Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.480 – 486

RESEARCH ARTICLE

A Comparative Study and Literature Survey on Privacy Preserving Data Mining Techniques

Sheryl Parmar[1], Mrs. Preeti Gupta[2], Ms. Pallavi Sharma[3]

[1]Amity University, Rajasthan, India
[2]Amity University, Rajasthan, India
[3]Amity University, Rajasthan, India
[1] parmarsheryl@gmail.com; [2] pgupta@jpr.amity.edu; [3] psharma1@jpr.amity.edu

Abstract- Privacy preserving data mining is concerned mainly about the hiding of the sensitive information and is a well-researched branch even today. In our daily life, the information sharing is more intuitive but it requires a secure protection so that the information is not revealed anywhere. The data mining process if carried out without security purposes, then it may lead to attacks, intrusions, disasters etc. kind of situations. In this paper, we represent the importance of privacy preservation for data mining and its various techniques to keep the data safe and secure and prevent it from different abuses.

Keywords— Privacy Preserving Data Mining, Signification, Data Perturbation, Sliding window Protocol, Cryptography

I. INTRODUCTION

Data Mining is the process of extracting knowledge from various available data sources and also has similar terms like data archaeology, data dredging, knowledge extraction and pattern analysis. The primary goal of Data Mining is to find out the task relevant data from the data sources using various techniques. Large information repositories, data warehouses and databases contain a wide range of data from which the knowledge has to be discovered. Various Data mining models and methods help to identify the necessary data which is then used for query processing, decision making and management of available information.

In the extensive amount where Data Mining is carried out in various applications, it becomes significant to not let the information known to everyone publicly. Privacy is the correct term which can be applied to the available data in order to gain security. In various applications, privacy preservation becomes important so as to not have any negative effects or any illegal access to the data or anyone's daily life. If the personal records of someone are leaked out, then there is not a single chance that the data could not be mutualized or misoperated. For example, in the banking industry the knowledge of name, address, gender etc. might not be dreadful for any person, but the credit card information may lead to an adversary actions that can result into unexpected consequences.

In this paper, we represent Privacy Preserving Data Mining and its different methods to have the security of the information and thereby a systematic management of information. There are some sensitive data that are supposed to be unrevealed that is the data needs to be protected to prevent the unfavourable conditions by applying various algorithms. The first step towards privacy preserving data mining is to extract the information from the data sources. The second step is mining the data by using various techniques or algorithms and the third and the final step is to apply privacy preserving methods for obtaining the desired results. Because of this procedure, the data quality is maintained according to the level of privacy achieved. The complexity of the algorithms depends upon the implementation and how efficiently the results are produced within the time duration taken by the particular algorithm.

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.[3][4] 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. [6] 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. [5]After mining privacy preservation techniques are used toprotect data from unauthorized access. Sensitive data of an individual can be prevented from being misused.

III.RESEARCH CHALLENGES

The research challenges have been identified in the field of privacy. In certain real world applications, there can be undesirable effects over data mining carried out without applying privacy on the data. Some of the examples are as follows:

