ISSN:0975-9646

Survey on Data Mining with Privacy Preservation

Sachin Janbandhu[1], Dr.S.M.Chaware[2]
[1,2]*Department of Computer Engineering*
TSSM'S Bhivarabai Sawant College of Engineering and Research, Narhe, Pune. India

Abstract: **Data mining is the process to look for hidden patterns and trends in data that is not immediately apparent from summarizing the data. Due to the rapid development in technology, it needs updated techniques to provide privacy for data mining. There are various methods and techniques which have been created for providing privacy to the process of data mining. Here in this paper we have done the survey of various privacy preserving data mining algorithms and various techniques for privacy preserving data mining along with their advantages and disadvantage. In last we have discussed present limitations and scope for future research in privacy preserving data mining.**

Keywords: **Privacy Preserving Data Mining, Privacy, Randomization, K-Anonymity, Secure Multiparty Computation, perturbation approach, Distributed privacy preserving.**

I. INTRODUCTON

Data mining is a process to extract the information or knowledge automatically and intelligently from a huge amount of data, here in the process of data mining sensitive information can be disclosed by compromising the individual's right to privacy [1]. Increasing demand of Privacy preservation in data mining gives us direction to research about privacy preserving data mining.

Considering the rapid development in technology such as internet, data storage, data processing methods, we need to pay equal attention towards privacy preserving data mining. For secured public system we not only need to take care about the trimming of data but also the data inference. There are number of privacy preserving data mining methods have been proposed [2-37], but most of these methods are having some drawbacks. This paper gives a survey of various privacy preserving data mining methods and analyses the representative methods for privacy preserving data mining, as well as points out their advantages and disadvantages.

The rest of this paper is organized as follows. In section 2, introduction of various methods of privacy preservation. In section 3, the randomization technique for privacy preserving on the original data is analyzed. In section 4, the anonymization technique is discussed. In section 5 perturbation approach is explained. The encryption method and distributed privacy preserving data mining will be discussed in section 6. Section 7 contains the conclusions and future work.

II. METHODHS OF PRIVACY PRESERVATION

Main aim of Privacy preservation data mining is to find out such solution which will minimise the risk of misuse the data used for method generation. There are number of effective methods for privacy preserving data mining which have been proposed after extensive study of data mining community in recent years. In order to perform the privacy preservation most of the technique uses some form of transformation on the original data. Here in this case it is important to maintain the benefits of privacy preservation even after the transformed dataset is made available for mining. We have classified such techniques as follow

A. Randomization

This is one of the popular techniques in privacy preserving data mining studies. By adding noise to the original data, the values of the records are created in this method The individual values of the records can no longer be recovered as noise added to original data is large enough to maintain the privacy. Randomization techniques achieve both, privacy preservation and knowledge discovery with the help of random-noise-based perturbation and Randomized Response scheme. Though this technique causes high information loss it is a more efficient method.

B. Anonymization

Using generalization and suppression anonymization technique creates indistinguishable records among group of records. K-anonymity is known as representative anonymization technique.

To identify records uniquely it considers quasi-identifiers which can be used in conjunction with public records. There are so many techniques which has been proposed such as (a, k)-anonymity, l-diversity, t-closeness, M-invariance, Personalized anonymity, p-sensitive k-anonymity and so on. Here anonymization technique results in loss of information to some extent but ensures the originality of data.

C. Encryption

Encryption technique solves the problem of data privacy easily. Use of encryption techniques makes easy to conduct data mining among mutual un-trusted parties, or even between competitors. In distributed data mining encryption technique is used due to its privacy concern. Neglecting the efficiency of Encryption, it is used in both approaches of distributed data mining i.e. horizontally partitioned data and that on vertically partitioned data.

III. RANDOMIZATION TECHNIQUE

Due to the independency of noise added to a given records and behaviour of other 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.

The model of randomization is shown in Figure 1.

Fig 1. The Model of Randomization

Representative randomization method is a combination of random-noise-based perturbation and Randomized Response scheme. One of the proposed a scheme for privacy preserving data mining using random perturbation and discussed how the reconstructed distributions may be used for data mining [4] is where In their randomization scheme, a random number is added to the value of a sensitive attribute. For example, if x_i is the value of a sensitive attribute, $x_i + r_i$, rather than x_i , will appear in the database, Where ri is a random noise drawn from some distribution. It is shown that given the distribution of random noises, reconstructing the distribution of the original data is possible. Subsequently, another approach to conduct privacy preserving association rule mining [5]. Where random matrix based spectral filtering technique is used to get the original data from perturbed data, proposed by Kargupta [6]. PCA-DR and MLE-DR, these two data reconstruction methods were proposed by Huang. With respect to different randomization operators [8-10], various distribution reconstruction algorithms have been proposed. To estimate the original data distribution based on randomization operators and the randomized data, Bayesian analysis is used as a basic idea. Such as expectation maximization algorithms generates reconstruction distribution which converges to the maximum likelihood estimate of the original distribution.

