Hybrid Approach for Privacy Preservation data Mining Using Random and Mod Techniques

Prashant Namdev
Dept. Computer Science and Engineering
L.N.C.T & S.Bhopal, INDIA
prashantnamdev555@gmail.com

Manoj Kumar
Dept. Computer Science and Engineering
L.N.C.T & SBhopal, INDIA
mannu175@yahoo.com

Abstract— In modern time, the privacy-preserving data-mining (PPDM) has been extremely analyzed, due to the vast eruption of the private information over the internet. Various algorithmic technologies have been observed for the privacy-preserving data-mining. This paper investigates a descriptive analysis of the various approaches for the privacy preserving data mining like Randomization, K-anonymization technique, Perturbation Technique, Cryptographic technique, etc, for the sharing of information and for its privacy. The basic target of the privacy preserving data-mining is to introduce several algorithms for changing the actual data and for protecting the information that are to be get misused, such that the sensitive data and the privacy of the knowledge still retains as it is after the process of mining. This concept is used to again repeat the various privacy-preserving data-mining techniques to secure the privacy of the private information and attaining the data clustering along with the less loss of information for the multiplicative attributes in the data-set.

Keywords- Privacy, Preservation, Randomization, Mod Operation, Data Security.

I. INTRODUCTION

In the topical years, the concept of data mining has been detected as a threat to the privacy as due to the extreme propagation of the digital-data that is organized by the institutes. This has directed to the more trouble associated with the privacy of the important data. In the current years, several technologies have been developed for altering or conversion of the data in a way such that as to protect privacy. An observation on few of the approached that are used for the privacy-preserving data-mining can be detected. In this paper, it will be described a general brief of the metaphorical analysis of the various technologies in the field of privacy-preserving data-mining. The concept of data-mining is used for extracting the intelligent information from the large databases. Privacy Preserving Data Mining (PPDM) is targeted on the process of how to introduce or design the algorithm for protecting the actual data, such that knowledge associated with the data must be private and secured after the completion of the mining process [1]. In the data-mining (DM), the users are allowed to use the data but not provided with the association rules and they are also independent for using their personal tools. Hence, it is compulsory to implement the privacy limitations on data earlier the phase of mining. Privacy preservation data mining approaches provides the publishing of data for purpose of mining with the protecting the private information of an organization. Various privacy preservation approaches are available for the private and confidential information but all these suffers various kinds of the attacks.

Commercial aspects are also associated with privacy problems. Various aspects are collected information regarding the persons for their personal specific requirements. Frequently, various departments in a company themselves can seems it is as mandatory to distribute the information. In such situation, every organization or the firm should confirm that privacy of an individual is not got affected or that of the private business detail is not disclosed. However various kinds of the protection of person's information have been evolved, and there are various ways for misleading these approaches. Techniques of data mining may have been evolved to retrieve the knowledge successfully to support a variance of fields like weather forecasting, marketing, medical diagnosis, and the national security. A dispute is still there to mine few types of the data without breaking the privacy of the data owners. As the data-mining have become more extensive, privacy aspects are raising. In this project, to protect the privacy of these kinds of details records may be de- identified earlier the details records are get distributed along with the other users without being breaking the person's privacy. This may be performed by removing the unique identity values like passport number, age, etc. Even if this type of information is removed then still there are other types of the information fields are present when the linked along with other fields present in the data-sets can recognize the individual. To enable the security for these kinds of breaches, there is a

requirement of the variation of the data mining algorithm. To enable and guarantee efficient data collection, it is vital to apply approached that decrease the risk of disclosure and increasing the analysis of mining results with a confirmation of the privacy. This paper, explains the various methods and technologies in the domain of the Privacy-Preserving Data-Mining (PPDM).

