need to protect private knowledge during a data mining process. This problem is called Privacy Preserving Data Mining (PPDM).

II. PRIVACY PRESERVING DATA MINING TECHNIQUES

A Distributed Data Mining (DDM) model assumes that the data sources are distributed across multiple sites. The challenge here is: how can we mine the data across the distributed sources securely or without either party disclosing its data to the others? Most of the algorithms developed in this field do not take privacy into account because the focus is on efficiency. A simple approach to mining private data over multiple sources is to run existing data mining tools at each site independently and combine the results.

Data mining deals with large database which can contain sensitive information. It requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. Advancement of efficient data mining technique has increased the disclosure

risks of sensitive data. A common way for this to occur is through data aggregation.

Data aggregation is when the data are accrued, possibly from various sources, and put together so that they can be analyzed. This is not data mining privacy, but a result of the preparation of data before and for the purposes of the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when originally the data were anonymous [2] and [3-5]. Association rule mining: Association rules are based on the popularity and confidence. Transactions are distributed across the sources. Non-categorical attributes and quantitative association rule mining are significantly more complex problem. [4] Random Rotation: Random Rotation is the approach used for data perturbation. This method works only for the vertically partitioned datasets, but not for the horizontally partitioned datasets.

This method does not develop distributed algorithms to preserve privacy[3]. Data Perturbation: Data Perturbation does not correspond to real-world record owners. The attacker cannot perform the sensitive linkages or recover sensitive information from the published data. Records released is synthetic i.e. it does not correspond to real world entities represented by the original data. The individual records in the perturbed data are meaningless to the human recipient as only statistical properties of the records are preserved. The perturbation method does not reconstruct the original values[2] ID3 Decision Tree Algorithm: Classification based ID3 Decision Tree algorithm does not accept nominal attributes and missing values. [1]

A. CLASSIFICATION OF PPDM

PPDM can be classified according to different categories. These are

Data Distribution- The PPDM algorithms can be first divided into two major categories, centralized and distributed data, based on the distribution of data. In a centralized database environment, data are all stored in a single database; while, in a distributed database environment, data are stored in different databases. Distributed data scenarios can be further classified into horizontal and vertical data distributions as shown in figure 1. Horizontal distributions refer to the cases where different records of the same data attributes are resided in different places.

While in a vertical data distribution, different attributes of the same record of data are resided in different places. Earlier research has been predominately focused on dealing with privacy preservation in a centralized database. The difficulties of applying PPDM algorithms to a distributed database can be attributed to: first, the data owners have privacy concerns so they may not willing to release their own data for others; second, even if they are willing to share data, the communication cost between the sites is too expensive [4] and [8].

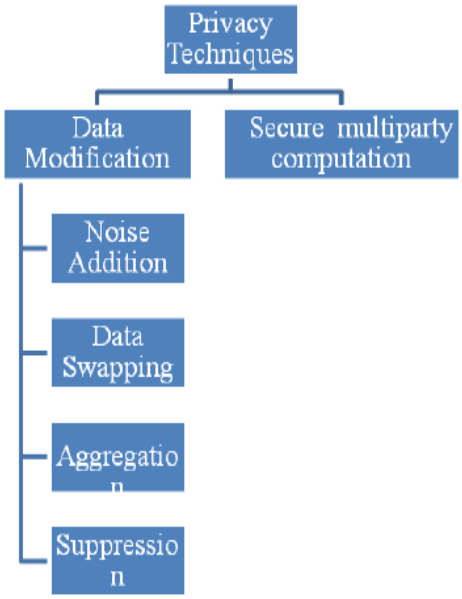

Figure 1: classification of privacy data mining

Hiding Purposes - The PPDM algorithms can be further classified into two types, data hiding and rule hiding, according to the purposes of hiding. Data hiding refers to the cases where the sensitive data from original database like identity, name, and address that can be linked, directly or indirectly, to an individual person are hided. In contrast, in rule hiding, the sensitive knowledge (rule) derived from original database after applying data mining algorithms is removed. Majority of the PPDM algorithms used data hiding techniques. Most PPDM algorithms hide sensitive patterns by modifying data [2] and [9].