Warner [11] has proposed the randomized response (RR). RR scheme was developed to maintain the privacy of sensitive information while collecting it from individuals as well as while processing the data. For mining association rules considering secrecy constraints MASK scheme is proposed by Rizvi and Harshita [12]. An approach to conduct privacy preserving decision tree building [13] was proposed by Du and Zhan. Various issues regarding accuracy provision with respect to various reconstructed measures in privacy preserving market basket data analysis are discussed by Guo [14]. In Short at the time of data collection randomization technique can be easily implemented. It is one of the useful techniques used to hide individual data in privacy preserving data mining. Though it results in high information loss, the randomization method is more efficient.

IV. ANONYMIZATION TECHNIQUE

To protect the privacy of individuals is the major concern of world considering the rapid growth in database, computing technologies, and networking. Increasing amount of sensitive data can be examined and analysed which leads to use of data mining tools for inferring trends and patterns. Removing key identifiers such as the name and social-security number from individual records from individual records, the data record are often made available. To identify individual records exactly, the combination of other attributes (quasi-identifiers) can be used. For example, attributes such as race, birth, sex, and zip are available in public records such as voter list. When these attributes are also available in a given data set such as medical data, they can be used to infer the identity of the corresponding individual with high probability by linking operation, as is shown in Figure 2.

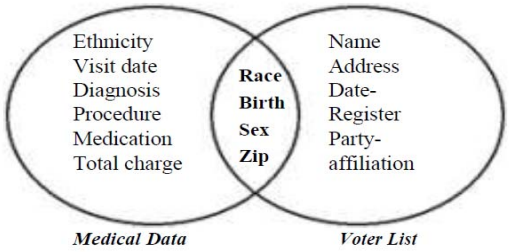

Figure 2. A Sample of Linking Attack

In order to preserve privacy, Sweeney [15] proposed the k-anonymity model which achieves k-anonymity using generalization and suppression, so that, any individual is indistinguishable from at least k-1 other ones with respect to the quasi-identifier attributes in the anonymized dataset. For example, Table 2 is an anonymous table of Table 1. Generalization involves replacing (or recoding) a value with a less specific but semantically consistent value. For example, the date of birth could be generalized to a range such as year of birth, so as to reduce the risk of identification. Suppression involves not releasing a value at all. It is clear that such methods reduce the risk of identification with the use of public records, while reducing the accuracy of applications on the transformed data

TABLE I. Original Data

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 II. Anonymization of Table 1

Race	Birth	Sex	Zip	Disease
Blank	1965	M	0214*	Flu
Blank	1965	M	0214*	Cancer
Blank	1966	M	0213*	Obesity
Blank	1966	M	0213*	Gastritis
White	1968	F	0213*	HIV
White	1968	F	0213*	Cancer

Number of algorithms has been proposed for k-anonymity implementation using generalization and suppression. Agrawal [16] has proposed an optimal algorithms starting with fully generalized and specializing the datasets in minimal k-anonymous table. LeFevre [17] has described an algorithm which uses a bottom up technique and priori computation. To make a table to be released k-anonymous a top down heuristic approach were presented [18]. Sweeney [19] has proved theoretically that the optimal k-anonymity is NP-Hard and provided approximation algorithms for optimal k-anonymity. Machanavajjhala [20] explained introduces the 1-diversity method, and focused on how a user may guess the sensitive values with high confidence when the sensitivity data is lack of diversity. At the same time models such as p-sensitive k-anonymity- closeness and M-invariance etc are discussed in the literature in respect to deal with the problem of k-anonymity. Neglecting the concrete needs the same amount of preservation for all individuals is the main focus of k-anonymity method. While applying extra privacy control to another subset, insufficient protection may be provided t a subset of people. Xiao and Tao [25] introduced a new generalization framework based on the concept of personalized anonymity. Here a technique used fulfils all demanding requirements which results in retention the largest amount of original data. K-anonymity solutions results in high information loss and low usability due to dependency on pre-defined generalization hierarchies or full order imposed on each attribute domain. Algorithms based on clustering technique [26-29] which reduces the amount of information loss.

