International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 3, March 2016

Survey on Privacy Preserving Data Mining: Techniques and Application

[1] Vivek Jawanjal, [2] Shubham Pawar, [3] Akshay Kamble, [4] Prof. V.S. Mahalle
[1][2][3] BE. Final Year CSE. [4] Computer Science & Engineering,
[1][2][3][4] Department of CSE, Shri Sant Gahanna Maharaj College of Engineering (SSGMCE) Shegaon Maharashtra, India
[1] vivekjawanjal3@gmail.com [2] shubhampawar006@gmail.com, [3] akshaykamble247@gmail.com
[4] vsmahalle@gmail.com

Abstract:-- **Data mining is the process of extraction of data from large amount of database. One of the most important topics is nowadays in research community is Privacy preserving data mining (PPDM). The goal of privacy preserving data mining is to extract relevant knowledge from large amounts of data while protecting at the same time sensitive information. To solve such problems there are number of methods and techniques have been proposed for protecting sensitive information. This paper provides a wide survey of different privacy preserving data mining algorithms and A tabular comparison of different technique is presented.**

***Keywords*—Data Mining; Privacy Preserving; Sensitive Data; Privacy Preserving Techniques;**

I. INTRODUCTION

Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories. By performing data mining, interesting knowledge, high-level information can be extracted from database.

The main goal of data mining is to extract knowledge and new patterns from large data sets in human understandable structure. For data mining computations we have to first collect data without much concern about privacy of data. Because of privacy concerns some people are not giving right information. Therefore Privacy preserving data mining has becoming important field of research. In order to make a publicly system secure, we must ensure that not only private sensitive data have been trimmed out, but also that certain inference channels should be blocked with respect to privacy.

The paper is organized as follows. In Section 1, we give the basic concept of data mining and privacy. In Section 2, we describe Privacy Preserving data mining with its framework. Section 3 provides some of the Application in this field. Section 4 contains different techniques with their limitations. A tabular comparison of different techniques of PPDM given by different authors is shown in section 5. And finally we conclude in Section 6.

II. PRIVACY PRESERVING DATA MINING

Privacy preserving data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. People have become well aware of the privacy intrusions on their personal data and are very reluctant to share their sensitive information.

In figure 1, framework for privacy preserving Data Mining is shown. Data from different data sources or operational systems are collected and are preprocessed using ETL tools. This transformed and clean data from Level 1 is stored in the data warehouse. Data in data warehouse is used for mining. In level 2, data mining algorithms are used to find patterns and discover knowledge from the historical data. After mining privacy preservation techniques are used to protect data from unauthorized access. Sensitive data of an individual can be prevented from being misused.

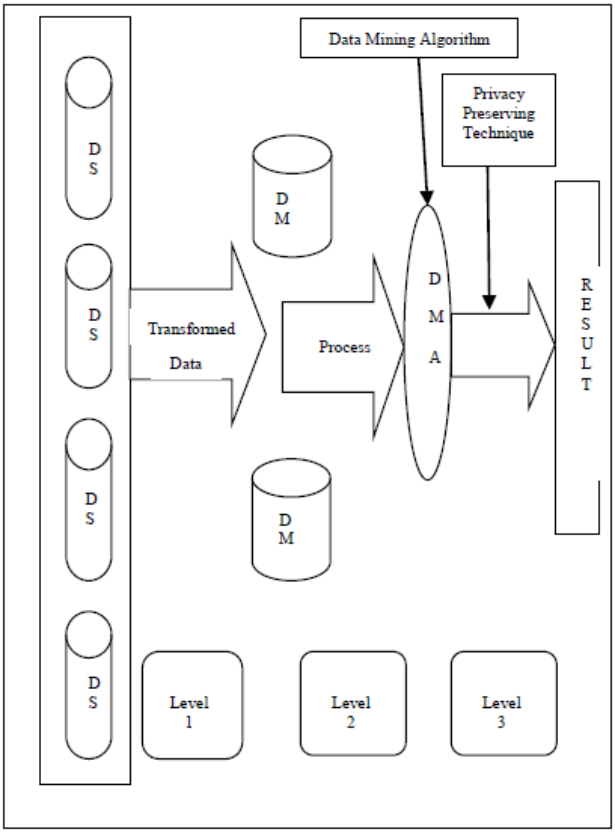

Fig. 1 Framework of privacy preserving data mining

III. APPLICATION

A. Medical Database:-Scrub & Data fly System. The scrub System design for identification of medical record which in the form of text data. Medical records are typically in the form of text which contains information about patient. Data fly was design to avoid identification of the subject of medical record.