B. DATA MINING TASKS / ALGORITHMS

Currently, the PPDM algorithms are mainly used on the tasks of classification, association rule and clustering. Association analysis involves the discovery of associated rules, showing attribute value and conditions that occur frequently in a given set of data. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Clustering Analysis concerns the problem of decomposing or partitioning a data set (usually multivariate) into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups.

C. PRIVACY PRESERVATION TECHNIQUES

PPDM algorithms can further be divided according to privacy preservation techniques used. Four techniques – sanitation, blocking, distort, and generalization -- have been used to hide data items for a centralized data distribution. The idea behind data sanitation is to remove or modify items in a database to reduce the support of some frequently used item sets such that sensitive patterns cannot be mined. The blocking approach replaces certain attributes of the data with a question mark. In this regard, the minimum support and confidence

IJIRAS

level will be altered into a minimum interval. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges, the confidentiality of data is expected to be protected. Also known as data perturbation or data randomization, data distort protects privacy for individual data records through modification of its original data, in which the original distribution of the data is reconstructed from the randomized data. These techniques aim to design distortion methods after which the true value of any individual record is difficult to ascertain, but "global" properties of the data remain largely unchanged. Generalization transforms and replaces each record value with a corresponding generalized value[7-9].

III. EXAMPLES OF DATA MINING ALGORITHMS

Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms. This is the natural trade-off between information loss and privacy. Some examples of such technique as:

- ✓ Randomization method: The randomization technique uses data distortion methods in order to create private representations of the records .In this which noise is added to the data in order to mask the attribute values of records In most cases, the individual records cannot be recovered, but only aggregate distributions can be recovered. These aggregate distributions can be used for data mining purposes. Data mining techniques can be developed in order to work with these aggregate distributions. Two kinds of perturbation are possible with the randomization method:
  - Additive Perturbation: In this case, randomized noise is added to the data records. The overall data distributions can be recovered from the randomized records. Data mining and management algorithms re designed to work with these data distributions.
  - Multiplicative Perturbation: In this case, the random projection or random rotation techniques are used in order to perturb the records [7].
- ✓ The k-anonymity model and l-diversity: The k-anonymity model was developed because of the possibility of indirect identification of records from public databases. This is because combinations of record attributes can be used to exactly identify individual records. In the k-anonymity method, the granularity of data representation is reduced with the use of techniques such as generalization and suppression. This granularity is reduced sufficiently that any given record maps onto at least k other records in the data. The l-diversity model was designed to handle some weaknesses in the kanonymity model since protecting identities to the level of k-individuals is not the same as protecting the corresponding sensitive values, especially when there is homogeneity of sensitive values within a group [3-6].
- ✓ Perturbation Technique: This technique used for the privacy preservation in which data are perturbed But this cannot reconstruct the original data and also not good for the large data.
- ✓ Condensation Technique: Instead of perturbed data, it works on the pseudo data. So it provide better privacy preservation than the techniques which use simply data modification on original data. But it does not give longer effect on data mining. Because it has the same format as the original data.
- ✓ Cryptographic Technique: It performs encryption of the sensitive data. There is also proper toolset for algorithm in the field of the data mining But this technique is difficult to scale when more parties are involved and also not good for large database.
- ✓ Blocking Based Technique: In this technique to provide privacy to the individual it replaces the unknown values to the sensitive transaction. Reconstruction of the original data is quite difficult.
- ✓ Combine strategy for the privacy preservation: In the combine strategy multiple strategies are used to obtain any privacy preserving methodology. In the given figure combine strategy is shown based on the data transformation and data encryption techniques. Due to the combining this various techniques robust security can be obtained. Sometime privacy preserving techniques may have some disadvantages or some limitations but that can be overcome in combine strategy. So security result would be more effective of combine strategy than a single privacy preserving techniques used [8]. In the given figure-2 original data are transformed after that transformed data are encrypted. So here data transformation and data encryption methods are used in this combine strategy.

