Privacy Preserving Data Mining

Divya Rana

University School of Information and Communication Technology, GGSIPU

Abstract: ***There is a tremendous increase in the research of data mining. Data mining is the process of extraction of data from large database. Knowledge Discovery in database (KDD) is another name of data mining. Privacy protection has become a necessary requirement in many data mining applications due to emerging privacy legislation and regulations. One of the most important topics in research community is Privacy Preserving Data Mining (PPDM). Privacy preserving data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. The Success of Privacy Preserving data mining algorithms is measured in terms of its performance, data utility, level of uncertainty or resistance to data mining algorithms etc. In this paper we will review on various privacy preserving techniques like Data perturbation, condensation etc.***

Keywords: Data Mining, Privacy Preserving, KDD

1. Introduction

Our society has substantially enhanced the potential to spawn and gather data from diverse sources [1]. The enormous amount of data is needed in our every aspect of lives. There is an urgent need for tools and techniques for transforming this vast amount of data into useful information and knowledge. This has led to the generation of a promising and flourishing end called Data Mining [2]. Data Mining is also referred to as Knowledge Discovery from Data (KDD). Knowledge Discovery is extraction of patterns and relationships not readily known to exist. Data Mining [3] is the process of discovering insightful, interesting and novel patterns as well as descriptive, understandable, and predictive models from large scale data. Data mining is the intelligent search for new knowledge in existing masses of data. Protecting private data is an important concern among users while accessing and sharing data but, at the same time, serious concerns have grown over individual privacy in data collection, processing and mining. Preserving privacy [4] is a serious issue in data mining. As data mining ascertain to efficiently locate valuable and refined information from large databases, is particularly vulnerable to misuse. The knowledge present in the data is extracted for use, the individual's privacy is protected and the data holder is protected against the misuse of the data. The goal of privacy preserving in data mining is to protect data from being misused. The goal of this paper is to review the privacy preserving techniques which are helpful in preserving security.

According to [5] data mining is categorized into 5 tasks:

- Exploratory data analysis (EDA). Typically interactive and visual, EDA techniques simply explore the data without any preconceived idea of what to look for.
- Descriptive modeling. A descriptive model should completely describe the data (or the process generating it); examples include models for the data's overall probability distribution (density estimation), partitions of the dimensional space into groups (cluster analysis and segmentation), and descriptions of the relationship between variables (dependency modeling).
- Predictive modeling: classification and regression. The goal here is to build a model that can predict the value of a single variable based on the values of the other variables. In classification, the variable being predicted is categorical, whereas in regression, it's quantitative.
- Discovering patterns and rules. Instead of building models, we can also look for patterns or rules. Association rules aim to find frequent associations among items or features, whereas outlier analysis or detection focuses on finding "outlying" records that differ significantly from the majority.
- Retrieval by content. Given a pattern, we try to find similar patterns from the data set.

2. PPDM Framework

In data mining, the raw material [6] is transactional data and data mining algorithm serves as filter which filters out valuable nuggets of information from huge amount of data. Data is collected from single or various organizations and stored at respective databases. For analytical purposes, the data is transformed into suitable format and then modified data is stored into the data warehouse and various data mining procedures/algorithms are applied for the generation of useful information. Privacy cannot be applied at one step, but needs to be applied at all levels. At level 1, raw data is gathered from diverse sources and is transformed into suitable appearance for systematic purposes and stored into data warehouse. Privacy techniques are applied at this stage also while collecting data.

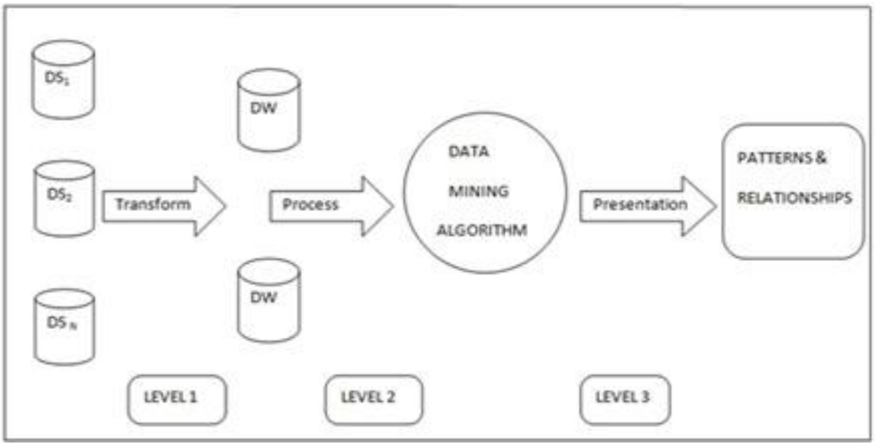

Figure 1: A Framework for PPDM [6]

At level two, in data warehouse, used for reporting and data analysis. They store current and historical data and are used for creating reports. Data from data warehouse is subjected to get through a number of processes. These processes are blocking, suppression, perturbation, modification, generalization, sampling etc. For the discovery of knowledge/information, data mining algorithms are applied to processed data. Even the data mining algorithms are modified for the purpose of protecting privacy without sacrificing the goals of data mining. At level three, the knowledge revealed by data mining algorithms are checked for its sensitiveness towards disclosure risks. Privacy algorithms are applied at all three levels.

