PRESERVING PRIVACY IN DATA MINING USING SEMMA METHODOLOGY

Vijaylaxmi
Student (FMIT)
Jamia Hamdard University
New Delhi, India
tyagi.vijaylaxmi@gmail.com

Gunjan Batra
Student (FMIT)
Jamia Hamdard University
New Delhi, India
gunjan_batra27@yahoo.co.in

Dr. M. Afshar Alam
Professor (FMIT)
Jamia Hamdard University
New Delhi, India
aalam@jamiahamdard.ac.in

***Abstract* - The huge amount of data available means that it is possible to learn a lot of information about individuals from public data. Here, this open data need to be sheltered from unlawful contact. The privacy-preserving data mining (PPDM) has thus become a significant subject in most recent years. Generally privacy means "keep information about person from being available to others" but, the real worry is that their information not be mishandle. The data mining techniques enable users to extract the hidden patterns which may lead to leakage of sensitive data. So the main concern is to secure the data mining result with the help of PPDM. This paper provides a framework to preserve privacy in data mining results by manipulating SEMMA analysis cycle.**

I. Introduction

Data mining key goal is to deliver information. The growth of data mining process began when company data was first stored on computers, unrelenting with improvements in data access. Data mining deals with bulky database which may hold insightful information. Data mining involves the extraction of hidden patterns, finding predictive information which may compromise confidentiality and privacy obligations. Today, the maturity of data mining techniques has amplified the discovery risks of this insightful data. The main hazard of data mining is extracting confidential information about individual like their personal information. Here, organizations require privacy by hiding the sensitive data from unauthorized access. Some perceptive information about individuals, businesses and organizations has to be concealed before it is public or available. Henceforth data mining is clubbed with security and named as privacy-preserving data mining. The main objective of PPDM is to achieve privacy without compromising accuracy of data mining results. Privacy preserving data mining has the potential to increase the reach and benefits of data mining technology. The term privacy-preserving data mining was introduced in the year 2000 by two different papers with same name by Agrawal and Srikant (2000) and Lindell and Pinkas (2000). Agrawal and Srikant devised a term called randomization that plays a vital role to gain faith and confidence of users in storing their private records for efficient centralized data mining while limiting the disclosure of their actual private values; Lindell and Pinkas invented a cryptographic protocol for decision tree construction over a data set horizontally partitioned between two parties. These methods were then reused and expanded by many researchers worldwide.

II. Privacy-preserving data mining (PPDM)

Privacy-preserving data mining (PPDM) refers to the technique that provides shelter to insightful information from unwanted or unauthorized revelation. Privacy preservation key objective is to protect data records from any type of disclosure. PPDM provides high quality information to end users without exposing personally

identifiable information. PPDM employs a variety of practices to alter either the original data or the data generated (calculated, derived) using data mining methods Organizations must consider different dimensions before designing their privacy policies and access control rules. These dimensions are:

- Purpose
- Data categories
- User groups
- Actions
- Obligations
- Data object context

To attain eminence outcome while preserving the privacy of the data proficiently, five facets [2] have to be included are:

(1) The sharing of the basic data
(2) How basic data are tailored
(3) Which mining method is being used
(4) If basic data or rules are to be unseen
(5) Which supplementary methods for privacy preservation are used

III. SEMMA METHODOLOGY

SAS Institute defines data mining as the process used to divulge expensive information and complex relationships that exist in huge amounts of facts. Data mining process divides into five stages that are Sample, Explore, Modify, Model, and Assess [3] which makes it simple for various business processes like select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy.

A. Benefits of SEMMA methodology:-

Data mining techniques can be enhanced by using the SEMMA. Here, are some examples where SEMMA helps data mining techniques to achieve their goals:

A. Clustering the customer's groups with their buying trends.
B. Identifying the most beneficial customers
C. Understanding product-customers relationships
D. Analyzing factors affecting purchasing patterns, payments and response rates
E. escalating earnings by advertising to those most likely to purchase
F. Uncovering risk factors
G. Acquiring new customers

B. Modeling PPDM using SEMMA methodology

Steps involved in modeling the PPDM using SEMMA methodology are figure 1.
The steps are explained below:

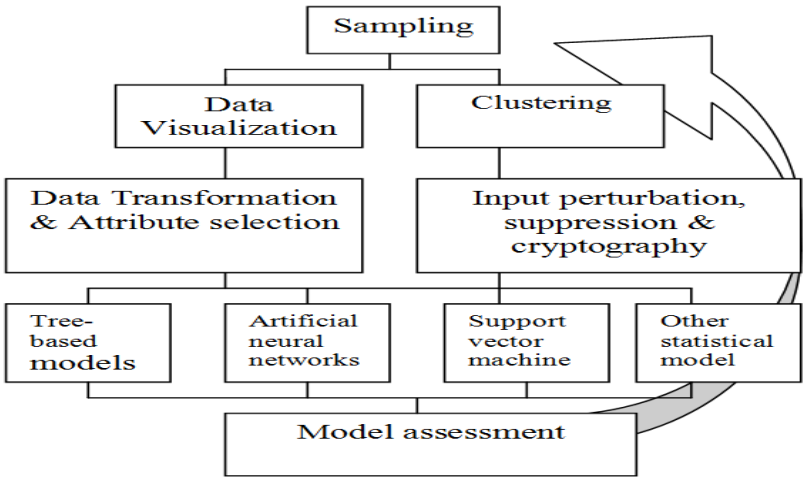

Figure 1: Modify Steps in SEMMA Methodology in PPDM techniques

a. SAMPLE: Analysis of some sample is easier, more efficient, and can be as accurate as exploring the entire database. This step includes the sampling of data from large database. The sample data can be created with the help of various tables like training, validation and test tables [4] for better evaluation. Sampling key purposes are speed and efficiency. Sampling can be done by simply selecting random datasets or by selecting N-th record from large database, this type of sampling is also called systematic sampling or clustering sampling and stratified sampling

b. EXPLORE: After sampling of data, next step is to explore the data for visually or analysis purposes. Also, explore step used to search the data for predictable associations, unexpected trends, and anomalies in order to provide better ideas to analysts. It also classifies attributes in three categories, identifier attributes, Key attributes and Confidential attributes. Identifier attributes are those attribute which give information that leads to a specific entity example one's name and ID No., key attributes (quasi-identifiers) are those attribute which may be known by an intruder example one's age and ZIP Code, last but most important are the confidential attributes which are assumed to be unknown to an intruder as they cause the maximum harm to one's confidentiality if known by others example one's income and medical diagnosis.

c. MODIFY: Organizations may need to manipulate the data that contain confidential or sensitive data. Privacy to data mining results can be provided by applying different data modifying techniques of PPDM. Modification of sensitive data can be done by transforming the values of attribute or normalizing the values of private attributes. Some of these techniques are explained in section 4 of this paper.

d. MODEL: After making data compatible for analysis or pattern discovery, the data can be model by applying various modeling approaches. Data mining approaches are artificial neural networks, decision trees, support vector machine, logistic models, and various statistical analyses such as memory-based reasoning, and principal component analysis. Different approaches have there own strengths and weaknesses like Decision trees are less suitable for inference tasks where the aim is to forecast the value of a continuous attribute but it supply a clear suggestion of which fields are most vital for prediction or classification.

e. ASSESS: Once models are generated they are interpreted according to the existing domain knowledge and data mining success criteria. In the final step, experts judge the results of models within domain context, while data miners apply data mining criterion (correctness, lift or gain tables, etc.). It is also advisable to test models on real problems (test) [5]. Moreover outcome directly associated to problem solving objectives it is advisable to assess other findings not necessarily related to the original objectives. This might reveal additional information, hints or suggestions for future modeling.

IV. Modify Step in SEMMA Methodology in PPDM techniques

The modify step can be done by transforming, creating and selecting the attributes to prepare the model for securing the sensitive data. The PPDM modification techniques are used in SEMMA method to provide security to data mining results.