B. Bioterrorism Application: - In Order to identify disease affected area it is necessary to track incidences of those common diseases . That Medical data would need to be report to public health agency. However there are some diseases that are not reportable disease by law. E.g. Common Respiratory Diseases.

C. Homeland Security Applications:-

- Credential validation problem
- Identity theft
- Web camera surveillance
- Vide-surveillance
- The watch –list problem

IV. PRIVACY PRESERVING DATA MINING TECHNIQUES

In this section, we focus on different privacy preserving data mining techniques such as a cryptographic technique, blocking based, hybrid technique etc.

A. Data Perturbation

Data Perturbation is a technique for modifying data using random process. In this technique by changing the sensitive data by adding, subtracting or any mathematical formula. Different data types are used: character type, Boolean type, classification type and integer. In this preprocess the original data set required. According to the preprocessing of data it classified into attribute coding and obtains coded data sets. Discrete formula prescribed is: A(max) – A(min)/n = length. Where A is continuous attribute, n is number of discrete and length is discrete interval length. Data perturbation cannot reconstruct the original data; it can reconstruct only distribution data.

It is important to secure the sensitive information therefore data perturbation play an role for preserving sensitive information. Distortion can done by different method applying such as a by adding unknown values, adding noise etc. In some technique it is difficult to protect the sensitive information, to overcome this problem new algorithm are developed which is able to reconstruct the distributed data.

An new approach is develop, It is based on singular value decomposition (SVD) and sparsified singular value distribution (SSVD) technique. In this matrices is introduced to compare the original datasets and distorted dataset. SSVD used for data utility, SVD method add noise to make perturbed data. Perturbation has drawback is that each data dimension is reconstructed independently.

B. Blocking based technique

In blocking based technique there is a sensitive classification rule, it is used to hiding the sensitive data from outsiders. In this technique two steps are used for protecting the privacy. First is to identify transaction of sensitive rule and second is to replace the known values to the unknown values (?). In this, there is a scanning of original database and then identifying the transactions supporting sensitive rule, and then for each record or transaction, algorithm replaces the sensitive information with an a unknown values. This techniques is used those applications that can save unknown values to the some

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 3, March 2016

kind of attribute. Generally we can hide the actual values by just replacing the '0' by '1' or '1' by '0' or with unknown values (?) in transaction. Blocking based techniques aim is that to preserve the sensitive data from unauthorized access. There are different sensitive rule for according requirement. For every sensitive information scanning of original database is needed. When the left side pair of rule is a subset of attribute values pair of transaction and the right hand side of the rule should be same as the attribute class of the transaction then only transaction supports any rule. These steps are continuing till the sensitive data are not hidden by the unknown values.

C. Cryptographic Technique

Cryptographic is a technique through which we can encrypt the sensitive data. It is used for preserving the data. This method is very popular because is provide security and safety for sensitive attributes. There are many algorithm are available for cryptographic technique. But this method has may drawbacks such as, it fails to protect the output of computation. It prevents privacy leakage of computation. This algorithm does not provide accurate results when two or more parties talks. It is very difficult to technique for large amount of database. Final result may break the privacy of individual record.

D. Randomization Technique

The Randomization method is a popular method in current privacy preserving data mining. In which noise is added to the data in order to mask the attribute values of records, this technique provides simple and effective way. This can be easily implemented at data collection phase for privacy preserving data mining by resisting user from learning sensitive data.

Randomization techniques consists the following steps:
1. Data providers randomize their data and transmit the randomized data to the data receiver.
2. Data receiver estimates the original distribution of the data by employing a distribution reconstruction algorithm.