3. PPDM Techniques

PPDM has become an important issue in data mining research [7]. A set of new approaches are provided for mining of data and at the same time without allowing the privacy of data to be violated [8]. Approaches can be classified into two main broad categories [9]:

- Procedures that save fragile information itself in the mining process
- Procedures that save fragile information mining results

The first category refers to techniques that apply perturbation, sampling, modification, generalization etc to original datasets in order to generate their correlative that can be revealed to un-trusted parties. The second category refers to techniques that proscribe the disclosure of delicate data which is derived through the use of data mining algorithms. PPDM techniques can be classified into [10]:

(1) Data distribution
(2) Data modification
(3) Data mining algorithms
(4) Data or rule hiding
(5) Privacy preservation

The first dimension refers to distribution of data which can be either Centralized or Distributed. Distributed data can be further classified into Horizontal or Vertical distribution. Horizontal distribution refers to cases where different records reside in different places whereas vertical distribution refers to cases where all values of different attributes reside in different places. The second dimension refers to the data modification. In data modification, original values are modified in the database and the altered values are released in public.

Methods of modification include:

- Perturbation, which is accomplished by the alteration of an attribute value by a new value (i.e., changing a 1-value to a 0-value, or adding noise),
- blocking, which is the replacement of an existing attribute value with a "?",
- Aggregation or merging which is the combination of several values into a coarser category,
- *S*wapping that refers to interchanging values of individual records, and
- Sampling, which refers to releasing data for only a sample of a population?

The third dimension refers to the data mining algorithms which are applied on transformed data to get patterns and relationships which were not readily known to exist. The fourth dimension refers to whether the raw or aggregated data should be hidden. The final dimension refers to the techniques that are used for protecting privacy.

Review of the Privacy Preserving Data Mining Techniques

Based on different dimensions the PPDM techniques are classified into five categories [11]:

(1) Randomized Response based PPDM
(2) Perturbation based PPDM
(3) Anonymization based PPDM
(4) Condensation approach based PPDM
(5) Cryptography based PPDM

We elaborate these in more detail in following subsections.

1. Randomized Response based PPDM

In Randomized Response, the data is scutter such that original source cannot tell with the probabilities better than a pre-defined threshold, whether the data from user contains truthful information or false information. The information received from individual user is scrambled and if the numbers of users are significantly large, the result information of these users can be evaluated with good amount of accuracy. Randomized Response based PPDM is used in decision tree classification. Data collection in randomized process is a two-step process [11].During first step; the data providers randomize their data and transmit the randomized data to the data receiver. In second step, the data receiver reconstructs the original distribution of the data by employing a distribution reconstruction algorithm. Randomization is very easy process and does not require the knowledge of location of records in data. It does not require the server to keep the original records in order to perform anonymization process [12]. It has disadvantage that it treats its all records equal irrespective of their local den

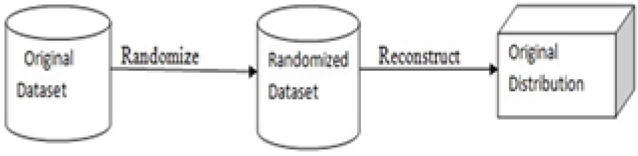

Figure 2: Randomization response model [6]

2. Perturbation based PPDM

In Perturbation [13] the original values are replaced with some duplicate data values so that the statistical information computed from the perturbed data does not differ from the statistical information computed from the original data to a larger extent. The perturbed data records do not correspond to real-world record owners, so the attacker cannot violate the privacy of derived data or recover sensitive information from the modified data. In perturbation approach, records released is synthetic i.e. it does not correspond to real world entities represented by the original data. Therefore the individual records in the perturbed data are meaningless to the human recipient as only statistical properties of the records are preserved. Perturbation can be done by using additive noise or data swapping or synthetic data generation [14]. Since the perturbation method does not reconstruct the original values but only the distributions, new algorithms are to be developed for mining of data. In perturbation approach, data mining algorithms treats each dimension independently.

