Study and Analysis of Privacy Preserving Data Mining Techniques in Current Scenario

Amrata Dixit
M. Tech. Scholar,
Dept. of Computer Science and Eng.
LNCT Jabalpur, M. P.

Sujeet Tiwari
Assistant Professor
Dept of Computer Science and Eng.
LNCT Jabalpur, M. P.

Naazish Rahim
Professor & Head, Dept of Computer Science and Eng.
LNCT Jabalpur, M. P.

***Abstract*—As with the increasing demand of the data mining techniques the privacy preserving is consider as the important factor. In this paper we discuss to provide the security during data mining technique without compromised the utilization of the data. Individuals are well familiar with the security threats and are averse to share their personal information on network. Because of this the outcome of data mining are negligent. By taking into consideration the privacy factor several techniques were proposed, but these methods are in the state of infancy.The fame of these PPDM techniques is based on the accuracy achieved by these algorithm and performance of the algorithm.Nevertheless there is no such algorithm exist which achieve accuracy as well as better performance. However algorithm those perform better may lack in accuracy factor or vice-versa. In this paper we discuss about various methods for ensuring security in data mining and also explore the direction of the future research work.**

Keywords—Data Mining, Privacy preserving data mining, privacy issue

I. INTRODUCTION

With the rapid development of the storage technology and understanding the importance of data collection enable any organization to aggregate a large volume of the data. From the observed analysis it has been found that the growth of the information doubles in every 20 months and this lead to an increasing size of the database [1].From this large database retrieving the useful information is one of our challenging tasks. However because of this gigantic size of the database the primitive tools and application of data analysis are not applicable. Thus for processing this large volume of data, data mining methods mingle the primitive tools of analysis with the ingenious algorithm. The essential material is the transactional data and the data mining algorithm have the ability to refine this enormous database and filter out the desired information [1]. It is the process of extracting useful information from the heap of the available database. These data mining methods make use of the computing devices for extracting the valuable, vital, understandable information from the large available database [2, 3]. The importance of data mining is not limited to one specific area, but it has its roots in different direction such as in banking, education, telecommunication, medical science, finance, commerce etc. Along with the usage of the retrieved information (managing a list of all available components along with the available quantity and prices, up to date profile of the customer and its purchasing) this extracted dataset when merged with the available data mining tools [4] are also helpful in obtaining the hidden knowledge that was unknown ahead of time. This retrieval of the hidden data knowledge is useful for market holders in making business discussion and planning different market strategies . Data mining evolve as one of the valuable method and is extending its roots from one sector to another. However along with it flourishing the privacy of the individual is a major concern during the aggregation, processing, mining of data [6]. Thus we says that the preserving an individual privacy is an important task during the data mining process. As in data mining method, retrieved valuable information are susceptible to different type of the attacks, misuse by unauthorized user, and others [7, 8]. Thus preserving privacy in the data mining technique is an important task for continuing the flourishing root of its.Actually there is no need fortrespass the security factor in data mining methods. The aim of the data mining technique is to make the general among population rather than disclosing the individual identity.The working procedure of the data mining method is the factor of introducing privacy, as it works by calculating individual information. Thus, emerges the need for protecting the individual privacy. The aim of introducing this privacy preserving algorithm is to decrease the risk of improper use of individual information, and generate the same result as it was generated before the application of this privacy preserving policies.

This paper presents the different issues of the privacy preserving data mining methods. This paper is categorized into 5 sections. Following the introductory section is the section 2 which described the framework of the PPDM method and section 3 illustrate the different classification method of the PPDM. In section 4 we discuss the various criteria upon which the performance of the algorithm is evaluated, and based upon these factors we evaluate the performance of some algorithm of PPDM. And in section 5 we conclude our work by presenting the direction of analysis for the future work.