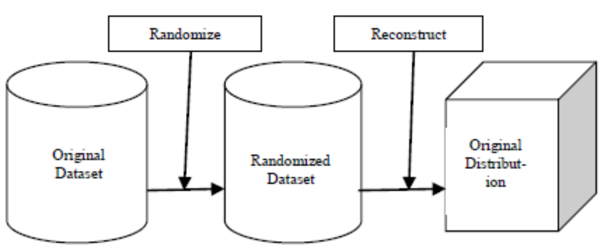

Figure 2. The Model of Randomization

E. Hybrid Technique

Hybrid is new technique in which we combine two or more privacy preserving data mining technique to preserve the data. A hybrid technique in which used randomization and generalization. In this approach first they randomize the original data and then generalized the modified or randomized data. This technique protect private data with better accuracy, also it can reconstruct original data and provide data with no information loss, makes usability of data and also data is reconstructed. Many other technique are used to make hybrid technique such as cryptographic technique, data perturbation etc.

V. COMPARISON TABLE

TABLE 1
COMPARISON OF DIFFERENT TECHNIQUES

S. No	Author	Year of Publication	Technique Used For PPDM	Approach	Result And Accuracy
1	Y.Lindell, B.Pinks[11]	2000	Cryptographic Technique	A technique through which sensitive data can be encrypted. There is also a proper toolset for algorithm of cryptography.	This approach is especially difficult to scale when more than a few parties are involved . Also it does to hold good for large database.
2	L.Sweeney[22]	2002	K-Anonymity	A record from a dataset cannot distinguished from at least k-1 records whose data is also in the data set.	K-Anonymity Approach is able to preserve privacy.
3	J.Vaidya, C Clifton[20]	2002	Association Rule	Distribution of data vertically into segment.	Distribution based Association Rule Data Mining provide privacy.
4	Hillol Kargupta, Souptick Datta, Qi Wing And Krishnamoorthy Sivakumar[7]	2003	Data Perturbation	They tried to preserve the privacy by adding random noise , while making sure that the random noise still preserve the signal from the data so that the pattern can still be accurately estimated.	Randomization based technique are used to generate random matrices.
5	CharuC.Aggrwa, Philip S. Yu[12]	2004	Condensation Approach	This approach works with pseudo-data rather than with modification of original data ,this helps in better preservation of privacy than techniques which simply use modification of the original data.	The use of pseudo data no longer necessitates the redesign of data mining algorithm, since they have the same format as the original data.
6	A. Machanavajjihala, J. Gehrke, D. Kifer and M. Venkitasubramaniam [24]	2006	L-Diversity Algorithm	If there are well represented values for sensitive attribute then that class is said to have L-Diversity	It is better than K-Anonymity .
7	Slava Kisilevich,Lior Rokach, Yuval Elovici, Bracha Shapira[21]	2010	Anonymization	Anonymization is technique for hiding individual, sensitive data from owner records. K-anonymity is used for generalization and suppression for data hiding.	Background knowledge and homogeneity attacks of K-Anonymity Algorithm do not preserve sensitivity of an individual.
8	P. Deivanai,J. Jesu Vedha Nayahi and V. Kavita[3]	2011	Hybrid Approach	Hybrid approach is combination of different technique which combine to give an integrated result.	It uses Anonymization and Suppression to preserve data.
9	George Mathew, Zoran Obradovic[25]	2011	Decision Tree	An approach which is technical methodology and should give judgmental knowledge.	A graph based framework for preserving patient's sensitive information.
10	Anita parmar, Udai Pratap Rao , Dhiren R. Patel[10]	2011	Blocking Based Technique	The sanitized dataset is generated from which sensitive classification rule are no longer mined.	Unknown value helps in preserving privacy but reconstruction of original dataset is quite difficult.
11	Sara Mumtaz, Azhar Rauf And Shah Khusro[16]	2011	Distortion Based Perturbation Technique in OLAP Data Cube	Data perturbation technique is also called which is also called uniformly adjusted distortion is proposed which initially distorts one cell of a cube and then distortion occurs in whole cube.	This distribution of distortion technique not only preserve ,but also provide utmost accuracy with range sum queries and high availability.
12	Hsiang cheh Huang Wai –chi Fangi[17]	2011	Histogram Based Reversible Data Hiding	A concept of reversibility which states that an original data can easily be hidden data can also be recovered perfectly . sensitive data is embedded into medical images which is very good technique for hiding secret data .	Histogram technique is basically used for X-ray or CT medical images and it has the potential to be integrated into database for managing the medical images in the hospital.
13	Jinfei Liu, Jun Luo and Joshua Zhexue Huange [5]	2011	Rating Based Privacy Preservation	A novel algorithm which overcomes the curse of dimensionality and provides privacy.	It is better than K-Anonymity and L-Diversity.
14	Khaled Alotaibi, V. j. Rayward-Smith,Wenjia Wang and Beatriz de la Iglesia[6]	2012	Multidimensional Scaling	A non linear dimensionality reduction technique used to project data on lower dimensional space.	The application of non metric MDS transformation works efficiently and hence produces better result.
15	Elahe Ghasemi Komishani and Mahdi Abadi[8]	2012	Trajectory data	Approach for privacy preservation in trajectory data publishing in which trajectories and sensitive attributes are generalized with respect to different privacy requirements of moving objects.	It is able to provide personalized privacy preservation in trajectory data publishing, but also it is resistant to all three identity linkage, attribute linkage and similarity attacks.
16	Tharveer Jahan, Dr G. Narsimha and Dr. C. V. Guru Rao[15]	2012	Data Perturbation Using SSVD	An analyzing system used to transfer original dataset into distorted data set using sparsified singular value Decomposition.	Use of Sparsified SVD than SVD is more successful.
17	D.Karthikeswarant, V.M.Sudha, V.M.Suresh and A.J.Sultan[19]	2012	Association Rule	Sanitize datasets using sliding window	A novel approach that modifies that modifies