3. Anonymization based PPDM

Anonymization refers to hiding the sensitive information or identity of record owners. Explicit identifiers, are removed in this approach, set of attributes containing information that identifies a record owner explicitly such as name, SS number etc.Explicit identifiers should be removed but still there is a danger of privacy intrusion [15] when quasi identifiers, set of attributes that could Potentially identify a record owner when combined with publicly available data, are linked to publicly available data. Such attacks are called as linking attacks. For example attributes such as DOB, Sex, Race, and Zip are available in public records such as voter list. Such records are available in medical records also, when linked, can be used to infer the identity of the corresponding individual with high probability. A value is replaced with less semantic consistent value called generalization and in suppression values are blocked. When such data is combined with publically available data, mining reduces risk of identification. Although the anonymization method ensures that the transformed data is true but suffers heavy information loss.

4. Condensation approach

Another approach used is Condensation approach [16]. It builds constrained clusters in the data set and after that produces pseudo-data. The basic concept of the method is to contract or condense the data into multiple groups of are maintained. This approach is used in dynamic data update such as stream problems. Each group has a size of at least ‘k’, which is referred to as the level of that privacy-preserving approach. The higher the level, the high is the amount of privacy. They use the statistics from each group in order to generate the corresponding pseudo-data. This is a simple privacy preservation approach but it is not efficient because it leads to loss of the information.

5. Cryptographic based PPDM

In cryptographic technique private data can be encrypted safely. This technique is used where two or more than two parties are involved in sharing their sensitive data but at the same time they are concerned for preserving privacy[17].Cryptographic techniques are ideally meant for such scenarios where multiple parties collaborate to compute results or share non sensitive mining results and thereby avoiding disclosure of sensitive information[12]. Cryptographic techniques are used in such scenarios because it provides models for privacy and for implementing privacy preserving data mining algorithms. This algorithm is not suitable when more parties are involved and for large databases.

4. Conclusion

In today’s world, privacy is the major concern to protect the sensitive data. People are very much concerned about their sensitive information which they don’t want to share. Our survey in this paper focuses on the existing literature present in the field of Privacy Preserving Data Mining. From our analysis, we have found that there is no single technique that is consistent in all domains. All methods perform in a different way depending on the type of data as well as the type of application or domain. This article represents a step towards defining the PPDM framework and Privacy Preserving Data Mining techniques.

References

[1] Ann Cavoukian, Information and Privacy Commissioner, Ontario, “Data Mining Staking a Claim on Your Privacy”, 1997 www.ipc.on.ca

[2] U. M. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy. “Advances in Knowledge Discovery and Data Mining”. AAAI/MIT Press, 1996.

[3] The Economist. “The End of Privacy”, May 1st, 1999. pp: 15

[4] K. Thearling, “Data Mining and Privacy: A Conflict in Making”, DS, November 1998

[5] David Hand, Heikki Mannila, Padhraic Smyth,” Principal of data Mining”, A Bradford book, The MIT Press, Cambridge Massachusetts, London, England

[6] Majid Bashir Malik, M. Asger Ghazi, Rashid Ali,” Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects”, Third International

Conference on Computer and Communication Technology2012

[7] R. Agrawal and R. Srikant. “Privacy Preserving Data Mining”, ACM SIGMOD Conference on Management of Data, pp: 439-450, 2000

[8] Y. Lindell and B. Pinkas, “Privacy Preserving Data Mining”, Journal of Cryptology, 15(3), pp.36-54, 2000.

[9] Aris Gkoulalas-Divanis and Vassilios S. Verikios, “An Overview of Privacy Preserving Data Mining”, Published by The ACM Student Magazine, 2010.

[10] S.V. Vassilios , B. Elisa, N.F. Igor, P.P. Loredana, S. Yucel and T. Yannis, 2004, “State of the Art in Privacy Preserving Data Mining” Published in SIGMOD Record, 33, 2004, pp: 50-57.

[11] Gayatri Nayak, Swagatika Devi, "A survey on Privacy Preserving Data Mining: Approaches and Techniques", International Journal of Engineering Science and Technology, Vol. 3 No. 3, 2127-2133, 2011

[12] Aggarwal C, Philip S Yu, "A General Survey of Privacy- Preserving Data Mining Models and Algorithms", Springer Magazine, XXII, 11-52, 2008.

[13] H. Kargupta and S. Datta, Q. Wang and K. Sivakumar, “On the Privacy Preserving Properties of Random Data Perturbation Techniques”, in proceedings of the Third IEEE International Conference on Data Mining, IEEE 2003.

[14] T. Jahan, G.Narsimha and C.V Guru Rao, “Data Perturbation and Features Selection in Preserving Privacy” in proceedings of 978- 1-4673-1989-8/12, IEEE 2012.

[15] Sweeney L, "Achieving k-Anonymity privacy protection using generalization and suppression" International journal of Uncertainty, Fuzziness and Knowledge based systems, 10(5), 571-588, 2002.

[16] Aggarwal C, Philip S Yu, "A condensation approach to privacy preserving data mining", EDBT, 183-199, 2004.

[17] Wang P, "Survey on Privacy preserving data mining", International Journal of Digital Content Technology and its Applications, Vol. 4, No. 9, 2010.