II. PRIVACY PRESERVATION DATA MINING

n Privacy Preserving Data Mining (PPDM) is an extremely new concept of the data-mining research disputes. It addressed to domains of the data-mining which contributes to prevent the private information from the revealing. The issue with the output of data-mining is that it may also leaks little information which is taken to be as private and sensitive. Easily access to these types of private data bears a risk to the privacy of the individuals. The real concern of the people is their personal details must not be got misused without their knowledge or information. The actual risk is that if the information is unlimited, it will be unrealistic to stop the misuse as given in paper [2]. There has also been a growing aspect regarding the chances of the abusing the private information under the concept of without the information to the actual owner of the data. By the definition, the privacy is a quality or the situation of being isolated from the availability or the aspects of the others. On associating the privacy with the data mining, the privacy indicates to maintain the information regarding the individual from being getting available to the others [3]. Privacy is the matter of the interest as it can have the inverse impacts on someone else's life. Privacy is not affected until the one feels that their private information is get used in a wrong way. Once the private information is leaked, the one may not protect it from being getting misused.

2.1 Data Mining

Data mining is one of the significant tools to retrieve patterns or the knowledge from the data. Data-mining mechanism may be utilized to mine repeated patterns, perform classification, find associations and done prediction, etc. The data needed for the process of data-mining can be recorded in the single database or in the shared resources. The traditional methods for the shared resources are data warehouse. Fig. 1 represents a usual distributed data-mining method for developing a data-warehouse consists of all the data. This needed the data-warehouse to be maintains and trusted the privacy of all the parties. As the data-warehouse already knows the source of the data, it learns the site-specific details also with global results.

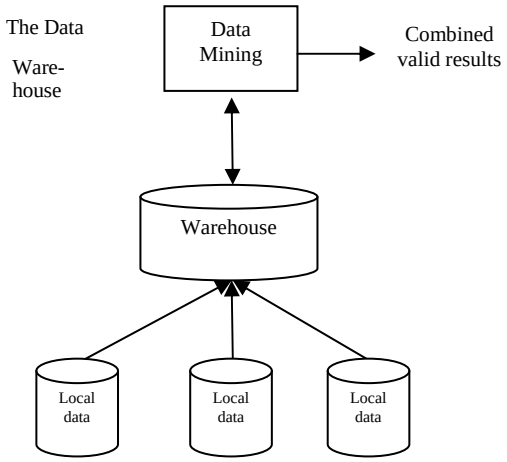

Fig. 1: Data warehouse approach to mining distributed sources

In the DM, users are supported with data and which is not the association-rules and are independent to use their personal tools therefore the limitation for the privacy has to be implemented on the data themselves earlier the phase of mining. Due to this reason, there is a requirement to create a technology which may direct to the new privacy control systems to transform a provided database into the new database in a way that to protect the normal rules to be mined from the actual database. The method of converting the source database into the new database that hides little sensitive patterns or the rules is referred as the sanitization process.

2.2 Privacy Preservation

Privacy preservation has a vital role in the domain of the data-mining as the large volume of the data has been gathered in various organizations for this type of data privacy preservation should be compulsory. These types of gathered data can be used by various organizations for doing the data mining works. Though, the private and personal details of the gathered data must be prevented. If any of the organization discloses or distributed their data or information then the preservation of the privacy is a vital issue.

The objective of the Data mining (DM) are to be taken out as a useful information through the various sources, while the objective of the privacy preservation in the data mining is to protected these data without any loss of the personal and the private information. Privacy preserving data mining (PPDM) is one of the new research areas of the data mining and the statistical databases [4]. Several kinds of advanced data mining algorithms are easily obtained the privacy of data. With these of two limitations the privacy preservation may be possible.

1) Raw private data or information such as the name, identifiers and the addresses must be changed in the actual database.

2) By the use of data mining algorithm, the sensitive information or knowledge may get mined from the database as the knowledge has a significant role for the privacy of the data.