				algorithm and preserves data.	the database to hide sensitive rules.
18	M.N.Kumbhar and R.Kharat[18]	2012	Association Rule By Horizontal and Vertical Distribution.	Different approach in the field of Association rule are reviewed.	The Performance of all models is analyzed in terms of privacy, security and communication.
19	Savita Lohiya and Lata Ragha[9]	2012	Hybrid Approach	A combination of K-Anonymity and Randomization.	It has better accuracy and original data can be reconstructed.
20	Martin Beck And Michel March Ofse[26]	2012	Anonymizing Demonstrator	Making a demonstrator with user friendly interface and performs Anonymization.	Swapping and Recording can be applied to enhance the utility.
21	Mi Wen, Rongxing Lu,Jinguhon Lei,Xiaohui Liang[27]	2013	(ECQ) Efficient Conjunctive Query	ECQ can protect the data confidentiality and integrity, as well as data and query privacy	Achieve Conjunctive query without data and query privacy leakage. ECQ is more efficient in terms users computation cost.
22	Mittal D, Kaur D, Aggarwal A[28]	2014	K-Means Algorithm	It Prevents Intermediate data leakage in the process of Computation while maintaining the correctness and validity of data mining process.	This method is used to solve the privacy issue of the cloud.

Algorithms are classified on the basis of performance, utility, cost, complexity, tolerance against data mining algorithms etc.

VI. CONCLUSION

In today's world, privacy is major concern to protect the sensitive data. People are concerned about that they do not share the sensitive information. Our survey is focuses on the on existing literature present in field of the privacy preserving data mining. From our analysis, we found that there is no single technique in all domain consistent. All method perform in a different way depending on the type of data and a type of application or domain. But still from analysis, we conclude that Cryptography and Random Data Perturbation method better than existing method and Cryptography technique is best for encryption of sensitive data. Perturbation helps to preserve data hence sensitivity is maintained.

REFERENCES

[1] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*", 2nd ed.,The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor 2006.

[2] M. B. Malik, M. A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", in *proceedings of Third International Conference on Computer and Communication Technology*, IEEE 2012.

[3] P.Deivanai, J. Jesu Vedha Nayahi and V.Kavitha," A Hybrid Data Anonymization integrated with Suppression for Preserving Privacy in mining multi party data" in *proceedings of International Conference on Recent Trends in Information Technology*, IEEE 2011.

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

[5] J. Liu, J. Luo and J. Z. Huang, "Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity requirements", in *proceedings of 11th IEEE International Conference on Data Mining Workshops*, IEEE 2011.

