ISSN:0975-9646

A Review on Privacy Preserving Data Mining: Techniques and Research Challenges

Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita

Dept. of Computer Science Engineering

Bhagwan Parshuram Institute of Technology, Indraprastha University, India.

***Abstract*- Privacy preserving data mining deals with hiding an individual's sensitive identity without sacrificing the usability of data. It has become a very important area of concern but still this branch of research is in its infancy .People today have become well aware of the privacy intrusions of their sensitive data and are very reluctant to share their information. The major area of concern is that non-sensitive data even may deliver sensitive information, including personal information, facts or patterns. Several techniques of privacy preserving data mining have been proposed in literature. In this paper, we have studied all these state of art techniques. A tabular comparison of work done by different authors is presented. In our future work we will work on a hybrid of these techniques to preserve the privacy of sensitive data.**

***Keywords*-- data mining; privacy preserving; sensitive attributes; privacy; privacy preserving techniques.**

I. INTRODUCTION

Data Mining [1] refers to extracting or "mining" knowledge from large amounts of data. Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. By performing data mining, interesting knowledge, regularities, or high-level information can be extracted from database and viewed or browsed from different angles. The discovered knowledge can be applied to decision making, process control, information management, and query processing. Data mining is considered one of the most important frontiers in database systems and one of the most promising interdisciplinary developments in the information industry. Data mining, with its promise to efficiently discover valuable, non-obvious information from large databases, is particularly vulnerable to misuse. So, there might be a conflict between data mining and privacy.

According to the definition, Privacy is the quality or condition of being secluded from the presence or view of others [23]. On relating privacy with data mining, privacy implies keep information about individual from being available to others [4]. Privacy is a matter of concern because it may have adverse affects on someone's life. Privacy is not violated till one feels his personal information is being used negatively. Once personal information is revealed, one cannot prevent it from being misused. Let us take an example, date of birth, mother's maiden name, or sex etc. may not become a threat for an individual, but if one more attribute like the unique identification number or voter ID are also known then it may cause a serious effect like identity theft.

In this paper, we discuss different approaches and techniques in the field of Privacy Preserving Data Mining (PPDM). 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 research challenges 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 [2] has originated as an important concern with reference to the success of the 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. This may lead to the inadvertent results of the data mining. Within the constraints of privacy, several methods have been proposed but still this branch of research is in its infancy.

In figure 1, framework for privacy preserving Data Mining is shown [2]. 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.

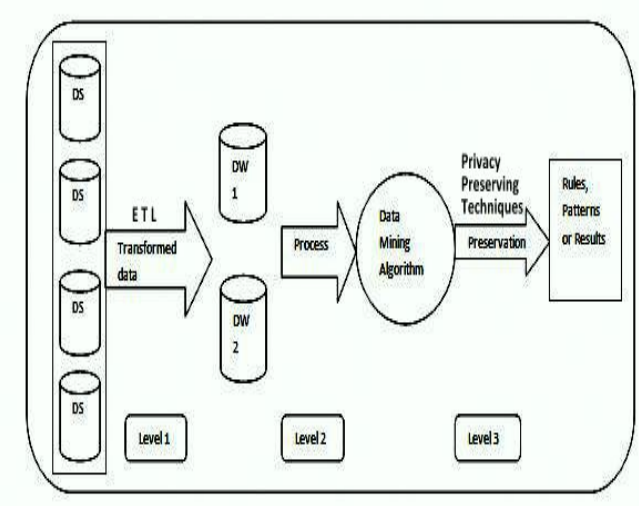

Fig. 1 Framework of privacy preserving data mining

III. RESEARCH CHALLENGES

Now- a-days, Data Mining is used in many applications. There are certain areas where data mining if used without privacy may cause serious affects .These areas are the main research challenges and are mentioned below.

A. Cyber Terrorism, Insider Threats, and External Attacks

One of the major threats people face today is Cyber Crime [4]. Since most of our information is stored on electronic media and a lot of data is also available on internet or networks. Attacks on such areas might be dangerous and devastating for an individual. For example, consider the Banking system. If hackers attack a bank's information system and empty the accounts, the bank could lose millions of dollars. Therefore security of information is a critical issue. There are two types of threats –Outsider or Insider. An attack on Information System from someone outside the organization is called outsider threat, such as hackers, hacking Bank's computer systems and causing havocs. A more critical problem is the insider threat. Insider threat can be due to an intruder present in the organization. Members of an organization have studied their policies and business practices and know every bit of the information so it can affect the organization's information assets.

B. Credit Card Fraud and Identity Theft

Another area which requires attention is detecting frauds and thefts. Frauds may be credit card frauds [4]. These can be detected by identifying purchases made of enormous amounts. A similar and a more serious theft is identity theft. Here one pretends to be an identity of another person by obtaining that person's personal information and carrying out all types of transactions under the other person's name. By the time, the owner finds out it is often far too late-the victims may already have lost millions of dollars due to identity theft.