This indicates to the privacy preservation approach which is used for some discriminating changes of data. The approaches used are as follows:

• Heuristic-based approach changes the selected values such that modifying few values of the data in the provided dataset from the actual value to another value.

• Reconstruction-based approach in which the actual distribution of data is constructed again. These types of algorithms are applied by first perturbing the data and then constructing again the distributions.

• Cryptography-based approached such as the secure-multiparty computation in which the computation is protected if no party knows anything at the end of computation except its own input and results.

In paper [4] the researches have been presented a novel based mechanism which strategically changes some of the transactions within database. It changes the values of support or the confidence for hiding the sensitive rules beyond generating various side effects. However, not required side-effects like falsely hidden and spurious rules, non-sensitive rules, falsely generated, can be generated in the process of rule hiding.

2.3 Privacy V/S Security

Some confusion presents within the people of industry regarding the security and the privacy[10]. Few people thought that the security and the privacy is the same, whereas the others thought that the privacy refers to some information will get hidden for someone else. The security is a significant tool for the privacy. Both security and the privacy are quite similar type of technologies, though, there are significant differences are there in between them:

a. Developer required understanding of these two types of technologies while developing the new system.

b. Developer must have the ability of understanding what the information is going to and what is coming from the database.

2.3.1 Privacy preservation combined strategy

Within the combine strategy various strategies are used to achieve any type of privacy preserving technique. Because of the combining these types of various technologies powerful security may be achieved. Some-time privacy preserving approached may have few disadvantages or few restrictions but which may be get overcome in the combined strategy. So the result of security will be more efficient of the combine strategy as compare to a single privacy preserving approach used [5]. Here explained the actual data that are to be transformed after this transformation the data are get encrypted. Therefore the data transformation and the data encryption approaches are used in the combine strategy.

2.3.2 The Relationship between Data Security and Data Privacy:

Organizations may have made the data security policy for guaranteed privacy of data of the information of their customers. Data privacy is a significant aspect for the organization for this they enables the security of their customers the data as this data is the big asset. The policy of data security is the means to the desired end, which is data privacy. Though, no security policy of data may be overcome the purposed sell or plead of consumer's data which was entrusted to an organization.

2.3.3 Classification of various Privacy Preservation Techniques:

Some of the Privacy preserving techniques are mentioned here [5].

a. Perturbation Technique

This technique is used for privacy preservation where the data are perturbed. But this never re-constructs the actual data and it is also not good for huge data.

b. Condensation Technique

In place of the perturbed data, this technique works on pseudo data. Hence it allows better privacy-preservation as compare to the techniques that uses simply the data-modification on the actual data. But it cannot provide a longer influence on the data-mining. As it has similar format as to the actual data.

c. Cryptographic Technique:

It carried out the encryption of private data. Also there is a proper tool-set for the algorithm in this field of data-mining. But this approach is complicated to scale at the time when more parties are included and also they are not good for the huge databases.

III. LITERATURE REVIEW

[6]Privacy-preservation is a significant application of the data-mining approach to offer efficiency and security of the security of data. Data-distortion is the major element for the privacy-preservation in the applications of security-related data-mining. This article suggests a perturbation dependent PDM approach for the data-distortion and compares it with the previous approach. The result of experiment presents that perturbation-based-PDM approach is much better in the comparison of the previous approach. In the current scenario the privacy-preserving in data-mining is a very significant topic of the research. Literature review of this paper clarify that various privacy-preserving approaches are available in the data-mining but still some demerits they have. Anonymity approach provides the privacy-protection and the usability of the data but it faced from the homogeneity and the background attack. Blocking approach suffers after the analysis, that is the output of suggested approach for the execution-time and the dissimilarity is further better.

[7] With the raise in sharing of private data by networks among the governments, businesses and the other parties, the privacy preserving has now become a very important problem in the data-mining and the knowledge-discovery. Privacy aspects can protect the parties from sharing directly data and few types of the information regarding the data. This paper suggested the solution for the securily computing the data-mining-classification algorithm for the horizontally-partitioned data without exposing any information regarding the data or the sources.

