IJCRT Publish Paper record is available at DOI: http://doi.one/10.1727/IJCRT.17181

Study and Review of Hybrid Approach for Privacy Preserving Data Mining

Gunjan S. Bonde

PG Student, Department of Computer Engineering,

SSBT's College Of Engineering and Technology,

Bambhori, Jalgaon [M.S],India

Akash D. Waghmare

Assistant Professor, Department of Computer Engineering,

SSBT's College Of Engineering and Technology,

Bambhori, Jalgaon[M.S],India

Abstract – **In today's era of growing technology the data collected by organizations has the requirement to preserve the privacy of the individuals. It needs to maintain privacy of the individuals because users sensitive data is stored online over the centralized repository. The techniques like anonymization, randomization are used to achieve the privacy. But anonymization leads to certain level of information loss while preserving privacy. To overcome this drawback, hybrid approach is used. The proposed system involves combination of two techniques i.e anatomization and perturbation techniques. The quasi-identifiers like zip code, age, gender of a person does not seem to be very important to protect but these fields when linked with some other attributes can expose the identity or sensitive information of an individual. The hybrid method focuses on the goal of preserving privacy by anatomizing and perturbing the quasi identifiers in the sensitive data of customers stored on centralized data repository without causing any loss to the information.**
***Keywords* – Anatomization, quasi-identifiers, perturbation**.

I .INTRODUCTION

Users sensitive data being collected from private as well as public organizations for various analysis or decision making purposes by data mining. It is necessary to maintain privacy of individual's data. Privacy here means identity of the person not being revealed while unveiling any sort of data or using the data for any research or business purposes. Thus Privacy Preserving Data Mining is a real challenge these days. The attributes can be divided into following categories:

1. Identifying attributes: These attributes are name, email-id which can explicitly reveals identity of person.
2. Quasi-identifiers attributes: The attributes like age, gender, zip code when linked with some other attributes can easily exposed a person's identity.
3. Sensitive attributes: This includes the data which should not be exposed or published against a person's identity. For e.g. while analyzing the sale of particular product in online shopping, the customer's identity should not be revealed against any product.
4. Non-Sensitive attributes: These are the fields which if disclosed publically do not lead to any problem.

Data hiding tries to remove confidential or private information from the data before its disclosure. In this case, many different methods have been addressed. The randomization method has been traditionally used which has less accuracy and high time complexity so new hybrid approach is used to overcome such problems. Such approach focuses on the preservation of the privacy of data with numerous SA with lesser information loss and better data utility. Anatomization approach is employed to minimize the information loss and perturbation techniques used to preserve privacy. In this case, the data miner does not know the raw data and also can get the similar result which is the key point for data miner is how to reconstruct the raw data distribution. The significant work are exists for preserving the privacy in database. But the work carried out on other document types is limited.

In order to achieve the main aim of privacy preserving data mining, hybrid approach is used to improve privacy. The proposed hybrid approach combining perturbation and anatomization with slicing for Privacy Preserving data mining. Online sensitive information of users is collected, then proposed approach is applied to anatomizes information by dissociating the quasi-identifier from SA and provides two tables, one for the QI attributes and the other for the SA. After dissociation perturbation techniques are applied to sensitive data to protect users sensitive data and also such methods applied to different documents types

such as ASCII documents. The main objective of project is to enhance the privacy of the data without changing the original data set and reduce the information loss and to increase performances and minimize time complexity.

The rest of the paper is organized as follows: Section II provides an related work. Section III presents the proposed solution while section IV finally concludes the paper.

II. LITERATURE SURVEY

Literature survey plays vital role in the software development process. Different time factor, economic factor are taken into consideration. The content of the paper focuses on the research and contributions of various sources. These include:

P.Usha and R. Shriram[3], proposed a method based non-homogeneous anonymization technique. This method categorized attributes into different groups. Then table is split into two tables with quasi-identifiers and sensitive attributes are in one table andthe remaining values in the other table. The goal of this method was to reduce the information loss caused because of homogenous anonymization. Result shows this method achieves high degree of data utility and data integrity.