Name	ID	ZipCode	Age	Income	Medical Diagnosis
Rehaan	121	110007	32	570000	Normal
Sahil	142	110081	30	480000	Tuberculosis
Garv	113	110062	28	720000	Diabetes
Avijt	145	110009	39	910000	Blood Pressure
Kanika	97	110009	35	780000	Blood Pressure
Ajay	23	110009	42	1000000	Diabetes

Figure 2: Sample data on which we will show ppdm techniques in next subsections

A. Input perturbation

In this technique we reveal the entire database, but after randomizing the entries. The randomization term was coined by Warner in year 1965. This is the process of distorting the input data so that the data values of individual entities are protected from revealing and one feels secure even after publishing his/her data. This perturbed database is then subjected to statistical analysis over data. Let's X is an attribute that needs to be protected from unauthorized access; X is a set of values i.e. $\{X_1, X_2 \ldots, X_n\}$, R is one of the randomization techniques that is applied over X in order to deviate the value, and then R(X) is the perturbed database. Several randomization techniques can be identified in privacy preserving data mining algorithms, including adding random numbers, generating random vectors, swapping, generalization, resampling and so on. Data is perturbed in two manners: generalization and value distortion. The generalization is a method tin which values of an attribute are divided into general; classes instead of specific value and the interval in which a value lies is returned instead of the original value. Whereas in value distortion method we distort a value by adding any random value y_i to each value x_i of an attribute.

Suppose there is a consolidated data mart of a company, and many customers as well as clients have access to this information. It protects the customers' data by letting them distort their records before sending them to the server, taking away some actual information and introducing some noise. At the server's side, statistical estimation over noisy data is employed to recover the aggregates needed for data mining. Noise can be introduced, for example, by adding or multiplying random values to numerical attributes (Agrawal & Srikant,

2000) or by deleting real items and adding bogus items to set-valued records (Evfimievski, Srikant, Agrawal, & Gehrke, 2002; Rizvi & Haritsa, 2002). Given the right choice of the method and the amount of randomization, it is sometimes possible to protect individual values while estimating the aggregate model with relatively high accuracy.

Name	ID	ZipCode	Age	Income	Medical Diagnosis
Rehaan	121	110007	32	500000	Normal
Sahil	142	110081	30	500000	Tuberculosis
Garv	113	110062	28	700000	Diabetes
Avijt	145	110009	39	900000	Blood Pressure
Karika	97	110009	35	700000	Blood Pressure
Ajay	23	110009	42	900000	Diabetes

Figure 3: Table after input perturbation

One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the database except the one that needs to be hidden from intruders.
Here, the idea is to modify data values such that reconstruction of the actual values is difficult, but the data mining results from the distorted data are still valid. Main threat of this technique is the ability to correctly guess the actual value.

B. Suppression

Privacy can be preserved by simply suppressing all confidential data before any disclosure or computation occurs. Given a database, we can suppress specific attributes in particular records as dictated by our privacy policy.
Here once we identify confidential, quasi and identifier attributes. We make decision about suppressing which of the attributes.
One can achieve privacy by suppressing either confidential attribute or identifier attribute. It is decided purely on the data mining result. Attributes that are more prone to analyze are remain intact. Rather than protecting the sensitive values of individual records, we may be interested in suppressing the identity (of a person) linked to a specific record because no one is interested in analyzing the identifier attribute.

ZipCode	Age	Income	Medical Diagnosis
110007	32	570000	Normal
110081	30	480000	Tuberculosis
110062	28	720000	Diabetes
110009	39	910000	Blood Pressure
110009	35	780000	Blood Pressure
110009	42	1000000	Diabetes

Figure 4: Table after Suppressing

In our example, we decide to suppress identifier attribute in order to hide individual's identity. Still this method is not very secure as one will get to know identities by some other means and quasi attributes helps intruders to make some guesses. As in above example if intruder knows the age of a person then it's very easy for him to know the income and diagnosis.

C. Cryptography

In cryptographic approach to PPDM parties exchange their data in an encrypted form. The first cryptographic technique to data mining was used by Lindell and Pinkas (2000) for the problem of decision tree construction over horizontally partitioned data then onwards several. Cryptography is an inefficient technique with large amount of data. Assume there are multiple parties as clients, servers, trusted and untrusted miners who all have some private inputs X_i , and they must compute some private outputs $Y_i = F_i$ (~x) etc are defined by the functionality we want to compute by data miners Build a cryptographic protocol that guarantees that after some rounds, the i_{th} party learns y_i and nothing else and this should hold even if some of the parties maliciously attempt to obtain more information. In particular, no party will know the inputs of the other parties. Many of the protocols based on encryption use the idea introduced by Yao. Generally speaking, secure multi-party computation is the branch of cryptography that deals with the realization of distributed tasks in a secure manner; in this case, the definition of security can have different flavors, such as preserving the privacy of the data or protecting the computation against malicious attacks.