The suggested approach (PPDM) merges the benefits of the RSA public-key-cryptosystem and the homomorphic-encryption model. Results of experiment presents that PPDM approach is the powerful in regards of the accuracy, privacy and the efficiency. The data-mining has been a famous area of research for more than a decade because of its huge spectrum of the applications. Though, the wide availability and popularity of the tools of data-mining has also increases the interest regarding the privacy of the individuals. The target of the researchers of privacy-preserving data-mining is to design a data-mining techniques approach which could be implemented on the databases without breaking the individual's privacy.

Privacy-preserving approaches for different models of data-mining have been suggested, originally for the classification on the centralized data after that for the association-rules in the distributed area. In this paper, The privacy-preserving distributed-KNN mining-algorithm has been described. As represented, the suggested algorithm is dependent on technology of homomorphism and the RSA-encryption that is semantically protected. Additionally, there was no global computations was done at centralized site were conducted but KNN-algorithm is calculated locally for every site and the local results are then transferred to a centralized site which needs to be compared. Results of experiment presented that the PPDM has the good ability of doing privacy preserving, providing efficiency and accuracy, and also relatively comparable to the traditional methods.

[8] Association-rule-mining describes the attractive relation-ship among the data. This paper is dependent on the concepts like condensation approach and the association-rule. The SMC (secure-multiparty-computation) is the transferring of data securely on network with the hiding-process that hides the sensitive association-rule that made the threat to the privacy. Currently the privacy-preserving data-mining has now become one of the very famous as it offers sharing of the sensitive or private data for the purposes of analysis. In this paper the system have proposed for increasing privacy and reducing the loss of information during sharing of data without leaking the person's identity and protect the privacy rules of the owners of database. For this approach condensation are applied for preserving the covariance details and preserving the privacy. Association-rule hiding along with the ISL and the DSR approach is utilized that is dependent on updating database-transaction such that confidence of the association-rules may be decreased.

[9] This technique offers the efficient accuracy within the reconstructing of frequent-item-sets along with the no impact on the support of the item-sets. Currently, the implementation of this technique is restricted to the local-environment that is used in companies. This work may be expanded further to work with the distributed-environment. The data is then distributed in between various system-side environments, so as for the data-mining, the mining-algorithm is needed through which the universal consequence is produced that offers the knowledge from distributed-database without being violation of the privacy.

IV. PROPOSED WORK

This section talks about the proposed work. In this work privacy is maintain by the means of Randomization and mod operation for perturbation. Algorithm of the proposed work is shown in figure2.

Input:

Original Data

Output:

Perturbed Data

Procedure:

Step 1: Read data set in A (m,n) m is number of element and n is number of attribute

Step 2: Divide data in to d1,d2 d1 having all the element but d1 is having qusi attribute and d2 having sensitive data

Step 3:for each column in d1 repeat 4 to 7

for i = 1:n

k=n-i+1;

Pick a random index from 0 to k in j

d1=swap(d1,k,j)// swap element

for i=1 to n repeat 9-10

for j=1 to noofcoulm(d1) repeat 10

d1(i,j)=mod(d1(i,j),5)

A= (d1,d2) //recombine

Figure 2: Proposed Work

The Architecture of the proposed work is shown in figure 3. As shown in below diagram.

V. RESULT ANALYSIS

The proposed work is evaluated on two parameters. First is Execution time and another is Privacy. This evaluation is done on following dataset and system configuration.

Dataset:

Title: Student Alcohol Consumption

Number of Instances: 1044

Number of Attributes: 32

System:

System: Dual Core

RAM: 4 GB

OS: 32-bits Windows 7.

Privacy: It is quantity which measures the privacy preservation of the dataset.