R.Mahesh and T. Meyyappan[4], proposed the method to achieve privacy preservation through generalization of quasi identifiers. Method focuses on setting up the attributes in the range and deleting the duplicate record. The values in the quasi-identifiers are generalized by suppressing the values by some character and using intervals for age field. This approach of duplicate record elimination helps in reducing information loss and gives better performance in terms of privacy gain when compared to k anonymity or l diversity. Result shows that method gives protection from the two types of attacks
record linkage and attribute-linkage.

M. Prakash, and G. Singaravel [5], proposed the method based on personal to preserve privacy while publishing the data. They used the method of Top down Greedy Algorithm for anonymization process. The process partitioned dataset by taking median as the split value. The values less than split value are taken as left hand side and greater are considered as right hand side. Both LHS and RHS are anonymized separately and combined. The values of 't' in LHS and RHS describe the level of privacy. Result shows this approach performs better performance.

Xiaolin Zhang and Hongjing Bi [6], proposed the method of random perturbation for Privacy Preservation Data Mining. The author proposed a way as data by replacing the attribute values with the code values(1,2,3,n) and arranging these values in a square matrix which were then randomly perturbed. The data is extracted from these matrices using if then rules after pruning it with post pruning algorithms. The method helps in improving accuracy and achieving better data mining results. Results of the method shows that the accuracy increases with the increase of data.

Xuyun Zhang, L. Yang and J. Chen [7], presented two phase top down specialization approach for anonymizing the data. The author used specialization approach instead of generalization to solve the privacy concerns and achieve the scalability requirement for the large datasets. The health dataset is used here for data analysis and other experimental purposes. The results and simulation proved that the scalability could be achieved with the two phase top down method.

Dilpreet Kaur, Divya Bansal and Sanjeev Sofat [8], proposed the comparative study of the anonymization techniques. After implementing the techniques on different data sets the author came with different inferences like as the number of attributes increases the information loss gets increased showing that the information loss is directly proportional to the number of attributes to be anonymized. Result shows that T-Closeness has less information loss than L-Diversity and K Anonymity
but these techniques still leads to extensive information loss.

Mebae Ushida and Kouichi Itoh [9], Proposed Privacy- Preserving using Data Aggregation for fulfilling the major requirements of cloud which are guarantee against leakage of stored information and providing aggregation results as per authority. So the data is masked by the user before storing on cloud and then the masked aggregation results are obtained. The user get the aggregation results by unmasking the masked results by their own secret private keys assigned as per the authority of the user.

Neha Gupta and l. Rajput [10], proposed a method to preserve privacy while data mining using the rotation and translation based perturbation. Translation based perturbation includes adding a particular noise to the data and in rotation based the data is rotated by the angle theta. For Rotation based find a value k as median of the number of records in the dataset. Taking k records at a time find a threshold value to rotate the data. For the translation based select a value for the noise and for each record if the attribute value is greater than the noise value the noise is added to the record otherwise it is subtracted to perturb the data. The method again is a victim of information loss because one the data is changed using a certain angle or noise it cannot be recovered.

Unil Yun and J. Kim [11], proposed the method of fast perturbation for large size databases and focused on minimizing the utility of the data to hide the original values of the data set by using a tree structure for perturbation. The author represented the data set in the form tree structure such that each node in the tree consist of six fields name, parent node, set of child nodes, count,

node link, utility information list. Results shows that method helps in decreasing the computational, access and search time for the records as each node has a separate path in tree structure.

Jianming Zhu [12], proposed the cryptographic method for privacy preserving data mining using Holomorphic encryption, ElGamal encryption and K-nearest neighbour. Holomorphic encryption is the best suited method for privacy preservation as the results for analysis and data mining as the Holomorphic encryption is the only cryptographic method which supports the application of operations on the encrypted data. Results shows that although the technique is successful in privacy preservation and reverse process but it has high time complexity and implementation complexity.

M. Suriyapriya and A. Joicy [13], proposed the privacy preserving method using the symmetric key encryption method. The data is being encrypted using the symmetric key i.e. same key for encryption and decryption generated by the . Nobody can guess the data once it is encrypted. The data is safe in this method only until the key is not known to the intruder. Once the key is found it can decrypt the encrypted data to its exact original values. Thus the data is safe only until the key is safe. Results shows that although there is not much information loss by using this method for preserving privacy but the method is not successful due to its complex working and performance inefficiency with increase in the size of data.