V. CONCLUSION

In this paper, we proposed a modeling method of PPDM by SEMMA methodology. PPDM techniques help in ensuring the quality delivery of data mining results to the end users without affecting the one's basic details. SEMMA methodology is an analysis cycle; it filters the datasets from large pool of records and also discloses hidden patterns that help in achieving data mining objectives. This paper provides an analysis model that help organizations' to implement data mining projects to ensure results quality and privacy. Applications of the model are as vast as that of data mining projects.

REFERENCES

[1] Y. Lindell and B. Pinkas, Privacy preserving data mining, Journal of Cryptology 15 (3), 177-206, (2002).

[2] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., Theodoridis, Y., 'State-of-the-art in Privacy Preserving Data Mining', SIGMOD Record, Vol. 33, No. 1, New York, March 2004. Download: http://dke.cti.gr/CODMINE/SIGREC_Verykios-et-al.pdf

[3] SAS Institute,"Data Mining and the Case for Sampling", www.ag.unr.edu/gf/dm/sasdm.pdf

[4] David Louis Olson, Dursun Delen, "Advanced data mining techniques"

[5] Rudjer Boskovic Institute website, http://dms.irb.hr/index.php

[6] lecture notes, http://www.mathcs.emory.edu/~lxiong/cs573/lect.html

[7] Evfimievski, A., Srikant, R., Agarwal, R., and Gehrke, J. "Privacy Preserving Mining of Association Rules," In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining (KDD'02), Edmonton, Alberta, Canada, July 2004.

[8] Agrawal, D. and Aggarwal, C.C. 2001. On the design and quantification of privacy preserving data mining algorithms. In Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART symposium on principle of database system, ACM, pp. 247–255.

[9] Agrawal, R. and Srikant, R. 2000. Privacy preserving data mining. In Proceeedings of the ACM SIGMOD Conference of management of data, ACM, pp. 439–450.

[10] E. Alexandre, and G. Tyrone. Privacy-Preserving Data Mining. 2009.

[11] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M.Y. Zhu, `Tools for privacy preserving data mining', ACM SIGKDD Explorations, 4(2), 28.34, (2002).

[12] Benny Pinkas, `Cryptographic techniques for privacy-preserving data mining', ACM SIGKDD Explorations, 4(2), 12.19, (2002).

[13] Vaidya, J., Clifton, C.: Privacy preserving association rule mining in vertically partitioned data. In: 8th ACM SIGKDD International Conference on Knowledge Discovery and Data
Mining

[14] D. Agrawal and C. C. Aggarwal. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In Proc. of ACM SIGMOD/PODS, pages 247–255, Santa Barbara, CA, May 2001.
S. J. Rizvi and J. R. Haritsa. Privacy-Preserving Association Rule Mining. In Proc. of the 28th International Conference on Very LargeDataBases,Hongko ng, China, August 2002.

AUTHORS PROFILE

1. Vijaylaxmi is an Assistant Professor of Computer Science department at Bharati Vidyapeeth College of Engineering, New Delhi. She obtained her bachelor's degree from Kurukshetra University, India in 2007. She is currently a M.Tech student and doing her dissertation under the supervision of Prof. Dr. M. Afshar Alam. Her research is centered on Data Mining. She has coauthored over four International publications.

2. Gunjan Batra is an Assistant Professor of Computer Science department at Bharati Vidyapeeth College of Engineering, New Delhi. She obtained her bachelor's degree from Kurukshetra University, India in 2007. She is currently M.Tech student and doing her dissertation under the supervision of Prof. Dr. M. Afshar Alam. Her research is centered on Data Mining. She has coauthored over four International publications.

3.  Dr. M. Afshar Alam is a professor in Department of Computer Science, Jamia Hamdard, New Delhi. He has Teaching experience of more than 17 years. He has authored 8 books and guided PHD research works. He has more than 30 publications in national/international journals. He has delivered special lectures as a resource person at various academic institutions and conferences. He is a member of expert comittees of UGC, AICTE and other national and international bodies. His research areas include software re-engineering, data mining, bioinformatics and fuzzy databases.