II. PPDM FRAMEWORK

Figure 1 represents the systematic diagram of framework of PPDM methods. In data mining techniques or in extraction knowledge pattern from database, process the data which is aggregated from the various organization and stored in the respective databases. After this stored data or the information is converted into a form which is appropriate for analysis, stored in data ware house, in which different data mining algorithm get operated for extracting the useful information or discovering knowledge. By taken into consideration the privacy protection different models have to be proposed.Ensuring privacy is not the method acquire in one step, even though it should be ensure in complete procedure from aggregation of information to the generation of knowledge pattern. The below diagram represent the three level of the security factor are taken into consideration. During the level1 the data are aggregated from single or the multiple data bases and it is transformed into the form so that they make applicable for the purpose of the analysis.At this stage we need the privacy factor to consider. Researcher's proposed different techniques operated at this stage, but a large part of them deal with converting these raw data into an analysis form.

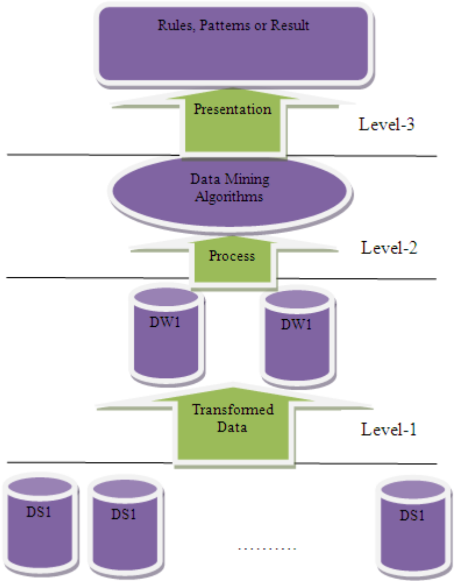
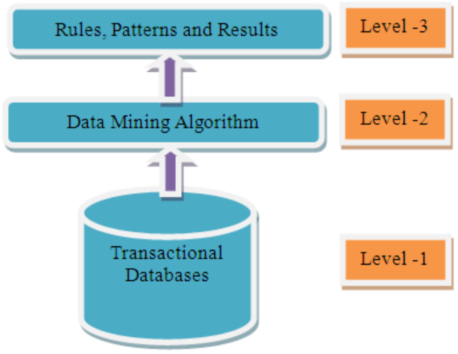


Fig. 1: A framework for PPDM

After the level 1 is the level 2 in this level the data from its warehouses are introduced with the different process that sanitized it so that these data are now disclosed to several unknown parties.Different methods applicable at this level are blocking, suppression,
perturbation, modification, generalization, sampling etc. After this the data mining techniques are applied so that we enable to retrieve the useful information and discover knowledge patterns from it Classification of privacy preserving.

III. TECHNIQUES

The three different levels which introduce the privacy preserving ability in the existing data mining techniques are as discussed below: [3].



At level 1, different methods [4], [5], [6], [7], [8], [9], [10], [11] discussed below are applicable on the available database or the raw data so to prevent the user from extracting the critical or sensitive data. After level 1 i.e level 2 we indulge both the privacy preserving methods [13-26] with the existing data mining algorithms thus ensuring individual privacy. At last the level 3 in which different researches proposed various techniques [31], [32] for the output obtained from the data mining methods.And this obtained output is shared among the parties.

2.1 Privacy Preserving At Level1 (Raw Data or Databases)

These are some way proposed by Clifton et al. [4] to ensure the privacy protection at level 1 are:

- Limiting the access
- Fuzz the data
- Eliminate the unnecessary data
- Augment the data
- Audit

In this level the two popular known methods for preserving privacy of an individual are:

- The randomization method
- The anonymization method

2.1.1 The randomization method

In this randomization method the actual or the original data is add up with some factor which may be a noise or any random number. The added noise factor is should be large enough so as to ensure that this would not be reconstructed by any unauthorized party.In this method the procedure for aggregating the data consist of the following step [4], thus the systematic representation of it is shown in figure 2

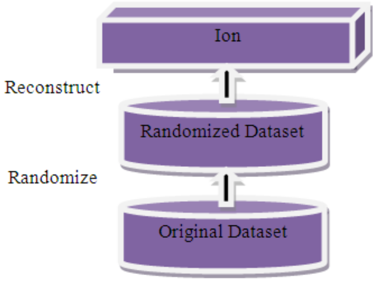


Fig.2.The model of Randomization

During first step the data providers randomized data and transmitted it to the receiver end. In next step data receiver by using the reconstruction algorithm, evaluate the actual distribution of the data. The two popular and well known techniques for randomization are the Random noise based and randomized response.