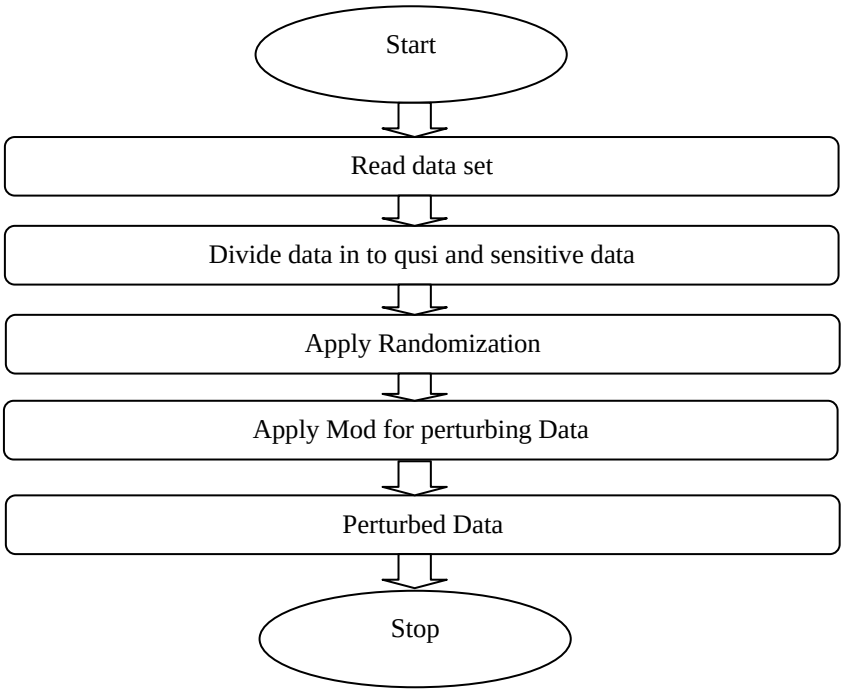

Figure 3: Architecture of the Proposed work

TABLE I: Privacy Comparison

Existing Work	Proposed Work
59.2098	118.859

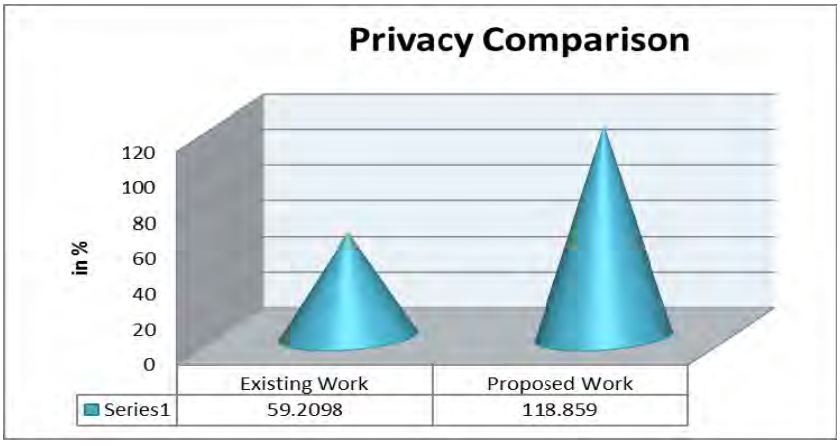

Figure 4: Privacy Comparison

Execution Time: It is a quantity which measures the execution time in seconds. !0 iterations are performed and average of those ten iteration is considered here for the comparison.