Data mining with K-Anonymity method is demanding area of research but there are so many points to be investigated such as combination of k-anonymity with other possible data mining techniques, detection and blocking k-anonymity violations. The anonymization method can ensure that the transformed data is true, but it also results in information loss in some extent.

V. PERTURBATION APPROACH

The perturbation approach restricts data services to learn or recover precise records. This restrictions leads to some challenges. As this method does not reconstruct the original data values excepting distribution, so here new algorithms require to be developed for reconstructed distributions to perform mining of the underlying data. For each problem like classification, association rule mining, or clustering, a new distribution based data mining algorithm need to be developed. A new distribution based data mining algorithm for classification problem is developed by Agrawal. Vaidya and Clifton and Rizvi and Haritsa develop methods for privacy-preserving association rule mining. Problems such as association rules and classification some clever approaches have been developed for distribution based mining of data. Using distribution instead of original records results in restricting the range of algorithmic technique that can be used on the data. The distribution of each data dimension reconstructed independently, in perturbation approach. Indirectly it means that all distribution based mining algorithm works under an implicit assumption to treat each dimension independently. Algorithms such as classification keeps relevant information hidden in inter attribute correlations. For example, the classification technique uses a distribution-based analogue of single-attribute split algorithm. However, other techniques such as multivariate decision tree algorithms cannot be accordingly modified to work with the perturbation approach. This is because of the independent treatment of the different attributes by the perturbation approach.

This means that distribution based data mining algorithms have an inherent disadvantage of loss of implicit information available in multidimensional records. Another branch of privacy preserving data mining which using cryptographic techniques was developed. This branch became hugely popular for two main reasons: Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast tool set of cryptographic algorithms and constructs to implement privacy -preserving data mining algorithms. However, recent work has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining.

VI. DISTRIBUTED PRIVACY PRESERVING DATA MINING AND ENCRYPTION METHOD

Due to the tremendous growth in internet it has produced more opportunities for distributed data mining, where people conduct mining task jointly with their private inputs. Such mining tasks can occur between mutual un-trusted parties, between competitors so it causes privacy leakage. So Distributed privacy preserving data mining algorithms require collaboration between parties to compute the results or share no-sensitive mining results, while provably leading to the disclosure of any sensitive information.

Distributed data mining consists of two ways such as horizontally partitioned data and vertically partitioned data. Horizontally partitioned data means that each site has complete information on a distinct set of entities, and an integrated dataset consists of the union of these datasets. In contrast, vertically partitioned data has different types of information at each site; each has partial information on the same set of entities.

Most of the privacy preserving distributed data mining algorithms focuses on to reveal the final result. Incorporating cryptographic techniques were designed by

Kantarcioglu and Clifton [30] to minimize the information shared while adding little overhead to the mining task. Lindell et al. [31] researched how to privately generate ID3 decision trees on horizontally partitioned data. The problem of privately mining association rules on vertically partitioned data was addressed in [32, 33]. Vaidya and Clifton [34] first studied how secure association rule mining can be done for vertically partitioned data by extending the Apriori algorithm. Du and Zhan [35] developed a solution for constructing ID3 on vertically partitioned data between two parties. Vaidya and Clifton [36] developed a Naive Bayes classifier for privacy preservation on vertically partitioned data and [37] proposed the first method for clustering over vertically partitioned data. All these methods are almost based on the special encryption protocol known as Secure Multiparty Computation (SMC) technology.

SMC originated with Yao's Millionaires' problem [38]. The basic problem is that two millionaires would like to know who is richer, with neither revealing their net worth. Abstractly, the problem is to simply compare two numbers, each held by one party, without either party revealing its number to the other. Two basic adversarial models were defined by SMC literature

A. Semi-Honest Model

Semi-honest adversaries follow the protocol faithfully, but can try to infer the secret information of the other parties from the data they see during the execution of the protocol.

B. Malicious Model

Malicious adversaries may do anything to infer secret information. They can abort the protocol at any time, send spurious messages, spoof messages, collude with other (malicious) parties, etc.

SMC technology used in distributed privacy preserving data mining areas mainly consists of a set of secure sub-protocols, such as, secure sum, secure comparison, dot product protocol, secure intersection, and secure set union and so on.

VII. CONCLUSION AND FUTURE WORK

Here in this paper a wide survey has been done on the various approaches for privacy preserving data mining and analysed the major algorithms for data mining with their drawbacks. In order to find out the perfect and efficient solution for privacy preservation the following issues should be widely studied:

1. Privacy should be achieved with accuracy. So application of various optimizations should be deeply researched.
2. Data sanitization process should be with minimum negative impact.
3. In distributed data mining, more efficient algorithms should be developed to balance all costs such as computation cost, communication cost, disclosure cost. Etc
4. Deployment of privacy preserving technique into practical is also need to be studied.