IV. PRIVACY PRESERVING DATA MINING TECHNIQUES

In this section, we focus on the different PPDM techniques which are developed like data perturbation, blocking based, cryptographic techniques etc.

A. Data Perturbation

Data Perturbation [5][7] is a technique for modifying data using random process. This technique apparently distorts sensitive data values by changing them by adding, subtracting or any other mathematical formula. This technique can handle different data types: character type, Boolean type, classification type and integer. In discrete data [5], it is required to preprocess the original data set. The preprocessing of data is classified into attribute coding and obtaining sets coded data set. The method of average region to disperse the continuous data is used here. Discrete formula prescribed by *Sativa Lohiya and Lata Ragha* [9] is: A (max) - A (min)/n = length. A is continuous attribute, n is number of discrete, and length is the length of the discrete interval. The technique does not reconstruct the original data values, it only reconstructs the distribution.

Data distortion or data noise are different names for data perturbation. It is very important and critical to secure the sensitive data and data perturbation plays an important role in preserving the sensitive data. Distortion is done by applying different methods such as adding noise, data transpose matrix, by adding unknown values etc[15]. In some perturbation approaches it is very difficult to preserve the original data . Some of these are distribution based techniques. In order to overcome this problem, new algorithm were developed which were able to reconstruct the distributions. This means that for every individual problem in classification, clustering, or association rule mining, a new distribution based data mining algorithm needs to be developed. In [13], develops a new distribution-based data mining algorithm for the classification problem, and Vaidya [20] and Rizvi [14] develop methods for privacy-preserving association rule mining.

A new approach in data perturbation was introduced by *Jahan*, G.Narsimha and C.V Guru Rao [15].It was based on singular value decompostion(SVD) and sparsified singular value distribution(SSVD) technique and having the feature of selection to reduce the feature space. In this method, different matrices have been introduced to compare or measure the difference between original dataset and distorted dataset.SSVD is efficient approch in keeping data utitlity, SVD also works better than other standard data distortion methods which add noise to the data to make it perturbed.

The perturbation approach has a drawback. The distribution of each data dimension is reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. In many cases, a lot of relevant information for data mining algorithms such as classification is hidden in inter-attribute correlations.

B. Blocking based technique

In blocking based technique [9][10], authors state that there is a sensitive classification rule which is used for hiding sensitive data from others. In this technique, there are two steps which are used for preserving privacy. First is to identify transactions of sensitive rule and second is to replace the known values to the unknown values (?). In this technique, there is scanning of original database and identifying the transactions supporting sensitive rule. And then for each transaction, algorithm replaces the sensitive data with unknown values. This technique is applicable to those applications in which one can save unknown values for some attributes. Authors in [9] want to hide the actual values , they replace '1' by '0' or '0' by '1' or with any unknown(?) values in a specific transaction. The replacement of these values does not depend on any specific rule. The main aim of this technique is to preserve the sensitive data from unauthorized access. There may be different sensitive rules according to the requirements. For every sensitive rule, the scanning of original database is done. When the left side of the pair of rule is a subset of attribute values pair of the 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. The algorithm replaces unknown values in the place of attribute for every transaction which supports that sensitive rule. These steps will continue till all the sensitive attributes are hidden by the unknown values.

C. Cryptographic Technique

Cryptography is a technique through which sensitive data can be encrypted. It is a good technique to preserve the data. In [11] , authors introduced cryptographic technique which is very popular because it provides security and safety of sensitive attributes . There are different algorithms of cryptography available .But this method has many disadvantages. It fails to protect the output of computation. It prevents privacy leakage of computation. This algorithm does not give fruitful results when it talks about more parties. It is very difficult to apply this algorithm for huge databases. Final data mining result may break the privacy of individual's record.

D. Condensation Approach

Another approach used is Condensation approach .It was introduced by *Charu C. Aggarwal and Philip [12]* which builds constrained clusters in the data set and after that produces pseudo-data. The basic concept of the method is to contract or condense the data into multiple groups of predefined size. For each group, certain statistics are maintained. This approach is used in dynamic data update such as stream problems. Each group has a size of at least '*k*', which is referred to as the level of that privacy-preserving approach. The higher the level, the high is the amount of privacy. They use the statistics from each group in order to generate the corresponding pseudo-data. This is a simple privacy preservation approach but it is not efficient because it leads to loss of the information.

E. Hybrid technique

Privacy preservation is a very huge field. Many algorithms have been proposed in order to secure the data. Hybrid technique is a new technique through which one can combine two or more techniques to preserve the data. *Sativa Lohiya and Lata Ragha* [9] proposed a hybrid technique in which they used randomization and generalization. In this approach first they randomize the data and then generalized the modified or randomized data. This technique protects private data with better accuracy; also it can reconstruct original data and provide data with no information loss. Many other techniques can also be combined to make a hybrid technique such as Data perturbation, Blocking based method, Cryptographic technique, Condensation approach etc.