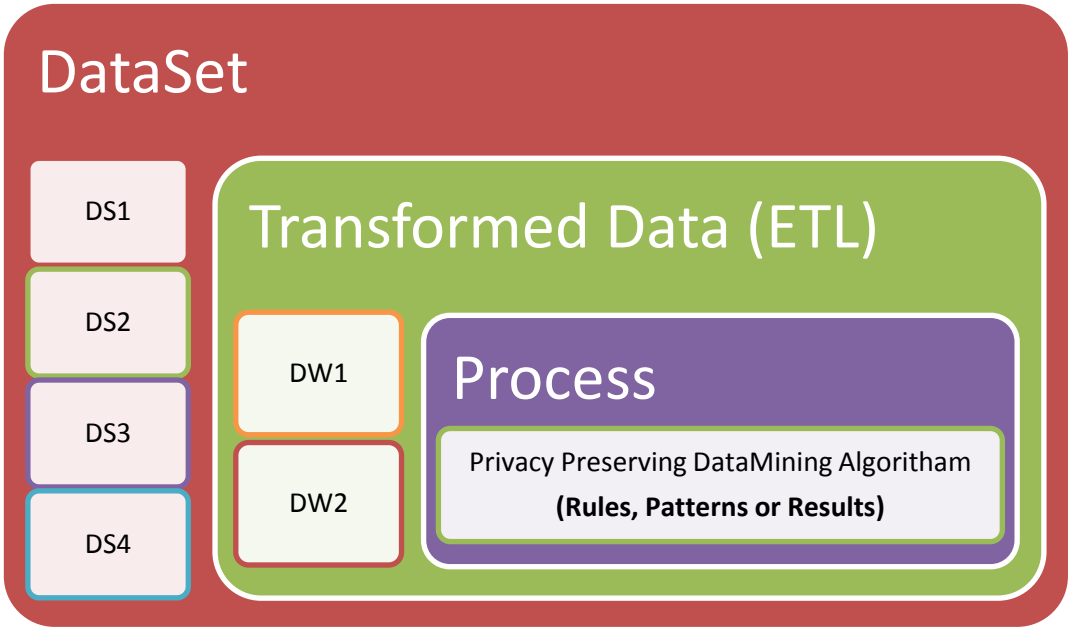

Fig. 1 Framework of privacy preserving data mining

A. In the social networking websites where millions of data are generated every day, the databases are required to be updated in the timely manner. [7] Also, the frequency of the usage of these kind of sites is much higher than expected. The attacks on these sites are very dangerous as the hackers hack the account of other users', misuse the private information and as a result the individual's identification becomes incorrect in these cases. This can also be considered as a theft because the knowledge of private information of a person, file transfers, chat logs etc. kind of functionalities can be utilized in a random fashion which is not permissible for an individual.[10] These type of attacks can be carried out by the members within the organization or outside the organization which are also known as intruders. Thus, the privacy policies are violated which needs to be revised by the authorities of the organization itself.

B. Medical organizations or hospitals also need to keep the data private so that the personal information of the individual cannot be identified very easily.[9][11] The medical data of a person can be known easily by just having some of the public data like cellular data list. The attributes of both the databases can be combined and therefore, it becomes very easy to extract the data of an individual by knowing the medical data and the identity of a person is inferred. The medical data of a person like visit date, procedure, disease, total amount can be linked with cellular data like name, phone number, address, zip code attributes so the entire identification of the person is recognized and this is known as linking attack. Medical data and cellular data combined can create threats like unnecessary calls and SMS, unpredictable warnings, fake medical reports, undiagnosed diseases' bills, etc. and many more. Therefore, there must be a strong policy of non-violation to prevent data from identity thefts kind of attacks.

IV. PRIVACY PRESERVING DATA MINING TECHNIQUES

SOME of the techniques of Privacy Preserving Data Mining are mentioned below:

A. Randomization

In the randomization approach, a random function is added to the original data so that the information is not revealed. But, the random function has to be much powerful in terms of providing security to the data. [12] The random function that can be added here is a sound function that is signification of data by setting some threshold value. When the characters n the data have a count of more than threshold, a longer sound wave needs to be added and if the count is less than the threshold, a shorter sound wave is to be added. [15] When the count of the characters is equal to the threshold, a fixed value of sound wave has to be added. Wecan use the parameter of sound in the environment tothe data columns to control the auditory icon. We may use the intuitive mapping between the sound and the events they indicate. The result will be the data added to a sinusoidal function and also here, a frequency of some hertz is a fixed value. Hence, the result finally would be the original data with $\sin(2\pi ft)$ where f is the frequency in hertz and t is the time in milliseconds. [18] The time is the count of characters of the data and it is compared with the threshold. With a 't' higher than threshold, the function to be added will be $\sin(2\pi ft+90^{o})$ and that with a 't' lesser than threshold, the function to be added will be $\sin(2\pi ft-90^{o})$ and in the case of count equal to threshold, the function remains the same that is $\sin(2\pi ft)$. Due to adding of different sound waves in the original information, now it would not be easier to obtain the original data.

B. Cryptographic technique

This technique indicates the encryption of the sensitive information available in the databases. The cryptographic approach includes the utilization of various encryption/decryption algorithms in order to secure the data so that no individual or no computer program can recognize that data. [20] Different algorithms support secure information transfer between parties sharing the data through certifications and authorizations. [21] The encryption keys

have the most popularized usage in this technique and these keys result into more secure information when it has to be sent to the other party. For example, Alice and Bob can communicate securely and safely with the help of encryption/decryption using the key and its information specified in it. But, this technique is no longer reliable when there are large numbers of parties sharing the data. [22] Also, these algorithms are impractical in the applications where millions of data are generated very often. It can break the individual's privacy and can lead to attacks.