REFERENCES

[1] Han Jiawei, M. Kamber, Data Mining: Concepts and Techniques, Beijing: China Machine Press, pp.1-40,2006.

[2] V.S.Verykios, E.Bertino, I.N.Fovino, L.P.Provenza, Y.Saygin, Y.Theodoridis, "State-of-the-art in Privacy Preserving Data Mining", New York, ACM SIGMOD Record, vol.33, no.2, pp.50-57,2004.

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

[4] R. Agrawal, R. Srikant, "Privacy-Preserving Data Mining", ACM SIGMOD Record, New York, vol.29, no.2, pp.439-450,2000.

[5] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, "Privacy Preserving Mining of Association Rules", Information System, vol.29, no.4, pp.343-364,2004.

[6] 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.

[7] 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.

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

[9] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrk, "Privacy Preserving Mining of Association Rules", In Proceedings the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, pp.217-228, 2002.

[10] S. Rizvi, J. Haritsa, "Maintaining Data Privacy in Association Rule Mining", In Proceedings the 28th International Conference on Very Large Data Bases, pp.682-693, 2002.

[11] 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.

[12] S.J. Rizvi, J.R. Haritsa, "Maintaining Data Privacy in Association Rule Mining", In Proceedings the 28th VLDB conference, pp.1-12, 2002.

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

[14] 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.

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

[16] R. Bayardo, R. Agrawal, "Data Privacy Through Optimal k-Anonymization", In Proceedings the 21st International Conference on Data Engineering, pp.217-228, 2005.

[17] K. Lefevre, J. Dewittd, R. Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity", In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp.49-60, 2005.

[18] B. Fung, K. Wang, P. Yu, "Top-down Specialization for Information and Privacy Preservation", In Proceedings of the 21st IEEE International Conference on Data Engineering, pp.205-216, 2005.

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

[20] A. Machanavajjhala, J. Gehrke, D. Kifer, "l-Diversity: Privacy Beyond k-Anonymity", ACM Transactions on Knowledge Discovery from Data, pp.24-35,2007.

[21] T. Truta, B. Vinay, "Privacy Protection: p-Sensitive k-Anonymity Property", In Proceedings of the 22nd International Conference on Data Engineering Workshops, pp. 94-103, 2006.

[22] R.C.Wong, J.Y.Li, A.W. Fu, "(a, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing", In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.754-759, 2006.

[23] 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.

[24] 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.

[25] X.K. Xiao, Y.F. Tao, "Personalized Privacy Preservation", In Proceedings of the ACM Conference on Management of Data (SIGMOD), pp.229-240, 2006.

[26] G. Loukides, J.H. Shao, "An Efficient Clustering Algorithm for k-Anonymisation", International Journal of Computer Science And Technology,vol.23, no.2, pp.188-202, 2008.

[27] J.L. Lin, M.C. Wei, "Genetic Algorithm-Based Clustering Approach for k-Anonymization", International Journal of Expert Systems with Applications,vol.36, no.6, pp.9784-9792, 2009.

[28] 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.

[29] 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.

[30] 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.

[31] Lindell, Yehuda, Pinkas, "Privacy preserving data mining", In Proceedings of the Advances in Cryptology–CRYPTO, pp.36–54,2000.

[32] J. Vaidya, 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, 2002.

[33] I. Ioannidis, A. Grama, M.J. Atallah, "A Secure Protocol for Computing Dot-Products in Clustered and Distributed Environments", In Proceedings of the 31st International Conference on Parallel Processing, pp.379-384, 2002.

[34] J. Vaidya, 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, 2002.

[35] W.L. Du, Z.J. Zhan, "Building Decision Tree Classifier on Private Data", In Proceedings of the IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining, pp.1-8, 2002.

[36] J. Vaidya, C. Clifton, "Privacy Preserving Naive Bayes Classifier for Vertically Partitioned Data", In Proceedings of the 2004 SIAM International Conference on Data Mining, pp.522–526, 2004.

[37] J. Vaidya, C. Clifton, "Privacy-Preserving k-Means Clustering over Vertically Partitioned Data", In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.206–215, 2003.

[38] Yao, C. Andrew, "How to Generate and Exchange Secrets", In Proceedings of the 27th IEEE Symposium on Foundations of Computer Science, pp.162-167, 1986.