V. COMPARISON BETWEEN DIFFERENT TECHNIQUES

There are many different techniques proposed in the field of Privacy Preserving Data Mining but one outperforms over other or vice versa on different criteria. Algorithms are classified on the basis of performance, utility, cost, complexity, tolerance against data mining algorithms etc. We have shown a tabular comparison (table 1) of the work done by different authors in a chronological order (from past to present). We have taken the parameters like technique used for PPDM, its approach, results and accuracy.

TABLE I
TABULAR COMPARISON OF DIFFERENT TECHNIQUES

S. No	Authors	Year of Publication	Technique Used for PPDM	Approach	Result and Accuracy
1.	Y.Lindell, B.Pinkas [11]	2000	Cryptographic Technique	A technique through which sensitive data can be encrypted. There is also a proper toolset for algorithms of cryptography.	This approach is especially difficult to scale when more than a few parties are involved. Also it does not hold good for large databases.
2.	L. Sweeney[22]	2002	K- Anonymity	A record from a dataset cannot be distinguished from at least k-1 records whose data is also in the dataset.	K- Anonymity Approach is able to preserve privacy.
3.	J. Vaidya and C. Clifton[20]	2002	Association Rule	Distribution of data vertically into segments.	Distribution Based Association Rule Data Mining provides privacy.
4.	Hillol Kargupta, Souptik Datta, Qi Wang and Krishnamoorthy Sivakumar[7]	2003	Data Perturbation	They tried to preserve data privacy by adding random noise, while making sure that the random noise still preserves the "signal" from the data so that the patterns can still be accurately estimated.	Randomization-based Techniques are used to generate random matrices.
5.	CharuC.Aggarwa, Philip S. Yu[12]	2004	Condensation Approach	This approach works with pseudo-data rather than with modifications of original data, this helps in better preservation of privacy than techniques which simply use modifications of the original data.	The use of pseudo-data no longer necessitates the redesign of data mining algorithms, since they have the same format as the original data.
6.	A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam [24]	2006	L-Diversity Algorithm	If there are 'l' 'well represented' values for sensitive attribute then that class is said to have L- Diversity.	It is better than K-Anonymity in preserving Data mining.
7.	Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira[21]	2010	Anonymization	Anonymization is a technique for hiding individual's sensitive data from owner's record. 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.Kavitha[3]	2011	Hybrid Approach	Hybrid Approach is a combination of different techniques 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, methodological 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	Finding sensitive attribute and then they replace known sensitive values with unknown values ("?"). Finally the sanitized dataset is generated from which sensitive classification rules are no longer mined.	Unknown Values help in preserving privacy but reconstruction of original data set is quite difficult.
11.	Sara Mumtaz, Azhar Rauf and Shah Khusro[16]	2011	Distortion Based Perturbation Technique in OLAP Data Cube	Data perturbation technique 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 preserves, but also provides utmost accuracy with range sum queries and high availability.

S. No	Authors	Year of Publication	Technique Used for PPDM	Approach	Result and Accuracy
12.	Hsiang-Cheh Huang, Wai-Chi Fang[17]	2011	Histogram Based Reversible Data Hiding	A concept of reversibility which states that an original data can easily be hidden and the 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 databases for managing the medical images in the hospital.
13.	Jinfei Liu, Jun Luo and Joshua Zhexue Huang[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	Multi-Dimensional 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 results.
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.	Thanveer Jahan, Dr. G.Narsimha and Dr. C.V Guru Rao[15]	2012	Data Perturbation Using SSVD	An analyzing system used to transform 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	Sanitizes datasets using Sliding Window Algorithm and preserves data.	A novel approach that modifies the database to hide sensitive rules.
18.	M. N. Kumbhar and R. Kharat[18]	2012	Association Rule By Horizontal and Vertical Distribution	Different approaches in the field of Association rule are reviewed.	The performance of all models is analyzed in terms of privacy, security and communications.
19.	Savita Lohiya and Lata Ragha[9]	2012	Hybrid Approach	A combination of K- Anonymity and Randomization.	It has a better accuracy and original data can b reconstructed.
20.	Martin Beck and Michael Marh¨ofer[26]	2012	Anonymizing Demonstrator	Making a demonstrator with user friendly interface and performs Anonymization.	Swapping and Recording can be applied to enhance the utility.

VI. CONCLUSION

In today's world, privacy is the major concern to protect the sensitive data. People are very much concerned about their sensitive information which they don't want to share. Our survey in this paper focuses on the existing literature present in the field of Privacy Preserving Data Mining. From our analysis, we have found that there is no single technique that is consistent in all domains. All methods perform in a different way depending on the type of data as well as the type of application or domain. But still from our analysis, we can conclude that Cryptography and Random Data Perturbation methods perform better than the other existing methods. Cryptography is best technique for encryption of sensitive data. On the other hand Data Perturbation will help to preserve data and hence sensitivity is maintained .In future, we want to propose a hybrid approach of these techniques.

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.