2.1.2 The anonymization method

In case of demographic analysis or in health research sector there is need for releasing the specific information about a person(also referred as micro data) by some organization such as the Health center or Government sector [12].Because of this, the condition may arise which accidently release the sensitive information about an individual. Thus this jeopardy situation of connection private information is handled with great care. For this we introduce different privacy preserving methods to protect and safeguard the individual sensitive information.

The three different types of micro data attributes are 1) identifiers (IDs) 2) quasi-identifiers (QIs), 3) sensitive attributes (SAs). The first type of the micro data -identifier (e.g., name, social security number and other) is utilized in identification of a tuple uniquely. Thus this type of identifier must be covered or secured.Quasi identifier (e.g., date of birth, zip code, gender) these micro data are combined together and can be treated as identifier in carriage of exterior information. SAs (e.g., salary, Criminal charge, diseases) this type of micro data are also covered so we have a surety that no one obtained its connection with other specific person.

TABLE 1.ORIGINAL TABLE [26]

Name	Race	Birth	Sex	Zip	Disease
Alice	Blank	1965-3-18	M	02141	Flu
Bob	Blank	1965-5-1	M	02142	Cancer
David	Blank	1966-6-10	M	02135	Obesity
Helen	Blank	1966-7-15	M	02137	Gastritis
Jane	White	1968-3-20	F	02139	HIV
Paul	White	1968-4-1	F	02138	Cancer

TABLE 2.e ANONYMIZATION OF TABLE 1[26]

Name	Race	Birth	Sex	Zip	Disease
Alice	Blank	1965	M	0214#	Flu
Bob	Blank	1965	M	0214#	Cancer
David	Blank	1966	M	0213#	Obesity
Helen	Blank	1966	M	0213#	Gastritis
Jane	White	1968	F	0213#	HIV
Paul	White	1968	F	0213#	Cancer

Bayardo and Agrawal [14] introduce an optimal technique, which starts from the completely generalized table and specialized this into a minimal k-anonymous table.

LeFevre et al. [15] introduced a technique which uses the bottom up strategy and pre-evaluation concept. Fung et al. [16] to obtain the k-anonymous table they suggest top-down strategy.

Sweeney [17] introduces approximation techniques for the K–anonymity and also considered the problem of k –anonymityas NP hard. Different models were proposed namely p-sensitivek-anonymity [18], t-closeness [19], and M-invariance [20] with the objective, to solve the problem of k –anonymity. Xiao and Tao [21] introduce a technique that fulfill every one requirement by performing minimum generalization and in turn preserve the large volume of data from the actual data set.References [22-25] introduced the clustered based methods which are helpful in reducing significantly the loss of the information.

K-anonymity as a major topic for the research and also present the various issues which are taken into considerationsuch as combining this with an existing techniques of the data mining. This ensures that the transformed data which we obtained is correct and reduces the loss of data during transformation.

2.2 Privacy Preserving At Level2 (Data mining algorithms and techniques)

With the rapid usage of an internet, individual are interested in performing the data mining activity in jointly manner. However, during the protected computation of different parties there is a chance of getting the critical information in the hand of an untrusted entity or even the competitors. Here we suggest the two popular representation of the distributed data mining such as horizontal and vertical data mining[26].In horizontal partition each site equipped with full information on different entities set. On the other hand in vertical partitioning different information is equipped in each different site.