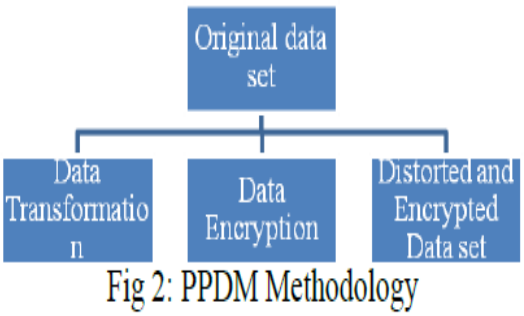

Fig 2: PPDM Methodology

Figure 2

IV. CONCLUSION

Privacy in data mining has evolved as a research field of great interest and need. Lots of research is being done in order to preserve the privacy of the individuals without sacrificing the quality of data and without adding to the complexity in the data mining process. But most of them suffer from information loss or computational complexity. In this paper we have demonstrated how to preserve privacy during data analysis and not at the cost of information loss.

REFERENCES

[1] Aldeen (2016), "A technique of data privacy perervation in deploying third party mining tools over the cloud using SVD and LSA", in Research Journal of Applied sciences, volume 11, issue-2, pp. 27-34.
[2] Agrawal Arpit (2013), "Security based Efficient Privacy Preserving Data Mining", in International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 2, Issue 7, ISSN 2319 – 4847, pp. 225-233.
[3] G. Vikas et.al (2015), "Dataless Data Mining: Association Rules-based Distributed Privacy-preserving Data Mining", 12th International Conference on Information Technology - New Generations in International Journal of IEEE computer society, pp.-615-620.
[4] Gokulnath et.al (2015), "Preservation of Privacy in Data Mining by using PCA Based Perturbation Technique", in IEEE based International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), pp.202-206.
[5] Hariharan et.al. (2016), "Enhancing Privacy Preservation in Data Mining using Cluster based Greedy Method in Hierarchical Approach", in Indian Journal of Science and Technology, Vol 9, issue-3, pp. 1-8.
[6] Jaideep Vaidya, Chris Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining in 2002
[7] Majid Bashir Malik, M. Asger Ghazi and Rashid Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", 2012 Third International Conference on Computer and Communication Technology
[8] Malik et.al (2015), "A model for Privacy Preserving in Data Mining using Soft Computing Techniques", 2nd International Conference on Computing for Sustainable Global Development (INDIACom) IEEE, pp.-181-186.
[9] M. Saravanan, A. M. Thoufeeq, S. Akshaya & V.L. Jayasre Manchari, "Exploring New Privacy Approaches in a ScalablenClassification Framework", Data Science and Advanced Analytics (DSAA), 2014 International Conference
[10] Patel et.al (2015), "Quasi & Sensitive Attribute Based Perturbation Technique for Privacy Preservation", in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 11, ISSN: 2277 128X, pp.450-455.
[11] Rajaei and Haghjoo (2015), "An improved Ambiguity Anonymization technique with enhanced data utility", in IEEE based International Conference on Information and knowledge Technology, pp. 136-142.
[12] Shah and Gulati (2016), "Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey" in International Journal of Computer Applications, Volume 137 – No.12, pp. 40-46.
[13] Vaghashia and Ganatra (2015), "A Survey: Privacy Preservation Techniques in Data Mining", in International Journal of Computer Applications (0975 – 8887) Volume 119 – No.4,pp. 20-26.
[14] Zhenmin Lin, Jie Wang, Lian Liu, Jun Zhang, "Generalized Random Rotation Perturbation for Vertically Partitioned Data Sets", Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium.
[15] Zuber et.al (2012), "An Empirical Study on Privacy Preserving Data Mining", in International Journal of Engineering Trends and Technology- Volume3, Issue6, pp. 687-693.

IJIRAS