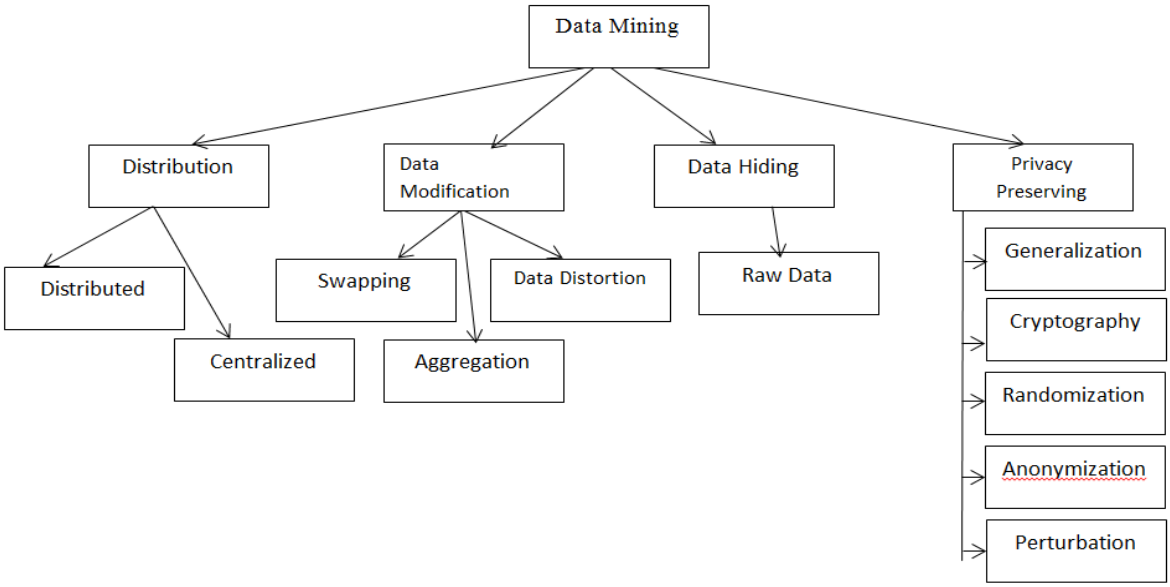

Figure 1. Structure of Literature Survey

III. PROPOSED SYSTEM

An enormous quantity of personal sensitive information is available in recent decades and tampering of any part of this information imposes a great risk, so hybrid approach is applied to sensitive data. The quasi identifiers are the crucial ones in the task of preserving privacy. The proposed method is an initiative to preserve privacy on the online centralized data repository with minimum information loss. Such method is the hybrid approach for Privacy Preserving Data Mining based on anatomization and perturbation implemented over the centralized server. The anatomization approach dissociates the correlation observed between the quasi-identifiers attributes and sensitive attributes (SA) and yields two separate tables with non-overlapping attributes. In the enhanced slicing algorithm, vertical partitioning does the grouping of the correlated SA in Sensitive Table together. Then perturbation based techniques applies which add noise to original sensitive attributes of the dataset which privatize sensitive attributes. Also such methods are applied to different document types such as ASCII documents.

Architecture

The goal of hybrid approach is to preserving privacy on the quasi-identifiers in the sensitive data of customers stored on centralized data repository without causing any loss to the information in the process. The architecture of the proposed system is shown in Figure 2.