The different techniques which are named above used a protocol for performing encryption operation as Secure Multiparty Computation (SMC) technology. The two basic model of SMC are as below,

- ❖ Semi-Honest model
- ❖ Malicious Model

2.3 Privacy Preserving At Level3 (output of data mining algorithm and techniques)

At level 3 more protection and security is provided by the privacy preserving data mining technique because at this level no database or raw data are shared among parties. At this level the obtained output of the data mining techniques are shared among the parties.

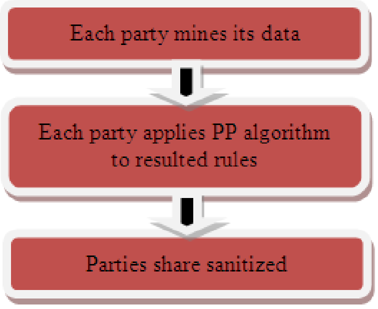

Each party mines its data
Each party applies PP algorithm to resulted rules
Parties share sanitized

Fig 3.Parties sharing rules

This systematic representation in fig.3 indicate that the parties shared the discovered knowledge from the databases, by eliminating the data which are sensitive , share the rule, while no parties known that which knowledge is belong to which intended party. The major challenge which introduce here is the releasing all the discovered patterns which are not critical or delicate.

IV. Evaluation Of Privacy Preserving Algorithms

In this section we evaluate the outcome of the privacy preserving data mining algorithm with the help of the following parameters as discussed below [19]:

- ❖ Performance: The performance factor of these algorithm is estimated as the time required by the algorithm to reach privacy criteria
- ❖ Data Utility: This parameter is the estimation, of the loss of information or loss of functionality in generating the result , which easily obtained in the absence of the PPDM techniques
- ❖ Uncertainty level: This is the estimation of the uncertainty level by which the critical or secret information which is covered or hidden can be forecast.

- Resistance: This is the measure of the tolerant factor which must have the PPDM algorithm against the various data mining techniques and its proposed models

V. Conclusion

In data mining privacy and accuracy shows denial between them. It is inauspicious if one of them is managed in regard to another. In view of this we revised a large amount of already used PPDM techniques. Decisively it is observed that there is not even a method for privacy preserving data mining technique for feasible principles which accomplishment, services, intricacy, charges and resilience contrary to are given mining algorithms. It is quite possible to achieve better algorithm in comparison to others.

References

[1] Ahmed HajYasien,"Preserving Privacy In Association Rule Mining", a Thesis submitted to Griffith University, June 2007.

[2] N.Zhang,"Privacy-preserving Data Mining", *Texas A&M University*, pp. 19-25, 2006.

[3] R.Agrawal, R.Srikant, "Privacy-Preserving Data Ming", *ACM SIGMOD Record*, New York, vol.29, no.2, Pp.439- 450, 2000.

[4] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", *In Proceedings of the 3rd International Conference on Data Mining*, pp.99-106, 2003.

[5] Z. Huang, W. Du, B. Chen, "Deriving Private Information from Randomized Data", *In Proceedings of the ACM SIGMOD Conference on Management of Data*, Baltimore, Maryland,USA, pp.37-48, 2005.

[6] J D. Agrawal, C.C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms", *In Proceedings of the 20th ACM SIGMODSIGACTSIGART Symposium on Principles of Database Systems*, pp.247-255, 2001.

[7] S.L.Warner, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", J. Am. Stat. Assoc., vol.60, no.309, pp.63-69, 1965.

I.

[8] W. Du, Z. Zhan, "Using Randomized Response Techniques for Privacy Preserving Data Mining", *In Proceedings 9th ACM SIG KDD International Conference on knowledge Discovery and Data Mining*, pp. 505-510, 2003.

[9] L.Guo, S.Guo X.Wu,"Privacy Preserving Market Basket Data Analysis", *in proceedings the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp.103-114, 2007.