[6] K. Alotaibi, V. J. Rayward-Smith, W. Wang and Beatriz de la Iglesia, "Non-linear Dimensionality Reduction for Privacy-Preserving Data Classification" in *proceedings of 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security,Risk and Trust*, IEEE 2012.

[7] 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.

[8] E. G. Komishani and M. Abadi,"A Generalization-Based Approach for Personalized Privacy Preservation in Trajectory Data Publishing", *in proceedings of 6'th International Symposium on Telecommunications (IST'2012)*, IEEE 2012.

[9] S. Lohiya and L. Ragha, "Privacy Preserving in Data Mining Using Hybrid Approach", in *proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks*, IEEE 2012.

[10] A. Parmar, U. P. Rao, D. R. Patel, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Database" , in *proceedings of International Symposium on Computer Science and Society*, IEEE 2011.

[11] Y. Lindell, B.Pinkas, "Privacy preserving data mining", in *proceedings of Journal of Cryptology*, 5(3), 2000.

[12] C. Aggarwal , P.S. Yu, "A condensation approach to privacy preserving data mining", in *proceedings of International Conference on Extending Database Technology (EDBT)*, pp. 183–199, 2004. 746

[13] R. Agrawal and A. Srikant, " Privacy-preserving data mining", in *proceedings of SIGMOD00*, pp. 439-450.

[14] Evfimievski, A.Srikant, R.Agrawal, and Gehrke , "Privacy preserving mining of association rules", in proceedings of KDD02, pp. 217-228.

[15] 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.

[16] S. Mumtaz, A. Rauf and S. Khusro, "*A Distortion Based Technique for Preserving Privacy in OLAP Data Cube*", in proceedings of 978-1-61284-941-6/11/$26.00, IEEE 2011.

[17] H.C. Huang, W.C. Fang, "*Integrity Preservation and Privacy Protection for Medical Images with Histogram-Based Reversible Data Hiding*", in proceedings of 978-1-4577-0422-2/11/$26.00_c, IEEE 2011.

[18] M. N. Kumbhar and R. Kharat, "*Privacy Preserving Mining of Association Rules on horizontally and Vertically Partitioned Data: A Review Paper*", in proceedings of 978-1-4673-5116-4/12/$31.00_c, IEEE 2012.

[19] D.Karthikeswarant, V.M.Sudha, V.M.Suresh and A.J. Sultan, "A Pattern based framework for privacy preservation through Association rule Mining" in *proceedings of International Conference On Advances In Engineering, Science And Management (ICAESM -2012)*, IEEE 2012.

[20] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data", in *The Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, CA, July 2002*, IEEE 2002.

[21] Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira, "Efficient Multi-Dimensional Suppression for K-Anonymity", *in proceedings of IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 3. (March 2010), pp. 334-347, IEEE 2010.

[22] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", in *proceedings of Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2002.

[23] The free dictionary.Homepage on Privacy [Online]. Available:http://www.thefreedictionary.com/privacy.

[24] A. Machanavajjhala, J.Gehrke, D. Kifer and M. Venkitasubramaniam, "I-Diversity: Privacy Beyond k-Anonymity", *Proc. Int'l Con! Data Eng. (ICDE)*, p. 24, 2006.

[25] G. Mathew, Z. Obradovic," A Privacy-Preserving Framework for Distributed Clinical Decision Support", *in proceedings of 978-1-61284-852-5/11/$26.00 ©2011 IEEE*.

[26] Martin Beck and Michael Marh¨ofer," Privacy-Preserving Data Mining Demonstrator", in *proceedings of 16th International Conference on Intelligence in Next Generation Networks*, IEEE 2012.

[27] Mi Wen, Rongxing Lu ; Jingshen Lei ; Xiaohui Liang , 2013,ECQ: An Efficient Conjunctive Query scheme over encrypted multidimensional data in smart grid, Global Communications Conference (GLOBECOM), 2013 IEEE, 796 – 801

[28] Mittal, D. ; Kaur, D. ; Aggarwal, A., 2014 , Secure Data Mining In Cloud Using Homomorphic Encryption IEEE International Conference on Cloud Computing in Emerging Markets CCEM),2014, pp : 1 – 7