C. Trajectory data

Privacy can be applied to the trajectories and sensitive attributes using generalization and suppression according to the requirements of the moving objects. The trajectory data are related to the location based services where the devices or objects are moving and their location based data get updated frequently. The attributes like name, age, address etc. make no effects to the privacy and hence, they can be eliminated simply. [23]The attributes like location_name, location_time, zipcode, etc. are the sensitive information and hence, they can be replaced with the help of a special character or a particular number. Here, the identities or the attributes cannot be linked with each other and hence, there are no chances of the similarity attacks. These kind of data are generally useful in the cellular devices where the GPS services are widely used to identify the location. The updating locations must be the correct ones irrespective of the maps and the directions and the time at which the actions or events take place must be associated with the similar attributes present in the database.

D. Hybrid approach

The combination of various methods can be useful here in order to carry out the privacy preserving data mining. The data perturbation approach and association rule can be combined where firstly the original dataset is converted into a distorted dataset using SSVD (Sparsified Singular Value Decomposition) and then sanitizing the data using Sliding window protocol that modifies the distorted values of the dataset to hide the sensitive rules. To distort the data, the original dataset is transformed into a data matrix. [24]The data matrix cannot be converted into the original data set unless the error factor i.e. the difference between the original dataset and data matrix is known. SSVD is the derivative of SVD (Singular Value Decomposition) which indicates the partitioning of the dataset into sub matrices. Once this is done and the distorted data is generated, sanitization of data is performed so that the content of the data becomes unreadable using Sliding Window Protocol by deciding the particular size of a window. Thus, hybrid approach can also be used to combine other different techniques to prevent data from several kinds of attacks.

V. COMPARISON BETWEEN DIFFERENT TECHNIQUES

Following are the methods or approaches proposed till now for the purpose of privacy preservation in data mining: 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. [12][16] 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.

Sr. No	Method name	Concept	Results
1	k-anonymity and l-diversity	Reduce the granularity of the data representation using generalization. This granularity is reduced sufficiently that any given record maps onto at least *k* other records in the data. The *l*-diversity model was designed to handle some weaknesses in the *k*-anonymity model.	K-anonymity is able to preserve the privacy but l-diversity is much more secure.
2	Association Rule	Distributing the data vertically into equivalent segments	Association rules are generated which are distribution based for data mining and hence privacy preserving.
3	Data Perturbation	Adding a random noise to the original data to preserve the patterns to be accurately estimated	Random matrices are generated to keep data private.
4	Decision Tree	A tree structure is generated to identify the best attribute and the tree grows and apply the privacy over them.	A graph-based framework to prevent sensitive information from attacks.
5	Generalization and suppression	Replacing the attribute values with some character , special character, number etc. to unidentified the original record	The resultant dataset will have protected values so that no intruder can overcome the original information.
6	Classification framework	Uses the anonymization and perturbation approaches and their sub-methods to modify the original data	Modifies the data using the particular technique and generates secure datasets.
7	Cryptographic technique	Use any encryption/decryption algorithm so as to keep the data in a secure manner.	The information provided is encrypted with the help of some key or certifications that provides original data only upon the appropriate decryption method.
8	Condensation approach	Just the original data are modified which better helps in the privacy and works for pseudo-data.	The pseudo-data is non-significant as it has the same format as the original data.
9	Blocking based technique	The sensitive values of the attributes are replaced with unknown values and the final dataset is no longer mined.	Unknown values preserve privacy but reconstruction of original data becomes difficult.
10	Value class membership and distorted value	The data is split into various classes of similar type and the values are distorted with the help of some class functions	The original data are generated in the form of intervals through discretization and adding random perturbations.
11	Distributed privacy preservation	Data sets are partitioned either horizontally or vertically across the individual entities which wish to derive aggregate results.	In horizontally partitioned data sets, the individual records are spread out across multiple entities, each of which have the same set of attributes. In vertical partitioning, the individual entities may have different attribute of the same set of records.
12	Hybrid approach	Combining the methods and as a result the privacy can be maintained.	It is more accurate and it can keep data secure than any other technique.

Table I : Privacy Preserving Data Mining techniques comparisons &survey results

VI. CONCLUSION

In daily life, privacy has become a topic of more concern as people do not want to share the sensitive information with other people or parties. In this paper, we represent a literature on the various privacy preserving techniques that are most popularly used. All the techniques work upon the type of the application. The type of data also matters a lot when it comes to maintaining private data. In future, we intend to combine various methods and algorithms to present a hybrid approach so that the privacy could be stronger and hence there would not be any kind of violation in the policies. 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 end.,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. JesuVedhaNayahi 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] SlavaKisilevich, LiorRokach, Yuval Elovici, BrachaShapira, "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, IEEE2012.