[10] Dimitris Sacharidis, Kyriakos Mouratidis, and Dimitris Papadias," k-Anonymity in the Presence of External Databases", *IEEE Transactions On Knowledge And Data Engineering*, vol. 22, no. 3, March 2010.

[11] L.Sweeney,"k-Anonymity: A Model for Protecting Privacy", *International Journal of Uncertainity.Fuzziness and Knowledge-based Systems*.vol.10, no.5, pp.557-570, 2002.

[12] R.Bayando, R.Agarwal,"Data Privacy through Optimal k-Anonymization", *In proceedings of the 21st International Conference on Data Engineering*, pp.217-228, 2005.

[13] K.Lefevre, J.Dewittd, R.Ramakrishnan,"Incognito: Efficient K-anonymity", *In Proceedings of the 2005 ACM SIGMOD International Conference on Management of data*,pp.49-60, 2005.

[14] B.Fung, K.Wang, P.Yu,"Top-down specialization for Information and Privacy preservation. *In proceedings of the 21st IEE International Conference on Data Engineering*, pp.205-216, 2005.

[15] L.Sweeney," Achieving k-Anonymity Privacy Protection Using Generalization and Suppression". *International Journal on Uncertainity.Fuzziness and Knowledge-based Systems*, vol.10, no.5, pp.571-588, 2002.

[16] T.Truta, B.Vinay,"Privacy Protection-Sensitive kAnonymity Property", *in proceedings of the 22nd International Conference on data Engineering Workshops*, pp.94-103, 2006.

[17] N.H.Li, T.C.Li,"t-Closeness: Privacy beyond k-Anonymity and l-Diversity*", in proceedings of the 23rd International Conference on data Engineering*, pp.106-115, 2007.

[18] X.K.Xiao, Y.F.Tao,"M-Invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets", *In proceedings of the ACM Conference on Management of data (SIGMOD)*, pp.689-700, 2007.

[19] X.K.Xiao, Y.F.Tao,"Personalized Privacy Preservation", *in proceedings of the ACM Conference on Management of data (SIGMOD)*, pp.229-240, 2006.

[20] G.Loukides, J.H.shao,"An Efficient Clustering Algorithm for k-Anonymization", *International Journal of Computer Science and Technology*, vol23, no.2, pp.188-202, 2008.

[21] J.L.Lin,M.C.Wei,"Genetic Algorithm-Based Clustering Approach fork-Anonymization",*International Journal of Expert Systems with Applications*,vol.26,no.6,pp.9784-9792,2009.

[22] L.J.Lu, X.J.Ye,"An Improved Weighted-Feature Clustering Algorithm for k-Anonymity", *In Proceedings of the 5th International conference on Information assurance and Security*, pp.415-419, 2009

[23] Z.H.Wang, J.Xu, W.Wang, B.L.shi,"Clustering Based Approach for Data Anonymization*", Journal of Software*, vol.21, no.4, pp.680-693, 2010..

[24] Pingshui WANG, "Survey on Privacy-Preserving Data mining", *International Journal of Digital Content Technology and its Application*.Volume4, Number 9, December2010.

[25] M.Kantarcioglu,C.Clifton,"Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data",*IEEE Transactions on Knowledge and Data Engineering*,vol.16,no.9,pp.1026-1037,2004.

[26] J.Vaidhya,C.Clifton,"Privacy Preserving Association Rule Mining in Vertically Partitioned Data", *In proceedings of the 8th ACM SIGKDD International Conference on knowledge Discovery and DataMining*,pp.639-644,2002.

Ioannidis, A.Grama, M.J.Atallah, "A Secure Protocol Computing Dot-Products in Clustered and Distributed Environments", *in proceedings of the 31st International Conference on Parallel Processing*, pp.379-384, 2002.

[27] J.Vaida,C.Clifton,"Privacy Preserving Association Rule Mining in Vertically partitioned Data", *In Proceedings of the 8th ACM SIGKDD International Conference on knowledge Discovery and Data Mining*,pp.639-644,200