TABLE II: Execution Time Comparison

Existing Work	Proposed Work
35.6715	7.44839

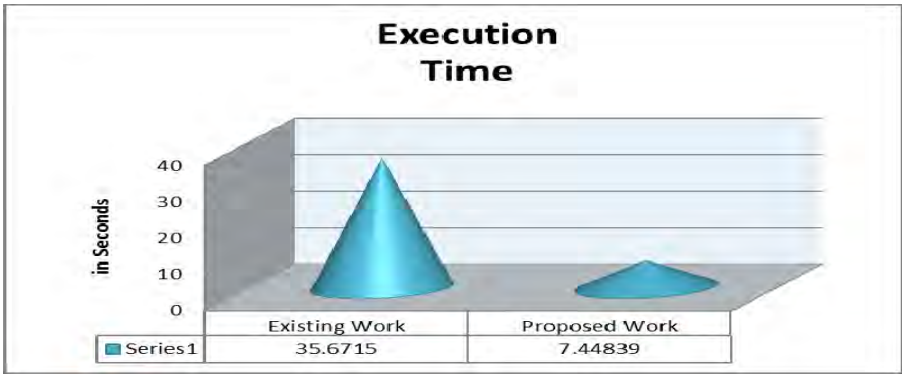

Figure 5: Execution Time Comparison

VI. CONCLUSION

In this paper, privacy issues of dataset have been described and the new approach for the privacy-preservation has also been suggested. The randomization and mod approaches are utilized to hide the sensitive data. Data-randomization approach is implemented such that the private details cannot be introduced by data-mining approaches. Vector-quantization is the recent technique for the privacy-preserving data-mining, upon implementing this encoding-procedure one may not leak the actual data therefore the privacy is got preserved. From figure 3 and 4 along with table I & II , it is clearly mentioned that the performance of the proposed work is more effective and better than existing work.

REFERENCES

[1] Manish Sharma, AtulChaudhary, Manish Mathuria, ShaliniChaudhary and Santosh Kumar, An Efficient Approach for Privacy Preserving in Data Mining, International Conference on Signal Propagation and Computer Technology (ICSPCT) IEEE 2014, pp 244-249.

[2] Archana Tomar, Vineet Richhariya, Mahendra Ku. Mishra," A Improved Privacy Preserving Algorithm Using Association Rule Mining in Centralized Database",International Journal of Advanced Technology & Engineering Research (IJATER) ISSN NO: 2250-3536 Volume 2, Issue 2, March 2012.

[3] M. Prakash, G. Singaravel, "A New Model for Privacy Preserving Sensitive Data Mining", in proceedings of ICCCNT Coimbatore, India, IEEE 2012.

[4] Anbazhagan, Dr. R. Sugumar, M. Mahendran and R. Natarajan, An Efficient Approach for Statistical Anonymization Techniques for Privacy Preserving Data Mining, International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 7, September 2012, pp. 482-485.

[5] Kun Liu, Hillol Kargupta, and Jessica Ryan, Random Projection-Based Multiplicative Data Perturbation for Privacy PreservingDistributed Data Mining, IEEE transactions on knowledge and data engineering, vol. 18, no. 1, January 2006, pp. 92-106.

[6] Neha Patel, Prof. Shrikant Lade, Prof. Ravindra Kumar Gupta, "Quasi & Sensitive Attribute Based Perturbation Technique for Privacy Preservation", Patel et al., International Journal of Advanced Research in Computer Science and Software Engineering 5(11), November- 2015, pp. 450-455.

[7] Mohasin Tamboli, Jayapal PC Bhalerao M., "Privacy Preserving Data Mining (PPDM) For Horizontally Partitioned Data", International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 4, April 2014.

[8] Supriya S. Borhade, Bipin B. Shinde, "Privacy Preserving Data Mining Using Association Rule With Condensation Approach", International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 3, March 2014.

[9] Kunwar Singh kushwah, Abhay Panwar, "A Privacy Preservation Technique Using Machine Learning Technique", International Journal of Engineering and Innovative Technology (IJEIT) Volume 4, Issue 8, February 2015.

[10] D.Karthikeswarant, V.M.Sudha, V.M.Suresh A.Javed sultan "A pattern based framework for privacy preservation through association rule mining", IEEEInternational Conference On Advances In Engineering, Science And Management (ICAESM - 2012) March 30, 31, 2012.