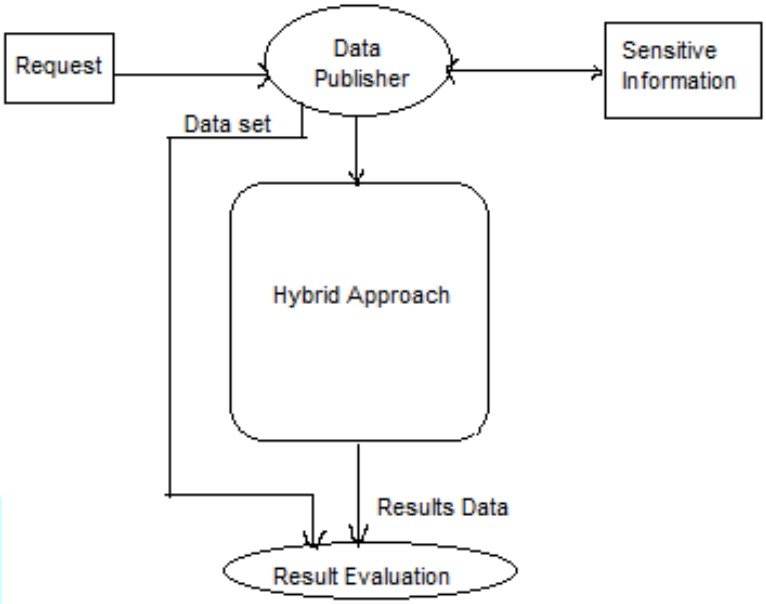

Figure 2. Architecture of the Propose System

Data publisher request to data owner of users data for analysis purpose. Data cannot be provide in original form but in perturbed form. At first, method identifies the quasi-identifiers and sensitive attributes from dataset which can reveal an individual's information. Then hybrid approach can be applied to original data to make them perturbed. Anatomization with slicing applied to sensitive attributes so that related data are grouped together based on their correlation. Then perturbation techniques applied which focuses on the data by the addition of noise. Query handler is accepting the query data from the client and process the query with the data base and fetching the datasets from the data base. At last, result is evaluated which is a process to find the error rate of different states in the data perturbation of original data and the perturbed data.

IV. CONCLUSION

Hybrid approach will used to preserving privacy without information loss. Existing techniques are used to achieve the goal. But unfortunately some leads to certain level of information loss while preserving privacy. But in the proposed system, hybrid approach will used to overcome such problem and try to a great extend in hiding the identity of customers and preserving their privacy without information loss. Less execution time for maintaining privacy will achieved by new hybrid approach. The proposed method will resolved the critical conflict between the privacy preservation and information loss.

REFERENCE

[1] Arshveer Kaur, "A Hybrid Approach of Privacy Preserving Data Mining using Suppression and Perturbation Techniques", International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) IEEE, 2017.
[2] Alpa Shah and Ravi Gulati, "Evaluating Applicability Of Perturbation Techniques For Privacy Preserving Data Mining By Descriptive Statistics", Intl. Conference on Advances in Computing, Communications and Informatics, Sept 2016.
[3] P. Usha, R. Shriram et. al., "sensitive attribute based non-homogeneous anonymization system", Information Communication and Embedded Systems (ICICES) International Conference on (pp. 1-5). IEEE.,Feb 2014.
[4] R.Mahesh and T. Meyyappan , "Anonymization technique through record elimination to preserve privacy of published data", International Conference on (pp. 328-332). IEEE., 21-22 Feb. 2013.
[5] Prakash, M., Singaravel, G., " An approach for prevention of privacy breach and information leakage in sensitive data mining", Computers and Electrical Engineering, 45, 134-140, (2015).
[6] Zhang, X., Bi, H. , "Research on privacy preserving classi_cation data mining based on random perturbation`", In Information Networking and Automation (ICINA), International Conference on (Vol. 1, pp. V1-173). IEEE., 2010, October.
[7] Zhang, X., Yang, L. T., Liu, C., Chen, J., "A scalable twophase top-down specialization approach for data anonymization using mapreduce on cloud`", Parallel and Distributed Systems, IEEE Transactions on, 25(2), 363-373, 2014.
[8] Arora, D. K., Bansal, D., Sofat, S., "Comparative Analysis of Anonymization Techniques", International Journal of Electronic and Electrical Engineering, ISSN 0974-2174 Volume 7, pp. 773-778, 2014.
[9] Ushdia,M., Itoh, K., Katayama, Y., Kozakura, F., Tsuda, H., "A Proposal of Privacy Preserving Data Aggregation on the Cloud Computing", Network-Based Information Systems (NBiS), 16th International Conference on (pp. 141-148). IEEE, September 2013.

[10] Gupta, N., Rajput, I., "Preserving Privacy Using Data Perturbation in Data Stream", International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), 2013.
[11] Yun, Kim, J., "A fast perturbation algorithm using tree structure for privacy preserving utility mining", Expert Systems with Applications, 42(3), 11491165, 2015.
[12] Zhu, J., "A new scheme to privacy-preserving collaborative data mining", Information Assurance and Security, IAS'09, Fifth International Conference on (Vol. 1, pp. 468-471). IEEE, 2009, August.
[13] M. Suriyapriya, A. Joicy, "Attribute Based Encryption with Privacy Preserving In Clouds", International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 2, February 2014.